Implementation of monocular depth estimation on CityScapes dataset

Research Project

submitted
by

Abhishek Lahiri

born 12.10.1997 in Barasat, India

Written at

Lehrstuhl für Mustererkennung (Informatik 5)
Department Informatik
Friedrich-Alexander-Universität Erlangen-Nürnberg.

Advisor: Prof. Hakan Calim

Started: 01.08.2024

Finished: 28.10.2025

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Erlangen, den 27. Oktober 2025

Abstract

The Sidewalk Environment Detection System for Assistive NavigaTION (SENSATION) is designed to enhance the everyday navigation experience for the Blind or visually impaired persons (BVIPs) in urban environments.This research focuses on applying an integrated framework of monocular depth estimation to produce an understanding of metric depth. Monocular depth prediction models typically generate relative depth, which limits their usability in real-world perception tasks. To address this, four recent depth estimation architectures—MiDaS, DPT, MonoDepth2, and ZoeDepth—are adapted and evaluated on the Cityscapes dataset. A scale–shift alignment procedure, informed by stereo-derived disparity maps, is employed to convert relative predictions into absolute metric depth. This is very relevant in the context of the actual goal of estimating various obstacles on the road, like bicycles, cars, posts, etc., for visually impaired people. Experimental evaluation demonstrates that relative depth can be reliably transformed into metric scale, using a scaling factor for the specific dataset, which is Cityscapes in this case. The framework provides a reproducible pipeline for benchmarking monocular depth estimation on Cityscapes and underscores its potential for advancing perception in order to enhance the mobility capabilities of BVIPs by ensuring safer and more efficient navigation on pedestrian pathways.

Contents

Chapter 1

Introduction

Depth estimation is one of the most fundamental challenges in the field of computer vision, as it has a wide range of applications in autonomous driving systems, augmented reality, 3D scene reconstruction, etc. While well-established methods using Light Detection and Ranging (LiDAR), stereo cameras, and structured-light sensors provide accurate measurements, they are expensive, with high power consumption, and environmentally constrained. Monocular depth estimation can be considered as a good alternative, as it provides the potential to extract depth information from single RGB images using computational methods and algorithms rather than specialized sensors. This approach has a clear advantage in terms of resource consumption, especially in the problem that we are trying to solve, catering to BVIPs. However, most of these methods produce relative depth rather than absolute metric depth, limiting their utility in applications requiring precise measurements such as navigation and obstacle avoidance.

This research addresses this challenge, to obtain the absolute metric depth from relative monocular depth data by combining neural architectures with semantic understanding. This study evaluates four state-of-the-art (SOTA) depth estimation models—MiDaS, Dense Prediction Transformer (DPT), MonoDepth2, and ZoeDepth—using the Cityscapes dataset, which provides a comprehensive analysis on how these models perform in urban street settings and sidewalks.

To establish the relation between relative depth predictions and ground truth metric measurements, a scale-shift alignment method is used based on disparity maps. Additionally, the research framework incorporates semantic segmentation capabilities to en-

able object-class-specific depth analysis, computing mean depth values for distinct objects/obstacles such as vehicles, pedestrians, and infrastructure elements. This approach provides valuable insights regarding metric depth estimation techniques.

This study evaluates model accuracy when adapted for metric estimation and analyzes trade-offs between computational complexity, inference time, and prediction quality. Through systematic experiments on Cityscapes, the research provides insights into the scalability and usability of monocular depth estimation for obstacle detection that can be beneficial for BVIPs using a stereo-camera image data.

Chapter 2

Depth Estimation Models

2.1 MonoDepth2

MonoDepth2 is a widely used approach for Relative Depth Estimation (RDE) that builds upon the idea of self-supervised learning. Instead of relying on large amounts of ground-truth depth data, which are expensive and difficult to collect, MonoDepth2 trains using stereo image pairs or consecutive frames from monocular videos. The key principle is image reconstruction: the model predicts the depth of a given image, and then, using camera intrinsics and estimated pose, reconstructs nearby views. The difference between the original image and the reconstructed image becomes the training signal. This allows the network to learn depth in a self-supervised fashion, making it adaptable to large-scale datasets without explicit depth labels [God+19].

The architecture of MonoDepth2 follows an encoder–decoder design. A ResNet-based encoder extracts rich features from the input image, while a decoder predicts disparity maps at different resolutions. These disparity values are later converted into depth using the known camera parameters. A well-known limitation of monocular training is scale ambiguity, meaning the model can only learn relative depth. To address this, MonoDepth2 includes a stereo-trained variant where a nominal baseline is assumed. On the KITTI dataset, for example, predictions are scaled by a factor to align with metric depth. This makes the framework versatile: efficient enough for real-time applications, yet robust enough to provide reliable relative depth estimates. [God+19]

In this research, Monodepth2 model is applied to the Cityscapes dataset, where the relative depth data are converted into metric or absolute depth values using a dataset-specific scaling factor. In order to achieve this, a scaling factor is calculated based on known scene geometry, such as the baseline and focal length of the camera. This implementation ensures that the predicted depths align with real-world distances, making the model's output suitable for identifying the distance of objects or obstacles for BVIPs in streets and sidewalks. Similar to the other depth estimation models considered in this work, the use of a scaling factor is crucial in bridging the gap between relative and absolute metric depth estimation.

2.2 MiDaS

MiDaS (Mixed Dataset Training for Depth Estimation) is one of the most widely recognized monocular depth estimation models. Unlike approaches that train on a single dataset, MiDaS was trained on a mixture of diverse datasets allowing it to generalize across a wide variety of street scenes. The model architecture is also based on an encoder–decoder design, where encoders such as ResNet or Vision Transformers extract features, and a decoder reconstructs dense depth predictions. MiDaS predicts relative depth maps, which capture the geometry of the scene accurately but do not inherently represent metric scale.[Las+19]

The robustness of the MiDaS model comes from its training on heterogeneous datasets. The predictions are consistent across various environments, making it versatile to different tasks like robotics, autonomous navigation, and, in our case for assisting in obstacle identification of BVIPs. However, its limitation of relative depth is countered with an additional alignment process. To convert these predictions into metric depth, alignment strategies are used that compute scale and shift-invariant losses from known ground-truth samples, following the methodology proposed in the MiDaS paper. [Las+19]

In this project, MiDaS is evaluated on the Cityscapes dataset. Since its predictions are relative, a scaling factor and shift are computed by aligning model outputs with the disparity data obtained from Cityscapes. This enables the conversion of relative depth into absolute metric depth, which can be useful in calculating the distance of obstacles in city street scenarios.

2.3 DPT (Dense Prediction Transformer)

DPT (Dense Prediction Transformer) is a transformer-based model that adapts the Vision Transformer (ViT) architecture for dense prediction tasks such as semantic segmentation and monocular depth estimation. Unlike convolutional architectures that capture local spatial features, DPT can produce finer-grained and more globally coherent predictions. This allows the network to better capture structural relationships between distant parts of a scene, which is highly beneficial for estimating depth. The architecture combines a transformer encoder with a lightweight convolutional decoder, ensuring accurate depth predictions while maintaining computational efficiency. [Ran+21]

Several variants of DPT have been developed, including DPT-Large and DPT-Hybrid, each designed to address different application requirements. The larger models demonstrate impressive performance on established benchmarks such as NYU Depth v2 and KITTI, often achieving state-of-the-art results, while the smaller variants prioritize inference speed to meet the demands of real-time applications. DPT can output relative or metric depth depending on the dataset used for training. For example, models fine-tuned on NYU or KITTI datasets can approximate metric depth directly, while more general-purpose models predict relative depth that requires scaling.[Ran+21]

In our implementation, DPT is applied to the Cityscapes dataset with depth predictions adjusted using dataset-specific scaling parameters. This scaling process is essential for converting the model's inherently relative depth outputs into absolute metric measurements that can be meaningfully interpreted and compared. Through this adjustment, we ensure that DPT operates on the same measurement scale as the other models in our study—MiDaS, MonoDepth2, and ZoeDepth—enabling fair and consistent evaluation across all architectures within our experimental framework.

2.4 ZoeDepth

ZoeDepth is a relatively new monocular depth estimation model that introduces scale-invariant training methodologies to directly predict metric depth across different domains. While earlier models could only capture relative geometry, ZoeDepth is capable of integrating a scale-aware representation that helps it generalise effectively to unseen datasets. ZoeDepth is built on the DPT encoder-decoder architecture, replacing the encoder with

more recent transformer-based backbones. This makes for a strong case for this model in real-world scenarios, where metric accuracy is essential, especially in the case of this particular research focus. [Bha+23]

Another key feature of ZoeDepth is its wide variety of models—such as ZoeD-N, ZoeD-K, and ZoeD-NK—trained on combinations of datasets like NYU, KITTI, and synthetic data. This training approach of mixing multiple datasets with different depth distributions helps ZoeDepth achieve improved cross-dataset generalisation compared to other SOTA models. DPT's ability to directly infer metric-scale depth from individual RGB images sets it apart from other models, which inherently require additional scaling techniques to obtain depth data. [Bha+23]

For this research, ZoeDepth is applied to the Cityscapes dataset to evaluate its outputs in terms of Metric Depth Estimation (MDE). As this model can already provide metric depth data, this helps in establishing a proper comparative analysis on how accurate does ZoeDepth perform relative to the other aligned models such as MiDaS, Monodepth2, and DPT.

Chapter 3

Experiment

3.1 Dataset

This research was conducted on the CityScapes Dataset with different monocular depth estimation models to find out the relative and absolute depth data. Cityscapes Dataset is a comprehensive, high-resolution image dataset designed to advance research in semantic understanding of complex urban street scenarios. It contains videos in stereo vision captured across 50 cities of Germany and nearby regions, showing different environmental conditions and surroundings. Each of the images is labeled at the pixel level with many object categories such as roads, vehicles, buildings, and pedestrians, etc. making it ideal for tasks such as semantic segmentation as well as depth estimation. The dataset consists of 5,000 finely annotated images and about 20,000 coarsely labeled ones, all in high resolution (1024×2048). Each annotated frame also comprises of stereo images, GPS data, and measurements like vehicle motion, which might be useful to develop models that better interpret real-world scenes.[Cor+16]

The dataset was introduced in 2016 by Marius Cordts and colleagues from Daimler AG and the Max Planck Institute for Informatics to fill a major gap in computer vision: there were few high-resolution, richly annotated datasets for urban environments. Over time, it has become a foundation for research on autonomous driving, smart cities, and urban scene understanding. It also expands to even more complex and details datasets like Cityscapes 3D that have added new possibilities for studying 3D object detection and spatial reasoning. Designed for easy use with tools like TensorFlow and PyTorch, Cityscapes continues to be one of the most widely used benchmarks for testing computer vision systems in real-world urban scenarios.

3.2 Methodology

The experimental pipeline is divided into two main stages: (1) Data Processing and Depth Prediction, handled by three separate scripts (`run_midas_depth.py`, `run_zoedepth.py`, `run_monodepth2.py`), and (2) standardized Evaluation, handled by a single script that takes the ground truth depth and predicted depth and generates the evaluation metrics (`depth_evaluation.py`).

3.2.1 Dataset and Ground-Truth Preparation

For this implementation, the Cityscapes `val` split (500 images) was utilised. The pipeline requires three components of the dataset: `leftImg8bit` (RGB inputs), `disparity` (ground truth), and `camera` (calibration data).

A critical prerequisite was the conversion of Cityscapes' disparity maps into metric depth. This was achieved via the following formula:

$$\text{Depth (meters)} = \frac{\text{baseline} \times \text{focal\_length}}{\text{disparity}} \tag{3.1}$$

(1) **Disparity:** The 16-bit disparity `.png` files were loaded, and pixel values p were converted to disparity units using the official formula: disparity $= (p - 1)/256.0$.

(2) **Calibration:** The corresponding `.json` file for each image was parsed to extract the stereo `baseline` (in meters) and the `focal_length` (in pixels).

(3) **Robustness:** The parsing logic was designed to handle inconsistencies in the camera JSON files, checking for focal length under the keys `fx`, `focalLength`, and `K[0][0]` (the first element of the intrinsic matrix).

(4) **Metric GT:** The formula was applied to all valid disparity pixels ($p > 0$) to generate the ground-truth metric depth map.

3.3 Implementation

In this project, the goal is to apply various SOTA models of monocular depth estimation on the above-mentioned CityScapes dataset and to potentially find out absolute depth data. For this evaluation, the following list of models was taken into consideration -

Table 3.1: List of RDE and MDE Models

Category	Models
MiDaS v3.1 (BEiT Backbone)	• dpt-beit-large-512 • dpt-beit-large-384 • dpt-beit-base-384
MiDaS v3.1 (SwinV2 Backbone)	• dpt-swinv2-large-384 • dpt-swinv2-base-384 • dpt-swinv2-tiny-256
MiDaS v3.0 (DPT)	• dpt-large
MiDaS v2.1 (Hybrid & Convolutional)	• dpt-hybrid-midas • midas-v21-384 • midas-v21-small-256
MonoDepth2	• mono+stereo_640x192
ZoeDepth	• zoedepth-nyu-kitti • zoedepth-kitti • zoedepth-nyu

3.3.1 Model Implementation and Scale-Shift Alignment

RDE Models (MiDaS & MonoDepth2)

These models require an alignment step to map their relative output to metric space. We implemented a **global scale-and-shift alignment** strategy.

(1) **Calibration Phase:** A calibration set of $N = 20$ images was sampled from the validation set. For each image, the model's relative prediction (`Pred_Rel`) and the metric ground truth (`GT_Metric`) were generated.

(2) **Least-Squares Fit:** A mask was created for valid pixels (0.1m $<$ `GT_Metric` $<$ 80.0m). Using these pixels, we solved the linear least-squares problem `GT` $= s \times$ `Pred_Rel` $+ t$ to find the optimal scale (s) and shift (t) for that single image.

(3) **Robust Aggregation:** After processing all 20 samples, this yielded 20 *scale* and 20 $shift$ values. To get a stable global factor and reject outliers (e.g., from failed predictions on dark or sky-filled images), the **median** of each list was computed:

- global_scale $=$ median$(s_1, \ldots, s_N)$
- global_shift $=$ median$(t_1, \ldots, t_N)$

(4) **Main Processing Phase:** This single `global_scale` and `global_shift` pair was then applied to *a subset of 100 images* in the validation set. The final metric prediction for every image was calculated as: `Pred_Metric` $=$ (`global_scale`$\times$`Pred_Rel`)$+$ `global_shift`

This method fairly tests how well a single, generalized alignment factor (derived from a sample) applies to the entire dataset.

Model-Specific Details:

- **MiDaS / DPT:** Loaded via the Hugging Face `transformers` library. The models output relative inverse depth, which was inverted (1.0/output) to get relative depth before alignment. [Las+19] [Bir+23] [Ran+21]
- **MonoDepth2:** Loaded from its official GitHub repository. This required custom `sys.path` manipulation and careful loading of the `state_dict` to remove non-weight keys (`height`, `width`) from the checkpoint file. The model outputs relative disparity, which was inverted (1.0/disparity) to get relative depth.

MDE Model (ZoeDepth)

The `ZoeDepth` model [Bha+23] was also loaded from Hugging Face. As it is trained to output metric depth directly, **no alignment was performed.** The raw output from the model was saved directly as the final metric prediction. This allows for a direct evaluation of its out-of-the-box metric accuracy.

3.4 Evaluation Metrics

A unified evaluation script (`evaluate_depth.py`) was developed to ensure all models were judged by the same criteria. The script accepts a directory path via a command-line argument (`--dir`), loads all ground-truth (`.npy`) and predicted (`.npy`) files, and computes aggregate metrics.

Evaluation Mask: Before calculation, all pixel data was masked. A pixel i was included in the evaluation *only if* it met the following criteria, as per standard Cityscapes evaluation protocols:

(1) `GT_Metric`$[i] > 0.1$ (Valid ground truth)

(2) `GT_Metric`$[i] < 80.0$ (Within valid range)

(3) `Pred_Metric`$[i] > 0.0$ (Valid prediction)

It aggregates all valid pixels from all files (where $0.1\text{m} < D_{gt} < 80\text{m}$) into two large vectors and computes a set of standard depth estimation metrics. [Eig+14]

The following standard metrics were computed on this valid pixel set:

Error Metrics:

- **AbsRel (Abs. Rel. Diff.):** $\frac{1}{|V|}\sum_{i\in V}\frac{|d_i-\hat{d}_i|}{\hat{d}_i}$
- **SqRel (Sq. Rel. Diff.):** $\frac{1}{|V|}\sum_{i\in V}\frac{||d_i-\hat{d}_i||^2}{\hat{d}_i}$
- **RMSE (Root Mean Sq. Error):** $\sqrt{\frac{1}{|V|}\sum_{i\in V}(d_i-\hat{d}_i)^2}$
- **RMSElog (RMSE log):** $\sqrt{\frac{1}{|V|}\sum_{i\in V}(\log(d_i)-\log(\hat{d}_i))^2}$
- **log10 (log10 Error):** $\frac{1}{|V|}\sum_{i\in V}|\log_{10}(d_i)-\log_{10}(\hat{d}_i)|$
- **SILog (Scale-Inv. log Error):** $\frac{1}{|V|}\sum_{i\in V}(\log(\hat{d}_i)-\log(d_i)+\alpha(d,\hat{d}))^2$

Accuracy Metrics:

- $\delta < 1.25$ **(d1):** % of pixels i where $\max(\frac{d_i}{\hat{d}_i},\frac{\hat{d}_i}{d_i}) < 1.25$
- $\delta < 1.25^2$ **(d2):** % of pixels i where $\max(\frac{d_i}{\hat{d}_i},\frac{\hat{d}_i}{d_i}) < 1.25^2$
- $\delta < 1.25^3$ **(d3):** % of pixels i where $\max(\frac{d_i}{\hat{d}_i},\frac{\hat{d}_i}{d_i}) < 1.25^3$

(Where V is the set of valid pixels, $\hat{d}$ is the ground truth depth, and d is the predicted depth.)

Chapter 4

Results

4.1 Results and Analysis

All models were successfully run on the Cityscapes `val` split according to the methodology described. The quantitative results, aggregated from 500 validation images, are presented in Table 4.1.

Table 4.1: Quantitative Comparison of Zero-Shot Monocular Depth Estimation on the Cityscapes `val` Split. Best-performing model in each column is in **bold**.

	Error Metrics (Lower is Better)						**Accuracy Metrics** (Higher is Better)		
Model	**AbsRel** ↓	**SqRel** ↓	**RMSE** ↓	**RMSElog** ↓	**log10** ↓	**SILog** ↓	**d1** ↑	**d2** ↑	**d3** ↑
Group 1: DPT/MiDaS (General-Purpose Transformers)									
dpt-beit-large-512	0.5891	6.4567	11.7136	1.2336	0.3189	150.7946	0.2189	0.4537	0.7441
dpt-beit-base-384	0.5454	5.7394	11.5191	1.3586	0.3419	177.4878	0.2331	0.4890	0.7753
dpt-swinv2-large-384	0.5397	5.4948	11.5466	1.3160	0.3315	167.6102	0.2287	0.4777	0.7817
dpt-swinv2-tiny-256	0.5464	5.7749	11.6785	1.2669	0.3206	156.6102	0.2311	0.4789	0.7805
dpt-large	0.6012	6.8429	11.9592	1.3420	0.3448	176.3917	0.2224	0.4649	0.7145
dpt-hybrid-midas	0.5672	6.1504	11.6980	1.3680	0.3465	180.8973	0.2309	0.4832	0.7422
Group 2: Domain-Specialized Models									
ZoeDepth (Metric)	0.3632	5.6101	15.1052	0.8020	0.2324	52.6385	0.4814	0.6705	0.7301
Monodepth2 (CNN)	**0.2142**	**3.6892**	**11.5411**	**0.3226**	**0.0935**	**10.2691**	**0.6724**	**0.8684**	**0.9483**

4.1.1 Analysis of Results

The quantitative results from Table 4.1 reveal a stark and conclusive finding: a model's performance in this zero-shot scenario is overwhelmingly dictated by its training data, not its architecture or size. The models naturally segregated into two distinct performance tiers.

Group 1: The Domain Gap (DPT/MiDaS)

The entire suite of DPT/MiDaS models, despite being newer and based on powerful transformer backbones (BEiT, SwinV2), performed exceptionally poorly.

- **Extremely Low Accuracy:** The primary accuracy metric, `d1`, hovered between 21.8% and 23.3% for all models in this group. This means that, even after a global alignment, over 75% of the predicted pixels were incorrect by a margin of 25% or more.
- **High Relative Error:** The Absolute Relative Error (`AbsRel`) was consistently high, ranging from 0.53 to 0.60. This confirms that the geometric structure of the predictions was poor.
- **Irrelevance of Scale:** Within this group, there was no clear winner. The largest model (`dpt-beit-large-512`) performed slightly worse than the smaller `dpt-beit-base-384`. The smallest model, `dpt-swinv2-tiny-256`, performed just as well as its larger counterparts.

This poor performance is a classic symptom of a large domain gap. These models are trained on a wide *mixture* of datasets, including indoor scenes. Their resulting features are too general and do not transfer effectively to the specific domain of outdoor urban driving scenarios present in Cityscapes.

Group 2: The Power of Specialization (Monodepth2 & ZoeDepth)

Monodepth2: The convolutional `Monodepth2` model was the undisputed winner of this experiment, outperforming all other models in 8 out of 9 metrics.

- **Superior Accuracy:** With a `d1` score of **67.24%**, it was nearly 3 times more accurate than the best DPT model. Its `d2` (86.8%) and `d3` (94.8%) scores show that the vast majority of its predictions were structurally sound.
- **Low Relative Error:** Its `AbsRel` (0.2142) was less than half that of the DPT group. Its `log10` (0.0935) and `SILog` (10.2691) were an order of magnitude better, further reinforcing its high-quality geometric output.
- **The KITTI Advantage:** This superior performance is almost certainly due to its training data. `Monodepth2` was trained on the KITTI dataset, which, like Cityscapes,

is an outdoor driving dataset. This small domain gap allowed its learned features to transfer almost seamlessly.

ZoeDepth: The `ZoeDepth` model, which was tested on its direct metric output, was the middleperformer.

- It was significantly better than the DPT group, with a `d1` accuracy of 48.14% (over 2x better).
- It was, however, significantly worse than `Monodepth2` (48.14% vs 67.24% `d1`). This is logical, as `ZoeDepth` was trained on a mix of KITTI (driving) and NYUv2 (indoor), which likely diluted its specialization compared to the pure KITTI training of `Monodepth2`.
- **RMSE Anomaly:** The one metric where `Monodepth2` did not win was `RMSE`. Its score (11.5411) was comparable to the DPT group. `ZoeDepth` had the worst `RMSE` (15.1052). This indicates that all models, even the best ones, struggle with making large absolute errors on distant objects (like buildings or sky), and `RMSE` heavily penalizes these outliers.

Chapter 5

Discussion and Conclusion

5.1 Conclusion

This research successfully designed and executed a rigorous pipeline to evaluate and compare the zero-shot generalization of several state-of-the-art monocular depth estimation models on the challenging Cityscapes dataset. By implementing a standardized framework—including meticulous metric ground-truth calculation and a fair, consistent alignment protocol—we were able to generate a clear and definitive set of results.

The findings are conclusive: **a model's zero-shot performance is overwhelmingly dictated by the similarity of its training data to the target domain, not by its architectural complexity or scale.**

The entire family of DPT/MiDaS transformer models, trained on broad, general-purpose dataset mixtures (e.g., MiDaS-dpt-beit-large-512), failed to generalize to the specific urban driving domain of Cityscapes. Their poor performance, characterized by an Absolute Relative Error (`AbsRel`) of over 53% and a `d1` accuracy of only ∼23%, renders them unreliable for this task in a zero-shot setting. Their predicted depth maps are not geometrically sound, and the choice between a large BEiT backbone or a tiny SwinV2 transformer made no meaningful difference, as all were equally affected by the fundamental domain gap.

Conversely, the older, convolutional `Monodepth2` model, which was trained on the KITTI driving dataset, was the unambiguous winner. Its `d1` accuracy of 67.2%—nearly three times higher than the DPT models—and its superior relative error metrics demonstrate that it produces a structurally coherent and largely correct representation of the scene. This is a direct consequence of its specialized training on a similar domain.

The `ZoeDepth` model provided a perfect middle groundresult, reinforcing this conclusion. Trained on a mix of KITTI (driving) and NYUv2 (indoor) data, its performance was diluted compared to the specialist `Monodepth2` but still vastly superior to the generalist DPT models.

For practical application, this experiment provides a critical insight: one cannot simply select the largest, newest model from a leaderboard and expect it to perform. For a safety-critical task like autonomous driving or in our case assisting BVIPs in obstacle avoidance in the streets, a model (even an older one) trained on relevant, in-domain data is the superior and more reliable choice. This work proves that without in-domain fine-tuning, even the most powerful general-purpose models are unsuitable for specialized, real-world deployment.

Building on this conclusion, several avenues for future work become clear. The most immediate step is to quantify the impact of **targeted fine-tuning**: by training the high-capacity DPT/MiDaS models on the Cityscapes `train` split, we could determine if their powerful transformer architectures can overcome the initial domain gap to ultimately surpass the specialized `Monodepth2`. Furthermore, this framework could be used to explore **unsupervised domain adaptation (UDA)** techniques, which aim to bridge the domain gap without requiring ground-truth labels from the target dataset. Finally, this experiment provides a robust baseline for evaluating newer, emerging architectures that are explicitly trained for the autonomous driving domain from the outset, as well as for cross-dataset robustness checks, such as evaluating these models on adverse weather or night-time driving datasets.

List of Abbreviations

RDE Relative Depth Estimation

MDE Metric Depth Estimation

SOTA state-of-the-art

BVIPs Blind or visually impaired persons

List of Figures

List of Tables

Bibliography

[Bha+23] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller. Zoedepth: zero-shot transfer by combining relative and metric depth, 2023. arXiv: 2302.12288 [cs.CV]. URL: https://arxiv.org/abs/2302.12288 (cited on pp. 6, 10).

[Bir+23] R. Birkl, D. Wofk, and M. Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation, 2023. arXiv: 2307.14460 [cs.CV]. URL: https://arxiv.org/abs/2307.14460 (cited on p. 10).

[Cor+16] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. DOI: 10.1109/CVPR.2016.350. URL: https://arxiv.org/abs/1604.01685 (cited on p. 7).

[Eig+14] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. *CoRR*, abs/1406.2283, 2014. arXiv: 1406.2283. URL: http://arxiv.org/abs/1406.2283 (cited on p. 11).

[God+19] C. Godard, O. Mac Aodha, M. Firman, and G. Brostow. Digging into self-supervised monocular depth estimation. *arXiv.org*, August 2019. URL: https://arxiv.org/abs/1806.01260 (cited on p. 3).

[Las+19] K. Lasinger, R. Ranftl, K. Schindler, and V. Koltun. Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer. *CoRR*, abs/1907.01341, 2019. arXiv: 1907.01341. URL: http://arxiv.org/abs/1907.01341 (cited on pp. 4, 10).

[Ran+21] R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. *CoRR*, abs/2103.13413, 2021. arXiv: 2103.13413. URL: https://arxiv.org/abs/2103.13413 (cited on pp. 5, 10).