CAB420 – ASSESSMENT 1B

Milly Chambers – n11318546

Question 1 – Person Re-Identification

Preprocessing

To ensure consistency between both the non-deep learning and deep learning methods, a uniform pre-processing pipeline was applied to all images in the dataset. The dataset consists of images sized at 128x64 pixels, which were resized to 64x32 pixels. In doing so, computational efficiency is balanced with the retention of sufficient spatial information for feature extraction. This allows for smaller resolutions to expedite processing without significantly compromising discriminative features like body proportions or clothing patterns.

For the non-deep learning approach, RGB images were converted to grayscale. This simplification aligns with the linear assumptions of PCA and LDA, as it reduces the feature space to a single channel, mitigating noise form lighting variations and focusing on structural patterns. The deep learning method retained the RGB channels, which allowed for the convolutional neural network (CNN) to leverage colour information as another key discriminative feature. This divergence in pre-processing reflects the differing capabilities of the methods, where traditional techniques benefit from simplified inputs whereas deep learning models can exploit richer and higher-dimensional data.

Method Selection and Justification

Non-Deep Learning Approach: PCA and LDA

This approach combined Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to project the high-dimensional image data into a lower-dimensional subspace for optimal classification. PCA was applied to first reduce dimensionality whilst preserving 95% of the variance. LDA was then used to further transform the data to maximise separation between different identities, allowing class labels to be leveraged for enhanced discriminability. Matching was performed using the Euclidean distance between gallery and prove embeddings, as it is effective for small to medium sized datasets.

This approach was used for its computational efficiency and interpretability. PCA and LDA are well-established techniques for biometric tasks, especially when training data is limited such as in this case where there is only 5,933 images. Their linear nature makes them computationally lightweight, which enables rapid prototyping and deployment in resource-constrained environments. However, due to their reliance on linear projections, it limits their ability to complex non-linear relationships in the data. This can potentially hinder performance in scenarios with significant variants in pose, lighting or occlusion.

Deep Learning Approach: Siamese Network with Contrastive Loss

The deep learning method was implemented through a Siamese network architecture, consisting of a shared CNN backbone that generated 128-dimensional embeddings for input images. Contrastive loss was used for training, minimising the distance between embeddings of the same identity whilst maximising the distance between different identities. Through this approach, the embedding space for similarity measurement is directly optimised, aligning well with the ranking-based objectives of person re-identification.

The Siamese network was chosen for its ability to learn non-linear feature representations, which are critical for handling the variability inherent in real-world re-identification scenarios. Unlike the PCA and LDA approach, the CNN backbone can extract hierarchical features automatically from edges and textures to high-level patterns such as clothing or accessories without relying on handcrafted preprocessing. Though this method requires greater computational resources and training time, its flexibility and robustness to intra-class variations often justify the trade-off, especially in cases with larger datasets or more complex environments.

Method Analysis

Quantitative Comparison

Method	Top 1 Accuracy	Top 5 Accuracy	Top 10 Accuracy
PCA+LDA	1.99%	8.64%	13.62%
Deep Learning	22.59%	50.5%	68.11%

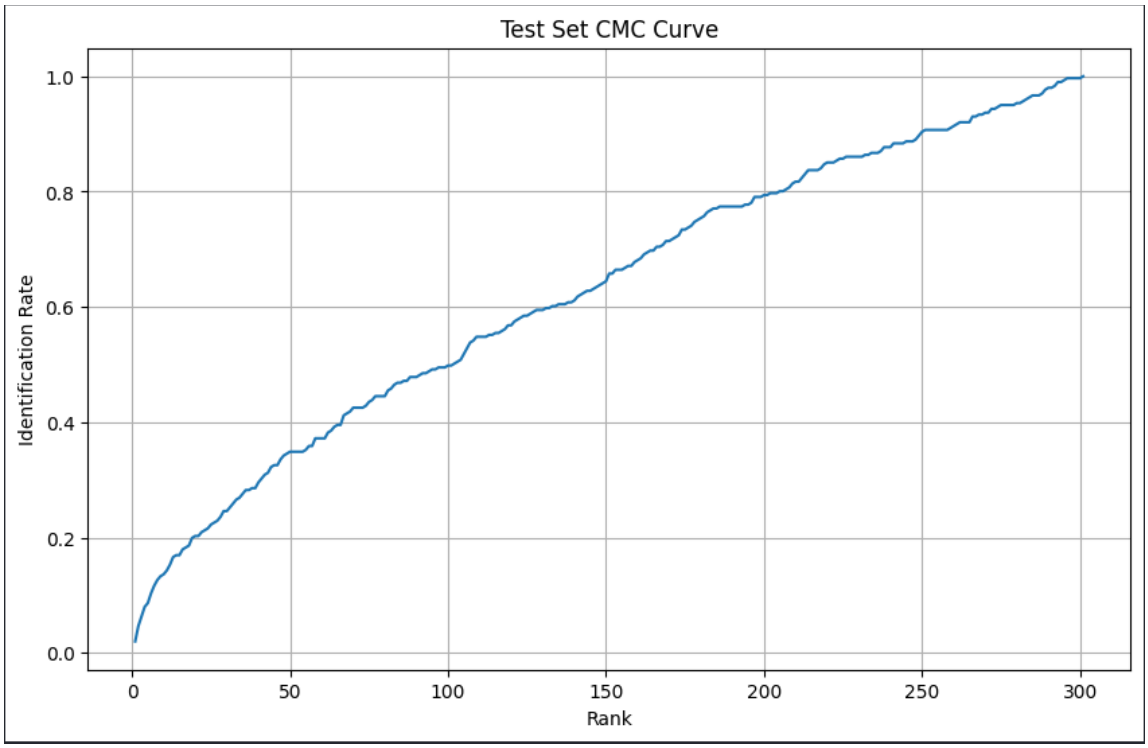

Figure 1: PCA+LDA CMC Curve

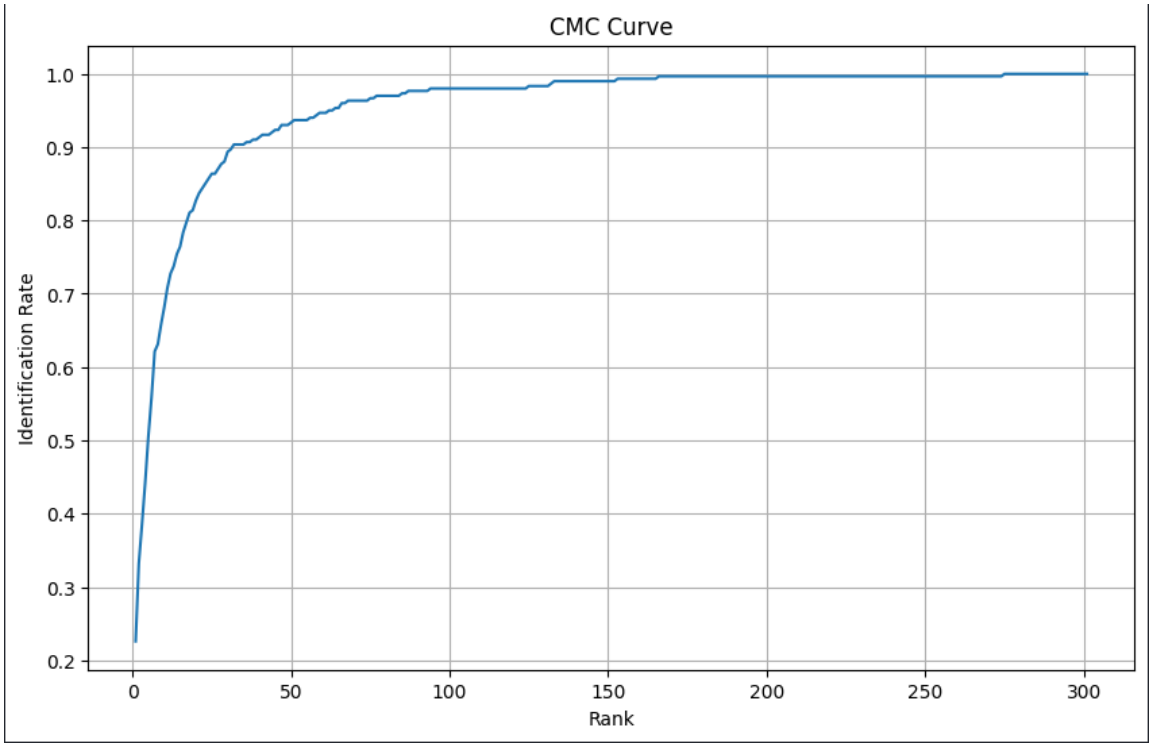

Figure 2: Deep Learning Approach CMC Curve

The deep learning model achieved more than a 10 times improvement in top 1 accuracy and outperforms the PCA+LDA method by considerable margins in both top 5 and top 10. The CMC curves further exemplify the disparity in performance, with the deep learning curve rising significantly steeper which indicates a high identification rate at low ranks and saturates quickly. The PCA+LDA approach's CMC curve rises slowly and fails to identify most matches even by rank 100, which is indicative of poor retrieval performance.

Qualitative Interpretation

The deep learning approach's superior performance can be attributed to its ability to learn discriminative feature embedding tailored to person similarity. Additionally, the model can leverage data augmentation, hard-negative mining, and large parameter spaces. In comparison, the PCA+LDA pipeline relies on reducing dimensionality based on variance and class separability in a linear subspace, which can cause issues with not being able to capture the visual complexity and intra-class variation typically found in pictures of people.

Performance Differences

A noteworthy difference is that the PCA+LDA approach occasionally yielded moderate performance on validation, with top 10 accuracy around 50%, but fails to generalise to the test set. This indicates potential overfitting to the identities it was trained on, lacking the capacity to generalise to unseen subjects. The deep learning method which was trained with metric learning techniques, was inherently better suited to generalisation across unsees identities which is a critical requirement for person re-identification.

Computational Efficiency

While the PCA+LDA method is highly computationally efficient and has quick training times, it suffers greatly in terms of accuracy. The deep learning method required a longer training time, access to a GPU, and more memory throughout training and inference. Despite the greater computational requirements, once the deep learning model is trained it can generate embeddings and perform a nearest-neighbour retrieval in a reasonable time and at scale.

Conclusion

The deep learning approach provides considerable retrieval accuracy gains as well as robustness, especially when generalising to unseen identities. Despite deep learning's higher computational demands, its far greater performance makes it a more viable solution for real-world implementations of person re-identification tasks. Despite classical methods such as the implemented PCA+LDA being efficient, they are only suitable for simple or resource-constrained scenarios where recognition accuracy is not critical.

Ethical Considerations

The development and deployment of person re-identification systems raise significant ethical concerns, particularly regarding personal data collection, application contexts, and inherent limitations.

The dataset being utilised, Market-1501, comprises of surveillance footage collected without explicit subject consent. Though anonymised, the potential for re-identification through linkage with other data sources poses a significant privacy risk. Additionally, biases in the dataset such as overrepresentation of certain demographics can perpetuate inequities. This could potentially lead to skewed performance across population groups.

Re-Identification technologies have legitimate use cases, such as secure authentication for personal devices or locating missing persons. However, its integration into mass surveillance systems risks enabling unwarranted tracking and erosion of civil liberties. A lack of transparency in deployment protocols can further exacerbate biases, disproportionately affecting marginalised communities.

Deep learning models, while powerful, operate as black boxes, which complicates efforts to audit their decision-making processes. This opacity can pose as problematic in legal or ethical discussions, where understanding model behaviour is crucial for accountability. Furthermore, re-identification systems often fail in challenging conditions such as low lighting or occlusions. This can lead to false matches with potentially severe consequences in high-risk applications such as law enforcement.

To help minimise these issues, ethical dataset curation must be prioritised, including explicit consent protocols and bias audits. Deployment of such identification systems should be governed by clear regulatory frameworks, helping to ensure transparency and accountability. Public discourse on the acceptable use of re-ID technology is equally critical to balance innovation with societal values.

Question 2 – Multi-Task Learning and Fine Tuning

Preprocessing

The preprocessing performed was designed to maximise computational efficiency whilst preserving sufficient image detail. All images were resized to 256x256 pixels, which was chosen to be a compromise between the original varying resolutions and the need for manageable GPU memory usage. Images were then normalised by scaling pixel values to [0,1], and training data was augmented with random horizontal flips to improve generalisation. The segmentation masks were converted to binary format (foreground = 1, background=0), which simplifies the segmentation task whilst maintaining the information needed for pet outline detection. A batch size of 32 was selected to maximise GPU utilisation while avoiding memory overflow, with consistent preprocessing applied to both the from-scratch and fine-tuned models to ensure a fair comparison is being made.

Implemented Methods

From-Scratch Model

The from-scratch convolutional neural network utilises a shared encoder with three progressively deeper convolutional blacks (32, 64, 128 filters), each using 3x3 kernels. Batch normalisation, and ReLU activation. This is an established designed for hierarchical feature extraction whilst maintaining computational efficiency. For classification, global average pooling reduces spatial dimensions before two dense layers (64 units then 37 class output). The segmentation decoder employs a symmetrical expansion path with transposed convolutions, gradually up-sampling from the bottleneck features the original 256x256px resolution. This architecture is inspired by U-Net, however, skip connections have been simplified to meet computational constraints. The model was trained end=to-end through a multi-task loss combining categorical cross-entropy for classification and binary cross-entropy for segmentation. This process was optimised with Adam and placed under a 15-minute training time limit to ensure the model didn't train for too long.

Fine-Tuned MobileNetV3Small

The pretrained MobileNetV3Small was adapted through removing its classification head and freezing the initial layers to preserve learned ImageNet features while allowing adaptation to the pet domain. The classification branch was rebuilt using global average pooling, followed by the same dense layer structure as the from-scratch model. For segmentation, a custom decoder was used in conjunction with the MobileNet feature extractor using 5 up-sampling blocks to reconstruct the 256x256px masks from the 8x8 bottleneck features. Utilising this design allows for MobileNet's efficient depth-wise convolutions to be utilised whilst addressing the resolution mismatch through learned up-sampling. The same multi-task loss and 15-minite time limit were applied for consistency.

Evaluation

Image Classification Performance

The classification performance difference between the two models was significant. The from-scratch achieved a test accuracy of 38.4% and an F1 score of 0.38, which is indicative of poor generalisation to unseen images. This was an expected outcome considering the relatively small size of the dataset and large number of output classes (37 breeds), which makes training a deep model from scratch prone to overfitting. In comparison, the fine-tuned MobileNet model resulted in a 83.3% accuracy and an F1 score of 0.83, which is double the performance of the from-scratch model across all measured performance metrics. These results highlight the strength of transfer learning, where pretrained features from a large general-purpose dataset such as ImageNet aid the model in capturing relevant patterns without the need to learn them from scratch.

Figures 4 and 7 further emphasize the performance gap between the two models. The MobileNet model shows a stronger diagonal dominance, which indicates greater correct predictions across most classes. In contrast, the from-scratch model's confusion matrix appears noisier with some considerable misclassifications, particularly in breeds with similar appearance. The improved class separation in the pretrained model is a result of its more robust feature extraction capabilities.

Semanti Segmentation Performance

In regard to semantic segmentation, both models performed quite well in comparison to classification. The from-Scratch DCNN achieved an F1 score of 0.86 and the MobileNet getting 0.92. the significantly smaller gap in performance between the two models suggests that segmentation is less dependent on high-level semantic discrimination and more forgiving of architectural limitations, which can be attributed to it being a pixel-level binary task (foreground vs background). However, the MobileNet model once again demonstrated better results, with higher precision, recall, and overall accuracy which can be seen in Table 2.

Looking at the segmentation confusion matrices in figure 4 and 7, both models effectively separate pet pixels from the background, though the MobileNet produces fewer false positives. Though minimal, this improvement could be crucial in real-world applications where precise boundaries are needed, such as medical imaging or automated cropping tools

Summary

Overall, the fine-tuned MobileNet significantly outperformed the from-scratch model in classification and segmentation tasks. The performance difference was especially considerable in the classification task, where pretrained features were essential for capturing high-level inter-class differences. With the segmentation task, although both models performed quite well, the pretrained approach still had measurable improvements in precision and recall. The results gathered from these two models clearly highlight the importance of transfer learning, especially with limited datasets and complex visual categories.

Task	Metric	From-Scratch DCNN	Fine-Tuned MobileNet
Classification	Accuracy	38.4%	83.33%
	Precision	0.45	0.84
	Recall	0.38	0.83
	F1 Score	0.38	0.83
Segmentation	Accuracy	87.11%	93.42%
	Precision	0.81	0.9
	Recall	0.91	0.95
	F1 Score	0.86	0.92

Table 2: Performance Metrics of From-Scratch DCNN and Fine-Tuned MobileNet

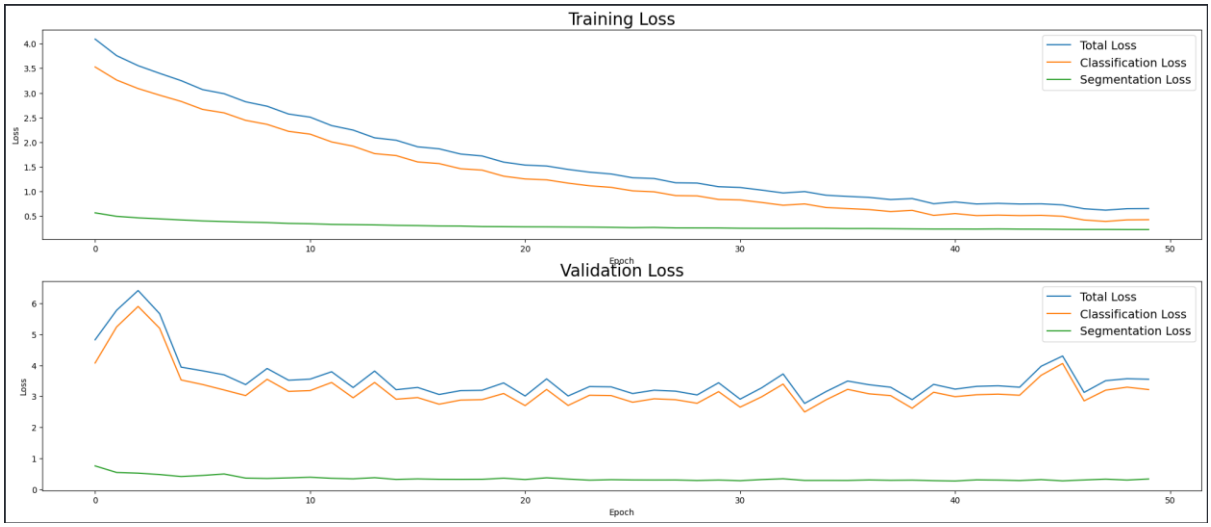

Figure 3: From-Scratch DCNN Training and Validation Loss

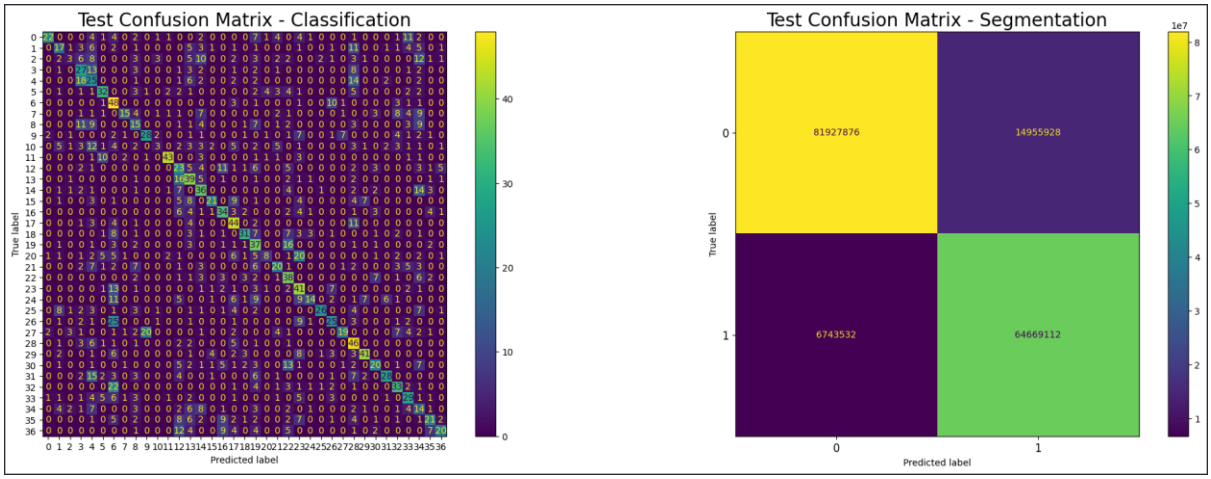

Figure 4: From-Scratch DCNN Classification and Validation Confusion Matrix

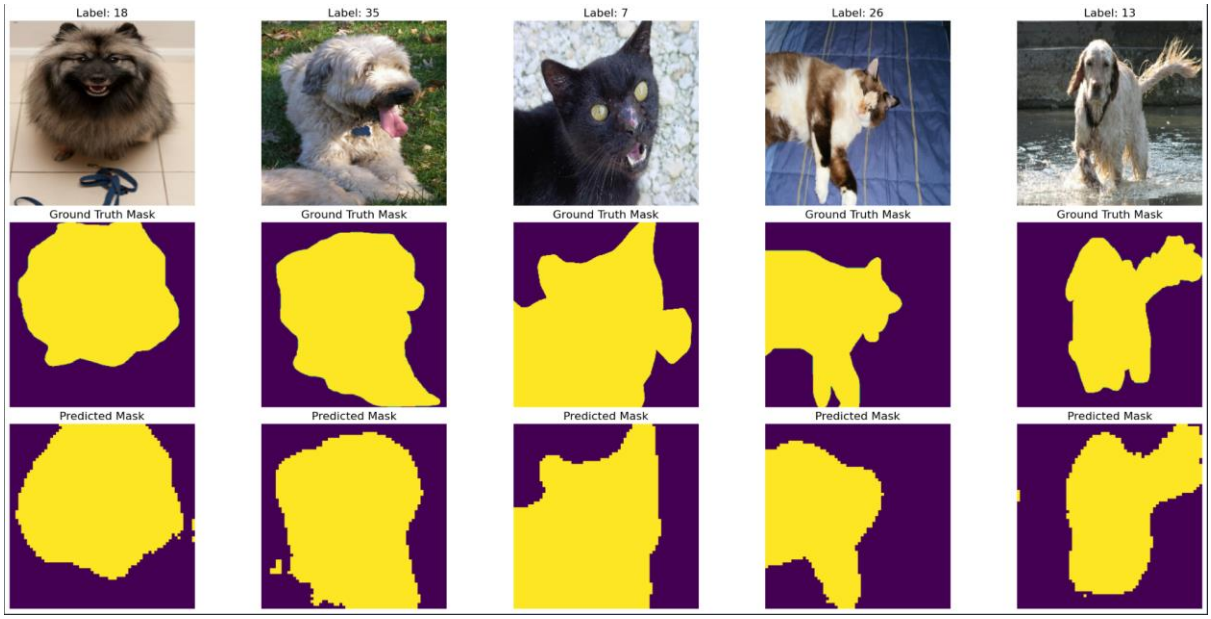

Figure 5: From-Scratch DCNN Masks

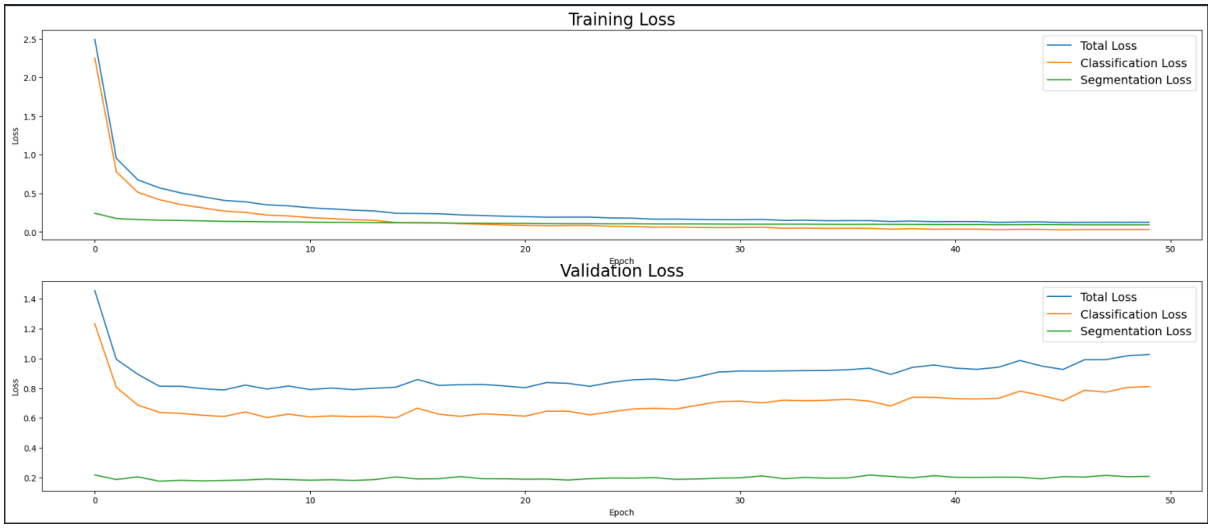

Figure 6: Fine-Tuned MobileNet Training and Validation Loss

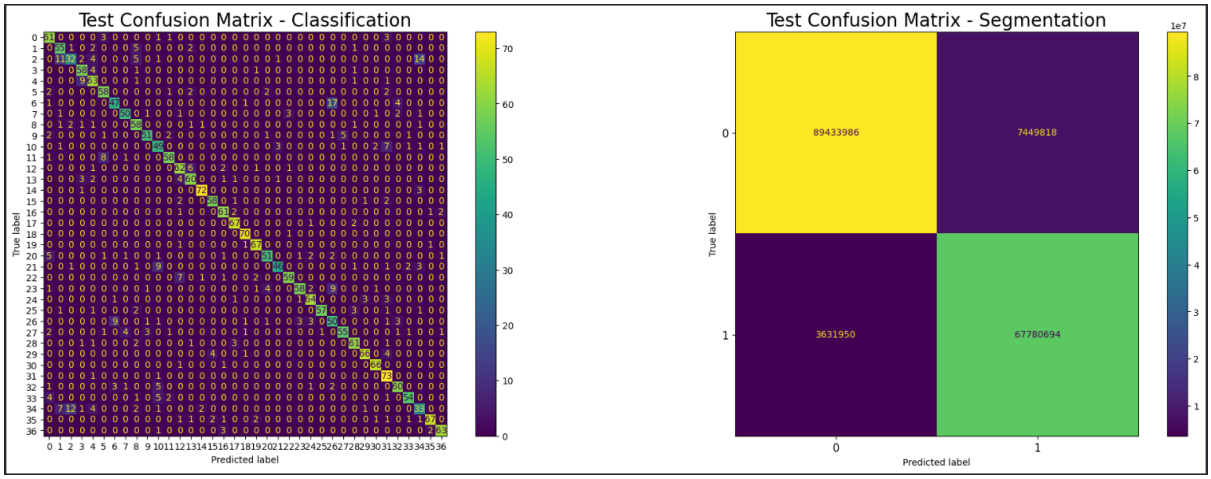

Figure 7: Fine-Tuned MobileNet Classification and Segmentation Confusion Matrix

Figure 8: Fine-Tuned MobileNet Mask

Improvement Strategies and Challenges

Several modifications to the architecture were explored for the sake of performance enhancement within computational limits. The from-scratch model was revised to use deeper pooling (reducing feature maps to 8x8 instead of 32x32) and additional up-sampling steps, which helped to improve feature richness but increased memory usage. For the MobileNetV3Small variant, the initial frozen backbone approach proved stable though it limited segmentation quality, which suggested the need for careful layer unfreezing in future work. Data augmentation remained basic due to concerns of over-regularisation with the small dataset, though additional transformations such as rotations or colour jitter could have helped. The most significant constraint was time available to train the model, which had to be limited to 15 minutes due to time constraints. These computational limitations particularly affected segmentation performance, as the models struggled to refine boundaries within the allotted training time. Future work would benefit from progressive resizing strategies or mixed-precision training to make better use of the available compute resources.