EE-584 Term Project
Vision Transformers for Dense Prediction
- Monocular Depth Estimation -

Okyanus Oral 2305134

September 23, 2023

Contents

1 Introduction

In the current literature, dense prediction tasks are predominantly tackled with convolutional neural networks (CNNs) within an encoder-decoder architecture. Although, convolutional architectures are dominant in vision, the downsampling operations to increase their receptive fields impose significant drawbacks for dense prediction [1]. Since, information loss on encoding stages cannot be recovered, low resolution features on deeper levels decrease the granularity of the predictions.

Recently proposed Vision Transformer [2] effectively maintains constant dimensionalality on its every level [1]. Accordingly the authors of [1] propose a dense prediction transformer (DPT) which utilizes ViT as its encoder architecture.

In this project I investigate the performance of DPT architecture for monocular depth estimation on DIODE dataset [3], which the architecture was not tested on before. Accordingly, this report provides a critical summary of DPT architecture and demonstrates the test results on DIODE dataset along with their analyses and discussions.

2 DPT - Development

DPT architecture utilizes ViT as its backbone. The idea behind using ViT as the encoder architecture is to benefit from global receptive field and the constant dimensionality of all processing stages. These allow the decoder to utilize high resolution features from multiple levels on dense reconstructions.

Note: The following explanations omit low level details since the background of the transformers are beyond the scope of the lectures.

2.1 Vision Transformer (ViT) Architecture

ViT introduced in [2] is inspired by the success of scaling of Transformers in natural language processing. In order to benefit vision tasks from a similar scaling the authors of [2] experiment with direct application of transformers to images. ViT architecture for image classification can be seen from Fig. 1.

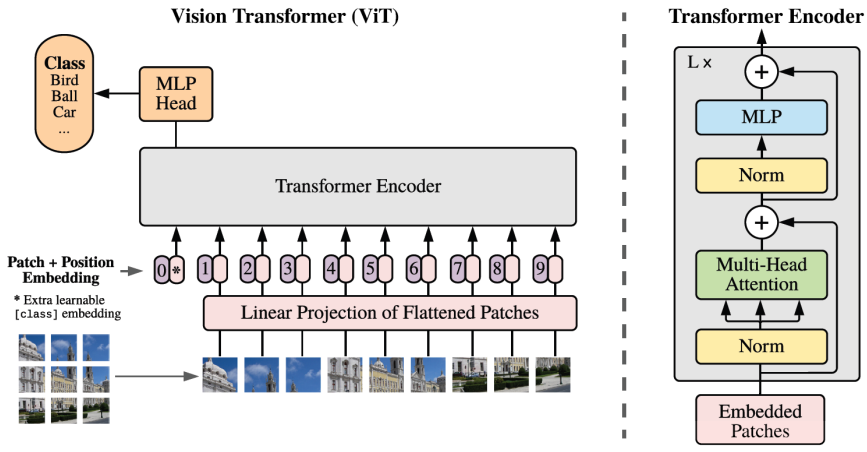

Figure 1: ViT Architecture

The ViT architecture presented in Fig. 1 can be summarized as follows. First images are split into patches of non-overlaping segments [2]. Then these 2D image patches are flattened and fed to a trainable linear projection [2]. Outputs of the projection operation are called patch embeddings [2]. In order to retain positional information of patches, trainable positional embeddings are then added to patch embeddings [2]. Here also a classification token is added to the sequence. The resulting sequence is then fed to Transformer encoder where the Transformer consists of alternating layers of multiheaded self attention and MLP blocks. [2].

Self attention can be though as learning an attention weighting in order to focus on important parts of the data.

To further elaborate on the intention behind each internal structure, representations of the data within ViT architecture are demonstrated with Fig. 2. From Fig. 2 it is evident that the linear projection of patches project images to a lower dimensional subspace. This projection operation therefore creates

a bag-of-words representation for the images [1, 4, 5]. A version of ViT utilizes a more sample-efficient architecture, ResNet50 [6], for the trainable projection operator.

Notice from Fig. 2 that learned position embeddings carry the position information of image patches to the Transformer. Moreover, attention distance is plotted with respect to network layers. Attention distance, which is computed as the average distance in the image space weighted by the attention weights, is analogous to receptive fields of CNNs [2]. It calculates the average distance in the image space which the information is integrated together [2]. As it can be seen from Fig. 2 some heads attend to most of the image even in the early layers. Therefore, ViT structure can be considered to have a global receptive field at its all layers [1, 2].

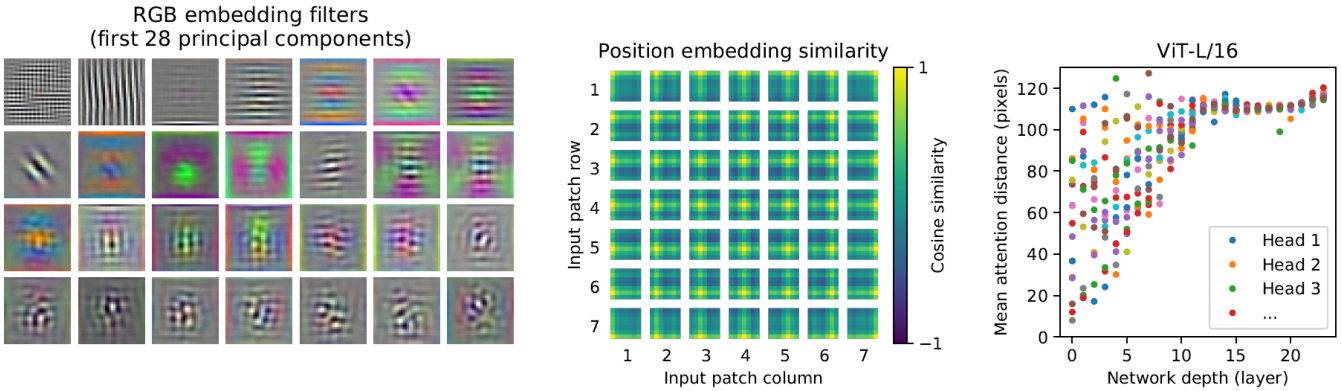

Figure 2: Internal Representations of Data. **Left:** Filters of the initial linear embedding of RGB values. **Center:** Similarity of patch embeddings to position embeddings. **Right:** Average distance in the image space which the information is integrated.

2.2 Dense Prediction Transformer (DPT) Architecture

The following is the summary of the DPT architecture, block diagram of the architecture can be seen in Fig. 3.

Encoder: As explained before, DPT uses ViT as its backbone encoder architecture. Critically, the ViT is utilized to obtain high resolution features from the tokens of Transformer layers [1]. Since Transformers maintain the same number of tokens, and the trained projection directly map image patches to their bag-of-word representations, ViT can be said to maintain the spatial resolution of image patches through all of its iterations[1].

Decoder: A convolutional decoder assembles the set of tokens from different transformer layers to image-like feature representations[1]. These image-like feature representations are fused together to obtain dense predictions[1]. The details of the reasambling operations for the fusing of features can be found in [1].

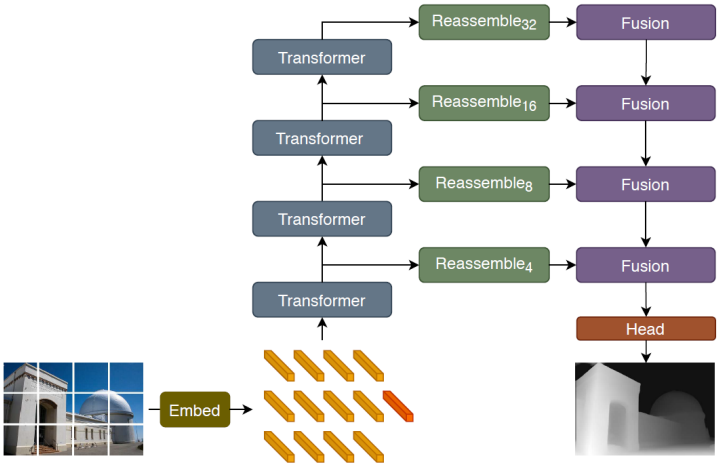

Figure 3: DPT Architecture

3 Prediction Alignment and Zero-shot Cross-dataset Transfer

Data for depth estimation are abundant, hence learning monocular depth estimation as a dense prediction task is well suited for DPT framework [1]. However, the abundance of data cannot be trivially benefited from due to the differences in forms (upto a scale depth, depth information as disparity maps, global disparity shifts etc.)[7]. In order to benefit from large corpus of data, the data should be converted to the same depth space. Accordingly, two allignment methods are proposed in [7]. These alignments are performed in disparity (inverse depth) space.

The first approach calculates the required translation and scaling of the predictions to match the ground truths using least squares criterion, see Eqn. 1.

$$(s,t) = \arg\min_{s,t} \sum_{i=1}^{M} (s\mathbf{d_i} + t - \mathbf{d_i^*})^2 \tag{1}$$

In Eqn. 1, s is the scaling and t is the translation coefficients, d^* corresponds to ground truth disparity and subscript i denotes image pixel indices, M denotes number of valid pixels. Then, aligned predictions in disparity space become,

$$\hat{\mathbf{d}} = s\mathbf{d} + t \tag{2}$$

where $\mathbf{d}$ and $\hat{\mathbf{d}}$ represent the initial and aligned predictions.

Noting that least squares criterion is susceptible to outliers, the second approach tries to increase the robustness of the alignments, see Eqn. 3. Accordingly, the aligned predictions, $\hat{\mathbf{d}}$ and the aligned ground truths, $\hat{\mathbf{d}}^*$, are formulated as in Eqn. 4

$$t(\mathbf{d}) = median(d), \qquad s(\mathbf{d}) = \frac{1}{M}\sum_{i=1}^{M} |d - t(d)| \tag{3}$$

$$\hat{\mathbf{d}} = \frac{\mathbf{d} - t(\mathbf{d})}{s(\mathbf{d})}, \qquad \hat{\mathbf{d}}^* = \frac{\mathbf{d} - t(\mathbf{d})}{s(\mathbf{d}^*)} \tag{4}$$

3.1 Monocular Depth Estimation Metrics with Zero-shot Data Transfer

Common monocular depth estimation metrics [8] can be utilized with these allignment methods. Some common metrics are formulated in Eqns. 5, 6, and 7 [8].

$$\text{Absolute Relative Error (AbsRel) between } x \text{ and } x^* \ : \frac{1}{M}\sum_{i=1}^{M} \frac{|x_i - {x^*}_i|}{|{x^*}_i|} \tag{5}$$

$$\text{Root Mean Squared Error (RMSE) between } x \text{ and } x^* \ : \sqrt{\frac{1}{M}\sum_{i=1}^{M}(x_i - {x^*}_i)^2} \tag{6}$$

$$\text{Inaccuracy over Threshold } (\delta > 1.25^k) \text{ between } x \text{ and } x^* \ : \ 100\frac{\#\text{ of pixels } \ (max\{\frac{x_i}{{x^*}_i}, \frac{{x^*}_i}{x_i}\} > 1.25^k)}{M} \tag{7}$$

Where x and x^* correspond to aligned prediction and ground truth images, subscript i denotes the i^{th} pixel and M is the number of pixels. x and x^* can be in distance or disparity space. For all these metrics lower values represent higher quality of predictions. AbsRel is a relative measure of infidelity from the ground truths. RMSE computes the mean squared deviation between predictions and ground truths. The inaccuracy over threshold can be considered as the percentage of predictions where the inaccuracy of pixel estimates (ratio of estimates and truths) exceed the set threshold 1.25^k.

4 DPT - Original Experiments

This section provides a brief summary of the experiments presented in the original paper.

4.1 DPT Models from ViT Variants

Three different variants of ViT architecture are utilized within DPT models[1]. These are ViT-Base, with 12 transformer layers; ViT-Large, with 24 transformer layers; and ViT-Hybrid, which uses ResNet50 as the projection operator followed by 12 transformer layers. The resulting DPT architectures will be called DPT-Base, DPT-Large, and DPTHybrid respectively. For all architectures patch size is selected as 16. Furthermore, although number of transformer layers are different for each backbone ViT, the number of layers collected to form the latent features is fixed to 4. The selection details of the respective 4 layers can be found in [1] for each architecture.

4.2 Dataset

In order to properly train DPT, a new meta dataset (the largest training set for monocular depth estimation upto that point) was gathered [1]. This new dataset contains about 1.4 million images and is named MIX6 [1]. It is built upon the MIX5 dataset [7], by expanding it with the datasets presented in [9, 10, 11, 12, 13]. For the exact contents of MIX5, the reader can refer to [7].

Furthermore, to demonstrate the fine tuning of DPT architectures on smaller datasets, KITTI[14] and NYUv2[15] datasets are utilized.

4.3 Training & Testing

As MIX6 is a diverse dataset, it contains data at different forms (up to a scale depth, depth information as disparity maps, global disparity shifts). For a sensible learning/testing scheme, these representations should be converted to the same depth space. Accordingly, the alignment methods mentioned in Section 3 are utilized. The training and testing of the architectures directly follows the procedures described in [7]. The performance metrics for monocular depth estimation are also selected to correspond to that of 3.

	Training set	DIW WHDR	ETH3D AbsRel	Sintel AbsRel	KITTI δ>1.25	NYU δ>1.25	TUM δ>1.25
DPT - Large	MIX 6	**10.82** (-13.2%)	**0.089** (-31.2%)	**0.270** (-17.5%)	**8.46** (-64.6%)	**8.32** (-12.9%)	**9.97** (-30.3%)
DPT - Hybrid	MIX 6	11.06 (-11.2%)	0.093 (-27.6%)	0.274 (-16.2%)	11.56 (-51.6%)	8.69 (-9.0%)	10.89 (-23.2%)
MiDaS	MIX 6	12.95 (+3.9%)	0.116 (-10.5%)	0.329 (+0.5%)	16.08 (-32.7%)	8.71 (-8.8%)	12.51 (-12.5%)
MiDaS	MIX 5	12.46	0.129	0.327	23.90	9.55	14.29

Table 1: Subset of the Original Test Results. Remember that smaller values of metrics indicate higher performance. For convenience the percentage of relative increase of values metrics relative performances are also provided as percentage decrease in with respect to the original MiDaS model.

The tests results are presented with respect to the MiDaS architecture [7]. MiDaS is a fully convolutional monocular depth estimation model that was the previous state of the art [1]. Furthermore, in order to make fair comparisons and discuss the architecture performances invariant of the training dataset sizes, a MiDaS model is trained on MIX6 dataset. As it can be seen from Tab. 1 DPT architectures exceed the performance of MiDaS on all test datasets. Moreover, increasing the dataset size does not seem to increase the performance of MiDaS architecture as can be seen from the tests of DIW and Sintel datasets. Or the increase in model performance is not as substantial as DPT architectures.

Sample results for qualitative assessment of depth estimation performances are given in Fig.4. As it can be seen from the figure, disparity estimates have higher resolution with DPT architectures. Moreover DPT architectures more accurately capture the sharp transitions between objects and their backgrounds.

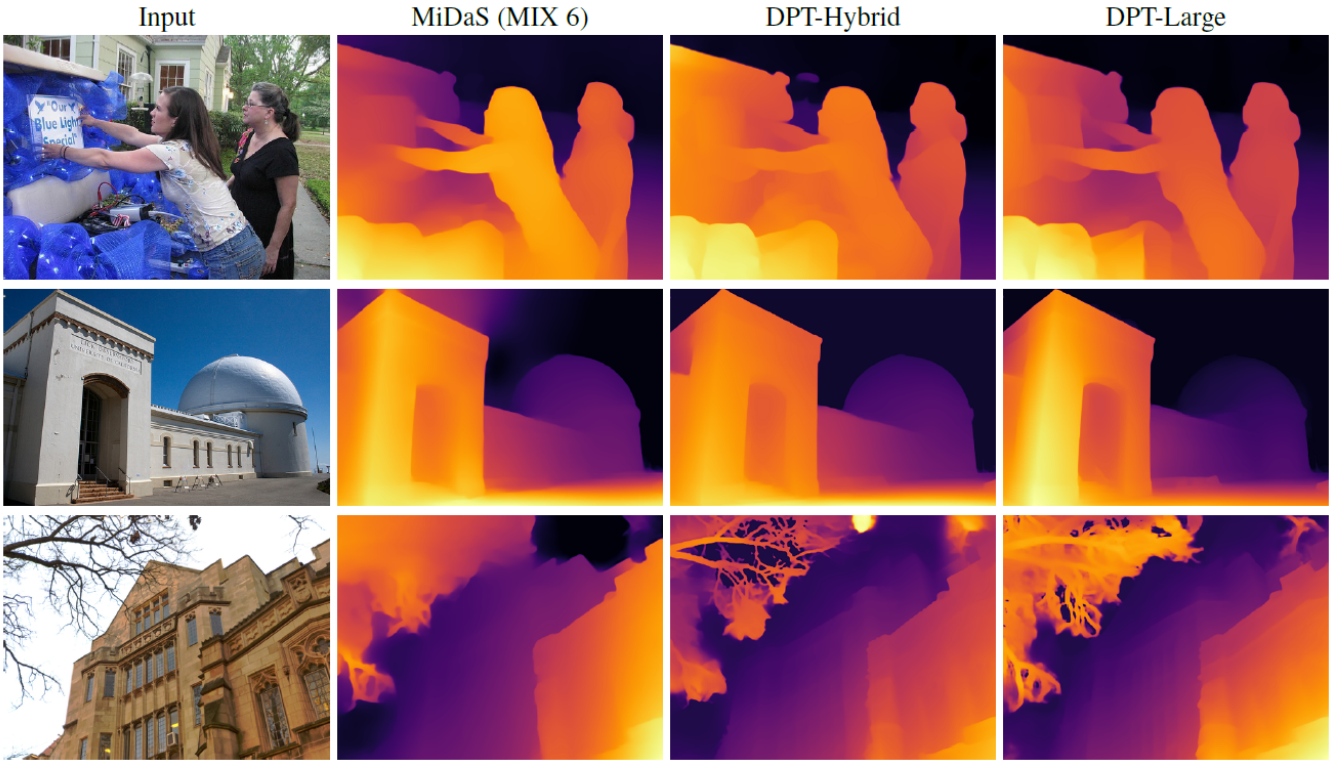

Figure 4: Sample results for monocular depth estimation. Compared to the fully-convolutional network used by MiDaS, DPT shows better global coherence (e.g., sky, second row) and finer-grained details (e.g., tree branches, last row) [1]

5 DPT - My Experiments

This section summarizes the test results of DPT and MiDaS architectures for the *validation set of DIODE dataset* [3] which was not utilized in any of the testing or training of the original work [1]s.

All codes relevant to the conducted tests (except for the trained models and the codes to download them) are written by me. The Dropbox link to access the codes can be found in Appendix.

5.1 DPT Models

In order to expand the discussions on the performances of DPT and MiDaS architectures, tests are conducted for DPT Large, DPT Hybrid and MiDaS (MIX5) models respectively [1, 7]. These models can be found in the following repository `https://github.com/isl-org/MiDaS`. Nevertheless, relevant functions are provided within my codes to automatically download the respective architectures.

5.2 Dataset

DIODE is an RGBD dataset composed of indoor and outdoor scenes with dense and accurate depth maps and validity masks [3]. Its depth readings are taken with a LIDAR-based scanner and the statistics of the measurements are shown in Tab. 2 [3].

	DIODE	NYUv2	KITTI	MAKE3d
Return Density (Empirical)	99.6%/66.9%	68%	16%	0.38%
# Images Indoor/Outdoor	8574/16884	1449/0	0/94000	0/534
Sensor Depth Precision	±1 mm	±1 cm	±2 cm	±3.5 cm
Sensor Angular Resolution	0.009°	0.09°	0.08°H, 0.4°V	0.25°
Sensor Max Range	350 m	5 m	120 m	80 m
Sensor Min Range	0.6 m	0.5 m	0.9 m	1 m

Table 2: Statistics of DIODE compared to other popular RGBD datasets [3]

The images in DIODE dataset have a resolution of 768×1024 pixels, and the dataset can be downloaded

from `https://diode-dataset.org`. Nevertheless, relevant functions are provided within my codes to automatically download its validation set.

DIODE validation set consists of 10 indoor scenes (325 samples) and 8 outdoor scenes (368 samples) of image/depth map/validity mask pairs. The scenes include offices, parks, residential buildings, lecture halls, and many others. Sample examples of RGBD images and the corresponding validity masks are shown in Fig.5.

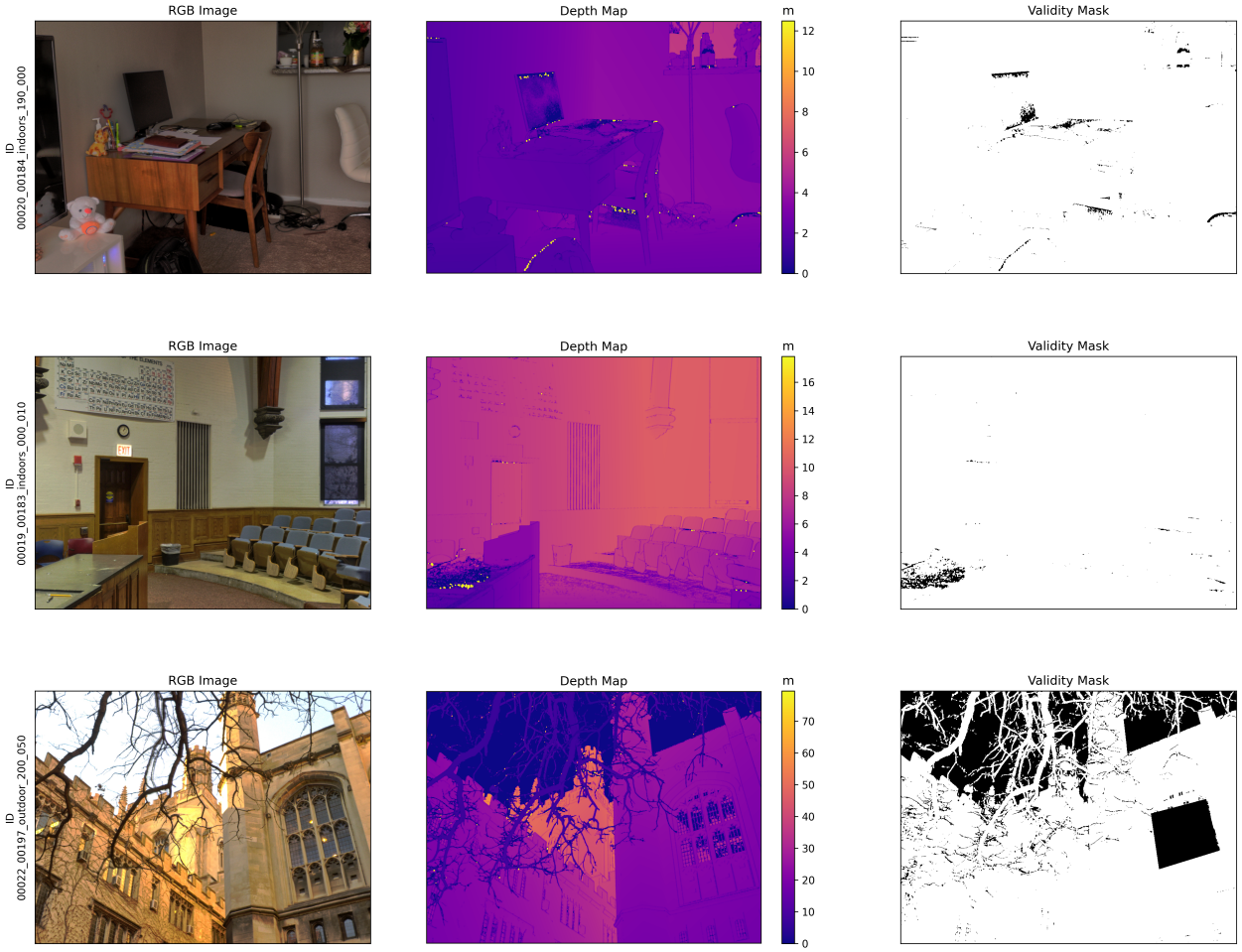

Figure 5: Sample Images, Depth Maps and Validity Masks from DIODE validation dataset. Left: RGB image. Middle: Dense Ground Truth Depth Map. Right: Validity Mask (depth measurements are valid for white, and invalid for black)

5.3 Testing

For appropriate quantitative assessment of architecture performances, I first aligned predictions and the ground truth depth measurements. For this I wrote the necessary codes according to the prediction formulations in Section 3.

Although, the selection of t and s with Eqn.4 is introduced as the robust implementation, when the respective paper [7] is inspected, one can notice that the authors also utilize least squared based selection of t and s parameters. Accordingly, when I utilized the 'robust' alignment method, the calculation of zero-shot cross-dataset transfer performances deviated significantly from the findings in the original paper [1]. Therefore I continued the assessment of performances with least squares based alignment method.

Performances of the DPT and MiDaS models are inspected with respect to AbsRel in depth space, RMSE in disparity space, ($\delta > 1.25$), ($\delta > 1.25^2$), and ($\delta > 1.25^3$) as is used in the original implementation

[1]. Moreover, since we are using least square based alignment, the predictions are matched to the ground truth depths. In that respect I found it interesting to include the RMSE metric in depth space. With the inclusion of this metric, the performances can be related to a physical measure.

Tab. 3 and Tab. 4 show the test results for the dataset. It is clear from Tab.3 that for indoors DPT models outperform MiDaS for each metric. When the restults for outdoors are inspected from Tab. 4 it can be deduced that MiDaS outperforms DPT models on AbsRel and depth RMSE.

indoors (325)	AbsRel	disparity RMSE	depth RMSE	$\delta > 1.25$	$\delta > 1.25^2$	$\delta > 1.25^3$
DPT Hybrid	0.1364 / 0.5663	0.0767 / 0.0624	**38.1306 / 417.1728**	9.7448 / 13.7699	2.9466 / 5.1572	0.9311 / 1.9497
DPT Large	**0.1476 / 0.58**	**0.0753 / 0.0613**	79.139 / 833.2194	**8.5391 / 10.4963**	**2.902 / 4.6556**	**0.9116 / 2.083**
MiDaS	0.1221 / 0.1636	0.0857 / 0.0783	11.4394 / 151.975	12.6531 / 16.7227	4.0461 / 6.7421	1.4172 / 2.7032

Table 3: DIODE Validation Set, computed performances, mean / std, for indoor scenes. For all metrics lower is better. Best performing model's mean / std are highlighted in bold.

outdoor (368)	AbsRel	disparity RMSE	depth RMSE	$\delta > 1.25$	$\delta > 1.25^2$	$\delta > 1.25^3$
DPT Hybrid	0.5088 / 1.0497	0.0906 / 0.0726	732.2222 / 4792.9042	43.9444 / 28.1566	**23.0857 / 22.081**	**13.2532 / 16.9308**
DPT Large	0.6337 / 2.5679	**0.0902 / 0.0721**	2545.2907 / 26005.0125	**43.4039 / 28.4907**	23.1867 / 22.4874	13.3975 / 17.08
MiDaS	**0.4718 / 0.6581**	0.0926 / 0.073	**438.418 / 4490.311**	46.8006 / 28.1977	26.6877 / 23.6239	15.8588 / 18.6376

Table 4: DIODE Validation Set, computed performances, mean / std, for outdoor scenes. For all metrics lower is better. Best performing model's mean / std are highlighted in bold.

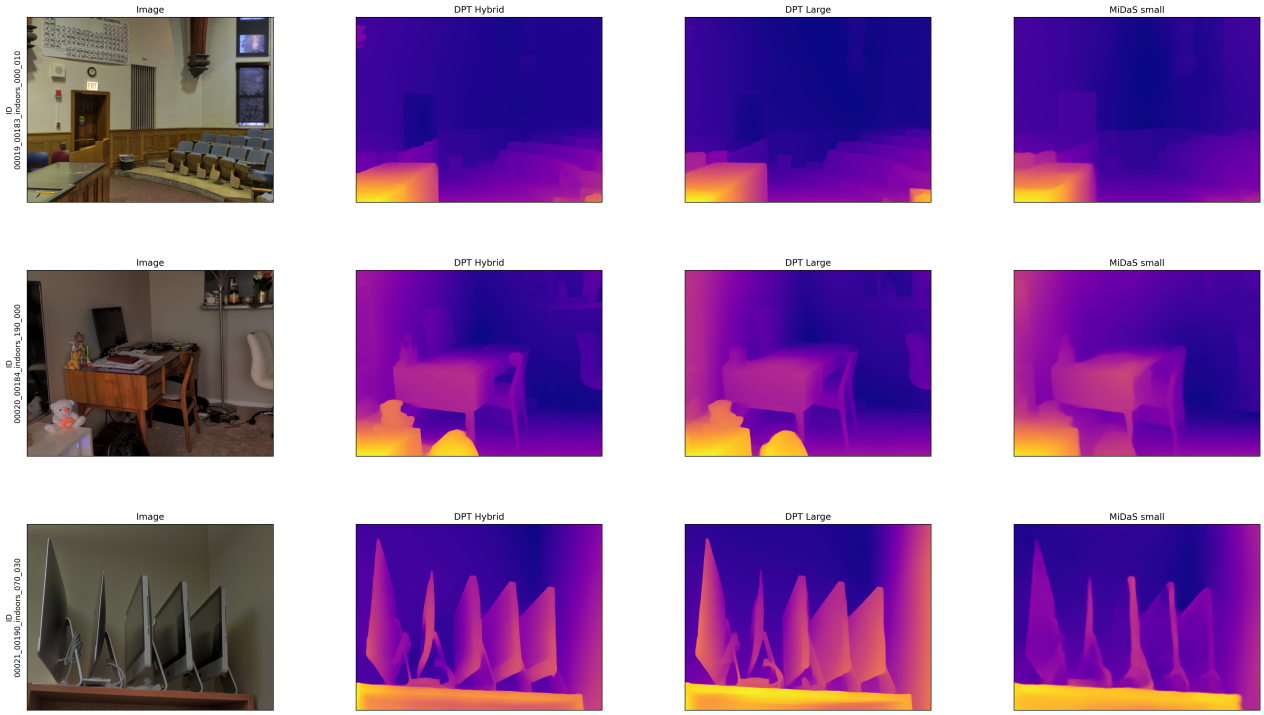

Figure 6: Visual Inspections - Indoors: From left to right, RGB image, DPT Hybrid disparity estimate, DPT Large disparity estimate, MiDaS disparity estimate.

5.4 Analyses

The test results on Tab. 3 are in unison, indicating the superior performance of DPT. In order to understand the factors that led to these results some indoor scenes and disparity map estimates are inspected

with a judicious manner.

In Fig. 6, 3 RGB images and the respective disparity estimates are shown. It is evident that DPT architectures can handle high frequency details better while MiDaS outputs smoothed disparity maps. The reason behind this is already explained in the sections related to the theoretical background.

It can also be noted from the second-row-results that DPT Large can better resolve finer details at longer distances compared to DPT Hybrid. On the other hand DPT Hybrid outputs has finer level of detail for closer objects than DPT Large. The reason for this may be addressed to utilization of ResNet50 (a CNN architecture) as the projection operator of ViT-Hybrid. Since the CNN architectures do not have global receptive fields, the effect of using CNN on bag of words representations may be decreasing the performance for far away objects. Note that the objects at the far away have smaller supports and effectively contain higher frequencies.

For the test results on Tab. 4, MiDaS performs better for AbsRel and depth RMSE. For both of these metrics note that the standard deviations of the results are higher than the means. Therefore, one should be careful of the use of 'mean $\pm$ std.' notation for a non-zero metric such as AbsRel. When the respective test results are inspected, it was seen that when standard deviations are more than the means, there are outliers in the data. Hence I computed the median of the respective metrics for the outdoor dataset. The medians of performances can be seen from Tab. 5

outdoor (368)	AbsRel	disparity RMSE	depth RMSE	$\delta > 1.25$	$\delta > 1.25^2$	$\delta > 1.25^3$
DPT Hybrid	**0.3106**	**0.0664**	**7.3037**	**45.2067**	15.6585	5.9096
DPT Large	0.3114	0.0676	7.4228	44.6895	**15.1265**	**5.8038**
MiDaS	0.3386	0.0686	8.0838	46.956	20.3664	8.4192

Table 5: DIODE Validation Set, computed performances, median, for outdoor scenes. For all metrics lower is better. Best performing model's median is highlighted in bold.

It is clear that when the effect of outlier data is eliminated, the DPT models outperform the MiDaS model. Furthermore, DPT-Hybrid seems to dominate DPT-Large on outdoor depth prediction.

Visual comparisons are also provided for the outdoor scenes in Fig. 7. From the images at the first row it is again observed that DPT Large can better resolve finer details at longer distances compared to DPT Hybrid. Notice that the railing at the backgound is not prominent with DPT Hybrid. Furthermore, MiDaS even fails to properly recover the railing at the foreground. DPT Hybrid outputs having a finer level of detail at the short range is also observed from the images at the second row, where DPT Large estimations are comparatively worse. Furthermore, for the image in second row, MiDaS fails (see the sky).

All the discussed effects can also be seen from third row images. Here the details of the depth information of the streetlamp is recovered best with DPT Hybrid while the fine details of the branches of the trees at the far back are captured with DPT Large. MiDaS on the other hand results in muddy estimations.

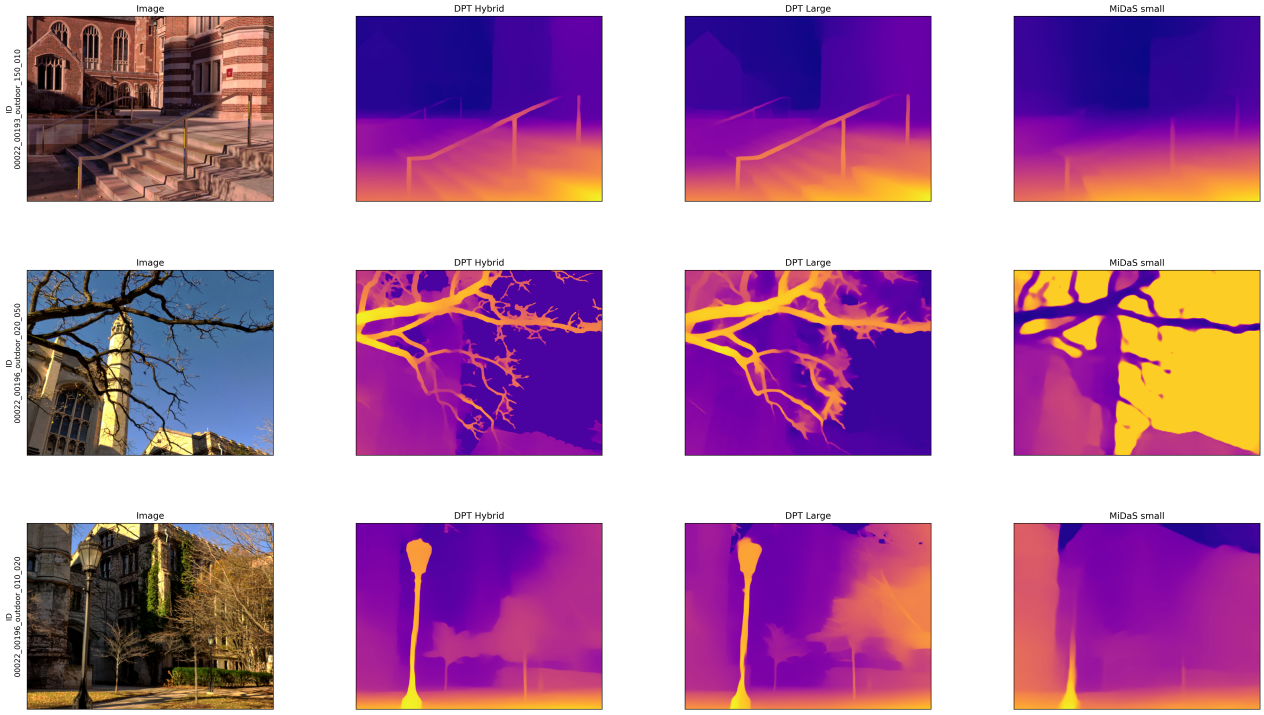

Figure 7: Visual Inspections - Outdoors: From left to right, RGB image, DPT Hybrid disparity estimate, DPT Large disparity estimate, MiDaS disparity estimate.

6 Conclusion

In this report, monocular depth estimation with vision transformers are explained. First the development stages of DPT architecture are presented, where the potential benefits of utilization of vision transformer (ViT) for dense prediction tasks are justified. Furthermore, the architectures of ViT and DPT are elaborated. Although some explanations lacked derivations, they tried to provide intuition.

Then the topic of prediction alignment and zero-shot data transfer is investigated. Here two prediction alignment methods are considered. Monocular depth estimation metrics are accordingly adapted to mixed dataset settings.

The training and testing settings mentioned in the original paper are presented to the reader. Here the original results are summarized.

On the last chapter experimentation with the respective architectures are elaborated. Here DIODE validation set is utilized in the assessment of model performances. The test results are presented and summarized for outdoor and indoor scenes. Here quantitative results are backed with qualitative discussions. Possible reasons for the performance differences between respective models are also discussed within analyses.

7 Appendix

- Dropbox link for my PowerPoint presentation: `https://www.dropbox.com/scl/fi/zgzc2h2qp0jkw6g7f34nc/OkyanusOral2305134TermProjectPresentation.pptx?dl=0&rlkey=d101dyw72he7md2osfjcz7dow`
- Dropbox link for all my codes: `https://www.dropbox.com/sh/9ezx9rdberqrtvl/AADZbQNXSyxSqMibQkmSAqp dl=0`

8 References

[1] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," *CoRR*, vol. abs/2103.13413, 2021. [Online]. Available: https://arxiv.org/abs/2103.13413

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[3] I. Vasiljevic, N. Kolkin, S. Zhang, R. Luo, H. Wang, F. Z. Dai, A. F. Daniele, M. Mostajabi, S. Basart, M. R. Walter, and G. Shakhnarovich, "Diode: A dense indoor and outdoor depth dataset," 2019. [Online]. Available: https://arxiv.org/abs/1908.00463

[4] Sivic and Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Proceedings Ninth IEEE International Conference on Computer Vision*, 2003, pp. 1470–1477 vol.2.

[5] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, no. 1-22. Prague, 2004, pp. 1–2.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[7] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," 2019. [Online]. Available: https://arxiv.org/abs/1907.01341

[8] C. Cadena, Y. Latif, and I. D. Reid, "Measuring the performance of single image depth estimation methods," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 4150–4157.

[9] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The ApolloScape open dataset for autonomous driving and its application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2702–2719, oct 2020. [Online]. Available: https://doi.org/10.1109%2Ftpami.2019.2926463

[10] Q. Wang, S. Zheng, Q. Yan, F. Deng, K. Zhao, and X. Chu, "Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation," 2019. [Online]. Available: https://arxiv.org/abs/1912.09678

[11] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, and S. Scherer, "Tartanair: A dataset to push the limits of visual slam," 2020. [Online]. Available: https://arxiv.org/abs/2003.14338

[12] K. Xian, J. Zhang, O. Wang, L. Mai, Z. Lin, and Z. Cao, "Structure-guided ranking loss for single image depth prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[13] Y. Yao, Z. Luo, S. Li, J. Zhang, Y. Ren, L. Zhou, T. Fang, and L. Quan, "Blendedmvs: A large-scale dataset for generalized multi-view stereo networks," 2019. [Online]. Available: https://arxiv.org/abs/1911.10127

[14] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[15] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 746–760.