Advancing Monocular Depth Estimation with Vision Transformers and Dynamic Post-Processing Techniques

https://github.com/Spospider/Advancing-Monocular-Depth-Estimation

Ali Eissa, Amr AbdelBaky, Omar Harb, Sara Mohamed
The American University in Cairo
AUC Avenue, New Cairo, Egypt
alieissa@aucegypt.edu, amrkhaled122@aucegypt.edu, omarharb@aucegypt.edu, sara_mohamed@aucegypt.edu

A. Abstract

Monocular depth estimation, the task of predicting dense depth maps from single RGB images, is pivotal for applications such as robotics, augmented reality, and 3D reconstruction. This work explores two complementary paths to improve accuracy: leveraging pre-trained Vision Transformers (ViT) and post-processing refinement techniques using traditional CV methods for the ZoeDepth framework. The ViT-based model serves as a proof-of-concept for leveraging transformer architectures in depth estimation, focusing on lightweight, low-resolution predictions suitable for resource-constrained environments. While its performance in terms of quality and inference time is limited, it provides valuable insights into the potential and challenges of using ViTs in this domain. For post-processing, we propose a series of refinement techniques for ZoeDepth, addressing its weaknesses in boundary precision and detail retention. These include an edge sharpening pipelines, superpixel segmentation for localized patch refinement, and high-frequency region enhancement to improve boundary precision and detail retention. With our approaches, evaluation on the NYU Depth V2 dataset demonstrate a notable improvement in metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Together, these approaches provide a diverse toolkit for advancing depth estimation methodologies.

B. Related Work

Monocular depth estimation has seen significant advancements in recent years, driven by the proliferation of deep learning techniques. Early works, such as those based on traditional computer vision methods like stereo matching, have been largely superseded by modern deep learning approaches.

One of the pioneering works in the field is MiDAS [5], which introduced a robust and efficient architecture for monocular depth estimation. MiDAS has been continuously improved, with versions like MiDAS v3.1 [6] achieving state-of-the-art performance on various benchmarks. However, MiDAS often struggles with spatial consistency and detail preservation.

To address these limitations, more recent works have explored innovative approaches. ZoeDepth [1] builds upon MiDAS, focusing on improving metric depth estimation and spatial consistency. While it achieves significant improvements over MiDAS, it still faces challenges in capturing fine details and handling complex scenes.

PatchFusion [4] introduces a tile-based framework for high-resolution depth estimation. By dividing the input image into patches and processing them independently, PatchFusion can capture finer details and improve overall accuracy. However, its computational cost is significantly higher compared to other methods.

In contrast, Marigold [**?**] leverages the power of diffusion models to generate high-quality depth maps. By training a diffusion model on a large dataset of images and depth maps, Marigold can produce highly detailed and realistic depth estimates. However, its computational complexity and potential for artifacts remain challenges.

As early approaches primarily relied on convolutional neural networks (CNNs), excelling at capturing local features but limited by their inability to model global context effectively. Recent innovations have introduced transformer-based architectures, such as Vision Transformers (ViTs) [2], which inherently possess a global receptive field, enabling them to capture long-range dependencies and complex scene structures.

Custom loss functions tailored for depth estimation, including the Structural Similarity Index Measure (SSIM) and gradient consistency loss, address key challenges like maintaining spatial coherence and preserving fine details. The

ZoeDepth models [3], built on ViT backbones, have demonstrated state-of-the-art performance in lightweight depth estimation. However, challenges remain in accurately delineating object boundaries and fine-grained details.

Our work aims to build upon these existing methods, exploring the feasibility of ViT-based architectures for depth estimation and content-aware post-processing techniques using traditional methods to achieve improved accuracy and efficiency compared to other ZoeDepth-based models.

C. Technical Approach

This project explores two complementary paths for depth estimation: a Vision Transformer (ViT)-based model and incorporating post-processing techniques on the existing ZoeDepth [1] framework with the aim of improving accuracy at a minimum processing cost as an alternative to other models also based on ZoeDepth.

C.1. Vision Transformer-Based Model

The `DepthEstimationModel` is built on a pre-trained ViT backbone (`vit-base-patch16-224`):

- **ViT Backbone**: Extracts global features from input images. The pre-trained parameters of ViT are partially frozen to leverage learned representations while fine-tuning the last transformer block.
- **Decoder Network**: A sequence of transposed convolutional layers upsamples the feature map into a single-channel depth map, ensuring high spatial resolution.
- **Upsampling Pipeline**: Reduces feature dimensions and reconstructs depth maps through transposed convolutions with ReLU activations and batch normalization.

C.1.1 Dataset and Pre-processing

- The **NYU Depth V2 dataset** was used for both of the approaches suggested in this paper, containing RGB images and corresponding depth maps. Images were resized to 224×224, normalized using ImageNet statistics, and augmented with color jittering to improve generalization.
- Depth maps were normalized to the range [0, 1] and interpolated to match the input resolution.

C.1.2 Training and Optimization

- **Optimizer**: Adam optimizer with a learning rate scheduler (`ReduceLROnPlateau`) to adjust learning rates based on validation loss.

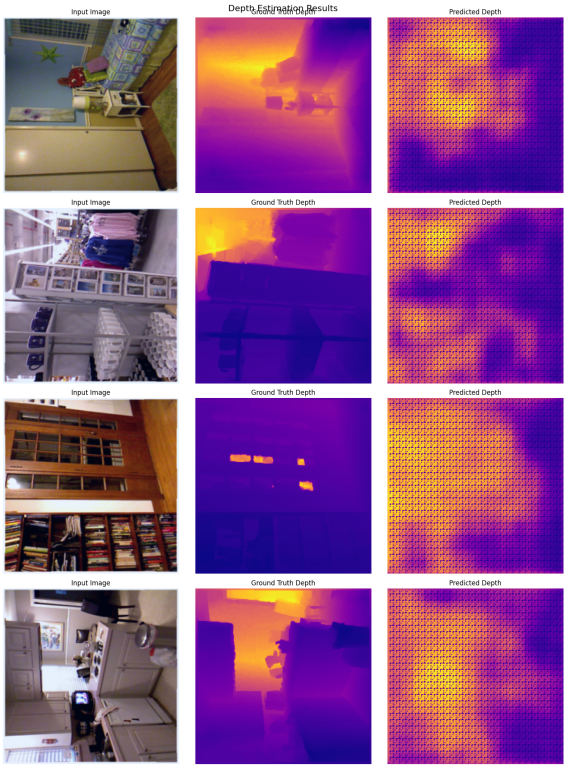

Figure 1. Depth maps predicted by the Vision Transformer-based model.

- **Training Loop**: Mixed precision training using PyTorch's AMP (Automatic Mixed Precision) for faster computations and lower memory usage.
- **Validation**: Performed after each epoch to monitor performance using the custom loss function.
- **Post-Processing for ZoeDepth**: Additional refinement steps, including edge detection, dilation, and blending, were applied post-inference.

C.1.3 Custom Loss Function

The `DepthLoss` function combines three components to ensure accurate depth prediction:

- **L1 Loss**: Minimizes pixel-wise differences between predicted and ground truth depths.
- **Gradient Loss**: Preserves depth gradient consistency for smooth transitions.
- **SSIM Loss**: Maintains structural similarity to the ground truth, crucial for preserving depth map details.

C.1.4 Evaluation Metrics

To evaluate the performance of the Vision-Transformer Based Model, two metrics were selected:

- **Average L1 Loss**: Otherwise known as MAE, this measures the average absolute difference between predicted pixel values and ground truth pixel values.
- **Average RMSE**: Otherwise known as MSE, this measures the square root of the average squared difference between predicted pixel values and ground truth pixel values.

Model	Average L1 Loss	RMSE
Vision-Transformer Model	0.2231	0.2898

Table 1. Evaluation Results for the Vision-Transformer Based Model

C.1.5 Insights

The Vision Transformer (ViT)-based approach for depth estimation demonstrated significant advantages in terms of speed and ease of training, particularly for the 224x224 model used in this study. The relatively small input size allowed for quick convergence during training, and the simplicity of the model architecture contributed to faster experimentation cycles. This made it ideal for applications where computational resources are limited, such as embedded systems or real-time robotics, where low-latency predictions are essential.

Looking to the future, the ViT-based model can be expanded in several ways. The current approach can be scaled up to handle higher-resolution images, which could further enhance depth estimation accuracy for more complex scenes while maintaining reasonable inference times. Additionally, adapting the ViT-based model for edge devices and low-power hardware is a promising direction for improving real-time performance without sacrificing quality.

A key avenue for future work is to leverage this approach for ultrafast, low-resolution depth predictions in embedded systems, such as drones, or wearable AR devices. Given the success of the ViT model in quick inference at lower resolutions, it could be particularly effective for applications that require fast, lightweight depth estimation while still capturing essential scene information. Future exploration could also focus on optimizing the model for diverse hardware platforms, enabling its deployment in resource-constrained environments while ensuring robust performance.

C.2. Refinement of the ZoeDepth Framework

The ZoeDepth framework employs a lightweight model for depth estimation. It builds upon the previously SoTA model *MiDAS* [5], which is considered the backbone of most depth estimation models today. MiDAS, despite its breakthrough at the time, produces relative values and not absolute depth estimation. ZoeDepth, on the other hand, measures depth in metric units [1]. However, it still suffers from issues pertaining to loss of sharp corners and details [6]. Aiming to address these issues without adding complexity, we implement some classical pre- and post-processing techniques.

1. **Post-Processing through Edge Sharpening Pipeline** To address the issue of "blobby" edges in ZoeDepth's output, we implement a pipeline for edge sharpening as a post-processing technique. It proceeds as follows:
   - **Basic Inference**: ZoeDepth predicts initial depth maps from monocular images (Fig. 2).
   - **Edge Detection and Masking**: Sobel operators identify object edges, which are dilated to create a mask for targeted refinement (Fig. 3).
   - **Min-Max Filtering**: A custom neighborhood filter enhances edge sharpness, improving structural details (Fig. 4).
   - **Blending**: Combines the filtered depth map with the original depth map using the edge mask, yielding enhanced boundary precision (Fig. 5).

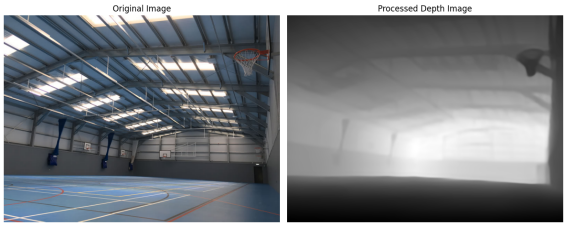

Figure 2. Initial depth map generated by ZoeDepth.

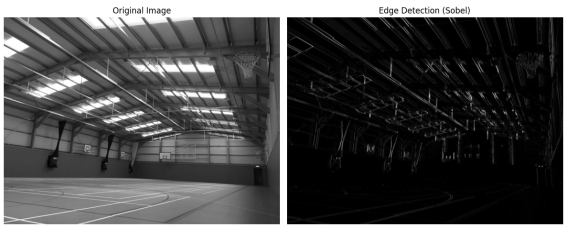

Figure 3. Edge detection results using Sobel filtering.

2. **Applying ZoeDepth on Superpixel-segmented Patches** The PatchFusion model [4] builds on

Figure 4. Dilated edge mask applied to enhance boundaries.

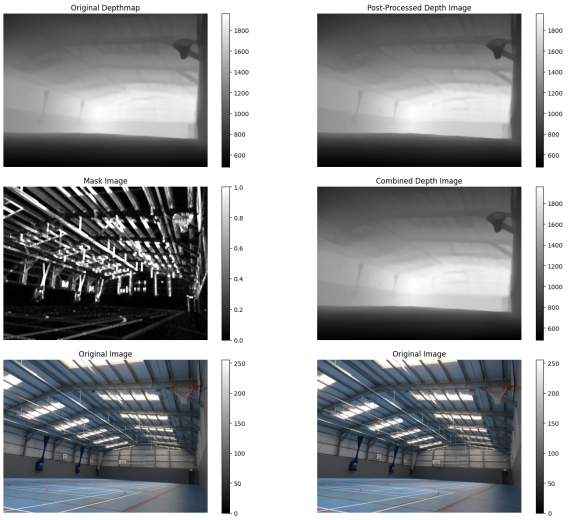

Figure 5. Combined depth map after edge sharpening pipeline.

ZoeDepth by applying it on 4x4 patches at a time rather than the entire image at once, then stitching the resulting depth maps. This yields more geometrically accurate results, at the cost of a computation time of 16x to 146x that of ZoeDepth. To this end, we attempt a different take on patches, by taking less patches, centered around blobs or clusters of the image, through superpixel segmentation.

- **Superpixel Segmentation**: The image is segmented into blobs based on superpixel similarity, which groups pixels together based on similarities in color and Euclidean distance (similar to blob detection). The SLIC algorithm in particular is used. 7
- **Applying ZoeDepth on Patches**: The segments are fed into ZoeDepth to yield separate depth maps for each segment.
- **Stitching it together**: The separate depth maps are stitched together to reveal the full image depth map.

Figure 6. Original image before undergoing patched ZoeDepth pipeline.

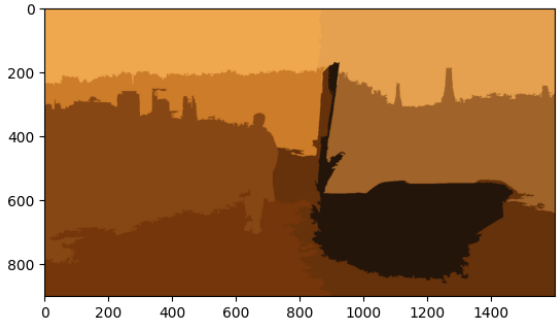

Figure 7. Image after SLIC segmentation.

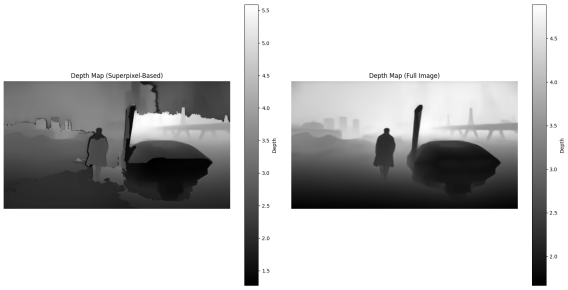

Figure 8. Stitched depth map compared to original output from ZoeDepth.

3. **Dynamic High-Frequency Grouped Patch Refinement** Following somewhat in the same vein as the previous experiment, this one also applies ZoeDepth on high-frequency clustered patches, integrating the result with the original ZoeDepth output for the whole image to yield more detailed results for high-frequency regions. This operates on the assumption that high-frequency regions correlate to subjects, details, or foregrounds of an image.
    - **Applying Sobel Filters**: The image is convolved with the two Sobel filters to extract edges and details, i.e. high-frequency regions.

- **Dynamically Selecting Patches**: Based on a threshold, the image then groups clusters of high-frequency pixels together. The patches are ensured to be non-overlapping.9
- **Applying ZoeDepth on Selected Patches**: ZoeDepth is applied on the patches extracted from the previous step.
- **Integrating Patch Results with Image Results**: The entire image is passed as usual to ZoeDepth, then combined with the results of the previous step to extract detailed depth map for detailed regions of the image 1011.

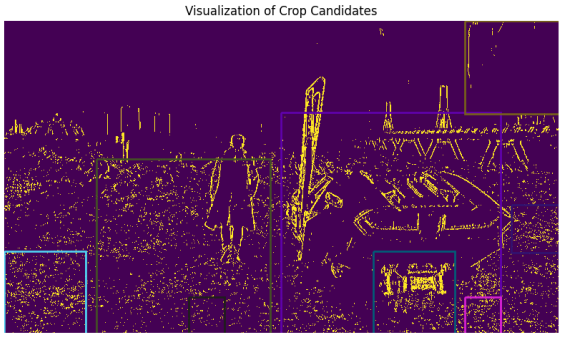

Figure 9. Clustered high-frequency patches.

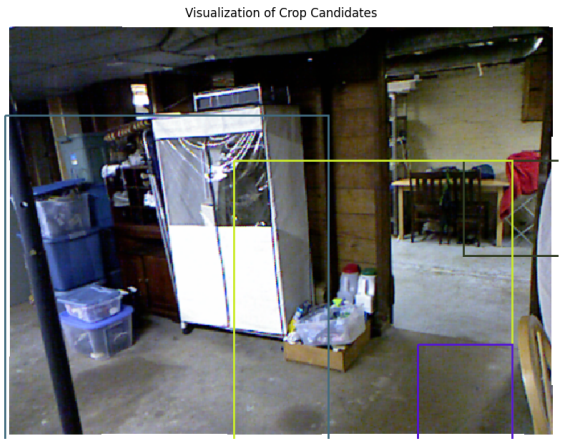

Figure 10. Patches selected from input image.

C.3. Insights from Depth Refinement Techniques

The ZoeDepth-based approach demonstrates a promising balance between accuracy and processing time through the integration of novel post-processing techniques and patch-based refinements. Each method contributes unique strengths and addresses specific limitations of the base ZoeDepth model:

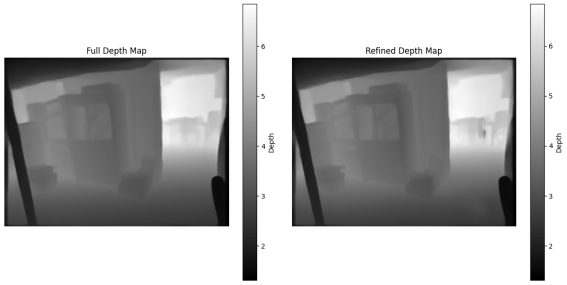

Figure 11. Results of high-frequency clustered patch refinement.

- **Post-Processing through Edge Sharpening Pipeline**: This technique effectively enhances the weak edge delineation inherent in ZoeDepth predictions. By leveraging the original input image to guide standard edge refinement techniques, we achieve improved boundary precision at a minimal computational cost. Additionally, this method is highly versatile, as it can be applied even when operating at lower resolutions. For example, predictions can be made on a downscaled image to reduce computational requirements, then upscaled back and post-processed using the original high-resolution input. This approach provides a low-cost solution to enhance inference results, especially for lightweight applications.
- **Applying ZoeDepth on Superpixel-Segmented Patches**: Inspired by the PatchFusion [4] methodology, this technique aimed to perform depth estimation on superpixel-segmented patches in a context-aware manner. However, despite the intent to improve local accuracy through spatially informed patch-based processing, this method did not produce satisfactory results. The inference process proved to be the slowest among all our experiments, and the predicted depth maps often lacked the desired consistency and precision, suggesting limitations in this approach for practical applications.
- **Dynamic High-Frequency Grouped Patch Refinement**: By dynamically identifying and processing high-frequency regions in the image, this method focuses computational resources on areas requiring greater detail. Patches containing high-frequency details are refined independently and then combined with the full-depth image result, enabling improved capture of finer details, particularly for regions further away in the scene. This approach strikes a balance between accuracy and efficiency, improving upon the inference time of PatchFusion while maintaining the ability to capture critical details. However, its reliance on predefined thresholds poses a challenge; future work should explore adaptive techniques to optimize these thresh-

olds based on the input image characteristics.

In summary, the ZoeDepth-based approach introduces a range of techniques with varying trade-offs in terms of accuracy and computational cost. While the edge sharpening pipeline offers a low-cost enhancement for general use, the dynamic high-frequency refinement method demonstrates the potential to balance detail preservation and efficiency. Future work should focus on optimizing parameters and further exploring adaptive methods to enhance the robustness and generalizability of these approaches.

C.4. Evaluation Metrics

The depth-enhancement techniques were applied to ZoeDepth "N" variant and evaluated on 50 images of the NYU Depth V2 dataset. Performance was evaluated using:

- **MAE (Mean Absolute Error)**: Measures average absolute error in depth predictions.
- **RMSE (Root Mean Square Error)**: Measures standard deviation of prediction errors, penalizing larger errors.

Model	MAE	RMSE
ZoeDepth Baseline	0.0973	0.1230
Edge Sharpening Pipeline	0.0983	0.1253
SLIC Segmentation Pipeline	0.1056	0.1373
High-Freq Patch Refinement	0.0959	0.1222

Table 2. Evaluation results for the processing techniques applied on ZoeDepth.

D. Conclusion and Future Work

This study explores two distinct approaches to monocular depth estimation, each with its own specific objectives and strengths. The first approach focuses on leveraging a Vision Transformer (ViT)-based model for lightweight, low-resolution depth estimation, suitable for fast, real-time applications like embedded systems and object avoidance. The second approach builds on the ZoeDepth model, refining depth prediction accuracy and enhancing spatial consistency through post-processing techniques and patch-based refinements. The depth-enhancement approach, augmented with edge sharpening, superpixel segmentation, and high-frequency patch refinement, focuse on improving depth map fidelity, particularly at object boundaries, while still maintaining manageable computational costs. These enhancements demonstrate that it is possible to achieve high accuracy without an extensive increase in processing time. Both approaches are designed to address different challenges, yet they complement each other by targeting various aspects of depth estimation.

The ViT-based approach provides a fast and efficient solution for depth estimation, emphasizing real-time performance at lower resolutions. This makes it ideal for applications where speed is crucial, such as autonomous navigation and robotics. In contrast, the ZoeDepth approach, augmented with edge sharpening, superpixel segmentation, and high-frequency patch refinement, focuses on improving depth map fidelity, particularly at object boundaries, while still maintaining manageable computational costs. These enhancements demonstrate that it is possible to achieve high accuracy without an extensive increase in processing time.

In conclusion, while the ViT-based approach excels in low-latency applications, and the ZoeDepth-based refinements enhance accuracy and detail in depth estimation, future work will focus on investigating hybrid models that combine the speed of ViT with the precision improvements of our explored depth-enhancement techniques. Additionally, further optimization of processing techniques, such as adapting thresholds for high-frequency patch refinement, will be explored to better suit different input images and scenes. These efforts aim to integrate the strengths of both approaches, enabling more efficient and accurate depth estimation across a variety of applications.

E. Statement of Individual Contribution

- **Ali Eissa**: Edge Sharpening Pipeline, Dynamic High-Frequency Grouped Patch Refinement.
- **Amr Abdelbaky**: ViT-based Depth Model, Surveying depth estimation model architectures.
- **Omar Harb**: ViT-based Depth Model, Surveying depth estimation model architectures.
- **Sara Mohamed**: Applying ZoeDepth on Superpixel-segmented Patches.

References

[1] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. 1, 2, 3

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. Available at https://arxiv.org/pdf/2010.11929. 1

[3] Andrey Ignatov, Grigory Malivenko, Radu Timofte, et al. Efficient single-image depth estimation on mobile devices, mobile ai & aim 2022 challenge: Report. *arXiv preprint arXiv:2211.04470*, 2022. Available at https://arxiv.org/pdf/2211.04470. 2

[4] Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patchfusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation, 2023. 1, 3, 5

[5] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer, 2020. 1, 3

[6] Patricio Gonzalez Vivo. The state of the art of depth estimation from single images. Medium article, 2023. Available at https://medium.com/@patriciogv/the-state-of-the-art-of-depth-estimation-from-single-images-9e245d51a315. 1, 3