Relative Depth Estimation from a Single Outdoor Image

1 Problem Definition

The goal of this project is to estimate a relative depth map $D(x, y)$ from a single high-resolution RGB image $I(x, y, 3)$ using self-supervised deep learning. The input is a coloured image $I \in \mathbb{R}^{H \times W \times 3}$, where each pixel has three values (red, green, and blue), and the output is a single-channel depth map $D \in \mathbb{R}^{H \times W}$, where each pixel's value represents its relative proximity rather than an absolute distance. The network learns to map the 3D spatial-colour input to a 2D spatial-depth output purely from appearance cues, without training through multiple images or ground-truth depth labels. This builds on the idea that depth can be inferred from monocular images [2], expanding the feasibility of understanding 3D scenes in scenarios such as robotics and AR/VR. A successful outcome would demonstrate clear separation between foreground and background based on visual structure, that is evaluated qualitatively through depth map visualization and quantitatively by consistent minimization of self-supervised losses during training.

2 Method

The method consists of three main components: input preprocessing, a patch-based supervision strategy, and a self-supervised loss-trained convolutional neural network (CNN). The design draws inspiration from previous work on monocular depth estimation [2], single-image generative learning [4], and self-supervised classification losses [1].

Preprocessing: The input image $I(x, y, 3)$ was first resized to 512×512 pixels to standardize dimensions and reduce computational cost. During training, random crops of size 480×480 were extracted from the resized image to introduce positional variability. In addition, colour jittering was applied with random adjustments to brightness, contrast, saturation, and hue. This allowed the model to experience slightly different views of the scene, mitigating overfitting.

Patch-Based Supervision: During training, random pairs of patches were extracted from the image. A weak geometric prior was applied, assuming that patches located lower in the image are closer to the camera. Each patch pair (i, j) was assigned a supervision label s_{ij}, where $s_{ij} = 1$ if patch i should be closer to patch j, and $s_{ij} = -1$ otherwise. This self-supervised strategy eliminates the need for ground-truth depth labels and is motivated by assumptions used in outdoor scene modeling [1].

Model Architecture: An encoder–decoder CNN was designed to predict a normalized single-channel depth map $D(x, y)$ from the RGB input. The encoder consists of three convolutional layers with ReLU activation and downscales feature maps through strided convolution, while the decoder uses transposed convolutions to upsample the feature maps back to the input resolution. The output depth map $D \in [0, 1]^{H \times W}$ represents relative distance, wherein higher values correspond to closer regions. The feasibility of learning complex structures from only a single input was demonstrated in SinGAN [4], motivating the design of a compact CNN in this project.

Loss Functions: The network was trained using a combination of ranking loss and smoothness loss. The ranking loss enforces correct depth ordering between patch pairs:

$$\mathcal{L}_{\text{rank}} = \sum_{(i,j)} \log(1 + \exp(-s_{ij}(d_i - d_j)))$$

where d_i and d_j are the predicted depths in the centers of the patches i and j. Additionally, an edge-aware smoothness loss encourages the depth map to be spatially coherent, particularly in regions with little image texture:

$$\mathcal{L}_{\text{smooth}} = \sum \|\nabla D\| \, e^{-\|\nabla I\|}$$

The final loss combines both terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rank}} + \lambda \mathcal{L}_{\text{smooth}}$$

where λ is a weighting factor set to 0.1 in this project.

3 Results

Over the course of 200 training epochs, the model demonstrated consistent improvement, with the total loss decreasing from 0.6959 at epoch 0 to 0.5132 at epoch 150.

The final predicted depth map (Figure 1) clearly separates foreground and background regions. Bright areas correspond to closer objects, such as the ground and the building base, while dark areas indicate more distant regions like the sky. Overlaying the predicted depth map onto the original RGB image (Figure 2) further illustrates the alignment between predicted depth transitions and actual scene structures. Additionally, edge comparisons (Figure 3) between the original image and the depth map show that depth discontinuities are well-aligned with important structural boundaries.

While supervised methods can produce very detailed monocular depth maps across different scenes, this project took a different approach by avoiding any external supervision, datasets, or pre-trained networks. As a result, the predicted depth map is understandably coarser and less detailed. However, given the extreme constraint of learning from just a single image, the network was still able to capture the overall scene structure, especially the relative layering between the sky, building, and ground. This shows that self-supervised depth estimation can work even when data is extremely limited. Based on how the losses were designed, removing either the ranking loss or the smoothness loss would likely cause the depth maps to become noisier or lose important boundaries.

4 Reflection

The model architecture and implementation were developed independently for this project. The loss functions and supervision strategies were inspired by prior work on self-supervised depth estimation [1] and single-image generative learning [4]. Previous evaluations of monocular depth estimation techniques [3] show that even weak self-supervision can yield meaningful scene layering, consistent with the results observed here. This project demonstrates that relative depth estimation can be achieved from a single image using self-supervised learning, expanding access to 3D scene understanding in low-data environments. By removing the dependency on large datasets or ground-truth labels, the method benefits users in resource-constrained settings. However, reliance on a vertical prior introduces bias, limiting generalization to indoor, aerial, or atypical scenes. The deployment of such systems without careful validation could result in navigational errors, especially in robotics and autonomous systems. Future improvements should focus on reducing such biases and developing more scene-adaptive self-supervised signals to ensure a broader and fairer application.

5 Conclusion

This project demonstrated that relative depth estimation can be achieved from a single outdoor image using self-supervised learning techniques. The results showed that meaningful depth relationships, such as separation between foreground and background regions, can emerge without any ground-truth labels. Future work could focus on applying the method to more diverse scenes by improving structural detail and developing self-supervised strategies that reduce reliance on geometric priors.

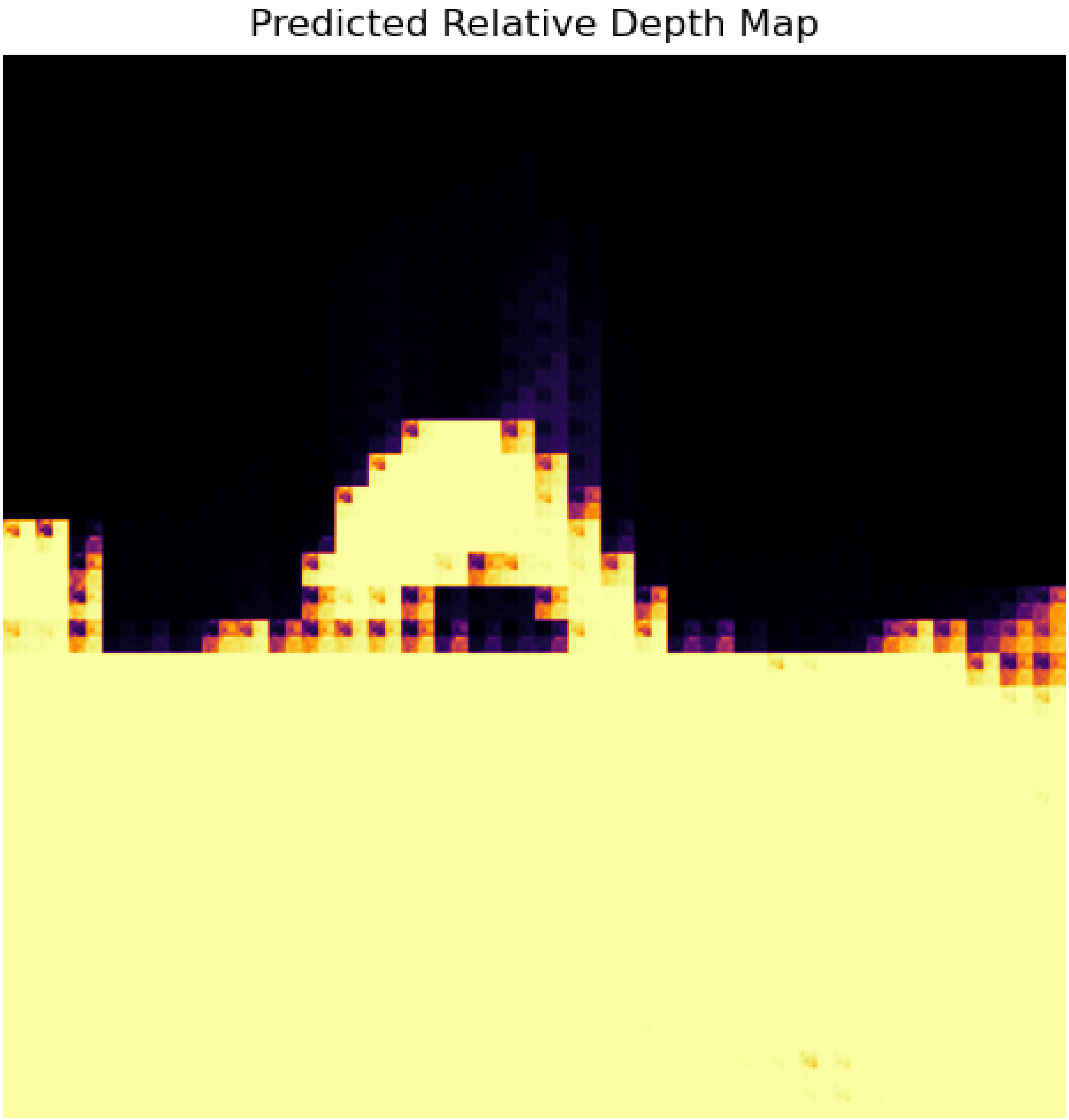

Figure 1: Predicted relative depth map generated after training. Brighter areas indicate closer regions, while darker areas correspond to regions further away.

References

[1] Clement Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[2] Lei He, Guanghui Wang, and Zhanyi Hu. Learning depth from single images with deep neural network embedding focal length. *IEEE Transactions on Image Processing*, 2018.

[3] N. Padkan, P. Trybala, R. Battisti, F. Remondino, and C. Bergeret. Evaluating monocular depth estimation methods. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2020.

[4] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

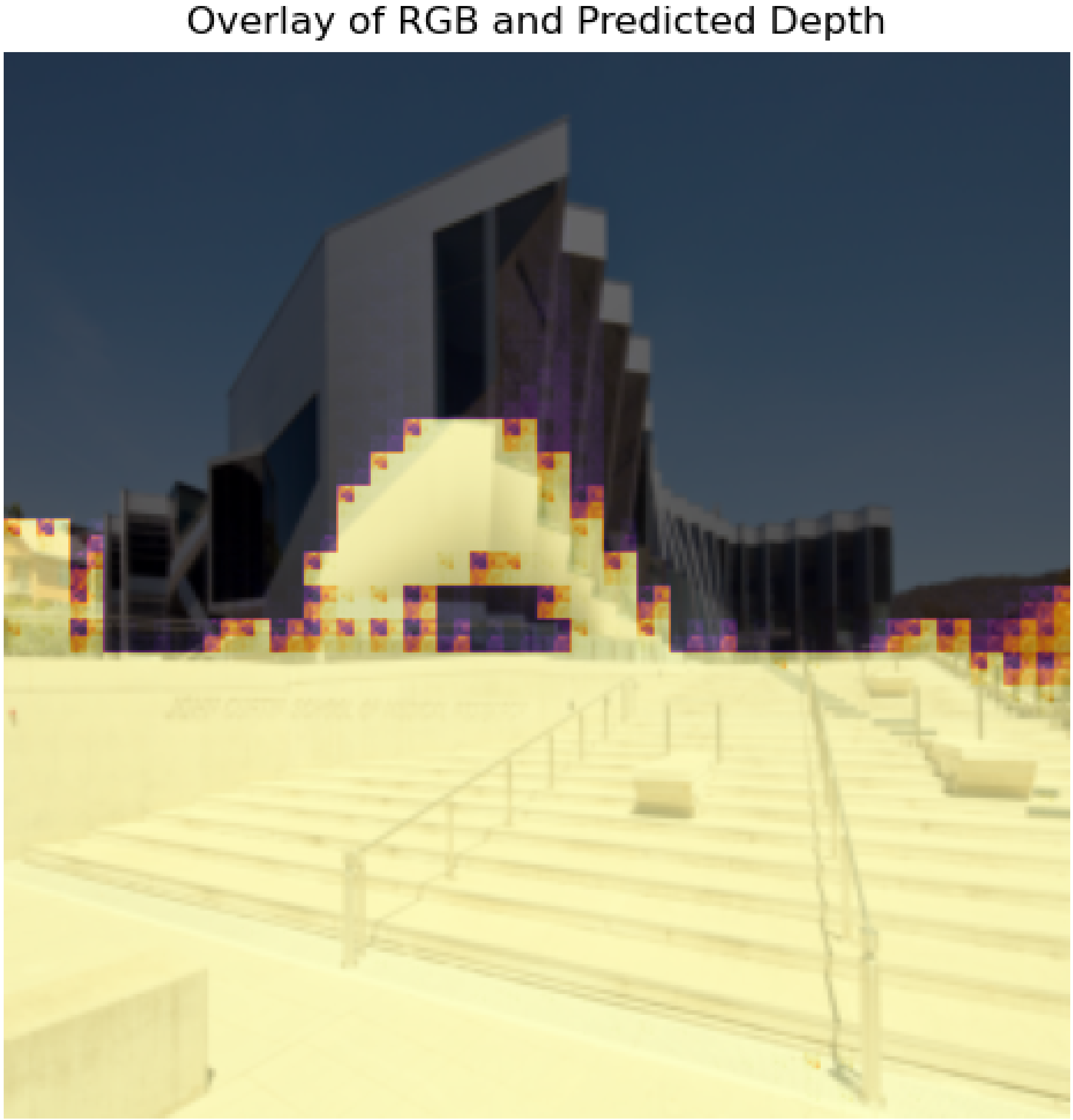

Figure 2: Overlay of the predicted depth map on the original input image, showing alignment between depth transitions and scene structures.

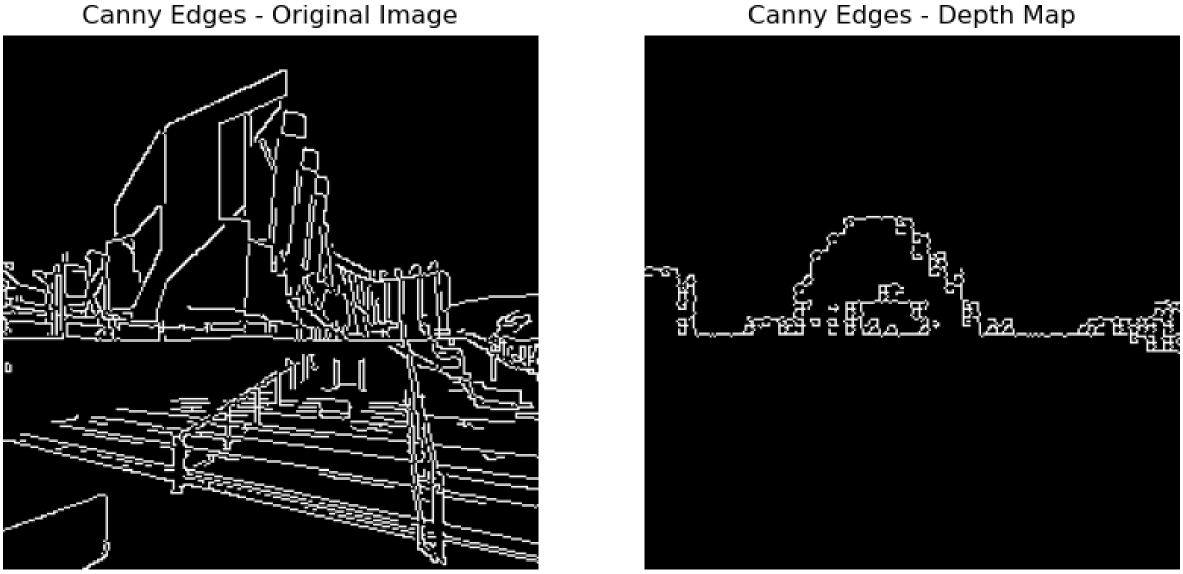

Figure 3: Edge comparison between the input image and predicted depth map, demonstrating that depth discontinuities align with important structural boundaries.