Robust Depth Estimation in Adverse Visual Conditions

Jiamin Sun
jiamins@stanford.edu

Wenfu (Norman) Lei
norman94@stanford.edu

Abstract

In computer vision, adverse visual conditions are challenging for monocular depth estimation. Both traditional CNN methods and recent transformer-based methods are impacted by the degraded input image quality resulting from insufficient lighting and blurry environments. In this project, we aim to study different methods to improve robustness and generalization based on a Dense Prediction Transformers (DPT) model, including fine-tuning using scale and shift invariant loss, customized consistency loss, architectural adjustments and regularization. We present thorough evaluations and discussions using both standard depth estimation evaluation metrics and visual metrics.

1. Introduction

Depth estimation from 2D images is a fundamental task for research areas including AR/VR, autonomous driving, and medical diagnosis. Despite promising progress in recent years, it remains challenging because of the inherent complexity and uncertainty of real-world scenarios, particularly under visually adverse conditions. Historically, CNN-based models have been widely used to solve this task [4]. Eigen et al. first proposed two deep network stacks to solve single image depth estimation [1]. In addition to model architectures, they introduced Scale-Invariant Error as a measurement metrics. In recent years, transformer-based models are gradually superseding CNN-based models on various image tasks. Ranftl et al. introduced dense prediction transformers (DPT) that achieved a significant performance improvement on depth estimation benchmarks [7].

Adverse visual conditions, such as reduced lighting and blurred imaging, may negatively impact the performance of transformer-based models in depth estimations. First, transformer is heavily based on self-attention layers. Reduced image quality directly leads to loss of information in attention maps. Second, for pre-trained models such as DPT-hybrid, general-purpose datasets such as ImageNet are often used [7]. When pretrained models are fine-tuned for depth estimation tasks, public datasets focusing on daytime outdoor and indoor scenes such as KITTI Depth and NYU Depth are primarily used [6][10][8]. These common datasets are not specifically collected under adverse visual conditions.

In this project, we aim to investigate robust depth estimation techniques based on DPT methods. The input for this task is a 2D 3-channel RGB image from a real scene. It is either physically captured in an adverse visual condition or simulates dark and blurred scenarios. Since major public depth estimation datasets are primarily collected during daytime, we plan to simulate dark and blurry environments using a selected original public dataset. The output for the task is a 2D 1-channel gray depth map. We aim to examine model performance under both original environments and corresponding simulated adverse visual conditions, experiment with different fine-tuning strategies, and study evaluation metrics for these specific visual conditions.

2. Related work

In the previous section, we briefly introduced the historical development of monocular depth estimation models. In this section, we present a more detailed overview of related work and evaluation metrics.

2.1. CNN models

Convolutional neural network (CNN) is powerful for depth estimation as it captures local and global image patterns efficiently. In the two deep network stacks proposed by Eigen et al., the first one is designed for global prediction, while the second one is for refinement [1]. Residual network is an effective enhancement to CNN architecture. Laina et al. leveraged residual learning to achieve an end-to-end state-of-the-art depth estimation method [4].

2.2. Transformer-based models

Dense prediction transformers (DPT) introduced by Ranftl et al. are based on vision transformer (ViT)[7]. There are different variations including DPT-large and DPT-hybrid, depending on the feature extraction mechanism. DPT-large extracted "non-overlapping patches followed by a linear projection", while DPT-hybrid leveraged a residual network [7]. They both demonstrated great improvement on

common depth estimation benchmarks evaluated with unseen data.

2.3. Standard evaluation

To quantitatively evaluate monocular depth estimation models, standard metrics from prior works are widely adopted, including the Absolute Relative Error (AbsRel), Root Mean Square Error (RMSE), and the accuracy under threshold ratios δ_1, δ_2 and δ_3, which measure the percentage of predicted pixels within increasing multiplicative tolerances of the ground truth depth. These metrics were introduced by Eigen et al. [1] and have become standards across datasets such as NYUv2 and KITTI.

2.4. Visual evaluation

In addition to standard evaluation metrics, to assess how well a model preserves structural details visually, particularly object boundaries, Koch et al. introduced Depth Boundary Error (DBE) metrics, specifically, DBE Accuracy and DBE Completeness [3]. These metrics evaluate the edge preservation between predicted and ground-truth depth maps. They are useful metrics for analyzing model robustness under adverse visual conditions.

3. Methods

We adopt a two-phase methodology for enhancing depth prediction under adverse visual conditions, beginning with a strong transformer-based baseline and progressively incorporating domain-specific adaptations. Our approach is grounded in the MiDaS 3.0 DPT-Hybrid-384 model and is guided by the goal of improving the structural and boundary-aware robustness in depth maps derived from images under adverse visual conditions.

3.1. Problem formulation

Given a single RGB image $I \in \mathbb{R}^{H\times W\times 3}$, the goal of monocular depth estimation is to predict a corresponding depth map $D \in \mathbb{R}^{H\times W}$. In this project, we predict relative depth and evaluate it against ground truth D_{gt} using both quantitative and qualitative metrics.

3.2. Baseline model: MiDaS DPT-hybrid-384

The baseline method uses pretrained MiDaS 3.0 DPT-Hybrid-384 model [9], which was trained on a mixture of indoor and outdoor datasets to estimate relative depth. This model serves as zero-shot performance on the NYU Depth Dataset V2, without fine-tuning. To better evaluate its robustness under adverse visual conditions, we apply controlled data augmentations to the NYU Depth Dataset V2. We will discuss more dataset details in the next section. Then we plan to run inference on both the original and augmented images and compare the predicted depth maps against the ground truth using different metrics. The comparison would help identify the limitation of the baseline model on both the original and augmented dataset simulating adverse conditions.

3.3. Supervised fine-tuning with scale-invariant loss

To adapt the model to the domain, we plant to fine-tune the model on NYU Depth Dataset V2 with both original and augmented training data, using Scale- and shift-invariant loss defined by Ranftl et al. [9]:

$$L_{\text{ssi}}(D, D_{gt}) = \frac{1}{2M}\sum_{i=1}^{M}\rho\left(D_i - D_{gt_i}\right)$$

D_i and D_{gt_i} refer to predicted depth and ground truth for a single sample. M is the number of pixels. $\rho(\cdot)$ is a specific loss function such as MSE. In practice, the full loss often includes further optimizations such as additional regularization term.

3.4. Consistency regularization for domain robustness

To improve model's robustness to domain shifts, we implement a consistency-based regularization strategy. Rather than using contrastive learning in the feature space, we regularize the model by enforcing consistency between the output depth maps of original and augmented image pairs. Given an input RGB image I and an augmented image $\hat{I}$ from training sets, we pass both through the model to obtain depth predictions D and $\hat{D}$. However, since the MiDaS model predicts relative depth, we have to apply the scale and shift alignment first to both outputs using the ground truth dpeth map D_{gt} and a foreground mask M, following the same strategy used in Ranftl et al. [9].

With computed scale s_{ori} and s_{aug} and shift t_{ori} and t_{aug}, this yields aligned predictions:

$$D_{aligned} = s_{ori}D + t_{ori}, \hat{D}_{aligned} = s_{aug}\hat{D} + t_{aug}$$

We then compute an L1 consistency loss:

$$L_{consistency} = ||D_{aligned} - \hat{D}_{aligned}||_1$$

The total loss with consistency regularization becomes:

$$L_{total} = L_{ssi} + \alpha L_{consistency}$$

where α is a regularization coefficient. This would encourage the model to produce consistent depth estimations across condition shifts. This method is inspired by contrasive adaptation approaches such as NightDepth [2], but implemented in the output depth map space rather than the feature space, making it align with supervision loss and be more interpretable.

3.5. Architectural adjustments and regularization

Encoder layer freezing. During fine-tuning, encoder layer freezing is commonly used for preserving the early feature extraction mechanism and improving training stability, especially with smaller datasets. For DPT models, Ranftl et al. discussed different downstream tasks and how they benefit from the pretrained encoder [7].

Decoder regularization. Another strategy to prevent overfitting and improve generalization is enhanced regularization on the decoder. Ranftl et al. discussed about adding dropout layer in image segmentation task before the final classification layer [7].

We aim to experiment with full encoder freezing and partial freezing of earlier layers. In addition, we experiment with adding dropout layers after CNN layers within the decoder as a strategy to improve model robustness and generalization.

4. Dataset and features

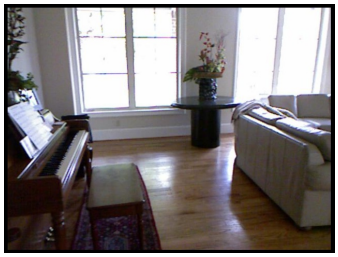
In this project, we leverage NYU Depth Dataset V2 with the official split for training and testing [6]. We downloaded the 2.8 GB labeled dataset from the official NYU Depth V2 Dataset website.

Data splitting and simulation. The original labeled dataset includes rgb input data and ground truth depth map data organized as one MATLAB file. We downloaded the official training and testing splits provided by Silberman el al. as another MATLAB file [6]. The RGB inputs were converted to JPEGs and the dpeth maps were converted to both gray PNGs and PFM formats for our evaluation compatibility. We organized the data into training and testing sets respectively by adapting the preparation code from Lee et al. [5]. There are 795 640x480 images in training set, 654 640x480 images in test set.

The photos inside labeled NYU Depth dataset are primarily daytime or well-lighted indoor scenes. To simulate the adverse visual conditions, we leveraged a vanilla image processing approach:

- Step 1: All RGB values are multiplied by 0.3.
- Step 2: B channel values are increased by 10.
- Step 3: A gaussian blur with std 10 is applied.

Figure 1 shows a sample input from NYU Depth V2 extracted from MATLAB and its augmented variation. The perturbation mutates important features such as texture, color and edges.



(a) Original input

(b) Augmented input

Figure 1: Example NYU Depth V2 living room input (a) and the corresponding augmented input (b)[6]

5. Experiments

We leveraged official MiDaS repository to conduct our experiments. This repo provides base inference script for us to modify and adapt. The baseline pretrained DPT-Hybrid-384 model weight was downloaded from MiDaS releases.

MiDaS repo does not release its official training code. We developed a training script to load `dpt_hybrid_384` model weight and prepare for finetuning. For Scale- and shift-invariant loss, we leveraged `ScaleAndShiftInvariantLoss` provided within MiDaS repo issues discussions, to essentially align the prediction using a learned linear transformation.

In our base fine-tuning setup, we used batch size of 4 according to our VM specifications. We employed Adam optimizer with weight decay. For learning rate and weight decay value selection, we conducted experimental training for 5 epochs. Through this process, we identified 5e-5 as our learning rate and 1e-4 as a reasonable weight decay as MiDaS is sensitive to larger values. Our models are primarily trained with 10 epochs, with a few variations: The **vanilla fine-tuning** and the **consistency loss + partial encoder freeze + decoder dropout** training used 15 epochs, and the **consistency loss + partial encoder freeze** training used 20 epochs. The major purpose is to observe the loss trend with more epochs when introducing major architectural changes or loss function changes. For vanilla finetuning, we implemented cross validation with 3 folds. To save computation resources, other experiments were conducted with a random train and validation data split without cross validation.

6. Results and discussion

In this section we discuss the quantitative and qualitative results from our experiments. Table 1 and Table 2 summarizes our experiments evaluation results on original test set and augmented test set. We also inspected DPT-hybrid-384 performance reported in [7]. DPT-hybrid-384 was pretrained on MIX 6, as a zero-shot result, for NYU set it achieved 8.69 $\delta > 1.25$, i.e., $(100 - \mathbf{8.69})/100 = \mathbf{0.9131}$ $\delta 1$. With fine-tuning using KITTI and NYU datasets, the evaluation on NYU V2 depth shows it matches or outperforms state-of-the-art performance [7]. While this states DPT-hybrid's ability to achieve strong performance on standard depth datasets, our experiments focus on evaluating robustness and generalization under augmented and adverse conditions.

6.1. Quantitative Results

Standard results. Table 1 summarizes the performance of various fine-tuning strategies on both original and augmented NYU Depth V2 test sets.

Compared to the baseline model, all fine-tuned variants significantly reduce error across all metrics - with AbsRel falling below 0.74 and δ_1 improving from < 0.21 to > 0.25 in general. Among all experiments, the fine-tuned model with **consistency + partial encoder freeze** strategy yields the most consistent and best overall performance across both data domains, suggesting that limited encoder adaptation along with alignment-based output regularization would strike a good balance between stability and domain-specific learning. Vanilla fine-tuning also improves accuracy compared to the baseline, but tends to underperform slightly relative to above strategy. Adding dropout to the decoder appears to mitigate overfitting marginally between original and augmented data, but high dropout rates (e.g., 0.5) would degrade performance due to excessive regularization.

Despite these improvements, the best models retain significant errors (e.g., AbsRel ≈ 0.66), and standard metrics like RMSE and MAE are not always sensitive to structural robustness under domain shifts. Additionally, the numerical differences between original and augmented inputs are small or ambiguous, making it hard to draw clear conclusions about robustness under adverse conditions using only standard metrics. To address this, we introduce Depth Boundary Error (DBE) metrics in the next part.

Visual evaluation results. As stated before in section 2.4, DBE focuses on visual structural details. It includes accuracy measure (DBE acc) and completeness measure (DBE comp). Accuracy measure essentially computes the accumulated distances from the prediction depth map edges to the ground truth map edges, where completeness focuses on the reverse direction to accumulate the distances from ground truth to prediction [3]. We leveraged `cv2` and `scipy.ndimage` for computing edge maps and distance maps. Table 2 and Table 3 summarize the performance of various fine-tuning strategies according to DBE metrics.

Compared with the baseline model, all models yield smaller accuracy errors and completeness errors on both original test set and augmented test set. Specifically, **Consistency + No freeze** model achieved 3.985 accuracy error and 7.563 completeness error on augmented set, indicating its potential to outperform other combinations to preserve structural depth details in adverse visual conditions.

We compared the differences between the error values on original test set and augmented test set in Table 3. The differences values provide additional insight on how a model performs and generalizes on different visual conditions. We noticed that **Consistency + No freeze** achieved the smallest accuracy error difference and a relatively low completeness error difference of 0.197. This indicates that this model is less sensitive to quality loss of the inputs, demonstrating its potential robustness to perform depth estimation for adverse visual conditions. Aside from **Consistency + No freeze** model, we also observed vanilla fine-tuning method yields stable performance based on DBE evaluation. It has a lower completeness error on augmented test set and the smallest completeness difference.

Our original hypothesis is that layer freezing and dropout may serve as a stronger regularization to prevent overfitting, specifically when using a targeted consistency loss minimizing the differences between the predictions on original and augmented data. In our experiments, Consistency + full encoder freeze has train loss of 0.1474, val loss of 0.1767. Consistency + full encoder freeze + 0.5 dropout has train loss of 0.1801, val loss of 0.1839, which indicates a smaller gap. However, the DBE analysis results do not fully support this speculation. While freezing the pretrained weights may preserve general purpose image features, it is also a trade-off as layer freezing also restricts the model's ability to learn new feature patterns. For dropout, in our experiments, the training image samples are limited. Therefore, while the dropout layers on the decoder mitigate the gaps between train and validation losses in experiments, aggressive regularization causes the model losing certain fine-grained features. The overall performance in terms of preserving visual structures may be degraded.

6.2. Qualitative Results

To complement above quantitative metrics, Figure 2 presents a visual comparison of predicted depth maps across several fine-tuning strategies: the baseline, vanilla fine-tuning, consistency loss without encoder freezing, and consistency loss with 0.5 dropout. Each strategy includes both original and augmented input predictions, alongside the ground truth depth maps, for two representative scenes.

In both scenes, the baseline model struggles to capture

Table 1: Standard evaluations from above experiments. For partial encoder freezing, in "consistency loss + partial model freezing", we freeze the early encoder layers until model.blocks.9 mlp. In "consistency loss + partial model freezing + dropout", we freeze early layers until model.blocks.7 mlp. The hypothesis is that with less freezing layers, the feature extraction process has more flexibility. Therefore, we are interested in whether regularization in decoder affects the performance. "Consistency" refers to "Consistency loss". The decimal number before "dropout" refers to the dropout rate. We provided model abbreviations to use in later results tables. Specifically, **C** denotes consistency loss, **Ff** denotes full freeze, **Fp** denotes partial freeze, **F0** denotes no freeze, **DO0.3** denotes dropout layers with 0.3 rate.

Model	**Dataset**	**AbsRel**	**RMSE**	**MAE**	δ_1	δ_2	δ_3
Baseline	Original	1.3178	4.0873	3.0803	0.1967	0.3602	0.4921
	Augmented	1.1272	3.4964	2.7089	0.2035	0.3684	0.5040
Vanilla finetuning (Vanilla)	Original	0.6646	2.3908	1.8801	0.2760	0.4784	0.6326
	Augmented	0.6584	2.3733	1.8672	0.2767	0.4820	0.6373
Consistency + Full encoder freeze (C+Ff)	Original	0.6701	2.4171	1.8854	0.2869	0.4964	0.6466
	Augmented	0.6476	2.3241	1.8203	0.2880	0.4934	0.6506
Consistency + Partial encoder freeze (C+Fp)	Original	0.6637	2.3799	1.8514	0.2902	0.4998	0.6513
	Augmented	0.6560	2.3465	1.8332	0.2893	0.4983	0.6534
Consistency + No freeze (C+F0)	Original	0.7330	2.5855	2.0314	0.2547	0.4509	0.5944
	Augmented	0.7298	2.5768	2.0238	0.2578	0.4557	0.5969
Consistency + Full encoder freeze + 0.3 dropout (C+Ff+DO0.3)	Original	0.7158	2.5755	2.0225	0.2694	0.4709	0.6180
	Augmented	0.6964	2.4855	1.9585	0.2688	0.4667	0.6194
Consistency + Full encoder freeze + 0.5 dropout (C+Ff+DO0.5)	Original	0.7325	2.6315	2.0714	0.2648	0.4625	0.6083
	Augmented	0.7061	2.5161	1.9852	0.2668	0.4630	0.6141
Consistency + Partial encoder freeze + 0.3 dropout (C+Fp+DO0.3)	Original	0.7257	2.5879	2.0280	0.2594	0.4568	0.6055
	Augmented	0.7203	2.5665	2.0081	0.2647	0.4634	0.6148

sharp object boundaries and underestimates depth variation, especially under augmented conditions. The predictions appear overly smoothed and lose geometric structure details, aligning with the lower DBE performances.

The **consistency + no freeze** model, which achieved the best DBE accuracy and completeness metrics, produces visibly shaper and more stable prediction across both scenes. For instance, in the shelving scene, the model accurately captures the vertical structure and depth gradients under both original and augmented conditions. Similarly, in the chair scene, the seat-backs and table edges are better preserved compared to the others. These results visually confirm the model's robustness and ability to retain structure details under domain shifts.

The **vanilla fine-tuning** model also produces reasonable outputs with relatively sharp transitions, though its predictions are slightly less stable, showing some smoothing region, like the seat-backs. This corresponds with its higher DBE accuracy, indicating the model is conservative in edge prediction, leading to lower completeness but at the cost of accuracy.

In contrast, the **consistency + 0.5 dropout** model suffers from noticeable over-smoothing and structural degradation, especially the background. Key edges and object contours are blurred out. This visually supports the observed tradeoff in DBE performance: While dropout might reduce overfitting, excessive regularization can diminish the model's ability to preserve some structure details for accurate depth prediction.

Overall, none of the strategies perfectly match the ground truth, which exhibits sharper edges and finer details in general. These qualitative results underscore the limits of the standard evaluation metrics, reinforcing the value of DBE metrics as the main diagnostic tool for structural fidelity and generalization under adverse conditions.

7. Conclusion and future work

In this project, we fine-tuned a pretrained Dense Prediction Transformer (DPT) model to study depth estimation for adverse visual conditions. Upon setting up a fine-tuning environment using a scale- and shift-invariant loss, we experimented with different techniques to improve model generalization and robustness, including an enhanced consistency

Table 2: Visual evaluations using DBE in pixels. "DBE acc" refers to accuracy error. "DBE comp" refers to completeness error.

Model	Dataset	DBE acc	DBE comp
Baseline	Original	6.665	8.399
	Augmented	6.953	8.888
Vanilla	Original	6.479	7.286
	Augmented	6.585	**7.433**
C+Ff	Original	**3.818**	7.785
	Augmented	4.028	8.216
C+Fp	Original	5.458	7.677
	Augmented	5.589	8.282
C+F0	Original	3.881	**7.366**
	Augmented	**3.985**	7.563
C+Ff+DO0.3	Original	4.013	7.701
	Augmented	4.403	8.101
C+Ff+DO0.5	Original	4.043	7.842
	Augmented	4.350	8.255
C+Fp+DO0.3	Original	3.822	7.569
	Augmented	4.131	7.926

Table 3: Accuracy and completeness error differences between original and augmented test sets. For both errors, the original test set value is extracted from the augmented test set value

Model	DBE acc difference	DBE comp difference
Baseline	0.288	0.489
Vanilla	0.106	**0.147**
C+Ff	0.210	0.431
C+Fp	0.131	0.605
C+F0	**0.104**	0.197
C+Ff+DO0.3	0.390	0.400
C+Ff+DO0.5	0.307	0.413
C+Fp+DO0.3	0.309	0.357

loss for domain robustness, architectural adjustments and regularization. We leveraged basic image processing methods to simulate blurred and nighttime environment againt original NYU Depth V2 data. This provided a foundation for the comparison between normal and degraded visual conditions.

We adopted both standard metrics including Absolute Relative Error (AbsRel), Root Mean Square Error (RMSE), and the accuracy under threshold ratios δ_1, δ_2 and δ_3, and visual metric Depth Boundary Error (DBE) for evaluation. Our results indicated that fine-tuning techniques consis-

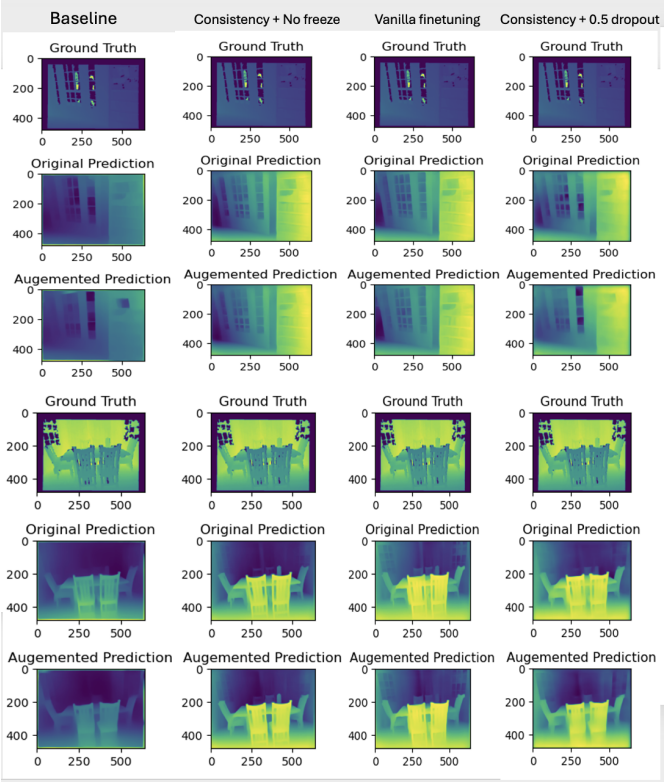


Figure 2: Comparison of depth maps. Inference results from baseline model, consistency loss + No freeze model, Vanilla Finetuning, Consistency loss + 0.5 dropout. We included 2 scenes from test set, under **dining room** class.

tently improve the overall performance of depth estimation. Specifically, in our context, consistency loss is a powerful method to minimize the differences between original and augmented prediction results. This effectively improves model robustness and generalization. On the other hand, incorporating encoder and decoder adjustments introduces tradeoffs. Encoder layer freezing preserves pretrained general purpose image features, however, for adverse visual conditions, more flexible feature extraction mechanism may be necessary. While decoder dropout layers mitigate the risk of overfitting, careful design is required to prevent feature loss caused by aggressive regularization.

Although transformers are powerful for image tasks such as depth estimation, because of its internal complexity, it remains challenging to interpret how an individual design change affects the overall training process and inference performance. For future work, in addition to progressively incorporate adaptations through trial and error, we want to continue exploring a more controlled experimentation design with isolated variables, and improved model selection using an integrated loss and evaluation mechanism. Further-

more, larger datasets with more simulated or real adverse visual data samples would facilitate the study in this field and help us further understand real-world vision tasks.

8. Contributions and acknowledgments

We contributed equally to the project, including dataset preprocessing, model implementation, experiments, evaluation, and report writing. All key decisions were jointly discussed and implemented collaboratively. For this project, we built upon the public official MiDaS repository provided by Intel. We used the **DPTDepthModel** and related inference code as our starting point for model fine-tunings. Additionally, we referenced and adapted depth evaluation utilities from this gist to facilitate evaluation and loss computations.

References

[1] 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.

[2] J. Gu, Z. Chen, and S. Song. Nightdepth: Nighttime monocular depth estimation via contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5703–5712, 2021.

[3] R. Koch, T. Batra, M. Bleyer, and C. Rother. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 41–50, 2018.

[4] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. *CoRR*, abs/1606.00373, 2016.

[5] J. H. Lee, M. Han, D. W. Ko, and I. H. Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *CoRR*, abs/1907.10326, 2019.

[6] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[7] R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. *CoRR*, abs/2103.13413, 2021.

[8] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[9] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer, 2020.

[10] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, 2017.