2024 IEEE/ACIS 22nd International Conference on Software Engineering Research, Management and Applications (SERA) | 979-8-3503-9134-3/24/$31.00 ©2024 IEEE | 10.1109/SERA61261.2024.10685601

A Robust Attention-based Convolutional Neural Network for Monocular Depth Estimation

1st Yuqi Song
University of Southern Maine
Portland, United States
yuqi.song@maine.edu

2nd Xin Zhang
University of Southern Maine
Portland, United States
xin.zhang@maine.edu

3rd Bokai Yang
University of Wisconsin - Eau Claire
Eau Claire, United States
yangboka@uwec.edu

4th Fei Zuo
University of Central Oklahoma
Oklahoma, United States
fzuo@uco.edu

5th Xianshan Qu
University of Central Oklahoma
Oklahoma, United States
xqu1@uco.edu

***Abstract*—In this study, we present a novel attention-based encoder-decoder model for monocular depth estimation, showcasing exceptional robustness and accuracy across standard and noise-injected variants of the KITTI dataset. By integrating Convolutional Block Attention Module (CBAM) and Squeeze-and-Excitation (SE) blocks, our approach significantly outperforms existing state-of-the-art methods, especially in environments affected by real-world noise such as changes in brightness, saturation, and RGB channels. The model's superior performance, validated through rigorous testing, marks a significant step forward in the field, offering promising applications in autonomous driving, augmented reality, and beyond. Our work demonstrates the potential of attention mechanisms in enhancing depth estimation models to reliably interpret complex scenes, paving the way for advancements in depth-dependent technologies operating in dynamic and challenging conditions.**

***Index Terms*—computer vision, monocular images, depth estimation**

I. Introduction

Monocular depth estimation [1], a cornerstone of computer vision, has garnered significant attention for its critical role in enabling a myriad of applications from autonomous driving to augmented reality. Unlike binocular depth estimation [2], [3], which mimics human vision by utilizing two viewpoints to infer depth, monocular depth estimation relies on a single image, posing a unique set of challenges and opportunities. This reliance on a solitary viewpoint is not just a limitation but an advantage in scenarios where space, cost, or computational resources are constrained, making it a more versatile solution compared to its binocular and sensor-based counterparts. Sensor-based methods, though accurate, often suffer from limitations such as high costs, environmental constraints, and the need for direct line-of-sight, highlighting the practical significance of monocular depth estimation.

Despite considerable advancements in monocular depth estimation, a significant portion of current research persistently frames the challenge as a purely mathematical regression problem [4]. This perspective, predominantly reliant on convolutional neural networks (CNNs), aims to predict depth information for each pixel with precision. However, this approach glosses over the nuanced complexities introduced by real-world conditions. Factors such as varying brightness levels, diverse image qualities, and other environmental influences are often sidelined, despite their profound impact on the accuracy of depth predictions. This oversight not only limits the robustness of depth estimation models but also restricts their applicability in dynamically changing real-world scenarios. It is within this context that our research introduces a paradigm shift, proposing an attention-based method that not only acknowledges but rigorously addresses these limitations head-on.

Addressing these challenges, our research introduces a novel attention-based method tailored for monocular depth estimation. Our approach not only acknowledges but rigorously analyzes the impact of noise on depth estimation accuracy. By employing an attention mechanism, our model dynamically focuses on salient features within the image, effectively filtering out noise and enhancing depth prediction. To comprehensively evaluate the resilience of our method to noise, we conducted experiments on the original image dataset (KITTI) [5] and several noise-included variants of these datasets. These variants were systematically generated to simulate common noise conditions in real-world scenarios, including alterations in RGB channels, brightness, and saturation.

Our experimental results unequivocally demonstrate that our method not only achieves superior performance on the original KITTI datasets but also maintains this leading edge across various noise-augmented versions of these datasets. This resilience to noise underscores the robustness of our approach, highlighting its practical applicability in diverse and challenging real-world conditions.

The contributions of our work are manifold:

- We propose an attention-based, lightweight CNN model designed to predict depth information from monocular images accurately.

979-8-3503-9134-3/24/$31.00 ©2024 IEEE

SERA 2024, May 30-June 1, 2024, Honolulu, USA
Authorized licensed use limited to: University of Central Oklahoma Chambers Library. Downloaded on September 27,2024 at 21:13:28 UTC from IEEE Xplore. Restrictions apply.

- Our method exhibits remarkable resistance to noise, whether it changes in brightness, saturation, or alterations in the RGB channels, establishing a new benchmark for robustness in monocular depth estimation.
- Through extensive experimentation, we validate the efficacy of our method, demonstrating its superior performance not only on standard benchmark datasets but also under varied noise conditions, thereby proving its versatility and reliability.

II. Proposed Method

A. Model Architecture

The core of our model is an attention-based encoder-decoder structure, which is augmented with both Convolutional Block Attention Module (CBAM) [6] and Squeeze-and-Excitation (SE) attention mechanisms [7]. These mechanisms are strategically placed to refine the feature representations by emphasizing relevant spatial and channel-wise features. The encoder-decoder architecture ensures that high-level semantic information is captured and effectively utilized for depth estimation.

Encoder-decoder Structure. For the encoder part, we employ DenseNet-169, excluding its final classification layer, as our encoder. This choice is motivated by DenseNet-169's efficiency in extracting high-resolution features and its capability to downsample the input image effectively. By leveraging a network pre-trained on the ImageNet dataset, we harness a rich feature representation learned from a diverse set of visual data, providing a solid foundation for depth estimation. The encoder's design, characterized by its dense connectivity, facilitates the flow of information throughout the network, ensuring that even the deepest layers have access to fine-grained details from the input image.

The decoder component of our network employs a straightforward upscaling scheme designed to match the encoder's output. For each level of the decoder, the process involves up-sampling the output of the previous layer to align with the size of the corresponding encoder layer's output after attention modulation through CBAM and SE mechanisms. We then concatenate these three outputs—direct encoder output, CBAM-modulated features, and SE-modulated features—to form a comprehensive feature map that integrates both spatial and channel-wise attention cues. Upon concatenation, a convolution operation is performed on the combined feature map to refine the features further and ensure that the resulting depth map is both detailed and accurate. This step is crucial for blending the different feature representations harmoniously and effectively translating the enriched feature set into a coherent depth estimation.

Integration of Attention Mechanisms. Our model's distinctiveness lies in the strategic placement of CBAM [6] and SE blocks [7] within the skip connections that link the encoder and decoder. This design choice ensures that the attention-modulated features directly inform the depth reconstruction process, significantly enhancing the model's focus on salient features. The process unfolds as follows for each encoder layer:

- Directly pass: The feature map from each encoder layer is directly passed to the corresponding decoder layer via skip connections. This direct transmission of information is crucial for preserving high-resolution details lost during downsampling.
- Next Layer Processing: Simultaneously, the feature map proceeds to the next encoder layer, ensuring the model captures increasingly abstract and semantic features at deeper levels.
- CBAM Application: On each encoder layer's output, we apply CBAM to generate a comprehensive attention map that encapsulates both spatial and channel-wise feature relevancies. This attention map guides the model to prioritize features that are most indicative of depth variations within the scene.
- SE Block Application: Alongside CBAM, we also apply the SE block on the same feature map to produce a channel-wise attention map. This map recalibrates the feature channels, emphasizing those that contribute most significantly to accurate depth perception.

The attention-based module within our architecture is integral to refining feature maps at every encoder layer. As each layer outputs a feature map of dimension $H \times W \times C$, the module applies both CBAM and SE operations to it. CBAM refines the feature map by emphasizing salient spatial and channel-wise features, outputting $(H \times W \times C)_{CBAM}$. Concurrently, the SE block recalibrates the feature channels based on global information, generating $(H \times W \times C)_{SE}$. A normal skip connection also carries the original feature map, $(H \times W \times C)_{\text{normal skip}}$, to preserve low-level details. These three variants of the feature map are then merged via a bit-wise addition to produce $(H \times W \times C)_{\text{att module}}$, a feature map that encapsulates both focused attention and original detail. This enriched feature map is then combined with the final layer's output from the decoder to form the depth map, ensuring that the reconstructed depth is informed by both detailed and abstract representations of the scene. The structure of our attention-based encoder-decoder model is shown in Fig. 1.

B. Loss Function Design

To optimize our model for robustness and precision in depth estimation, we employ a composite loss function that integrates three key components: BerHu loss, Scale-Invariant Mean Squared Error (SIMSE), and Gradient loss. The composite loss function is a weighted sum of the three components, tailored to enhance depth prediction accuracy and robustness, as Equation. 1.

$$\mathcal{L}_{\text{composite}} = \alpha \mathcal{L}_{\text{BerHu}} + \beta \mathcal{L}_{\text{SIMSE}} + \gamma \mathcal{L}_{\text{Gradient}}, \tag{1}$$

where α, β, and γ are weights for each loss component, determined through extensive experimentation to optimize performance over the original test sets and several variants.

979-8-3503-9134-3/24/$31.00 ©2024 IEEE

SERA 2024, May 30-June 1, 2024, Honolulu, USA
Authorized licensed use limited to: University of Central Oklahoma Chambers Library. Downloaded on September 27,2024 at 21:13:28 UTC from IEEE Xplore. Restrictions apply.

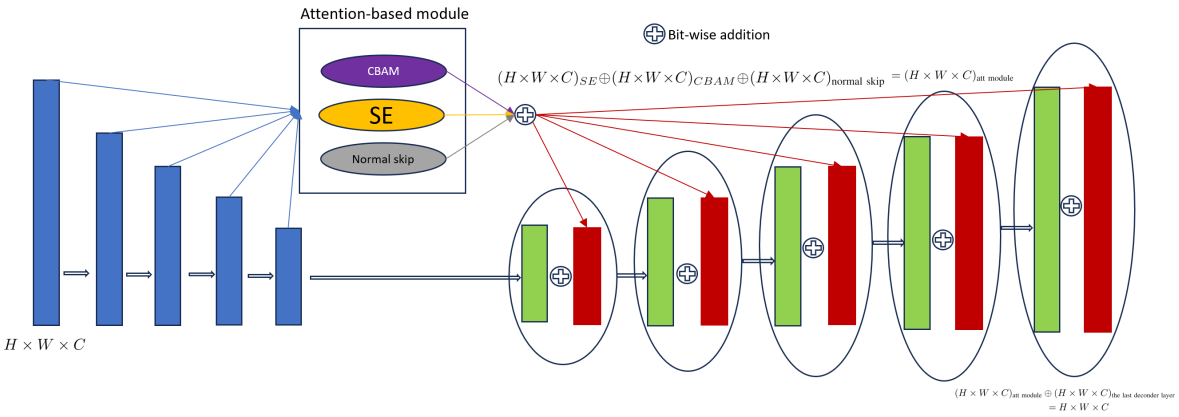

Fig. 1. Our model employs an attention-augmented encoder-decoder architecture designed for precise depth estimation from monocular images. The process begins as the first layer of the encoder produces a feature map with dimensions $H \times W \times C$. This feature map is then concurrently processed by an attention-based module, where it undergoes three distinct operations: application of the CBAM, the Squeeze-and-Excitation (SE) block, and a normal skip connection. Each operation transforms the feature map into corresponding attention-focused maps: $(H \times W \times C)_{SE}$, $(H \times W \times C)_{CBAM}$, and $(H \times W \times C)_{\text{normal skip}}$. These maps are then combined through a bit-wise addition to form a unified feature map $(H \times W \times C)_{\text{att module}}$. In the final stage of the decoder, this attention-enriched feature map is merged with the output of the last decoder layer using a bit-wise addition, leading to the generation of the final depth map.

Each component is chosen for its unique contribution to the model's performance, addressing specific challenges encountered in monocular depth estimation.

BerHu Loss. The BerHu loss [8] is defined as Equation. 2.

$$\mathcal{L}_{\text{BerHu}} = \begin{cases} |y-\hat{y}| & \text{for } |y-\hat{y}| \le c, \\ \frac{(y-\hat{y})^2+c^2}{2c} & \text{otherwise.} \end{cases} \quad (2)$$

where y denotes the Ground-truth depth value, $\hat{y}$ stands for the predicted depth value and c stands for the threshold that determines the switch between the L1 loss and L2 loss. In the following experiment, we treat c as a hyperparameter and determine the value for c based on extensive experiments. We adopt the BerHu loss to achieve robustness against outliers and ensure a balanced treatment of prediction errors. This loss function is particularly effective in monocular depth estimation as it combines the best attributes of L1 and L2 losses, providing precision for small errors while being forgiving towards large discrepancies, which are common due to occlusions or reflective surfaces in images. The ability of BerHu loss to adaptively switch between L1 and L2 loss behaviors based on the error magnitude helps in handling the diverse range of depth values encountered in real-world scenes, enhancing the model's overall predictive accuracy.

Scale-Invariant Mean Squared Error (SIMSE) Loss. The SIMSE loss function [9] is defined as Equation. 3.

$$\mathcal{L}_{\text{SIMSE}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 - \frac{\lambda}{n^2}\left(\sum_{i=1}^{n}(\hat{y}_i - y_i)\right)^2. \quad (3)$$

where y and $\hat{y}$ denote the Ground-truth depth value and the predicted depth value, respectively, n stands for the number of pixels in the image, and λ stands for the regularization parameter that controls the importance of the scale-invariant term. We integrate SIMSE to ensure the model's predictions are invariant to the absolute scale of depth, focusing instead on capturing relative depth differences accurately. This aspect is crucial in monocular depth estimation, where absolute depth cues are not directly available, and the model must infer depth from visual features and context. The scale-invariant property of SIMSE aligns with the goal of monocular depth estimation to accurately perceive the spatial arrangement and distances between objects in a scene, irrespective of their absolute distances from the camera.

Gardient Loss. The Gardient loss [10], [11] is defined as Equation. 4.

$$\mathcal{L}_{\text{Gradient}} = \sum_{i=1}^{n}\left(|\nabla_X\hat{y}_i - \nabla_X y_i| + |\nabla_Y\hat{y}_i - \nabla_Y y_i|\right), \quad (4)$$

where y and $\hat{y}$ denote the Ground-truth depth value and the predicted depth value, respectively, ∇_X stands for the gradients in the horizontal directions, and ∇_Y stands for the gradients in the vertical directions. The inclusion of Gradient loss is aimed at preserving the edge information and enhancing the structural integrity of the estimated depth maps. In the context of monocular depth estimation, maintaining sharp boundaries between objects and their backgrounds is essential for producing visually coherent depth maps that closely mimic the real-world structure of scenes. By emphasizing the gradients within the depth map, this loss component directly contributes to the model's ability to delineate object contours and surface transitions, which are pivotal for interpreting and navigating complex environments.

Combining these three loss functions together, we propose our final loss function as Equation. 1.

C. Noise-injection Methods

To evaluate our model's robustness and performance under varied real-world conditions, we introduce noise-injection

979-8-3503-9134-3/24/$31.00 ©2024 IEEE

SERA 2024, May 30-June 1, 2024, Honolulu, USA
Authorized licensed use limited to: University of Central Oklahoma Chambers Library. Downloaded on September 27,2024 at 21:13:28 UTC from IEEE Xplore. Restrictions apply.

Fig. 2. An image from the KITTI dataset before and after brightness adjustment. (a), (b), and (c) stand for the original image, $b_{1.5}$ variant, and $b_{0.5}$ variant, respectively. It is obvious that the brightness of the original image is adjusted.

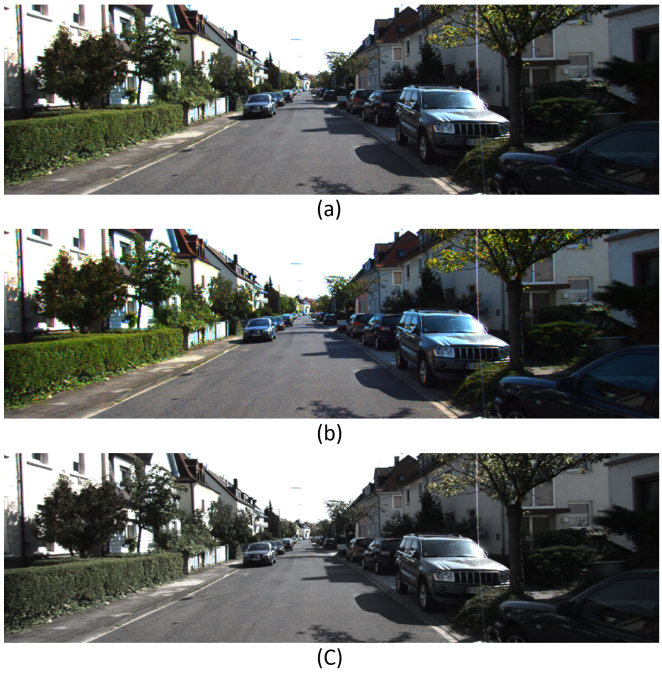

Fig. 3. An image from the KITTI dataset before and after saturation adjustment. (a), (b), and (c) stand for the original image, $s_{1.5}$ variant, and $s_{0.5}$ variant, respectively. The saturation of the original image is adjusted obviously.

Fig. 4. An image from the KITTI dataset before and after RGB channel alteration. (a), (b), (c), (d) represent the original image, the GRB version, the RBG version and the BGR version.

methods that simulate common environmental challenges such as changes in brightness, saturation, and color space.

Brightness Adjustment. Utilizing '**adjust_brightness**' function of PyTorch, we create two brightness variants of the original dataset, with factors of 0.5 and 1.5, resulting in datasets labeled as $b_{0.5}$ and $b_{1.5}$, respectively, as shown as Fig. 2.

Saturation Adjustment. Similar to the brightness adjustment, the '**adjust_saturation**' is employed to adjust the image saturation, with factors of 0.5 and 1.5, producing variants $s_{0.5}$ and $s_{1.5}$, respectively, as shown as Fig. 3.

RGB Channel Alteration. To simulate alterations in the RGB channels, we use PyTorch's '**permute**' function, rearranging the channels in the dataset images. This method tests the model's capacity to handle changes in color information, which is crucial for depth perception in varied lighting conditions. To be specific, we manipulate the color channels to create three new variants of the dataset, as shown as Fig. 4:

- GRB variant: The red and green channels are swapped.
- RBG variant: The green and blue channels are switched.
- BGR variant: The blue and red channels are exchanged.

These alterations test the model's ability to maintain depth estimation accuracy despite changes in color information, which can significantly affect the perception of depth in real-world scenes.

Then, by evaluating our model's performance on the original dataset and the seven generated variants (including adjustments for brightness, saturation, and RGB channel alterations), we comprehensively assess its robustness to a variety of real-world conditions. The overall performance is calculated as Equation. 5. This extensive testing ensures our model's adaptability and reliability across different environmental factors, demonstrating its effectiveness in monocular depth estimation tasks under varied conditions.

$$
\begin{aligned}
P_b &= \frac{P_{b_{0.5}} + P_{b_{1.5}}}{2}, \\
P_s &= \frac{P_{s_{0.5}} + P_{s_{1.5}}}{2}, \\
P_{RGB} &= \frac{P_{GRB} + P_{RBG} + P_{BGR}}{3}, \\
P_{\text{overall}} &= \frac{P_b + P_s + R_{RGB} + P_{\text{original}}}{4}
\end{aligned}
\tag{5}
$$

Authorized licensed use limited to: University of Central Oklahoma Chambers Library. Downloaded on September 27,2024 at 21:13:28 UTC from IEEE Xplore. Restrictions apply.

III. Experiments

A. Datasets

Our proposed method is evaluated on KITTI [5] datasets. **KITTI** is an outdoor dataset for monocular deep estimation and object detection and tracking based on deep learning, which is captured through a car equipped with 2 high-resolution color cameras, 2 gray-scale cameras, a laser scanner, a global positioning system (GPS) and contains 93,000 training samples. The original image size is around 1,242 $\times$ 375. Similar to NYU-V2, we also execute an inpainting method to fill in missing depth values. We use the training/ testing sets split of Eigen et al. [4], which is the most standard method for KITTI splitting.

B. Experimental Results and Analysis

Effectiveness. In Table I, we compare the proposed algorithm with the recent SOTA algorithms [11]–[17] and pioneering work [4] in monocular depth estimation on KITTI dataset quantitatively, which is able to fully demonstrate the effectiveness of our method.

Across all the evaluated metrics, our approach consistently outperforms the competing algorithms, cementing its effectiveness in predicting depth from single images. Our method achieves the highest accuracy in threshold metrics (δ_1, δ_2, and δ_3), which is indicative of its superior ability to estimate depth within stringent error margins. Notably, the δ_1 accuracy reaches an impressive 0.950, which means that 95% of our depth predictions are within 25% of the ground truth values. The RMSE is markedly lower for our method at 2.211, compared to the nearest competitor at 2.548. This demonstrates our model's precision and its capability to predict accurate depth across the dataset's varied scenarios. In terms of relative errors, both logarithmic (log. rel) and absolute (abs. rel), our method again leads with the lowest error rates. This superiority speaks to our model's consistency and its robustness in handling the diverse and complex nature of real-world scenes captured in the KITTI dataset. The squared relative error (sq. rel) further reinforces our method's leading position with the lowest value of 0.261. This outcome highlights the method's adeptness at minimizing large errors, which is crucial for depth-sensitive applications.

Robustness. To prove the robustness [19] and resistance to noise, we generate several variants and use our model to predict the depth of information on them. The experimental results are summarized in Table. II. We then conduct extensive experiments to prove that our method is able to resist noise and maintain accuracy in challenging conditions. To be specific, we perform our method and several comparisons over the original KITTI test set and its seven noise-injected variants, as we described in the previous section. Then we calculate their average score as the evaluation metrics. We summarize the experimental results in Table. III. In the critical metric of δ_1, which indicates the proportion of pixels where the predicted depth is within 25% of the ground truth, our method achieves an average of 0.926. This score not only surpasses the other methods by a significant margin but also highlights the algorithm's ability to maintain high accuracy in the presence of noise. For the δ_2 and δ_3 metrics, our algorithm similarly outperforms the competition with averages of 0.961 and 0.974, respectively. These metrics, which are less strict than δ_1, still require a high degree of precision, and our algorithm's leading scores reflect its ability to accurately predict depth across a broader range of pixel values. The RMSE, a direct measure of the model's error magnitude, further emphasizes the superiority of our approach. With an average RMSE of 3.029, our method demonstrates a lower average deviation from the actual depth values compared to the other methods, which exhibit higher errors. When it comes to relative errors, both logarithmic (log. rel) and absolute (abs. rel), our model achieves the lowest averages, 0.147 and 0.061, respectively. These results are indicative of the model's effectiveness in estimating depth without being swayed by outlier values or scale discrepancies. The squared relative error (sq. rel) average of 0.444 for our model is another testament to its precision, especially when it comes to the estimation of larger depths, which can be more prone to errors due to scaling factors.

Across all these metrics, the performance of our algorithm remains consistently superior, even when subjected to the challenging conditions of altered brightness, saturation, and RGB channels. This consistency is a strong indicator of the model's resilience to common real-world disturbances that can affect the visual appearance of a scene.

IV. Conclusion

Our research introduces a robust attention-based encoder-decoder model for monocular depth estimation, demonstrating superior performance on the KITTI dataset and its noise-augmented variants. Through integrating CBAM and SE attention mechanisms, our model excels in accuracy and resilience to real-world noise, outperforming state-of-the-art methods under varied conditions. This breakthrough not only progresses the theoretical understanding of depth estimation but also opens doors to impactful real-world uses, ranging from self-driving cars to immersive augmented reality experiences. The demonstrated efficacy of our model underscores its capability to boost the performance of technologies reliant on depth information, even in fluctuating conditions.

References

[1] X. Zhang, R. Abdelfattah, Y. Song, S. A. Dauchert, and X. Wang, "Depth monocular estimation with attention-based encoder-decoder network from single image," in *2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*. IEEE, 2022, pp. 1795–1800.

[2] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

979-8-3503-9134-3/24/$31.00 ©2024 IEEE

SERA 2024, May 30-June 1, 2024, Honolulu, USA

Authorized licensed use limited to: University of Central Oklahoma Chambers Library. Downloaded on September 27,2024 at 21:13:28 UTC from IEEE Xplore. Restrictions apply.

TABLE I

COMPARISONS OF DIFFERENT METHODS ON KITTI. FOR EACH COLUMN, WE HIGHLIGHT THE BEST-PERFORMANCE ALGORITHM.

Methods	δ_1	δ_2	δ_3	RMSE	log. rel	abs. rel	sq. rel
	(Higher is better)			(Lower is better)			
Godard et al. [16]	0.861	0.949	0.976	4.935	0.206	0.114	0.898
Eigen at al. [4]	0.692	0.899	0.967	7.156	0.270	0.190	1.515
Kuznietsov et al. [18]	0.862	0.960	0.986	4.621	0.189	0.113	0.741
Alhashim et al. [11]	0.886	0.965	0.986	4.170	0.171	0.093	0.589
Fu et al. [12]	0.932	0.984	0.994	2.727	0.120	0.072	0.307
Zhang et.al [1]	0.947	0.989	0.996	2.548	0.113	0.061	0.297
Ours	**0.950**	**0.992**	**0.999**	**2.211**	**0.097**	**0.052**	**0.261**

TABLE II

ROBUSTNESS EVALUATION OF OUR METHOD ON KITTI DATASET VARIANTS.

Dataset Variant	δ_1	δ_2	δ_3	RMSE	log. rel	abs. rel	sq. rel
Original	0.950	0.992	0.999	2.211	0.097	0.052	0.261
Brightness 0.5 (b0.5)	0.933	0.974	0.980	2.987	0.152	0.068	0.401
Brightness 1.5 (b1.5)	0.929	0.970	0.973	3.000	0.153	0.069	0.398
Saturation 0.5 (s0.5)	0.922	0.953	0.969	3.056	0.160	0.070	0.418
Saturation 1.5 (s1.5)	0.928	0.958	0.971	3.013	0.155	0.068	0.411
GRB Variant	0.901	0.922	0.950	3.887	0.180	0.099	0.698
RBG Variant	0.903	0.930	0.952	3.881	0.182	0.102	0.703
BGR Variant	0.889	0.921	0.947	3.864	0.179	0.100	0.701
Average	0.926	0.961	0.974	3.029	0.147	0.061	0.444

TABLE III

COMPARISONS OF DIFFERENT METHODS ON KITTI AND ITS VARIANTS. FOR EACH COLUMN, THE PERFORMANCE SCORE IS THE AVERAGE VALUE OVER THE ORIGINAL KITTI TEST SETS AND ITS 7 VARIANTS.

Methods	δ_1	δ_2	δ_3	RMSE	log. rel	abs. rel	sq. rel
	(Higher is better)			(Lower is better)			
Godard et al. [16]	0.829	0.902	0.947	5.847	0.259	0.132	1.023
Kuznietsov et al. [18]	0.831	0.922	0.954	5.432	0.247	0.125	0.901
Alhashim et al. [11]	0.851	0.929	0.961	4.908	0.239	0.103	0.798
Fu et al. [12]	0.907	0.950	0.967	3.609	0.192	0.079	0.498
Ours	**0.926**	**0.961**	**0.974**	**3.029**	**0.147**	**0.061**	**0.444**

[3] H. Javidnia and P. Corcoran, "Accurate depth map estimation from small motions," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2453–2461.

[4] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[5] 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*. IEEE, 2012, pp. 3354–3361.

[6] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[7] Z. Zhong, Z. Q. Lin, R. Bidart, X. Hu, I. B. Daya, Z. Li, W.-S. Zheng, J. Li, and A. Wong, "Squeeze-and-attention networks for semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 065–13 074.

[8] L. Zwald and S. Lambert-Lacroix, "The berhu penalty and the grouped effect," *arXiv preprint arXiv:1207.6868*, 2012.

[9] T. L. Da Silveira, L. P. Dal'Aqua, and C. R. Jung, "Indoor depth estimation from single spherical images," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2935–2939.

[10] L. Huynh, P. Nguyen-Ha, J. Matas, E. Rahtu, and J. Heikkilä, "Guiding monocular depth estimation using depth-attention volume," in *European Conference on Computer Vision*. Springer, 2020, pp. 581–597.

[11] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *arXiv preprint arXiv:1812.11941*, 2018.

[12] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2002–2011.

[13] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 239–248.

[14] Z. Hao, Y. Li, S. You, and F. Lu, "Detail preserving depth estimation from a single image using attention guided networks," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 304–313.

[15] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5354–5362.

[16] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 270–279.

[17] J. Zhang, Q. Su, C. Wang, and H. Gu, "Monocular 3d vehicle detection with multi-instance depth and geometry reasoning for autonomous driving," *Neurocomputing*, vol. 403, pp. 182–192, 2020.

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

[19] X. Zhang, Y. Song, X. Wang, and F. Zuo, "D-score: A white-box diagnosis score for cnns based on mutation operators," *arXiv preprint arXiv:2304.00697*, 2023.

Authorized licensed use limited to: University of Central Oklahoma Chambers Library. Downloaded on September 27,2024 at 21:13:28 UTC from IEEE Xplore. Restrictions apply.