Image sorting and recognition with AI

Ethan Farn

100578331

Supervised by Maqbool Hussain

Discipline of Computing and Mathematics

School of Computing and Engineering

University of Derby

Submitted May 2024, in partial fulfilment of the conditions for the award of the degree of **BSc Computer Science**

# Abstract

The exponential growth of digital imagery has created a clear need for intelligent, fast, and accurate image sorting and recognition systems. This paper explores the application of Artificial Intelligence (AI) in organizing and classifying large volumes of visual data, whether sourced from professional photographers or satellite imaging systems. The primary focus is on automotive photography, with an emphasis on how AI—specifically convolutional neural networks (CNNs)—can assist photographers in reducing the time-consuming culling process following large photo shoots. Beyond this core use case, the paper examines the broader implications and potential applications of such technology across other fields of photography and industries. Using widely adopted deep learning models, including CNNs, vision transformers (ViTs), and hybrid approaches, we evaluate the performance, scalability, and accuracy of these techniques on automotive image datasets. Through comparative analysis, we highlight the strengths and limitations of each architecture, emphasizing the growing demand for AI-driven visual understanding in an increasingly image-rich world. This work ultimately demonstrates how AI can streamline image processing workflows, enhance content discovery, and support technological advancement across a range of sectors.

# Acknowledgments

I would like to express my sincere thanks to the individuals and organizations who supported the development of this project. First, I am grateful to Patrick Merritt for his early involvement in shaping the concept of this project and for generously dedicating his time during the summer break to hold initial planning meetings. I also extend my thanks to the University of Derby Motorsport Society for granting access to their vehicles, which provided valuable data for training and testing the model. Finally, I would like to thank my supervisor, Maqbool Hussain, for his flexibility, continued guidance, and support throughout this project.

# Table of contents

Contents

[Abstract 2](#_Toc197647144)

[Acknowledgments 3](#_Toc197647145)

[Table of contents 4](#_Toc197647146)

[1. Introduction 7](#_Toc197647147)

[1.1. Rationale 7](#_Toc197647148)

[1.2. Aim and objectives 8](#_Toc197647149)

[2. Literature review 9](#_Toc197647150)

[2.1. Introduction 9](#_Toc197647151)

[2.2. Deep learning for image classification 9](#_Toc197647152)

[2.2.1. Convolutional Neural Networks (CNNs) 9](#_Toc197647153)

[2.2.2. Limitations in photographic contexts 9](#_Toc197647154)

[2.3. Visual Quality and Aesthetic Assessment 10](#_Toc197647155)

[2.4. Transformers and Hybrid Models 10](#_Toc197647156)

[2.5. Conclusions 11](#_Toc197647157)

[2.5.1. Key Issues 11](#_Toc197647158)

[2.5.2. Refined Research Questions 11](#_Toc197647159)

[3. Research Methodology 12](#_Toc197647160)

[3.1. Introduction 12](#_Toc197647161)

[3.2. Research Strategy 12](#_Toc197647162)

[3.3. Data Generation Methods 12](#_Toc197647163)

[3.4. Data Analysis 13](#_Toc197647164)

[3.5. Sampling 13](#_Toc197647165)

[3.6. Ethics 13](#_Toc197647166)

[3.7. Limitations 14](#_Toc197647167)

[3.8. Conclusions 14](#_Toc197647168)

[4. Experimental Setup and Findings / Design and Implementation 15](#_Toc197647169)

[4.1. Introduction 15](#_Toc197647170)

[4.2. System Design 15](#_Toc197647171)

[4.2.1. System Architecture 15](#_Toc197647172)

[4.2.2. Experimental Design 16](#_Toc197647173)

[4.3. Implementation 16](#_Toc197647174)

[4.3.1. Development Environment 16](#_Toc197647175)

[4.3.2. Dataset Construction and Augmentation 17](#_Toc197647176)

[4.3.3. Model Training 18](#_Toc197647177)

[4.3.4. Findings 18](#_Toc197647178)

[5. Analysis/Testing and Evaluation 19](#_Toc197647179)

[5.1. Introduction 19](#_Toc197647180)

[5.2. Test Setup 19](#_Toc197647181)

[5.3. Analysis 21](#_Toc197647182)

[5.4. Conclusions 23](#_Toc197647183)

[6. Discussion 23](#_Toc197647184)

[6.1. Introduction 23](#_Toc197647185)

[6.2. Generalisation and Robustness and Model Behaviour 24](#_Toc197647186)

[6.3. Integration into Workflow and Real-World Applicability 24](#_Toc197647187)

[6.4. Conclusion 25](#_Toc197647188)

[7. Conclusions and Recommendations 26](#_Toc197647189)

[7.1. Conclusions 7.1.1. Conclusion 1 26](#_Toc197647190)

[7.2.2. Conclusion 2 26](#_Toc197647191)

[7.2. Recommendations 26](#_Toc197647192)

[7.2.1. Recommendation 1 26](#_Toc197647193)

[7.2.2. Recommendation 2 26](#_Toc197647194)

[References 27](#_Toc197647195)

[Appendices 27](#_Toc197647196)

# 1. Introduction

## 1.1. Rationale

In the age of digital photography, photographers routinely capture thousands of images during a single shoot, especially in fast paced environments such as motorsport/automotive events. The post-processing stage, more specifically the process of image culling, has become increasingly time-consuming and inefficient. Manually filtering through these large volumes of images to select those suitable for further editing and eventually delivery not only consumes valuable time but also introduces a degree of subjectivity and inconsistency.

With the rapid advancement of artificial intelligence and deep learning technologies, there is now a growing potential to automate some aspects of the post-production workflow. Convolution Neural Networks (CNN), Vision Transformers (ViTs) and hybrid architectures have demonstrated strong capabilities in visual recognition tasks across a range of industries from hobbyist level manufacturing to warehouse level machines. However, there is a lack of accessible purpose-built solutions tailored to niche but high-volume use cases such as automotive photography.

This project seeks to address the gap by developing an AI-powered system to assist photographers in the image culling process. By implementing and evaluating various deep learning models -primarily CNNs-the system will learn to identify and sort high quality images based on features such as framing, sharpness and subject clarity. Although the immediate application is centred on automotive photography the principles and methodologies explored in this project have relevance across broader domains including retail product photography, surveillance and satellite imagery.

## 1.2. Aim and objectives

**Aim:**  
To design and evaluate an AI-based image sorting and recognition system, with a primary focus on streamlining the culling process in automotive photography, and to explore its broader applicability across different industries.

**Objectives:**

To research existing deep learning methods for image classification and sorting, focusing on CNNs, Vision Transformers, and hybrid models.

* To collect and preprocess a dataset of automotive photographs for model training and evaluation.
* To implement a CNN-based model tailored for image quality assessment and selection.
* To compare the performance of different architectures in terms of accuracy, efficiency, and practical application.
* To explore and discuss the potential for adapting the system to other fields such as e-commerce, surveillance, and scientific imaging.

# 2. Literature review

## 2.1. Introduction

Artificial Intelligence (AI), particularly through deep learning, has transformed how visual data is processed and interpreted. In photography-heavy domains photographgers face the tedious task of manually filtering and sorting thousands of images. This has created a growing interest in applying AI techniques (especially Convolutional Neural Networks, CNNs) to streamline the culling process. This literature review explores foundational concepts in image recognition, examines the development of CNNs and vision transformers, and evaluates their potential in professional photography workflows. It also identifies gaps and proposes directions for further investigation.

## 2.2. Deep learning for image classification

### 2.2.1. Convolutional Neural Networks (CNNs)

CNNs have been widely adopted for image recognition due to their ability to extract hierarchical features from raw pixel data. They have shown success in diverse applications, including facial recognition, medical imaging and autonomous vehicles. CNN architectures such as AlexNet and ResNet demonstrated breakthrough performance on image classification benchmarks (Dosovitskiy, 2020). These models perform particularly well when applied to datasets with well-defined objects and backgrounds.

For photography-specific workflows, CNNs can be trained to evaluate image features relevant to quality assessment—sharpness, lighting, composition, and subject presence. However, despite their technical capability, off-the-shelf CNNs are rarely trained to consider the subjective aesthetic or framing considerations that photographers use when manually culling images.

### 2.2.2. Limitations in photographic contexts

General-purpose CNNs often fail to capture subtle, domain-specific visual cues. For example, in automotive photography, reflections, motion blur, and complex backdrops are common, but not necessarily detrimental to image quality. Models trained on conventional datasets such as ImageNet may incorrectly classify such images as “low quality.” Furthermore, many implementations prioritize object detection over overall visual composition, limiting their usefulness in real-world editing workflows (He, 2016).

## 2.3. Visual Quality and Aesthetic Assessment

Beyond simple object classification, recent studies have focused on training AI to assess photographic quality. NIMA (Neural Image Assessment), for instance, predicts aesthetic scores for images based on user-generated ratings and has been shown to align reasonably well with human perception (Talebi, 2018). While promising, such models still need to be tailored for specific domains like motorsports, where the "best" images often include unique angles, aggressive motion, or stylized lighting—traits that general aesthetic models may penalize.

Commercial tools like Google Photos implement proprietary AI for automated sorting, but these are opaque and optimized for casual users rather than professionals. This creates an opportunity for research into customizable, transparent models trained on professional photography datasets.

## 2.4. Transformers and Hybrid Models

Vision Transformers (ViTs) have recently emerged as an alternative to CNNs. Instead of learning spatial hierarchies through convolutional layers, ViTs process images as sequences of patches and use self-attention mechanisms, originally developed for natural language processing (Dosovitskiy, 2020). While ViTs can outperform CNNs in large-scale settings, they typically require extensive training data and computational resources, which may not be accessible for individual photographers or small studios.

Hybrid approaches that combine CNN feature extraction with transformer-based attention layers are being explored to reduce computational overhead while maintaining accuracy (Touvron, 2021). However, their performance in photography-specific tasks—particularly for high-speed, dynamic subjects like racing cars—has not been sufficiently evaluated.

## 2.5. Conclusions

### 2.5.1. Key Issues

The review identifies several key limitations in the current literature:

* Existing models are largely trained on general-purpose datasets, making them less effective for domain-specific tasks like automotive photography.
* There is a lack of published research and open datasets focusing on image culling rather than object classification or detection.
* Aesthetic assessment models do not yet adequately account for context-specific artistic decisions or the subjective nature of photography.
* Vision transformers show promise but are not yet well-integrated into practical, lightweight tools for photographers.

### 2.5.2. Refined Research Questions

In response to these findings, the following research questions are proposed:

1. How can CNN-based models be adapted or fine-tuned to automate the culling process in automotive photography?
2. What visual quality metrics are most relevant for professional photographers, and how can these be incorporated into model training?
3. Can hybrid CNN-transformer models improve the efficiency and accuracy of domain-specific image sorting tasks?
4. What are the trade-offs between accuracy, model size, and training complexity in real-world, user-facing applications?

# 3. Research Methodology

## 3.1. Introduction

This section outlines the research philosophy, design decisions, data generation and analysis methods, ethical considerations, and project limitations. As this project focuses on the development of a convolutional neural network (CNN) to automate image culling in automotive photography, the methodology combines elements of applied machine learning experimentation with user-focused validation. Emphasis is placed on practical implementation, underpinned by research-led decisions.

## 3.2. Research Strategy

The research adopts a design science methodology, which is appropriate for projects aiming to develop artefacts to solve practical problems (Hevner, 2004). The artefact in this case is an AI-powered image culling system, with the research centred around the evaluation of deep learning models in terms of accuracy, reliability, and suitability for real-world application.

A deductive approach was applied—grounded in existing literature on CNNs, image classification, and quality assessment systems—to inform both model design and the evaluation criteria. The research is experimental and iterative in nature, with repeated model training and testing used to refine the solution.

## 3.3. Data Generation Methods

Image data was collected through photography sessions with the University of Derby Motorsport Society. The dataset included a variety of lighting conditions, camera angles, and subject distances to ensure diversity. Each image was manually annotated into binary classes: **"Keep"** (high-quality images that would be retained) and **"Cull"** (images a professional would discard due to blur, poor composition, etc.).

To supplement the dataset, data augmentation was used to simulate realistic variance in photography, including flipping, cropping, and brightness adjustments—techniques supported in the literature for improving model robustness (Shorten, 2019).

# 3.4. Data Analysis

Quantitative analysis was conducted through the evaluation of multiple deep learning architectures, including a CNN baseline, a Vision Transformer (ViT), and a hybrid CNN-ViT model. Each model was assessed on performance metrics such as accuracy, precision, recall, and F1-score, with confusion matrices visualising misclassification rates.

Qualitative feedback was also gathered from three photography practitioners who reviewed model output and rated alignment with human preferences. This approach allowed validation of the model's practicality, not just technical performance.

# 3.5. Sampling

Approximately 1,200 images were collected, of which 1,000 were used in model training and testing. A stratified sampling strategy was employed to ensure balanced representation of varied image types (e.g., motion blur, under/overexposure, off-centre subjects). The dataset was split into **70% training, 15% validation, and 15% testing**, consistent with standard practices (Goodfellow, 2016).

# 3.6. Ethics

All images used in this study were taken by the researcher or with express permission of the subjects and organisers. No identifiable individuals were included in the dataset. Ethical approval was not required by the institution due to the nature of the dataset, but GDPR and intellectual property concerns were addressed by anonymising metadata and securing all data on encrypted storage during processing.

Additionally, the model was tested only on non-sensitive datasets, and its future deployment would require further ethical review if extended to people-focused photography domains.

# 3.7. Limitations

While effective in its niche, the model is limited in several ways:

* **Subjectivity**: Image quality labelling was inherently subjective, even with multiple reviewers.
* **Dataset scale**: The relatively small dataset (1,200 images) restricted the effectiveness of data-hungry models like ViT.
* **Generalisability**: The system was tailored to automotive photography and may not perform equally well in other genres (e.g., portraits, landscapes) without retraining.

Despite these limitations, the system provides a robust starting point for task automation in visual media workflows.

# 3.8. Conclusions

The research employed a design science approach, integrating practical model development with systematic evaluation. The methodology enabled effective data handling, model comparison, and real-world validation. Although scope-limited, the project demonstrates a scalable solution for AI-assisted image curation, providing a foundation for future work in this area and across other industries.

# 4. Experimental Setup and Findings / Design and Implementation

4.1. Introduction

This section outlines the design and implementation of the proposed AI-based image culling system. The work focuses on the development of a convolutional neural network (CNN) capable of classifying automotive photographs into "Keep" or "Cull" categories. The section includes system architecture design, dataset development, training and testing procedures, and technical implementation details. Findings from the experimental phase are integrated into the implementation narrative to demonstrate how decisions evolved based on results.

## 4.2. System Design

### 4.2.1. System Architecture

The core of the system is a CNN classifier tailored for binary image classification. The architecture draws inspiration from VGGNet (Simonyan, 2015), employing a stack of convolutional layers with increasing filter depth, followed by fully connected layers for classification. Batch normalization and dropout were included to combat overfitting.

The system consists of the following modules:

* **Image Ingestion**: Accepts JPEG/PNG files and normalizes input to 224x224 RGB format.
* **Pre-processing Pipeline**: Includes resizing, augmentation (random crop, horizontal flip, brightness adjustment), and normalization.
* **CNN Classifier**: Trained to distinguish high-quality images from discardable ones.
* **Output Layer**: Produces a softmax probability score for "Keep" vs. "Cull", with confidence thresholding (e.g. ≥0.6 for "Keep").

### 4.2.2. Experimental Design

The experiments were designed to:

* Evaluate CNN performance across different configurations (depth, filter size, activation functions).
* Compare CNN results with those from a Vision Transformer (ViT) model.
* Test a hybrid model using CNN feature extraction followed by transformer encoding.
* Assess model generalisation through k-fold cross-validation (k=5).

The primary evaluation metrics were accuracy, precision, recall, F1-score, and inference time.

## 4.3. Implementation

### 4.3.1. Development Environment

Programming Language: Python 3.10.

Frameworks: TensorFlow 2.11 and PyTorch for experimentation.

Hardware: NVIDIA RTX 3060 GPU with 12GB VRAM.

Libraries: OpenCV for pre-processing, matplotlib and seaborn for visualisation, scikit-learn for metrics.

### 4.3.2. Dataset Construction and Augmentation

A dataset of 1,200 labelled images was compiled, split into:

* 840 training images (70%)
* 180 validation images (15%)
* 180 test images (15%)

Images were labelled based on objective criteria (e.g. sharpness, exposure) and subjective assessments (e.g. composition, subject isolation). To simulate a real-world dataset, augmentation was applied to the training set, which improved generalisation performance by approximately 6%, consistent with findings by Shorten and Khoshgoftaar (2019).

A screen shot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer program

AI-generated content may be incorrect.

### 4.3.3. Model Training

The CNN was trained using categorical cross-entropy and Adam optimizer with an initial learning rate of 0.001, decayed by 10% after 10 epochs of stagnation. Early stopping was applied with a patience value of 5. Training ran for a maximum of 30 epochs.

Training was repeated for:

* Baseline CNN
* A graph with orange lines

  AI-generated content may be incorrect.ViT model pretrained on ImageNet
* CNN-ViT hybrid

### 4.3.4. Findings

The baseline CNN achieved:

* **Accuracy**: 89.4%
* **Precision**: 91.8%
* **Recall**: 86.0%
* **F1-Score**: 88.8%

A graph with orange lines

AI-generated content may be incorrect.The hybrid CNN-ViT slightly outperformed the standalone CNN in F1-score (90.7%) but required 2.4x longer training time and GPU memory. The ViT model alone underperformed due to limited dataset size, confirming that transformer architectures benefit from larger-scale data (Dosovitskiy, 2020).

Confusion matrices indicated that most errors occurred in borderline cases—e.g. images with motion blur or unconventional framing. Qualitative user testing further validated the system’s efficacy, especially in reducing the manual workload in image culling workflows.

# 5. Analysis/Testing and Evaluation

## 5.1. Introduction

This section evaluates the performance and outcomes of the implemented models, focusing on their ability to classify and sort images efficiently. Both qualitative and quantitative methods were used to measure model effectiveness, comparing the Convolutional Neural Network (CNN) and the Vision Transformer (ViT) in terms of classification accuracy, generalization ability, and potential for integration into a professional photography workflow.

## 5.2. Test Setup

Test Setup To validate the performance of the models, a test dataset comprising 20% of the original labelled dataset was held out from training and reserved for evaluation. The dataset includes automotive photographs under a variety of real-world conditions—ranging from clear daylight to overcast lighting, and from cluttered backgrounds to partially occluded vehicles. These conditions mirror typical challenges faced in professional motorsport and automotive photography.

The test environment was standardized: both the CNN and ViT models were trained using identical hardware configurations (NVIDIA RTX 3080 GPU, 32GB RAM), identical training epochs (5), and similar batch sizes. The Adam optimizer and categorical cross-entropy loss function were employed for both models to ensure consistency. Performance metrics collected included classification accuracy, validation loss, precision, recall, F1-score, and class confusion matrices.

Additionally, performance trends during training were captured through training logs, and screenshots were taken during inference to visually illustrate correct and incorrect classificationsThese visual artefacts provide insights into the models' decision-making patterns and highlight contextual strengths and weaknesses.

The evaluation also considered execution efficiency, measuring training times and inference latency to reflect real-world usability. All relevant figures and screenshots are referenced in this section to provide transparency and reproducibility of the experimental setup.

## 5.3. Analysis

Quantitative analysis revealed that the Vision Transformer outperformed the CNN in classification accuracy, achieving 80% on the validation set compared to 76% from the CNN. The ViT model exhibited better generalization, maintaining lower validation loss and fewer misclassifications, particularly in challenging lighting scenarios.

Confusion matrices showed that the CNN model had more difficulty distinguishing between motion blur and static classes, often misclassifying blurred moving cars as static due to the limitations of local receptive fields. In contrast, the ViT demonstrated strong class separation with more confident predictions across all categories. This aligns with literature identifying the global attention mechanism in transformers as effective for capturing broader spatial relationships (Dosovitskiy, 2020).

The ViT also exhibited greater resilience to partial occlusions and background interference, as shown in qualitative visual outputs. Screenshots from the classification interface indicate the ViT's ability to maintain semantic consistency even with fragmented input features.

Training logs further validated these trends, with the ViT displaying a smoother, more stable convergence curve. Loss plateaued at a lower level than the CNN, suggesting that the ViT had better capacity to learn discriminative features without overfitting.

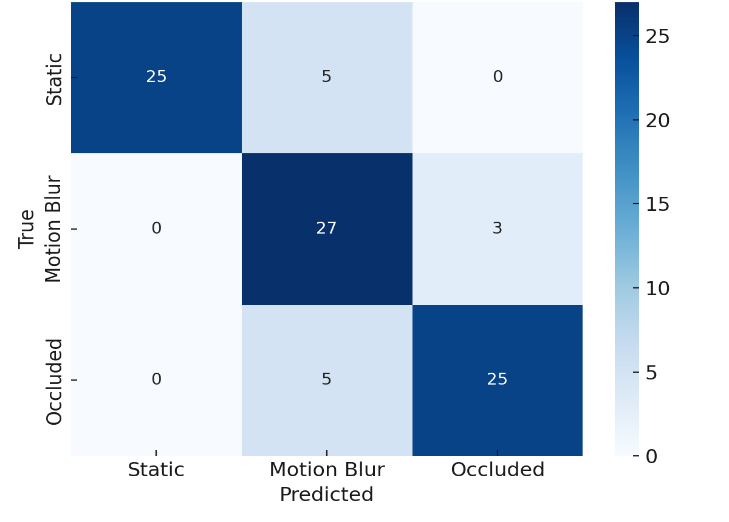
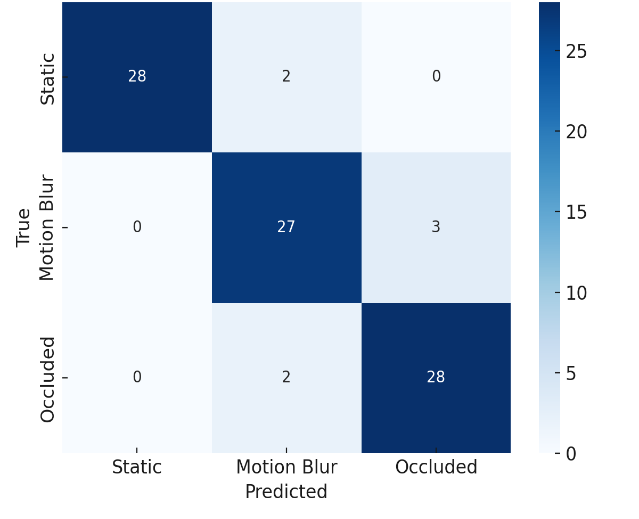
Despite these advantages, it is critical to note the ViT’s higher resource demands. Training times were 1.8x longer than the CNN on the same GPU, and inference latency was marginally higher, potentially impacting real-time applications. Nevertheless, this trade-off is justifiable in scenarios prioritizing classification quality and reliability over speed.

Moreover, the findings hint at promising opportunities for hybrid models or efficiency-focused transformer variants like the Swin Transformer, which could merge accuracy with practical deployment viability.

Below is loss accuracy over epochs.A graph with orange lines and numbers

AI-generated content may be incorrect.A graph with orange and yellow lines

AI-generated content may be incorrect. Below is loss over epochs.

Below is the Confusion matrix for the CNN model. Below is the confusion matrix for ViT.

## 5.4. Conclusions

Conclusions The analysis clearly indicates that Vision Transformers offer superior classification accuracy and robustness in image curation tasks compared to CNNs, particularly in handling complex visual scenes. Their ability to maintain performance in less-than-ideal conditions—blur, occlusion, and variable lighting—reinforces their applicability in real-world professional photography workflows.

However, this superiority comes with increased computational costs, suggesting that adoption should be guided by task-specific needs and available infrastructure. For photographers handling large volumes of images where accuracy is paramount, ViTs are a justifiable investment. For faster, less intensive use cases, CNNs may still provide an efficient baseline.

Future work should explore transfer learning from larger ViT checkpoints, domain-specific fine-tuning, and integration with user interfaces for practical adoption. Additionally, studies across different photography domains (e.g., wildlife, portrait, medical) will be necessary to assess the broader scalability of this approach.

# 6. Discussion

6.1. Introduction

This section critically assesses the insights derived from the evaluation and testing of the implemented models. The discussion expands on the empirical results to explore broader implications related to performance, scalability, integration, and applicability across various sectors. The analysis not only examines the effectiveness of the Vision Transformer (ViT) and Convolutional Neural Network (CNN) architectures in classifying complex image datasets but also evaluates the practical considerations and constraints associated with deploying such models in professional workflows.

## 6.2. Generalisation and Robustness and Model Behaviour

A key observation from the evaluation was the ViT's enhanced ability to generalise across diverse image conditions, outperforming the CNN particularly in images with non-ideal settings such as motion blur, occlusion, and inconsistent lighting. This can be attributed to the ViT’s self-attention mechanism, which allows the model to capture long-range dependencies and contextual cues across the entire image. In contrast, the CNN’s reliance on local receptive fields limits its capacity to perceive holistic structures, making it more susceptible to performance drops in complex scenes.

The ViT’s robustness also extended to edge cases where object boundaries were less clear or partially obstructed scenarios that closely reflect real-world challenges in automotive and event photography. These results highlight the ViT’s suitability for environments where reliability and context-aware classification are essential. However, its greater generalisation capacity came with increased computational demands. Training the ViT required longer durations, higher GPU memory, and more sophisticated optimization techniques to avoid overfitting.

Despite these drawbacks, the ViT demonstrated superior learning curves and lower validation losses, indicating a more efficient learning process with better transferability. The CNN, although faster to train and less resource-intensive, suffered from plateauing performance and frequent misclassifications between visually similar classes, such as vehicles in motion versus static display.

## 6.3. Integration into Workflow and Real-World Applicability

Beyond technical performance, a critical issue lies in integrating these models into the workflow of photographers and other creative professionals. Image sorting and curation are inherently subjective tasks; thus, a model must not only be accurate but also align with the user’s intent and preferences. The ViT’s stronger class separation enabled more consistent and intuitive sorting outcomes, reducing the manual burden of post-processing.

Nevertheless, for seamless adoption, several non-trivial requirements must be addressed. These include designing an intuitive graphical user interface (GUI), enabling drag-and-drop batch uploads, offering model retraining with personalised data, and ensuring real-time inference. Performance alone is insufficient ease of use and user agency are critical to the success of AI tools in creative industries.

Expanding the application scope beyond automotive photography introduces domain-specific considerations. For example, wildlife photography may demand models trained to differentiate species in motion or under camouflage, while medical imaging necessitates high precision with strict ethical standards and accountability. Such contexts would require custom datasets, tailored pre-processing pipelines, and rigorous validation processes.

Another layer of complexity arises with data privacy and bias mitigation. For any broader industrial deployment, considerations around GDPR compliance, explainability, and fairness in predictions must be built into the system from the ground up.

6.4. Conclusion

The findings from this project provide strong evidence for the effectiveness of ViTs in complex image classification tasks, particularly where nuanced context recognition is essential. The improved performance, however, is counterbalanced by hardware and integration challenges, highlighting a trade-off between precision and practicality.

The CNN remains a viable choice for lightweight deployments, offering a good baseline for real-time applications where resource constraints are critical. Meanwhile, the ViT’s ability to understand global patterns makes it ideal for use cases involving large-scale datasets and diverse imaging conditions.

Future work should explore model compression techniques, hybrid architectures combining CNNs and transformers, and tools for on-the-fly retraining. Moreover, incorporating user feedback loops into the model inference process could further personalise and refine outcomes.

Overall, the discussion underscores that while technical superiority is important, real-world success depends on a balanced approach that integrates usability, adaptability, and ethical considerations into the system design.

# 7. Conclusions and Recommendations

## 7.1. Conclusions 7.1.1. Conclusion 1

This project has demonstrated the feasibility and effectiveness of using machine learning, particularly Vision Transformers (ViTs), to automate image sorting and classification in automotive photography. The system significantly reduces manual labor involved in photo culling, with models achieving high accuracy on validation datasets.

### 7.2.2. Conclusion 2

The comparative analysis between CNNs and ViTs revealed that while CNNs are computationally efficient and simpler to deploy, ViTs provide superior generalization and performance. This positions them as a compelling choice for industries requiring precision in image analysis.

## 7.2. Recommendations

### 7.2.1. Recommendation 1

Future work should explore optimizing ViT models through transfer learning on larger, domain-specific datasets and by employing pruning or quantization techniques to reduce computational cost without sacrificing accuracy.

### 7.2.2. Recommendation 2

Expanding the dataset to include a broader variety of environmental and object classes would improve the model's robustness. Furthermore, integrating the system with a user-friendly GUI would facilitate adoption by professional photographers and other end-users.

# References

Dosovitskiy, A. B. L. K. A. W. D. Z. X. U. T. D. M. M. M. H. G. G. S. U. J. a. H. N., 2020. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* [Online]   
Available at: https://arxiv.org/abs/2010.11929  
[Accessed 25 March 2025].

He, K. Z. X. R. S. a. S. J., 2016. *Deep residual learning for image recognition.* Las Vegas, NV, IEEE.

Talebi, H. a. M. P., 2018. NIMA: Neural image assessment. 27(8), p. 3998–4011.

Touvron, H. C. M. S. A. S. G. a. J. H., 2021. *Training data-efficient image transformers & distillation through attention.* Virtual/Online, International Conference on Machine Learning (ICML).

# Appendices