Predicting relative depth from a single image by training a CNN with self-supervised losses

1 Problem Definition

This project tackles the fundamental problem of recovering the three-dimensional structure of a scene from a single two-dimensional RGB image by predicting a relative depth map that ranks scene elements by their distance. Monocular depth estimation is vital for applications such as autonomous navigation, augmented reality placement, and robotic obstacle avoidance, yet obtaining dense ground-truth depth is costly or impractical. Here, we train a compact convolutional neural network(CNN) in a pseudo-supervised student–teacher framework, a pre-trained MiDaS(Monocular Depth Estimation via a Multi-Scale Network) [4] model generates synthetic depth labels on heavily augmented versions of the input image, and the student network learns to reproduce those relative depth cues using multi-scale reconstruction and edge-aware smoothness losses. We consider the task successful when the student's predictions closely match the teacher's outputs in both visual quality and standard depth-estimation metrics (RMSE, MAE, δ-accuracy), demonstrating that meaningful depth structure can be learned from a single image without any real depth annotations.

2 Related Work [Postgraduate-Only]

Monocular depth estimation has significantly evolved with the rise of deep learning [9, 7], reducing the reliance on traditional methods that require ground-truth depth maps [6]. The focus later moved toward self-supervised methods that leverage various sources, such as stereo image pairs [1] and video sequences [2]. Depth estimation from single RGB images has seen significant advances with the development of large-scale datasets and self-supervised methods. Among the most prominent models is MiDaS [4], which combines information from multiple datasets and training paradigms to predict relative depth by aligning scene structures from stereo and multi-view sources. MiDaS employs a scale-invariant loss and geometric alignment techniques to learn consistent depth cues. Its lightweight variant, MiDaS_small, is suitable for real-time or constrained applications, which we used in this project to provide supervision in the form of pseudo-depth labels.

For dense prediction tasks such as depth estimation, encoder-decoder architectures are commonly used. U-Net [5] remains a popular baseline due to its skip connections, which help preserve spatial details critical for fine-grained predictions. More recent models like DPT [3] incorporate vision transformers to better capture long-range dependencies. However, in low-data settings, smaller CNN's offer better training stability and efficiency. This motivates our use of a minimal U-Net-based architecture as the student model in the proposed pseudo-supervised framework.

3 Method

Our approach trains a compact CNN to predict relative depth from a single RGB image with help of pseudo-supervision. Input image was augmented 300 (enough diversity and fast training) times using random flips, rotations, color jitter, and resizing to 256×256 to simulate scene variability. For each augmented image, a corresponding pseudo-depth map was obtained using MiDaS_small, a pre-trained monocular depth estimation model [4]. The dataset was split into 80% training and 20% validation subsets. The model architecture, termed StudentNet, is a lightweight encoder-decoder with skip connections, inspired by U-Net [5]. The encoder contains two convolutional blocks, the first processes the input RGB image with two consecutive convolutional layers expanding to 32 channels, and the second downscales using max pooling followed by two more convolutions expanding to 64 channels. The decoder includes a transposed convolution to upsample the feature map back to 32 channels, and a final convolutional layer outputs a single-channel depth map. This is trained using a combined loss function, a multi-scale mean squared error (MSE) that compares predicted and target depth at multiple resolutions, and an edge-aware smoothness term that encourages local consistency in homogeneous regions.

Training was performed for 7, 10 epochs using the Adam optimizer with a learning rate of $1\mathrm{e}{-4}$. Final predictions were evaluated against MiDaS outputs using standard depth metrics, root mean square error (RMSE), mean absolute error (MAE), and threshold accuracy ($\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$), $\delta < 1.25$ measures the percentage of predictions close to the teacher (MiDaS) model (within 1.25 × error margin). Figure 1 shows the basic flowchart of the implemented model.

4 Results

Our model was evaluated using standard depth metrics compared to the MiDaS teacher output. Table 1 summarizes the evaluation metrics of the trained model for 7 and 10 epochs. Low δ-accuracy reflects noise in the MiDaS pseudo-labels. Although 10-epoch training yields slightly better metrics, the 7-epoch model is visually sharper and more coherent.

Figure 2 and Figure 3 illustrate the predicted depth maps for the 7- and 10-epoch models, respectively, alongside the teacher (MiDaS) output. The 7-epoch model, though quantitatively inferior, maintains crisper object boundaries and exhibits fewer artifacts, suggesting that the model begins to overfit by the 10th epoch, smoothing out fine details. Also, a blue tinted artifact is present on the depth map of 10 epochs model which depicts the amplification of noise from pseudo depth maps of augmented dataset through MiDaS model.

To test whether constraining the network output to [0,1] would help, we replaced the plain convolutional output with a sigmoid activation (all other settings unchanged). Figure 4 compares the student prediction with sigmoid against the MiDaS teacher after 7 epochs. But the depth map seems, Washed-out, low-contrast “yellow” map.Because our pseudo-labels (from MiDaS) are not normalized to [0,1], forcing the network through a sigmoid causes a scale mismatch, large depth variations are compressed, and relative depth cues are lost. We therefore omit sigmoid in the final model.

Table 2 shows steady training loss reduction from 564k to 537k, with similar validation trends, indicating mild but consistent learning. MiDaS generates relative depth maps that are scale-invariant, but they are not normalized to a consistent range, which lead to disproportionately large loss values when raw depth is used for training. This lead to large absolute errors even when the predicted structure appears correct. These results highlights, standard error metrics do not always align with perceived structural quality, especially in self-supervised settings where supervision is approximate and has a restriction on training data.

5 Reflection

This project shows that depth estimation from a single image is feasible using self-supervised learning by avoiding ground-truth labels, but it also highlights key limitations. Since the model trains entirely on MiDaS (teacher) outputs, any noise or bias in the teacher is inherited by the student. Augmentations sometimes amplified these issues, especially near edges and fine details related to building. Although the method showcases an efficient way to enable depth learning in data-scarce settings, it raises concerns about reliability and potential misuse in privacy-sensitive contexts.

6 Conclusion

This work shows that a CNN can learn to estimate relative depth from a single image using pseudo-supervision and augmentations. Future improvements could include normalizing MiDaS depth maps or adding photometric consistency losses, and incorporating geometric priors such as symmetry [8] to enhance robustness and reduce dependence on the teacher model.

Table 1: Student model performance at different training epochs. Metrics are evaluated against pseudo-depths from MiDaS_small. RMSE and MAE are in depth units; δ-thresholds show the percentage of pixels within specific relative error bounds. Bold indicates better numerical performance.

Epochs	RMSE ↓	MAE ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
7	428.43	303.98	0.10%	0.19%	0.28%
10	**413.24**	**293.10**	**0.21%**	**0.41%**	**0.69%**

Table 2: Training and validation loss values over 7 epochs

Epoch	Training Loss	Validation Loss
1	564317	581360
2	564251	548459
3	561197	563069
4	564968	556613
5	554512	568898
6	546885	565955
7	537003	543945

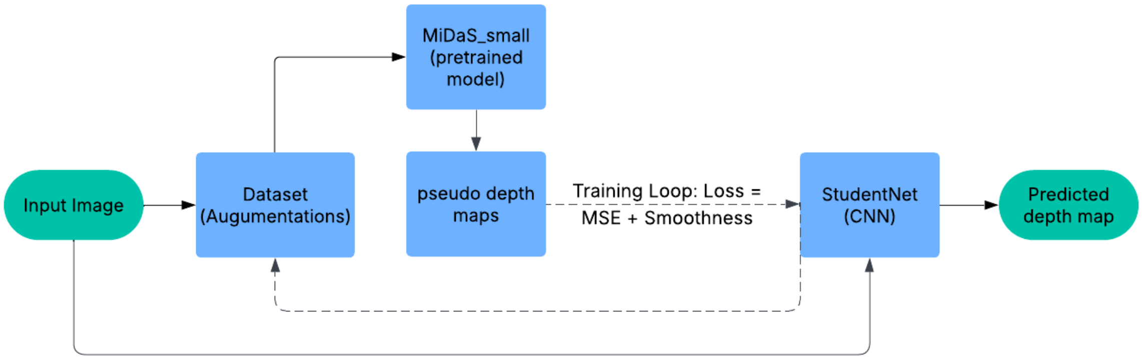

Figure 1: Flowchart of the model implemented to predict a depth image of input image

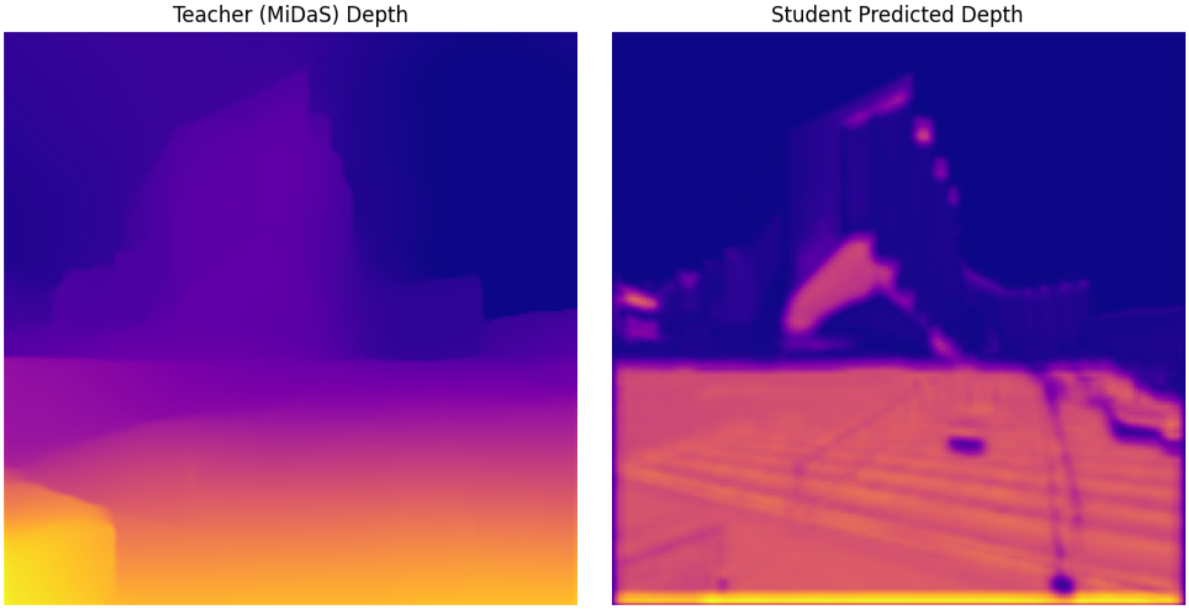

Figure 2: Depth maps comparison from a pretrained MiDaS model and the trained CNN model here for 7 epochs

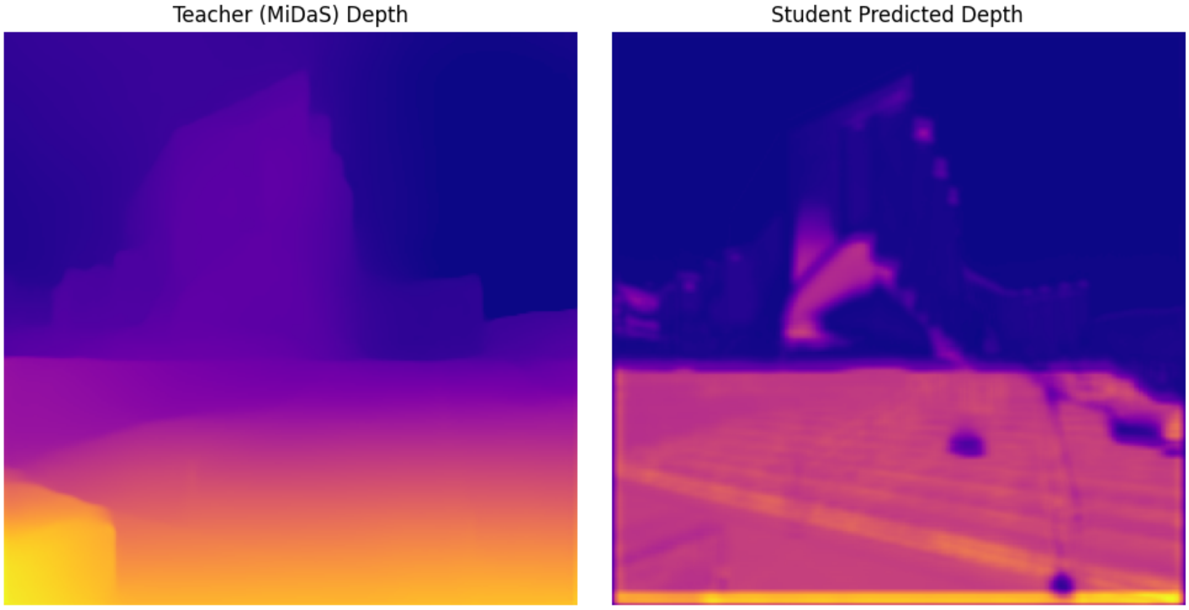

Figure 3: Depth maps comparison from a pretrained MiDaS model and the trained CNN model here for 10 epochs

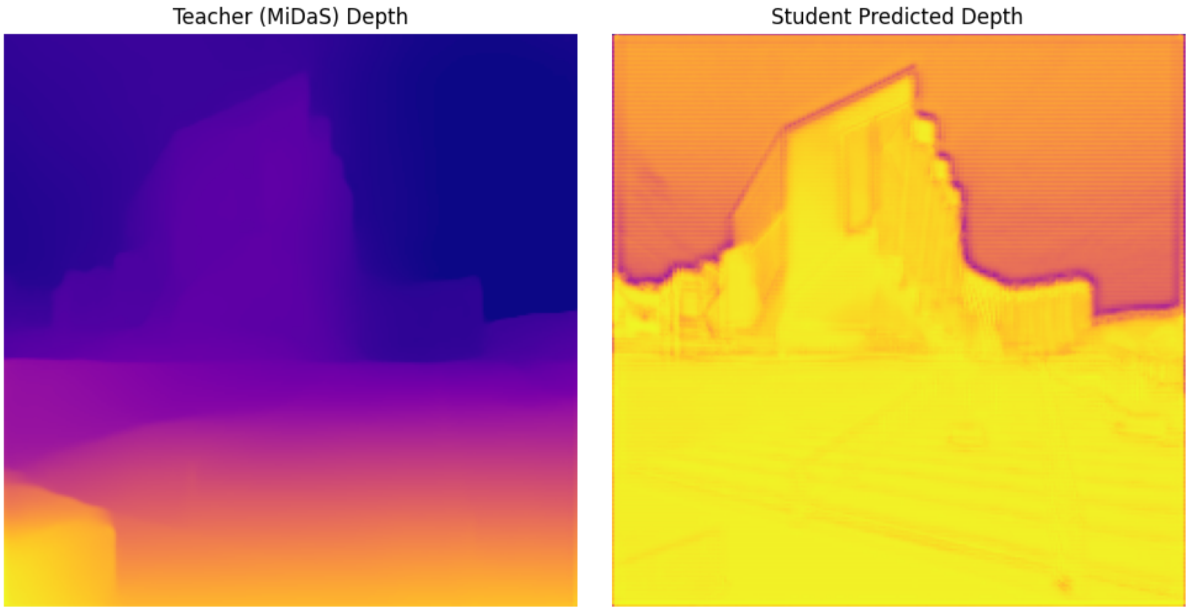

Figure 4: Depth maps comparison from a pretrained MiDaS model and the trained CNN model with sigmoid activation function in the final convolution layer.

References

[1] C. Godard, O. Mac Aodha, and G.J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

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

[3] R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. In *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, page ICCV, 2021.

[4] Ranftl René, Lasinger Katrin, Hafner David, Schindler Konrad, and Koltun Vladlen. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page TPAMI, 2022.

[5] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[6] A. Saxena, S. Chung, and A. Ng. Learning depth from single monocular images. In *Advances in neural information processing systems*, page 18, 2005.

[7] Fabio Tosi, Pierluigi Zama Ramirez, and Matteo Poggi. Diffusion models for monocular depth estimation: Overcoming challenging conditions. In *University of Bologna, Bologna, Italy*, 2024.

[8] S. Tsogkas, I. Kokkinos, and A. Yuille. Symmetry-aware depth estimation using deep neural networks. In *arXiv preprint*, 2016.

[9] C. Zhao, Q. Sun, C. Zhang, Y. Tang, and F Qian. Monocular depth estimation based on deep learning: An overview. In *Science China Technological Sciences*, page 1612–1627, 2020.