Development of Depth Maps for Outdoor Images

Devashish Tripathi
21093
devashish21@iiserb.ac.in

Aditya Kishore
21011
adityak21@iiserb.ac.in

Snehal Mahajan
20161
mahajan20@iiserb.ac.in

November 26, 2024

Abstract

Depth images have emerged as a significant addition to the process of producing new views of real-world scenes. Applications include Autonomous Driving, Augmented Reality and 3D Reconstruction. Depth images are generally derived using a two-camera system, measuring depth using the discrepancy between the two images, and include a texture map and a depth map, which represent the colour and distance information per pixel, respectively. Monocular depth estimation, which utilises a single-camera system, is gaining popularity thanks to the abundant single-view images, such as the recent DepthAnything papers. The task of outdoor depth estimation is crucial for areas such as Autonomous vehicles and surveillance, yet there are issues of noise, varying lighting and weather, especially in nighttime images from cheap but abundant sources such as StreetView or through on-site cameras. In order to counteract this, this report concerns itself with two approaches, one based on finetuning the DepthAnything model, while the other concerns itself with increasing the resolution of the nighttime images. The experiments were conducted on a dataset collected in Bhopal, India. The results, while not on par, hold promise that future work can be done in this direction.

1 Introduction

Depth estimation is essential for perceiving and reconstructing the real world from images with depth information that is composed of texture and depth maps containing colour and distance information, respectively. Such images lay the basis for various applications, such as autonomous driving, augmented reality, and 3D scene reconstruction, enabling spatial contexts within their decision-making processes.

Reconstruction of outdoor scenes, especially with Street View and pictures from on-site cameras, provides a relatively affordable way to reconstruct the exterior for purposes such as infrastructure mapping and security. However, the challenges of changing illumination and weather conditions, along with diverse environments, make this a sophisticated problem. A primary issue that arises in Outdoor scene reconstruction is the resolution of the input images. Oftentimes, the images sourced from the on-site cameras and StreetView do not have a sufficiently high resolution to map all the details of the scene. This issue becomes even more prominent during night-time, which decreases the resolution even further and may cause additional noise to the images. This can be seen in Figure 1

Monocular Depth Estimation involves using only one image to estimate depth, a task which traditionally involves using two images. This is beneficial, as there are abundant sources of non-stereo images. However, this approach is also expensive due to the amount of learning involved. Some novel works done include the Depth Anything family([5], which applied convolutions to work on broad generalization, and [4], which used multi-scale feature fusion for better accuracy). [1] discusses several algorithms to enhance the resolution of nighttime images, while [3] discusses an approach specifically for nighttime depth imaging.

The DepthAnything models work extremely well for Day-time images, however, they suffer in night-time images due to the lower resolution and the presence of noise. Thus, to counteract this, this report describes two approaches. The first mixes the approach followed in [3] with the Depth Anything model, and the second utilises the various algorithms described in [1] to enhance the resolution of the night-time images before creating a depth map. The report follows the following layout. Section 2 describes the prior works

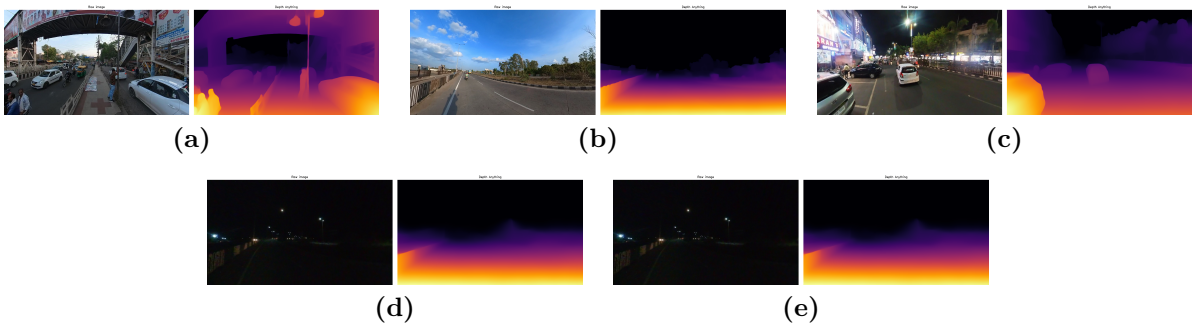


Figure 1: *Outdoor Depth Maps using DepthAnythingv1 for various scenes.* ***(a)*** *describes a busy daytime scene.* ***(b)*** *describes an empty daytime scene.* ***(c)*** *describes a busy nighttime scene.* ***(d)*** *describes an empty nighttime scene.* ***(e)*** *describes a noisy nighttime image. Notice the lack of details in the depth maps of nighttime images as compared to the daytime images, even in the case of busy scenes.*

mentioned here in detail, Section 3 describes the methodology used, Section 4 presents the results, and we conclude with Section 5.

2 Prior Work

We discuss the works discussed in Section 1 in detail here.

2.1 Depth Anything v1: Unleashing the Power of Large-Scale Unlabeled Data

Depth Anything v1 (DAv1)[5] is a very effective strategy for reliable monocular depth estimation. It was designed as a simple but influential foundational model that can be tuned to almost any kind of image. This involves the use of a data engine to gather and annotate a massive dataset containing about 62 million images, thus broadening coverage and improving generalization significantly. Major strategies are advanced methods of data augmentation which force the model to learn robust representations and additional supervision enabling the use of semantic priors from pre-trained encoders. Detailed zero-shot evaluations conducted on publicly available datasets and realistic images demonstrate the outstanding capability of DAv1 towards generalization. Fine-tuning with NYUv2 and KITTI datasets also resulted in achieving state-of-the-art performance, and its combination with depth-conditioned ControlNet brought significant improvements in depth estimation applications.

2.2 Depth Anything V2

Depth Anything v2 (DAv2)[4] is just an extension of the groundwork laid by the previous version since it emphasizes some essential methodologies that tend to improve upon monocular depth estimation without resorting to complex techniques. DAv2 is more accurate and robust than DAv1 since it replaces all the labeled real images with synthetic ones, thus increasing the potential of its teacher model, and trains student models extensively with pseudo-labeled real images. These updates make DAv2 significantly more efficient (over 10x faster) and accurate compared to models based on Stable Diffusion. It provides models across a wide range of scales, from 25M to 1.3B parameters, that support a wide range of applications while preserving strong generalization capabilities. The models achieve precise depth estimation through fine-tuning with metric depth labels. Thus, DAv2 provides a diverse benchmark for evaluation along with correct annotations to address the flaws found in existing test sets, thereby fostering further advancement within the field.

2.3 Nighttime Image Enhancement: A Review of Topical Concepts

[1] provides a detailed review of improvements in making night images better. Encountering problems such as bad lighting, uneven brightness, low contrast, noise, and strange colors to cover improvement in making nighttime images even better, the paper closely looks over twelve advanced algorithms explaining their main ideas, methods, and processing abilities. This also considers three performance measures, including the time it takes to process, in order to take into account these algorithms. That is information that can be used in telling the strengths and weaknesses. Thus, this paper provides a robust analysis of various techniques that can be used for nighttime resolution enhancement.

2.4 Unsupervised Monocular Depth Estimation for Night-time Images using Adversarial Domain Feature Adaptation

[3] deals with the problem of depth estimation for nighttime images. Proposing the gap between the daytime and the nighttime depth maps as a domain adaptation problem, they attribute this to the absence of a uniform light source in nighttime images. To solve this, they trained an encoder specifically for nighttime images using a PatchGAN-based adversarial discriminative learning method. The encoder is trained to generate nighttime features indistinguishable from those obtained from daytime images. The approach performed well on the Oxford night driving dataset, showing its efficiency.

3 Methodology

We propose two methodologies to solve the issues in nighttime depth estimation. They are described as follows.

3.1 Domain Adaptation based

Following [3], which considers the problem as a domain adaptation problem, we also decided to follow a similar approach for our first methodology. [3] first train an encoder-decoder type network (F_d, G_d) on daytime depth estimation. Then, they train a new encoder F_n with night-time images using adversarial learning that uses F_d as the generator and a PatchGAN discriminator. The final step involved using the new encoder F_n with the day-time decoder G_d for direct depth estimation for nighttime images. The overall workflow can be seen in Figure 2

To adopt the approach for our task, we decided to use DepthAnything. As the model already predicts day-time depth images robustly, we did not have to train or fine-tune it separately for the same. We made two copies of the encoder of the DepthAnything model. We trained one copy from scratch while fine-tuning the other one. We then used the existing decoder for the depth map generation tasks.

Considering the dataset, this approach was beneficial for two reasons:

1. The approach does not demand the dataset to have paired day and night images, as it aims to learn nighttime feature maps f_n generated from F_n, indistinguishable from daytime feature maps f_d generated from F_d. As our dataset is not paired, we were able to readily implement this approach without any modifications to the dataset.

2. We focus mainly on the second step of the approach, i.e., training the nighttime encoder F_n. It is an unsupervised approach, as it focuses primarily on bringing the invariant day and night feature spaces close. Thus, even with an unlabeled dataset, we were able to implement this approach.

A rough overview of approach 1 is given in Figure 3.

3.2 Nighttime Image Enhancement based

Our second approach aimed to improve the accuracy of depth map predictions for nighttime images by first enhancing the quality of these images. Nighttime images often suffer from low resolution, poor visibility, and

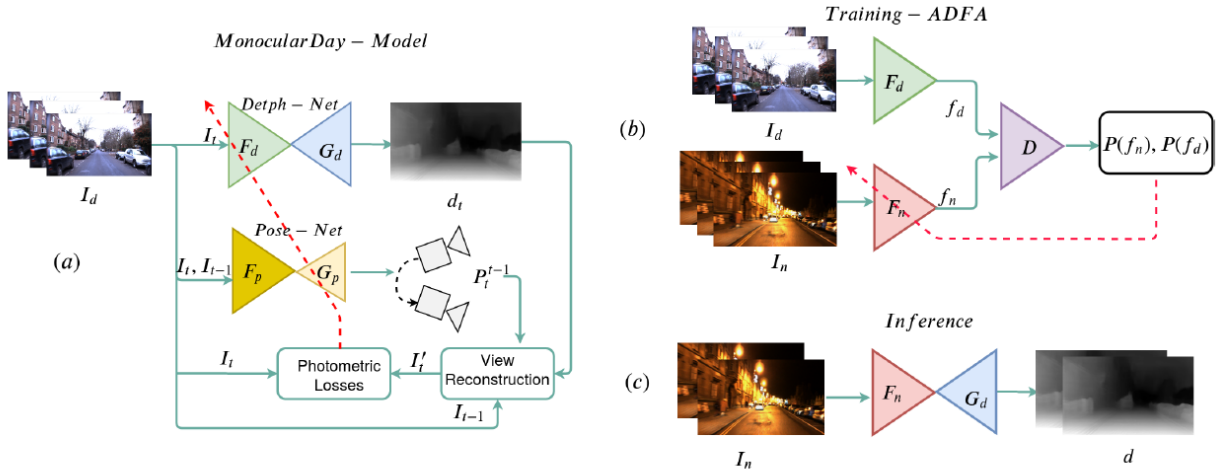


Figure 2: *Workflow followed by [3]. (a) describes training the model for monocular depth estimation for daytime images. (b) describes the adversarial training for the nighttime encoder, while (c) describes the inference process*

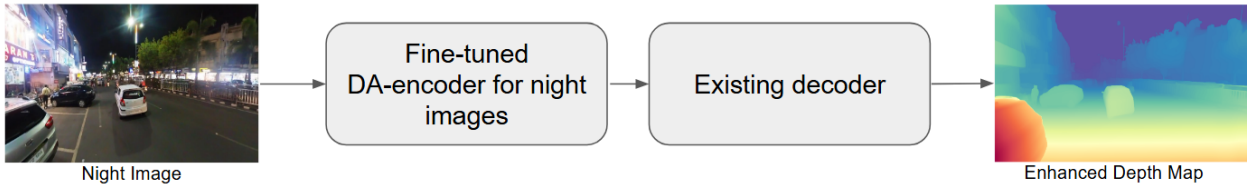


Figure 3: *Approach 1 Outline.*

lack of detail, which makes it difficult to generate accurate depth maps. To address this, we used several image enhancement algorithms described in [1] to process the images before feeding them into the depth prediction model.

The approach worked in two steps. First, the nighttime images were passed through various enhancement algorithms, which improved their brightness, contrast, and overall clarity. These enhanced images looked much clearer and contained more visible details compared to the original versions. Next, the enhanced images were sent to the DepthAnything model, which then predicted the depth maps.

This method had two significant advantages. First, it allowed us to evaluate multiple enhancement algorithms in the specific context of depth map generation, helping us understand which methods worked best. Second, it avoided the need for a labeled dataset, as the process relied only on the enhanced images and the depth prediction model, making it straightforward and efficient to implement.

A rough overview of approach 2 is given in Figure 4.

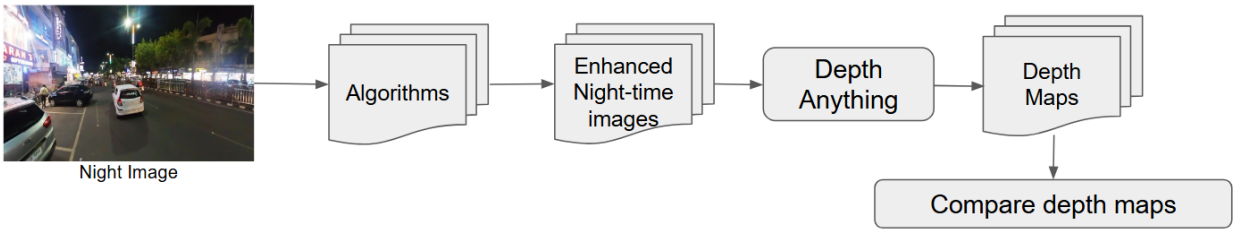


Figure 4: *Approach 2 Outline.*

3.3 Implementation Details

3.3.1 First Approach

For the first approach, the 'small' version of Depth Anything v1 was used. For the second approach, the 'small' version of Depth Anything v2 was used. Both models were sourced from HuggingFace. For both approaches, the same dataset was used. The dataset consists of day-night images collected from Bhopal, India, covering various urban environments during both daytime and nighttime. It contains around 10,000 images in total.

For the first approach, the encoder was trained for 15 epochs and fine-tuned for 40 epochs, with a batch size of 32. Uniform Xavier initialisation was used to initialise the encoder weights in case of training. The learning rate was set at 10^{-4} for both training and finetuning.

3.3.2 Second Approach

The enhancement stage involves applying a series of algorithms, each with unique mechanisms for improving image attributes. For instance, the Naturalness Preserved Enhancement (NPE) algorithm enhances images by refining non-uniform illumination while preserving their natural appearance. Similarly, the Fusion-Based Enhancement (FBE) algorithm uses morphological operations and adaptive histogram equalization to enhance contrast, while the Low-Light Image Enhancement (LIME) method relies on illumination maps created by identifying the brightest pixel values across color channels. Other algorithms, like the Bio-Inspired Multi-Exposure Fusion (BIMEF) and various Retinex-based approaches, further contribute by addressing different aspects such as noise reduction, exposure balance, and color preservation.

Once enhanced, the images are processed by a depth estimation model, such as Depth Anything V2, which generates depth maps for each scene. These depth maps represent the scene geometry, and their quality is expected to improve due to the superior input from the enhancement stage. To evaluate the effectiveness of this pipeline, the depth maps generated using the enhanced images are analyzed. By integrating these state-of-the-art enhancement techniques, the implementation not only improves the visual quality of nighttime images but also enhances the accuracy and reliability of the depth maps, making this a robust solution for nighttime depth estimation tasks.

4 Results and Discussion

The results achieved using each approach, and the relevant discussions are mentioned as follows.

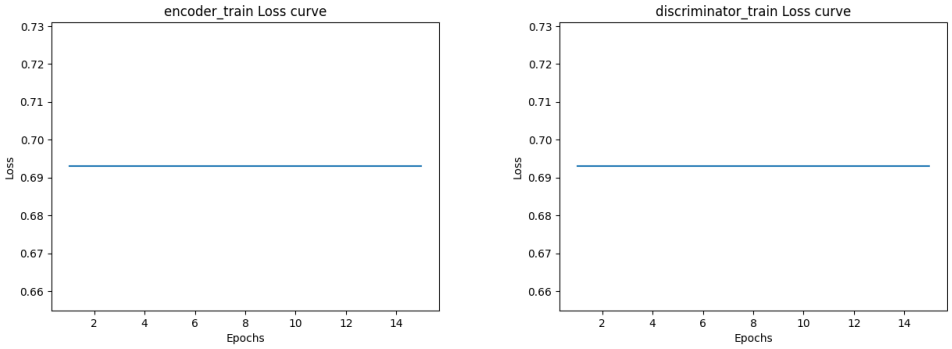
4.1 Domain Adaptation based

The loss curves for when the encoder was trained are in Figure 5, and when finetuned, are in Figure 6.

The results of using the trained encoder are in Figure 8a, and of using the fine-tuned encoder are in Figure 8b. For both of them, we present the results on a busy night scene. We do not present the other results as they are similar for the respective approach.

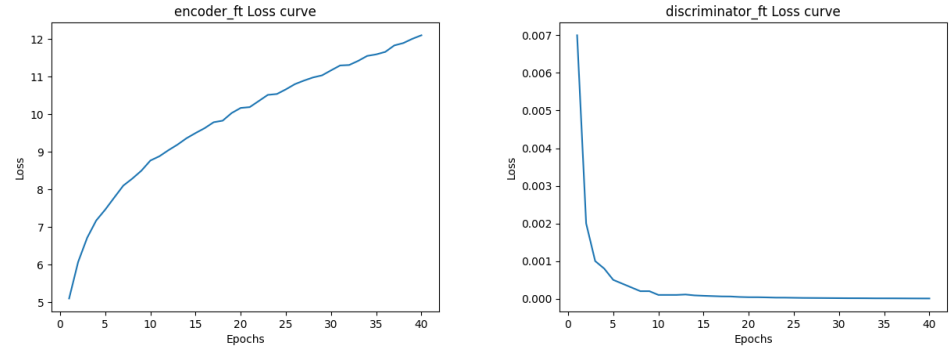
Discussion: The approach was not capable to produce any meaningful results. Using the fine-tuned encoder, we get no changes in the results, while the encoder trained from scratch fails to show any meaningful results. For the fine-tuned approach, the encoder loss increases while the discriminator loss continuously decreases. A possible reason for this is that the discriminator is too strong for the encoder. For the trained-from-scratch approach, both the encoder and the discriminator losses are stuck at the same value of 0.69, which is about ln(2), implying both of them are stuck against each other.

This behaviour may be explained using the size of the dataset and the encoder's architecture. The Depth Anything models use a DINOv2 backbone [2] as its encoder. DINOv2 was trained on several large datasets, such as LVD142M, ImageNet and a vast amount of curated web images. Our dataset, on the other hand, is relatively very small, which may help explain the encoder's performance. As it is not able to scale well on the small dataset, the discriminator is easily able to detect the difference between the invariant features of the daytime and the nighttime images, which cascades into the discriminator continuously improving and the encoder lagging behind. In the case of the training-from-scratch approach, the small size of the dataset



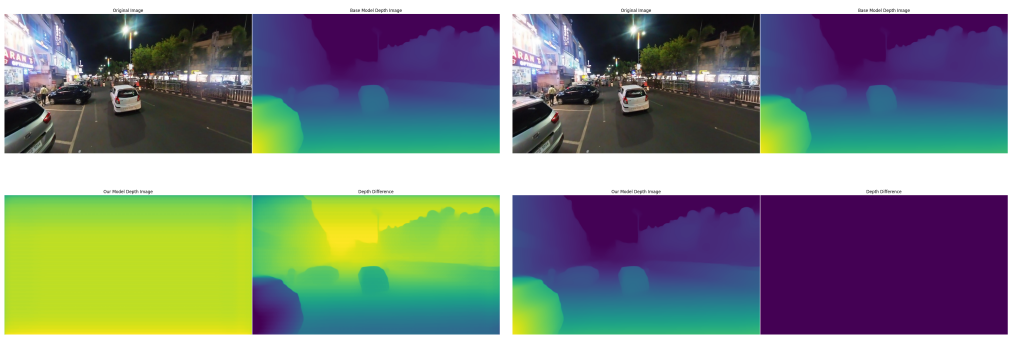
(a) *Encoder Loss Curve* (b) *Discriminator Loss Curve*

Figure 5: *Loss Curves for training the encoder*



(a) *Encoder Loss Curve* (b) *Discriminator Loss Curve*

Figure 6: *Loss Curves for finetuning the encoder*



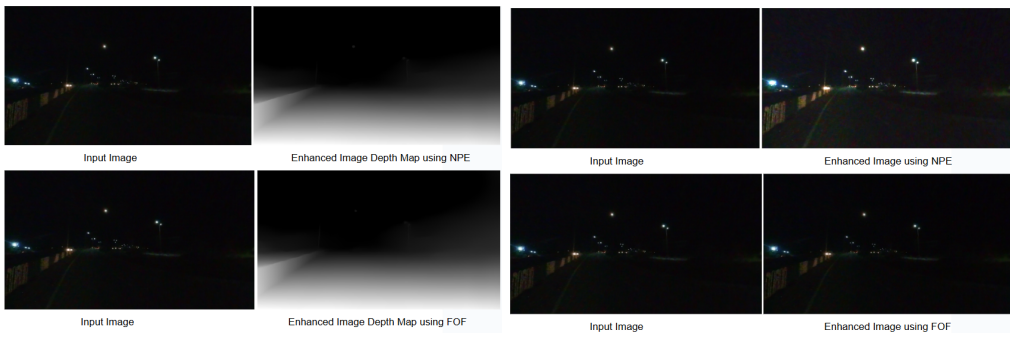
(a) *Trained Encoder* (b) *Fine-tuned encoder*

Figure 7: *Results on a busy night-time scene. Clockwise from top left: Actual scene, Depth map from Depth Anything v1, The difference in depth maps, Depth map from our encoder*

may have caused the stalemate between the encoder and discriminator, as neither of them is able to have a gain on the other due to the lower gradient.

4.2 Nighttime Image Enhancement based

Discussion: The Approach implements a diverse set of image enhancement algorithms, each employing unique methodologies to improve the visibility and detail of low-light images, particularly useful in nighttime photography and low-light scenarios. The **Naturalness Preserved Enhancement (NPE)** algorithm uses Gaussian blur to estimate illumination and logarithmic transformations to enhance both illumination and reflectance components for a balanced, natural appearance. **Low-Light Image Enhancement (LIME)** operates in the YCrCb color space, applying histogram equalization to the luminance channel to enhance brightness effectively. **Fusion-Based Enhancement (FBE)** combines morphological illumination estimation with CLAHE (Contrast Limited Adaptive Histogram Equalization) for detailed reflectance enhancement. The **Bio-Inspired Multi-Exposure Fusion (BIMEF)** algorithm blends the original image with a blurred version to enhance brightness and contrast harmoniously. The **Robust Retinex Model (RRM)** separates an image into illumination and reflectance components, reconstructing it through weighted combinations for improved clarity. Similarly, **Sequential Decomposition (SD)** employs Gaussian blur on the grayscale version for illumination estimation. **Fractional-Order Fusion (FOF)** enhances the image by combining denoising techniques with illumination subtraction. **Illumination Boost (IB)** applies logarithmic transformations to amplify the illumination, while **Adaptive Image Enhancement (AIE)** enhances the brightness channel through histogram equalization in the HSV color space. The **Camera Response Model (CR)** simulates higher exposure by scaling pixel values, and **Semi-Decoupled Decomposition (SDD)** builds upon SD with additional contrast preservation measures. Finally, the **Retinex-Based Multiphase (RBMP)** algorithm highlights details by subtracting estimated illumination from the image. Collectively, these algorithms provide a comprehensive toolkit for improving visibility in dark areas, recovering details lost in poor lighting, and facilitating comparative evaluations in academic research or practical applications. The results we claim on this approach are satisfactory to meet the primary purpose of enhancing the night-time images and generating a corresponding depth map associated with the input image.The best results were obtained on 3 Algorithms out of 12 algorithms implemented, **NPE**, **FOF** and **CR** algorithms showing off a significant enhancement in the highly noised night-time images, where as the other 9 algorithms were producing the results but not making up-to the expected results,all the results are based upon the Visual Turing of the outputs.



(a) *Input image vs enhanced image depth map* (b) *Input image vs enhanced image*

Figure 8: *Results on highly dark noisy night-time image along with the generated depth map of enhanced images of input image using various algorithms for enhancement*

5 Limitations and Future Work

Our implementation had certain limitations.

- For approach 1, we could have tried using the larger Depth Anything versions available on HuggingFace. We chose the small version due to lower training requirements.

- The evaluation was done based on perceptual differences between the generated depth maps, which are subjective in nature.
- For the first approach, an initial approach could have been to use a dataset with available ground truth depth maps and then finetune that approach for the provided dataset.
- More experiments could have been performed with hyperparameters, such as learning rates and batch size etc. For the second approach, more algorithms could have been tested out.
- For the second approach the various limitations we faced was majorly due to the rawness of the dataset. Amplification of Noise Explanation: Reflectance-based algorithms (e.g., RRM, RBMP) amplify noise during reconstruction, especially in low-illumination regions. Impact: Leads to grainy outputs and reduced image clarity. Example: Gaussian noise becomes more prominent after logarithmic transformations.
- In the 2nd approach due to noisiness of the dataset we encountered loss of Texture and Detail Explanation: Gaussian blur and morphological operations (e.g., SD, FBE) may blur fine textures, mistaking them for noise. Impact: Critical details, such as edges, are degraded or lost. Example: Light posts in a night sky may vanish during smoothing.
- While working over approach 2, the problem of Over-Saturation and Color Artifacts also arises due to low brightness contained images in dataset. Explanation: Brightness-enhancing algorithms (e.g., CR, AIE) can overexpose regions, causing color distortion and saturation. Impact: Results in unnatural colors and halos. Example: Noise-induced halos around light sources in dark images.

Future work can be focused on countering these limitations and, at the same time, combining both of these approaches to build a robust model capable of generating high-quality depth maps for outdoor scenes in both daytime and nighttime, which can then be followed by works on generating depth maps for adverse outdoor conditions such as extreme weather. We believe that a combination of these approaches if trained on a larger dataset and modeled in a more efficient manner, will certainly achieve satisfactory results.

References

[1] Ola Basheer and Zohair Al-Ameen. "Nighttime Image Enhancement: A Review of Topical Concepts". In: *SISTEMASI* 13 (May 2024), p. 1073. DOI: 10.32520/stmsi.v13i3.3938.

[2] Maxime Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. 2024. arXiv: 2304.07193 [cs.CV]. URL: https://arxiv.org/abs/2304.07193.

[3] Madhu Vankadari et al. *Unsupervised Monocular Depth Estimation for Night-time Images using Adversarial Domain Feature Adaptation*. 2020. arXiv: 2010.01402 [cs.RO]. URL: https://arxiv.org/abs/2010.01402.

[4] Lihe Yang et al. "Depth Anything V2". In: *arXiv:2406.09414* (2024).

[5] Lihe Yang et al. "Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data". In: *CVPR*. 2024.