Approximating Classical Feature Detectors using CNN Trained on a Single Image

1 Problem Definition

This project explores the feasibility of training Convolutional Neural Networks (CNNS) to replicate the functionality of classical computer vision algorithms, specifically Harris Corner Detection and Canny Edge Detection, under the constraint of using only a single image for training. Given patches extracted from one image of the John Curtin School of Medical Research, the goal was to train a CNN to predict corresponding Harris corner likelihood heat-maps and Canny edge probability maps. This investigation tests the limits of deep learning with minimal data, examining the model's ability to approximate complex functions versus achieving generalizable feature understanding. Success was defined by the visual similarity of the CNN outputs to the classical algorithms' results on the source image, quantitative metrics reflecting this similarity, and clear evidence of learning (and over-fitting) demonstrated by the training loss curves.

2 Method

The process involved ground truth generation, data preparation from the single image, model architecture definition and training.

- **Ground Truth Generation**

The Canny edge map was generated using OpenCV's 'cv2.Canny' function with thresholds selected to capture prominent edges. For Harris corners the code from lab 3 was used which consisted a manual implementation of Sobel filters which computed image gradients (Ix, Iy), the structure tensor components (Ix^2, Iy^2, Ixy) were calculated and smoothed using Gaussian filtering (sigma=1.0), and the Harris response (R) was computed using $R = \det(M) - k \times (\text{trace}(M))^2$ with k= 0.04. Non-maximum suppression (NMS) was applied to the response map to identify keypoint locations (shown in Figure 1). For CNN training, the raw Harris response map was clipped and normalized to [0,1] to serve as the target heat-map, while the Canny map was converted to a binary float map 0.0, 1.0. Figure 2 shows the input image and generated ground truth maps.

- **Data Preparation**

The single grayscale input image (figure 2a) and its corresponding Harris and Canny ground truth maps were divided into overlapping patches. To artificially increase the dataset size, geometric augmentations (random rotations and flips) were applied, ensuring the same transformation was applied to an input patch and its corresponding ground truth patches (see Figure 3). The resulting pool of patches was randomly split into training (80%) and validation (20%) sets. Custom PyTorch 'Dataset' and 'DataLoader' classes were implemented to serve batches of (input, Harris ground truth, Canny ground truth) tensors during training.

- **Model Architecture and Training**

Two separate U-net models, a standard encoder-decoder architecture with skip connections suitable for image-to-image tasks, were employed - one for Harris prediction and one for Canny prediction. Both models took single channel grayscale patches as input and outputted single channel maps via a Sigmoid activation. The models were trained using the Adam optimizer for 100 epochs. Mean Squared Error loss was used for the Harris heat-map regression task, and Binary Cross-Entropy loss was used for the Canny edge probability prediction. Model weights yielding the lowest validation loss for each task were saved to mitigate overfitting.

3 Results

The training and validation loss curves for both model are shown in Figure 4. Both models exhibited rapid initial learning, with losses decreasing significantly in early epochs. However, clear signs of overfitting started to emerge. The Harris model's validation loss plateaued after around 40 epochs, while the Canny model's validation loss began to consistently increase after approximately 30 epochs, indicating that continued training was degrading performance on unseen patches from the source image. The best recorded validation loss for Harris was approximately 0.0001 and for Canny was approximately 0.015.

Figure 5 presents the final predictions on the full image compared to the ground truths. The CNN's predicted Harris heat-map (see Figure 6) visually differs significantly from the classical ground truth. The CNN prediction activates strongly along edges and structural lines, creating a response more akin to an edge detector than a sparse corner map. Consequently, applying the same NMS procedure used on the classical map to this CNN heatmap failed to yield meaningful, sparse keypoints. The final MSE of 0.0795 reflects the visual difference between the predicted heatmap and the target ground truth heatmap.

In contrast, the Canny CNN prediction, thresholded at 0.5, provides a remarkably clean and complete edge map, visually resembling a standard Canny output much more closely. The visual success if further corroborated by high quantitative metrics as shown in Table 1. These metrics, while only valid for this single image, suggests the Canny CNN successfully learned to approximate the target function for this specific input.

4 Reflection

This project highlights the profound impact of data limitations in deep learning. While the CNNs, particularly the Canny model demonstrated a strong ability to approximate the target functions on the source image, this represents highly specialized memorization rather than general understanding. The models learned patterns specific to the John Curtin School building's architecture, viewpoint, and lighting conditions present in the single training image. The Harris CNN, specifically learned a representation emphasizing edges rather than distinct corner peaks suitable for NMS, diverging from the classical algorithm's typical characteristic. Consequently, these models possess no capability to generalize to other images or objects. The primary ethical consideration is avoiding misrepresentation of these results as demonstrating general feature detection. This is primarily serving as an education tool to demonstrate overfitting. The key takeaway is the stark difference between function approximation on seen data and generalizable learning, which requires diverse datasets.

5 Conclusion

Convolutional Neural Networks were successfully trained to approximate Harris Corner and Canny Edge Detection using only patches derived from a single image. The Canny edge prediction model achieved high visual fidelity and quantitative scores on the source image, while the attempt to replicate Harris corner detection resulted in a CNN that learned structural features but produced an edge-like heat-map unsuitable for standard corner extraction via NMS. Both models exhibited significant overfitting, demonstrating the critical limitation imposed by the single-image training constraint and the highlighting that the models learned image-specific approximations rather than generalizable feature detectors.

Table 1: Canny quantitative metrics.

Method	Pixel Acc	Edge IoU	Precision	Recall	F1-score
Canny	**0.9982**	**0.9577**	**0.9795**	**0.9772**	**0.9784**

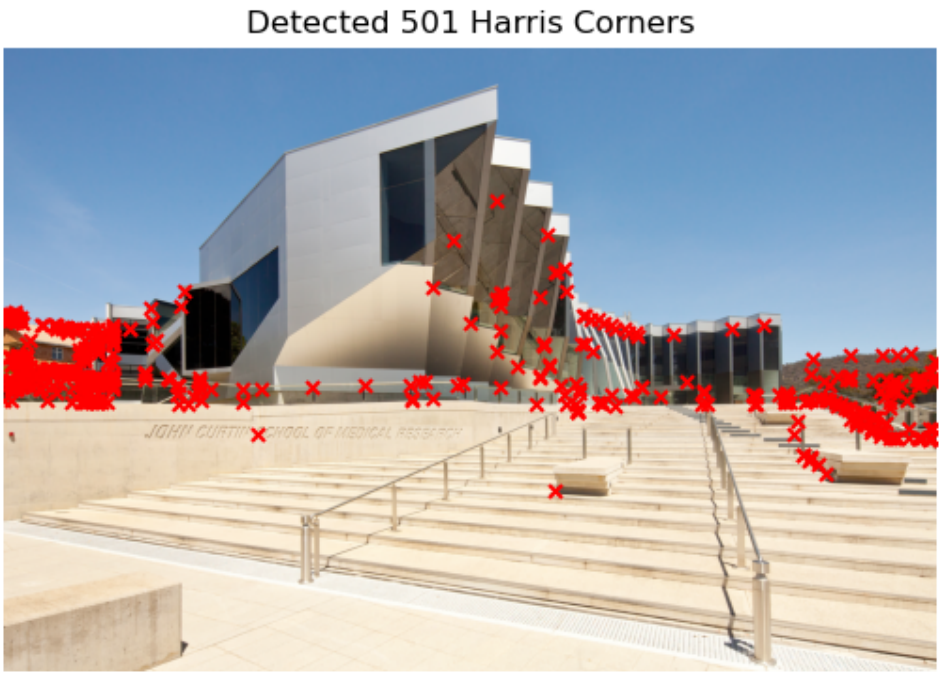

Figure 1: Harris corners detected.

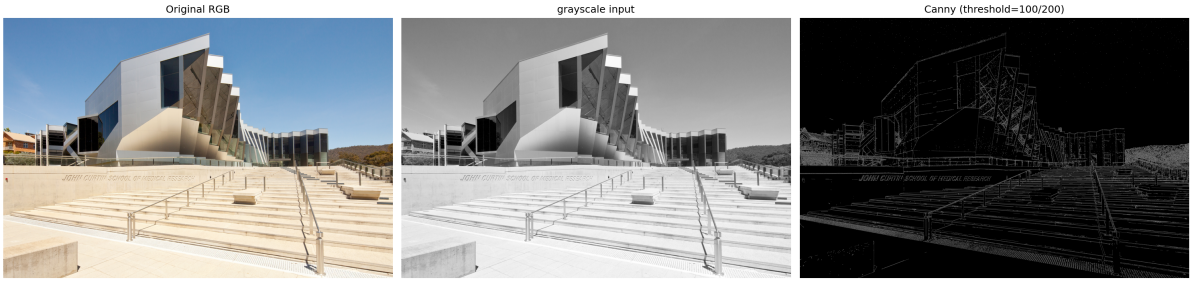

Figure 2: Canny edge ground truth.

[1]

References

[1] Philipp Fischer Olaf Ronneberger and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Computer Science Department and BIOSS Centre for Biological Signalling Studies, University of Freiburg, Germany*.

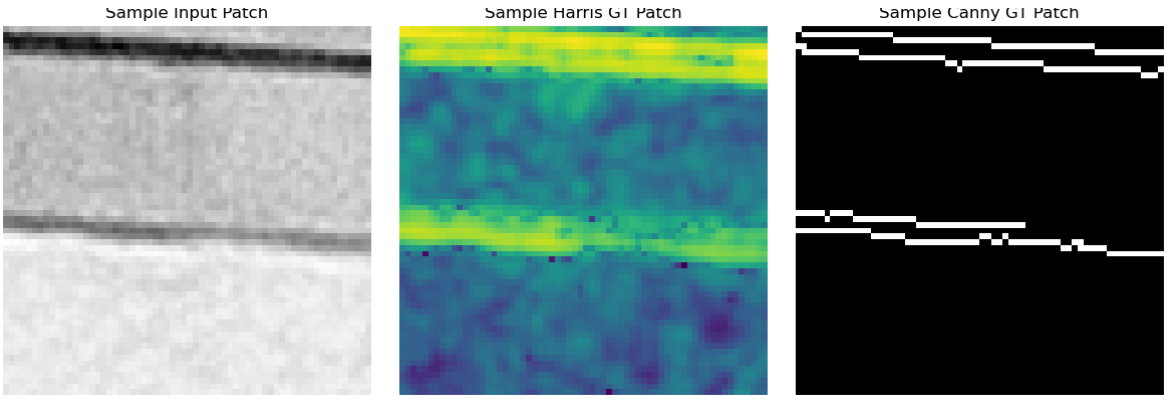

Figure 3: Sample batches.

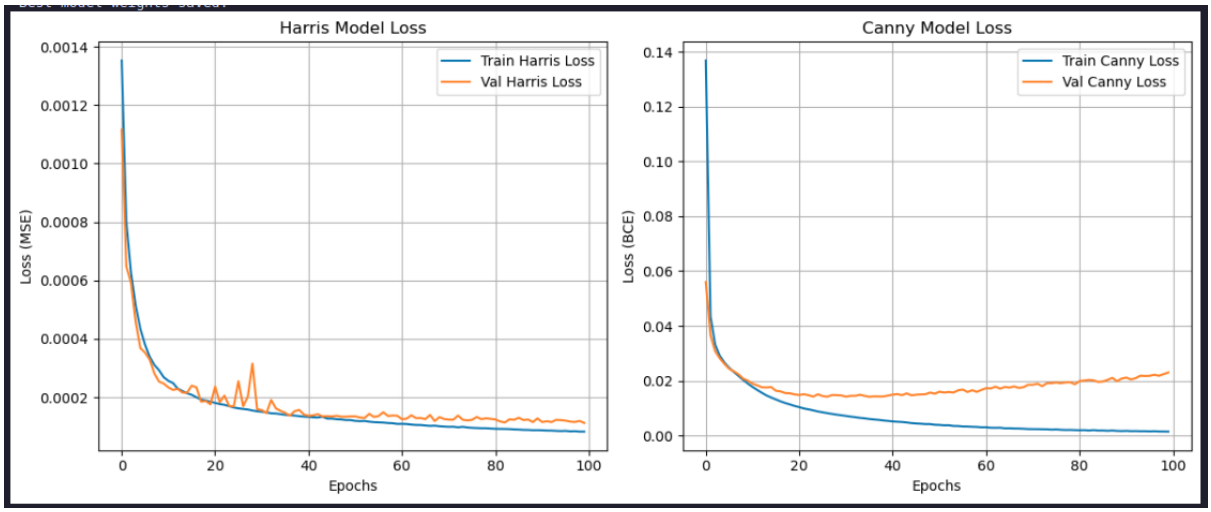

Figure 4: training and validation loss curves.

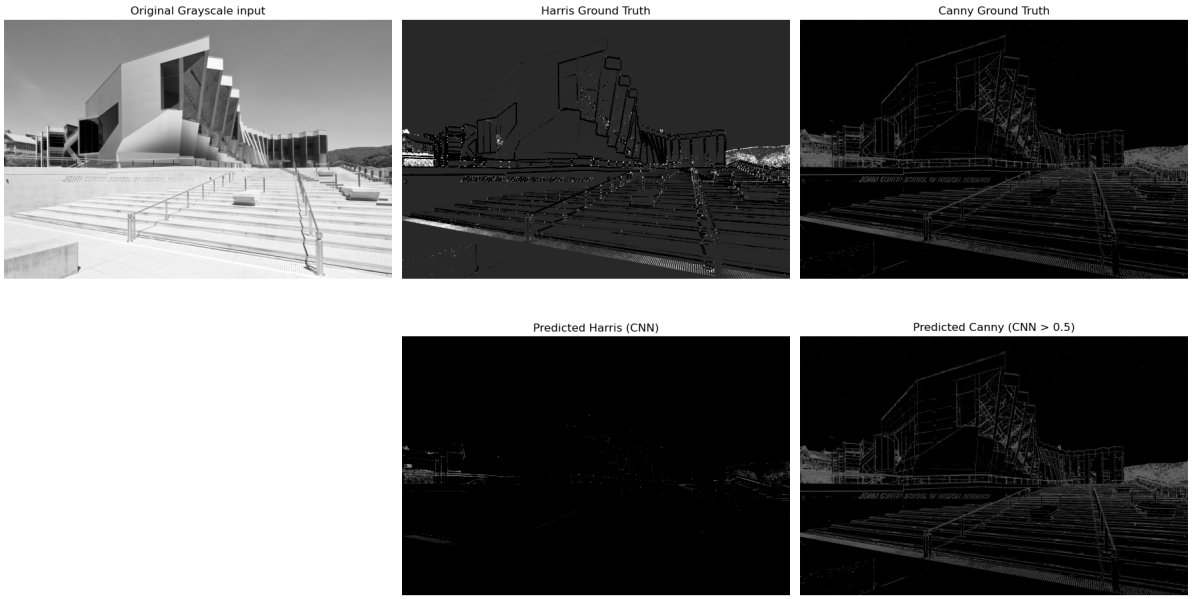

Figure 5: .

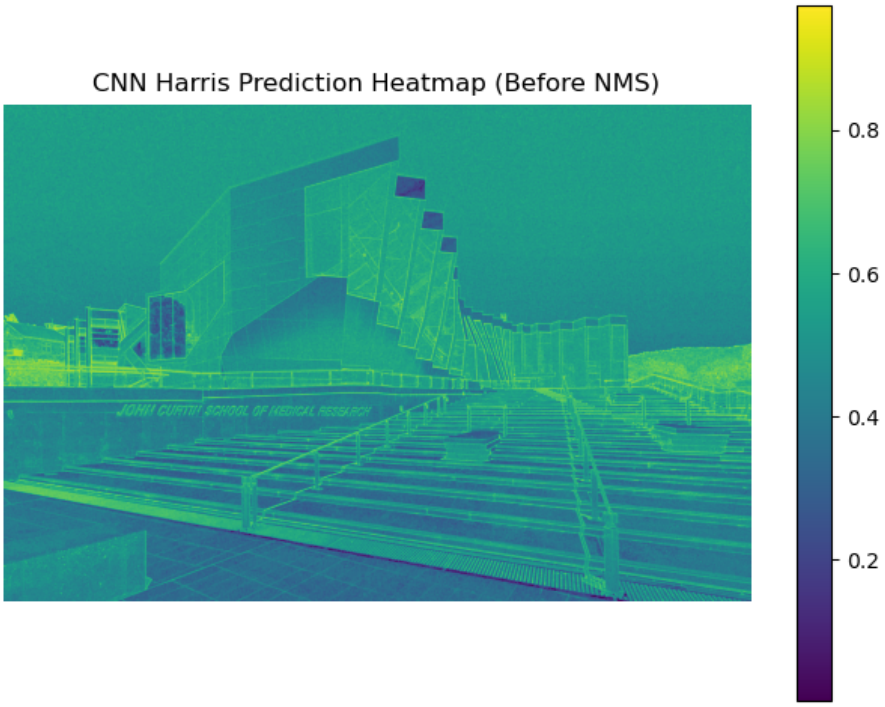

Figure 6: CNN Harris Prediction Heatmap (before nms)

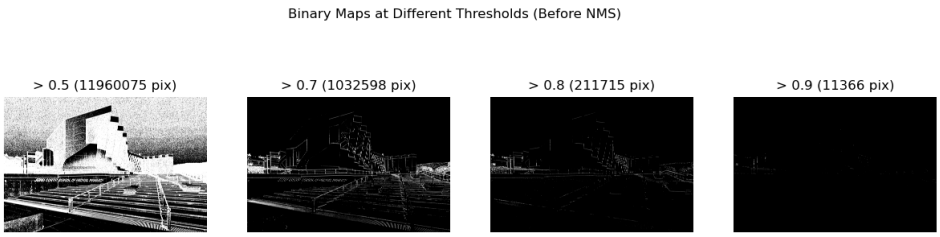

Figure 7: Binary maps at different thresholds (before nms)