Sapienza Università di Roma

Advanced Machine Learning

Evaluating Scene Classification with Vision Transformers and Hybrid Models on Places365

Group Members:

Laura Thoft Cesario – 1852596
Diego Ciciani – 2140394
Livia Oddi – 1846085
Damian Zeller – 2118831

Professor:

Fabio Galasso

December 28, 2024

Contents

1 Abstract

This study investigates the efficacy of Vision Transformers (ViTs), hybrid CNN-Transformer architectures, and advanced CNN models for scene classification using a subset of the Places365 dataset. By fine-tuning pre-trained models, we evaluate their adaptability to a domain that prioritizes global spatial relationships over object-centric features [1]. Our analysis highlights the strengths of hybrid architectures, such as R50 + ViT-B/16, which integrate the local feature extraction of CNNs with the global reasoning of Transformers, achieving superior performance. Results also underscore the limitations of certain architectures, such as EfficientNet-B7, in addressing scene-level tasks due to restricted depth and simplicity. Comparative insights from confusion matrices and attention maps reveal unique challenges, such as the misclassification of visually similar scene categories.

2 Introduction

Scene classification poses unique challenges compared to object recognition, as it requires understanding global spatial relationships rather than focusing on isolated objects. The Places365 dataset, introduced by Zhou et al.[2] [3] offers a benchmark for evaluating models on scene-centric data, pushing the boundaries of representation learning for broader visual contexts.
This project builds on the foundational research of Dosovitskiy et al.(ICLR 2021) [1], who introduced Vision Transformers for image recognition. While their work focused on general image recognition, our study adapts ViTs, Hybrid model, and CNNs for scene classification. The primary objective is to assess fine-tuning performance, computational efficiency, and model adaptability to scene-level tasks.

2.1 Dataset and Benchmark

Due to computational constraints, a subset of 4,000 images across 20 classes was extracted from the Places365 dataset using stratified subsampling, and then further divided into 60%, 20% and 20% for training, testing and validation. The classes were selected to ensure diversity, covering indoor, outdoor, natural, and artificial scenes, as well as functional and recreational spaces. Reducing the dataset to 4,000 images ensured computational feasibility while preserving class diversity. This approach allowed for efficient training on GPUs while maintaining meaningful representation of scene categories.

3 Model selection and setup

3.1 Models

The project investigates three distinct types of models after fine-tuning.

The **Vision Transformers** model *ViT-B/32* and *ViT-B/16* were selected for their ability to capture global spatial relationships using self-attention. ViT-B/32 serves as a computationally efficient baseline, while ViT-B/16, with its smaller batch size, enables finer-grained feature extraction. Both models, pre-trained on ImageNet21k, are well-suited for transfer learning.

The **Hybrid Model** *R50 + ViT-B/16* combines the local feature extraction capabilities of ResNet-50 with the global reasoning power of Vision Transformers. This model was chosen to assess whether integrating CNNs' inductive biases with the flexibility of Transformers enhances scene classification performance.

For the **CNN Models**, *EfficientNet-B7* was selected for its scalable architecture, making it a strong baseline for evaluating computational efficiency.
BiT-L (ResNet152x4), pre-trained on the large JFT-300M dataset, was chosen for its ability to transfer object-centric knowledge to scene-level tasks. As a traditional CNN-based model, it serves as a key reference for assessing CNNs' progress in handling broader spatial contexts like those in Places365.

3.2 Training Setup:

We took the hyperparameters from the research paper when applicable, and we also did automatic hyperparameter tuning with the *Optuna* library. All models were fine-tuned with Stochastic Gradient Descent (SGD)

with momentum as the the optimizer, cosine learning rate decay schedule with warm-up, a batch size of 32, 10 epochs with gradient clipping and the Cross-entropy loss.
Key enhancements included gradient clipping for robust training and smaller batch sizes to adapt to hardware constraints, ensuring computational feasibility without compromising model performance.

4 Experimental Results

4.1 Accuracy and fine-tuning costs

Model	Overall Accuracy	Best-Performing Class(es)
R50 + ViT-B/16	87.75%	Bedroom (100%)
ResNet152x4	86.00%	Bedroom (98%), Library indoor (98%)
ViT-B/16	85.25%	Beach (98%), Skyscraper (98%), Stadium soccer (98%)
ViT-B/32	82.75%	Bedroom (95%), Aquarium (95%), Swimming pool outdoor (95%)
EfficientNet-B7	8.25%	Canyon (25%)

Table 1: Model Performance Summary

The best performing model is the R50 + ViT-B/16. This can be explained by the dataset characteristics: local and global feature extraction have to be handled simultaneously, which this model with its hybrid architecture does best. Further in small fine-tuning sets local inductive biases are more effective, which can also explain why the ResNet152x4 does better than the ViT models. Looking at the confusion matrices (Section 7) the strengths of the different models become visible: the ViT models have superior global context understanding, seen in their ability to distinguish complex scenes such as ”Bedroom” and ”Swimming Pool outdoor”. The ResNet152x4 on the other hand excels in capturing local spatial patterns and therefor achieves high accuracy in most classes, particularly in classes like ”Library indoor” and ”Ballroom”.

Contrary to the findings of the research paper [1] EfficientNet-B7 performs terribly and is not competitive at all, which can be attributed to its simpler architecture and limited depth, which is struggling with transfer learning. The research paper also indicates that the ViT models achieve significant performance gains when pre-trained on JFT-300M, which may be another reason why the ResNet152x outperformed them.

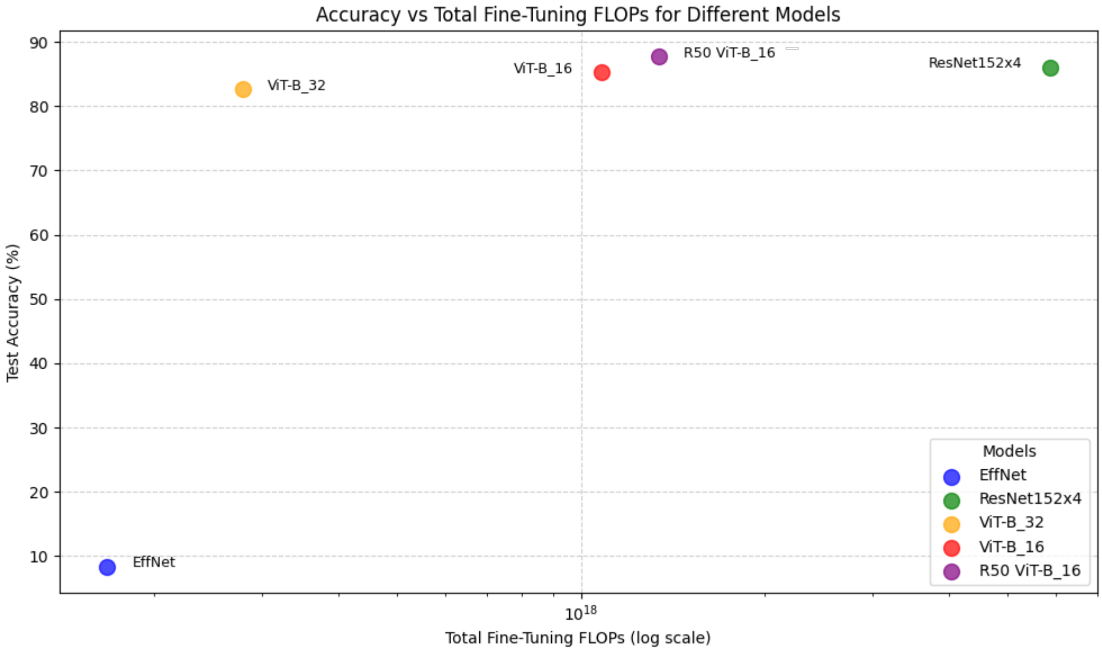


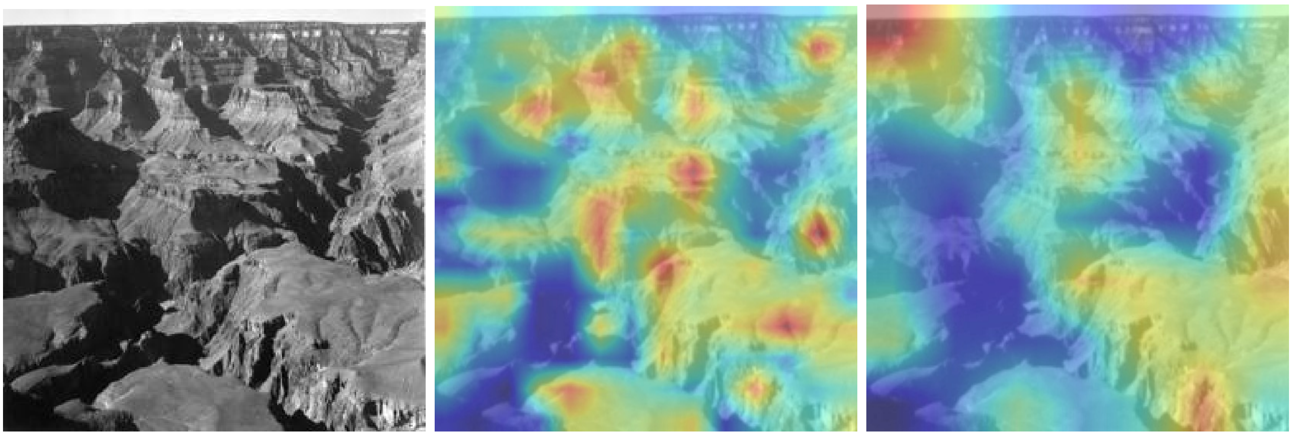
Figure 1: Accuracy vs FLOPs

One surprising finding is that certain classes like "Canyon" and "Castle" challenge all models. We will look deeper into this with the help of some attention maps in the following section.

When comparing the overall accuracy of the different models with the computational cost for fine-tuning (*Figure 1*), it can be seen that the EfficientNet-B7 is a lot "cheaper", but due to its bad accuracy not competitive. The other models have similar accuracy, while differing strength in terms of cost.

4.2 Missclassified Attention maps

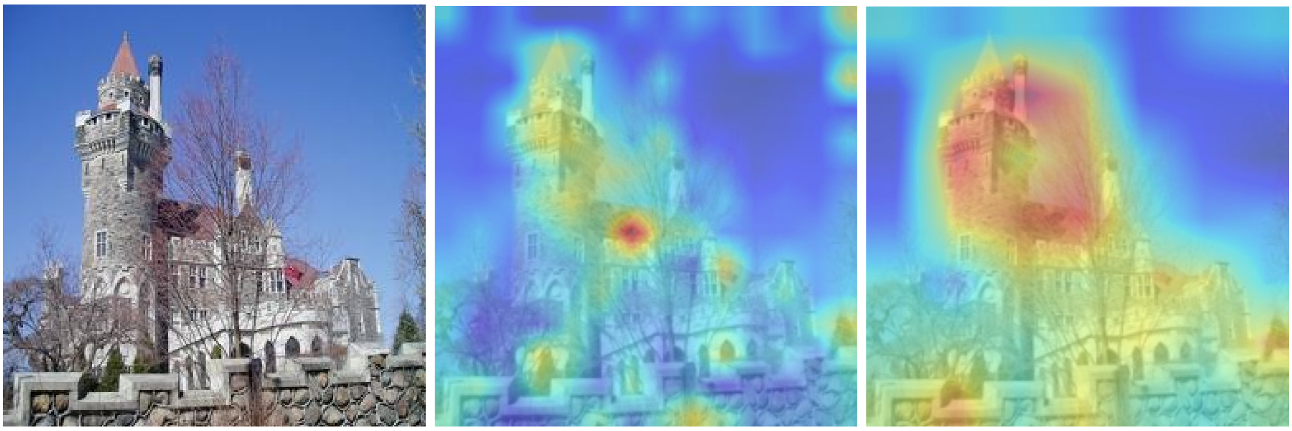
As described in the previous section certain classes pose problems to all models. In order to evaluate why this is the case we extracted images from the classes pair canyon and mountains, and church_outdoor and castle and looked at the correspondent attention maps for the ViTs and hybrid model (ViT_B16, ViT_B32, R50_ViT_B16). As we will see from the images 2, 3, 4, 5, the models *ViT_B16* and *ViT_B32* demonstrate contrasting performance, that likely arise from their patch sizes and attention mechanisms. *ViT_B16* uses 16×16 patches, focusing on fine-grained details, while *ViT_B32* uses larger 32×32 patches, which capture broader but less detailed features.



(a) Original Image (b) ViT_B_16 (c) ViT_B_32

Figure 2: Attention maps for canyon image using the ViT model

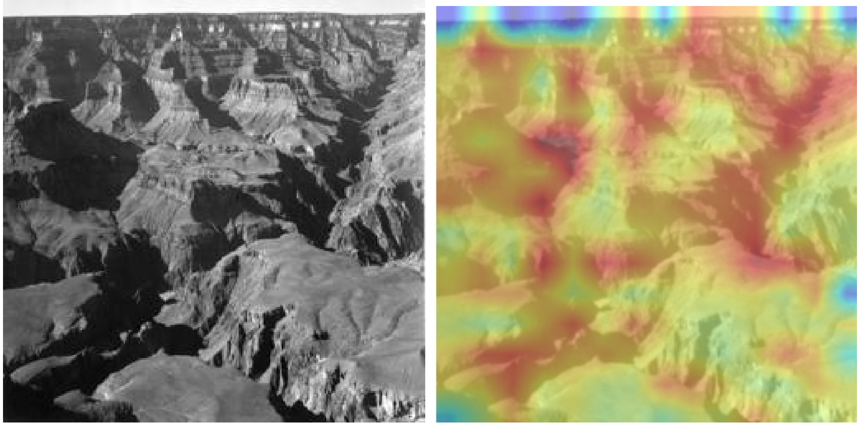
In the "canyon" scenario (*Figure 2*), *ViT_B16* correctly classifies the image by trying to identify the intricate patterns, the depth of the image, and steep ridges. On the other hand, *ViT_B32* misclassifies it as a "mountain" due to its coarser patch size, which overlooks the depth and unique contours of the canyon.



(a) Original Image (b) ViT_B_16 (c) ViT_B_32

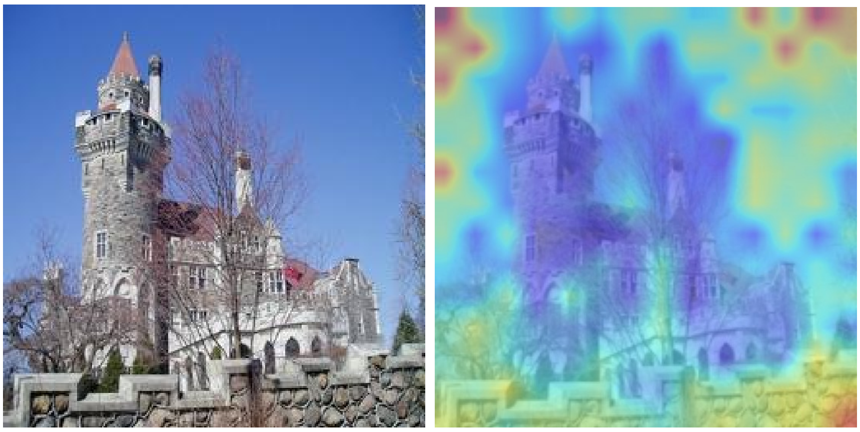
Figure 3: Attention maps for castle image using the ViT model

In the "castle" scenario (*Figure 3*), their behaviors are reversed. *ViT_B16* misclassifies it as a "church_outdoor" overly focusing on architectural details like the spire and stone walls. On the other hand, *ViT_B32* correctly identifies it, leveraging its broader attention to recognize the open surroundings that distinguish a castle.



(a) Original Image (b) R50+ViT_B_16

Figure 4: Attention maps for canyon image using the Hybrid model



(a) Original Image (b) R50+ViT_B_16

Figure 5: Attention maps for castle image using the Hybrid model

If we look at the hybrid model (*R50+ViT_B16*), we can see that it achieves mixed results. It correctly classifies the "canyon" (*Figure 4*) benefiting from its combination of fine detail recognition and contextual understanding. However, it also misclassifies the "castle" (*Figure 5*) as a "church_outdoor", prioritizing architectural elements while neglecting the broader scene. These findings highlight the trade-offs between local detail and global context. Smaller patches like those in *ViT_B16* excel in capturing fine-grained features but may miss larger patterns, while coarser patches in *ViT_B32* offer better contextual awareness but struggle with subtle distinctions. The hybrid approach shows promise but still faces challenges in balancing these aspects effectively.

5 Conclusions and Future Work

The study highlights the strengths and limitations of different architectures in scene classification: except for the EfficientNet-B7, restricted by its simpler architecture, all models perform competitively in terms of accuracy, but differ in fine-tuning cost. The attention maps provide a nuanced understanding of model behavior.

Experiments could be extended to other datasets for broader generalization, and self-supervised pre-training for ViTs on Places 365 can be explored. Other ideas involve investigating efficient ViTs for resource-constrained scenarios, and inquire into domain-specific pre-training on Places365 for improved performance.

References

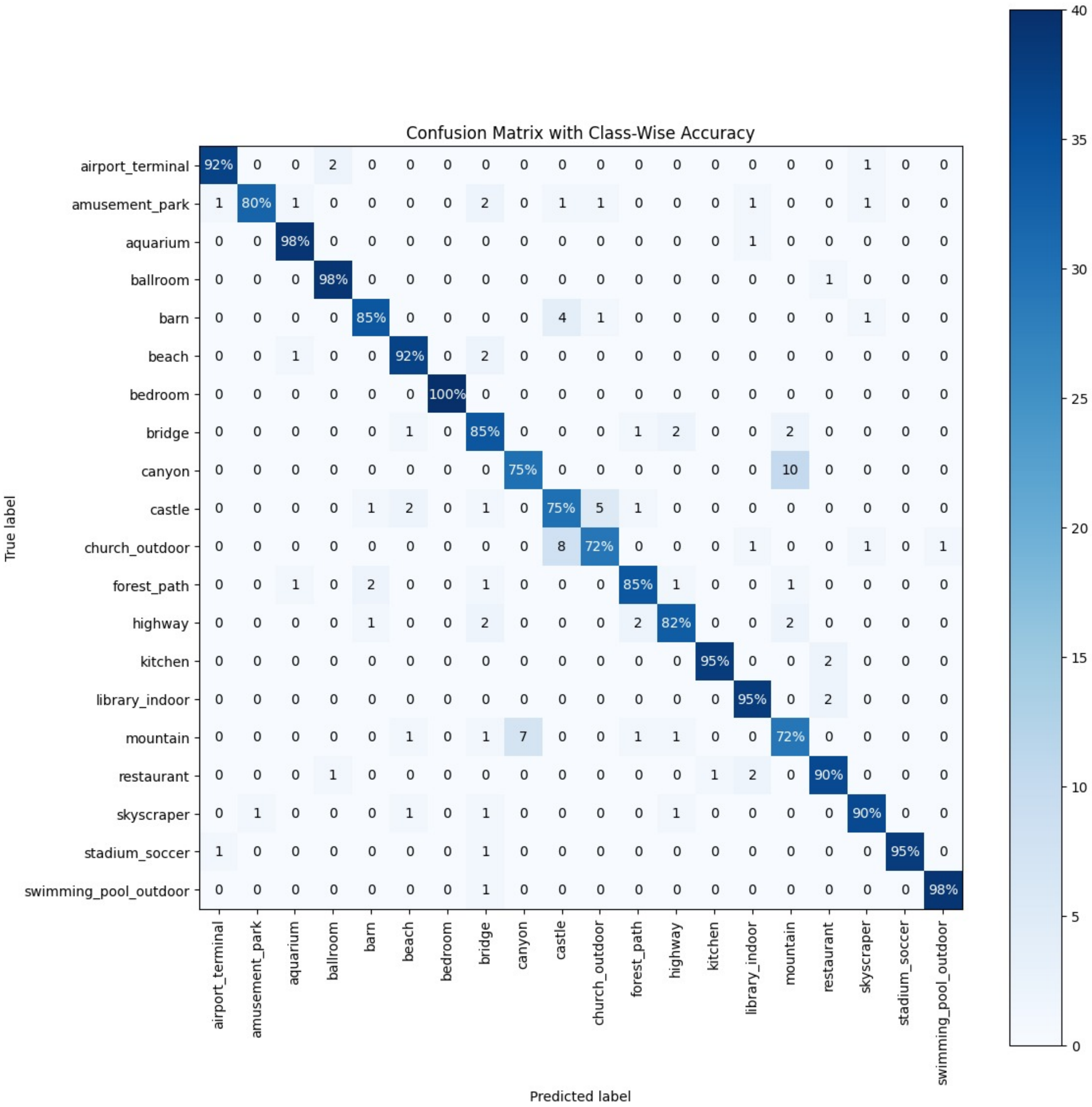
[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.

[2] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. 2016.

[3] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017.

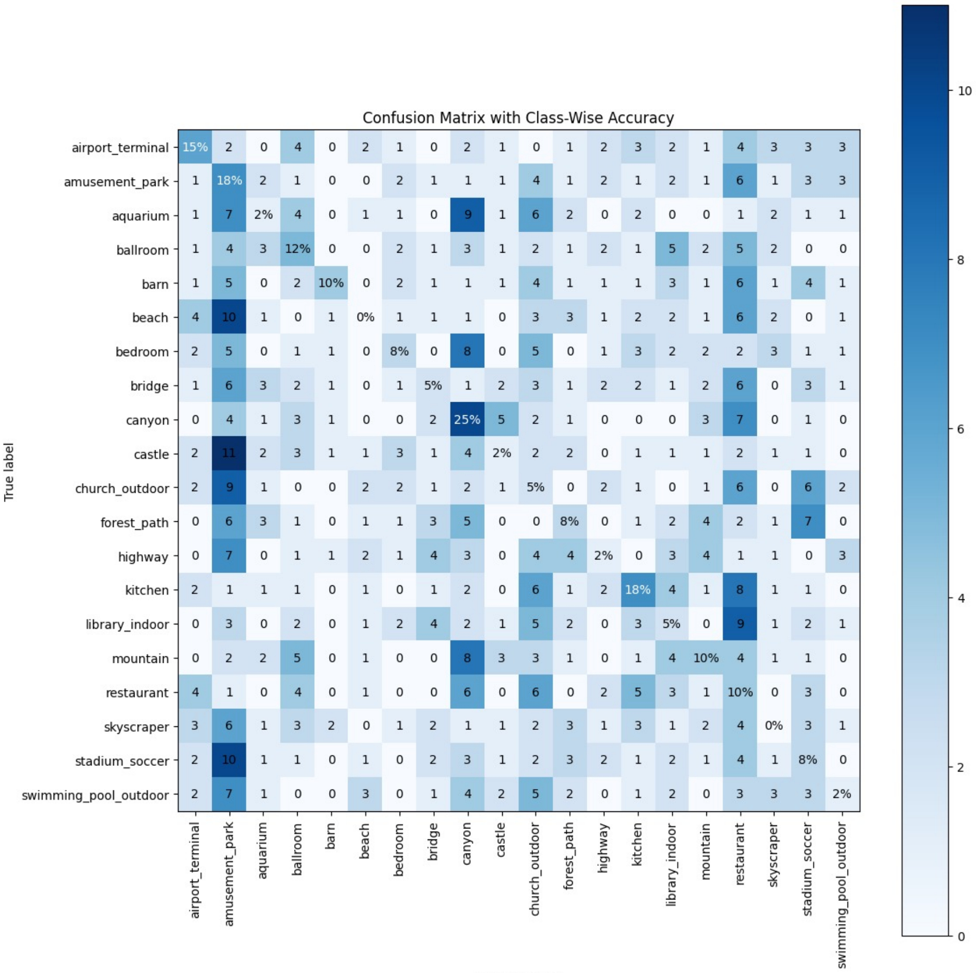
7 Appendix

7.1 Confusion Matrices



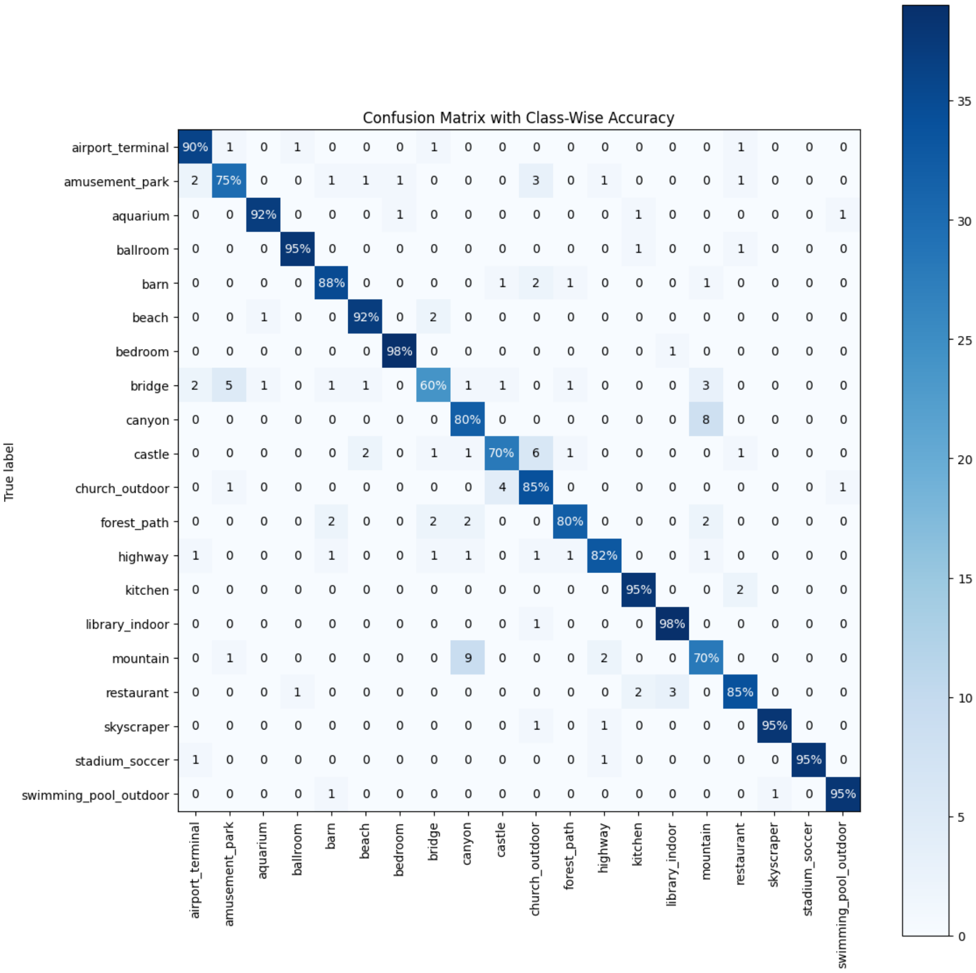
True label	airport_terminal	amusement_park	aquarium	ballroom	barn	beach	bedroom	bridge	canyon	castle	church_outdoor	forest_path	highway	kitchen	library_indoor	mountain	restaurant	skyscraper	stadium_soccer	swimming_pool_outdoor
airport_terminal	92%	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
amusement_park	1	80%	1	0	0	0	0	2	0	1	1	0	0	0	1	0	0	1	0	0
aquarium	0	0	98%	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ballroom	0	0	0	98%	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
barn	0	0	0	0	85%	0	0	0	0	4	1	0	0	0	0	0	0	1	0	0
beach	0	0	1	0	0	92%	0	2	0	0	0	0	0	0	0	0	0	0	0	0
bedroom	0	0	0	0	0	0	100%	0	0	0	0	0	0	0	0	0	0	0	0	0
bridge	0	0	0	0	0	1	0	85%	0	0	0	1	2	0	0	2	0	0	0	0
canyon	0	0	0	0	0	0	0	0	75%	0	0	0	0	0	0	10	0	0	0	0
castle	0	0	0	0	1	2	0	1	0	75%	5	1	0	0	0	0	0	0	0	0
church_outdoor	0	0	0	0	0	0	0	0	0	8	72%	0	0	0	1	0	0	1	0	1
forest_path	0	0	1	0	2	0	0	1	0	0	0	85%	1	0	0	1	0	0	0	0
highway	0	0	0	0	1	0	0	2	0	0	0	2	82%	0	0	2	0	0	0	0
kitchen	0	0	0	0	0	0	0	0	0	0	0	0	0	95%	0	0	2	0	0	0
library_indoor	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95%	0	2	0	0	0
mountain	0	0	0	0	0	1	0	1	7	0	0	1	1	0	0	72%	0	0	0	0
restaurant	0	0	0	1	0	0	0	0	0	0	0	0	0	1	2	0	90%	0	0	0
skyscraper	0	1	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	90%	0	0
stadium_soccer	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	95%	0
swimming_pool_outdoor	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	98%

Figure 6: Hybrid model confusion matrix



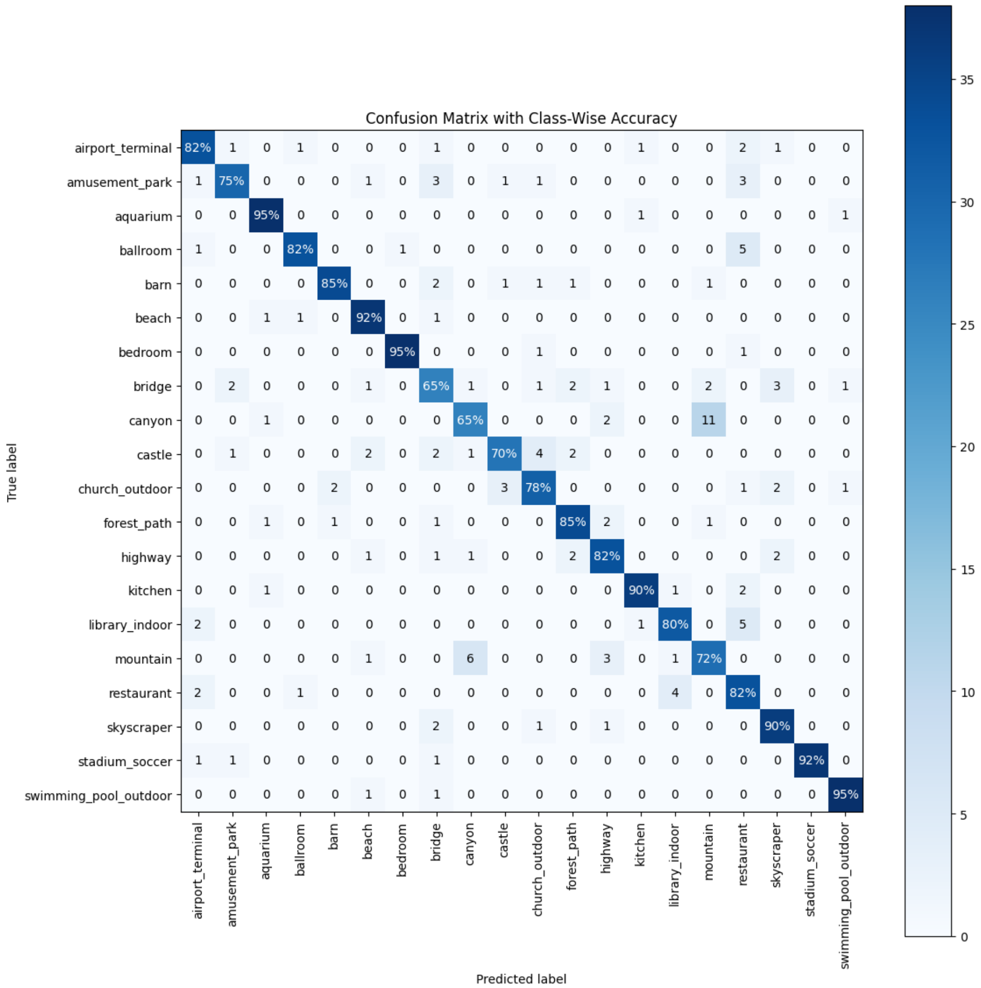
	airport_terminal	amusement_park	aquarium	ballroom	barn	beach	bedroom	bridge	canyon	castle	church_outdoor	forest_path	highway	kitchen	library_indoor	mountain	restaurant	skyscraper	stadium_soccer	swimming_pool_outdoor
airport_terminal	15%	2	0	4	0	2	1	0	2	1	0	1	2	3	2	1	4	3	3	3
amusement_park	1	18%	2	1	0	0	2	1	1	1	4	1	2	1	2	1	6	1	3	3
aquarium	1	7	2%	4	0	1	1	0	9	1	6	2	0	2	0	0	1	2	1	1
ballroom	1	4	3	12%	0	0	2	1	3	1	2	1	2	1	5	2	5	2	0	0
barn	1	5	0	2	10%	0	2	1	1	1	4	1	1	1	3	1	6	1	4	1
beach	4	10	1	0	1	0%	1	1	1	0	3	3	1	2	2	1	6	2	0	1
bedroom	2	5	0	1	1	0	8%	0	8	0	5	0	1	3	2	2	2	3	1	1
bridge	1	6	3	2	1	0	1	5%	1	2	3	1	2	2	1	2	6	0	3	1
canyon	0	4	1	3	1	0	0	2	25%	5	2	1	0	0	0	3	7	0	1	0
castle	2	11	2	3	1	1	3	1	4	2%	2	2	0	1	1	1	2	1	1	0
church_outdoor	2	9	1	0	0	2	2	1	2	1	5%	0	2	1	0	1	6	0	6	2
forest_path	0	6	3	1	0	1	1	3	5	0	0	8%	0	1	2	4	2	1	7	0
highway	0	7	0	1	1	2	1	4	3	0	4	4	2%	0	3	4	1	1	0	3
kitchen	2	1	1	1	0	1	0	1	2	0	6	1	2	18%	4	1	8	1	1	0
library_indoor	0	3	0	2	0	1	2	4	2	1	5	2	0	3	5%	0	9	1	2	1
mountain	0	2	2	5	0	1	0	0	8	3	3	1	0	1	4	10%	4	1	1	0
restaurant	4	1	0	4	0	1	0	0	6	0	6	0	2	5	3	1	10%	0	3	0
skyscraper	3	6	1	3	2	0	1	2	1	1	2	3	1	3	1	2	4	0%	3	1
stadium_soccer	2	10	1	1	0	1	0	2	3	1	2	3	2	1	2	1	4	1	8%	0
swimming_pool_outdoor	2	7	1	0	0	3	0	1	4	2	5	2	0	1	2	0	3	3	3	2%

(a) Effnet model confusion matrix



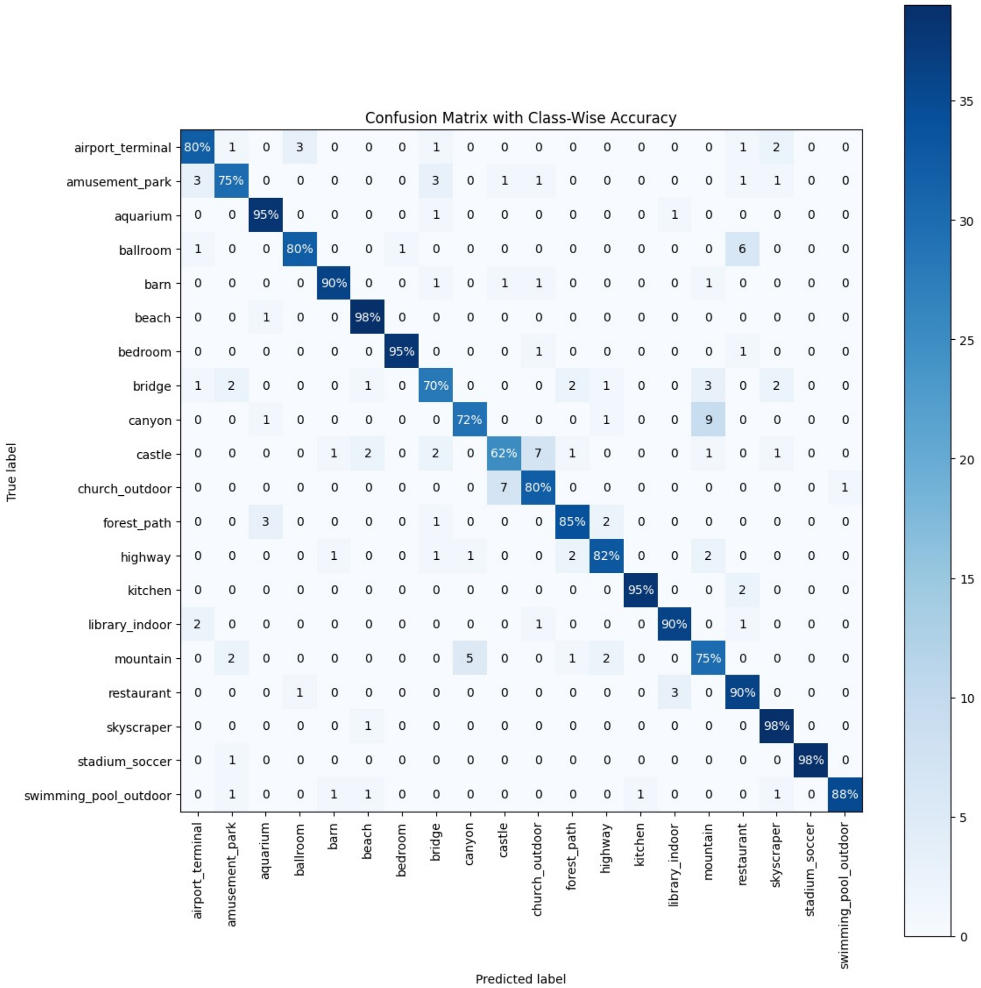
	airport_terminal	amusement_park	aquarium	ballroom	barn	beach	bedroom	bridge	canyon	castle	church_outdoor	forest_path	highway	kitchen	library_indoor	mountain	restaurant	skyscraper	stadium_soccer	swimming_pool_outdoor
airport_terminal	90%	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
amusement_park	2	75%	0	0	1	1	1	0	0	0	3	0	1	0	0	0	1	0	0	0
aquarium	0	0	92%	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1
ballroom	0	0	0	95%	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
barn	0	0	0	0	88%	0	0	0	0	1	2	1	0	0	0	1	0	0	0	0
beach	0	0	1	0	0	92%	0	2	0	0	0	0	0	0	0	0	0	0	0	0
bedroom	0	0	0	0	0	0	98%	0	0	0	0	0	0	0	1	0	0	0	0	0
bridge	2	5	1	0	1	1	0	60%	1	1	0	1	0	0	0	3	0	0	0	0
canyon	0	0	0	0	0	0	0	0	80%	0	0	0	0	0	0	8	0	0	0	0
castle	0	0	0	0	0	2	0	1	1	70%	6	1	0	0	0	0	1	0	0	0
church_outdoor	0	1	0	0	0	0	0	0	0	4	85%	0	0	0	0	0	0	0	0	1
forest_path	0	0	0	0	2	0	0	2	2	0	0	80%	0	0	0	2	0	0	0	0
highway	1	0	0	0	1	0	0	1	1	0	1	1	82%	0	0	1	0	0	0	0
kitchen	0	0	0	0	0	0	0	0	0	0	0	0	0	95%	0	0	2	0	0	0
library_indoor	0	0	0	0	0	0	0	0	0	0	1	0	0	0	98%	0	0	0	0	0
mountain	0	1	0	0	0	0	0	0	9	0	0	0	2	0	0	70%	0	0	0	0
restaurant	0	0	0	1	0	0	0	0	0	0	0	0	0	2	3	0	85%	0	0	0
skyscraper	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	95%	0	0
stadium_soccer	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	95%	0
swimming_pool_outdoor	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	95%

(b) ResNet model confusion matrix



	airport_terminal	amusement_park	aquarium	ballroom	barn	beach	bedroom	bridge	canyon	castle	church_outdoor	forest_path	highway	kitchen	library_indoor	mountain	restaurant	skyscraper	stadium_soccer	swimming_pool_outdoor
airport_terminal	82%	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	2	1	0	0
amusement_park	1	75%	0	0	0	1	0	3	0	1	1	0	0	0	0	0	3	0	0	0
aquarium	0	0	95%	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
ballroom	1	0	0	82%	0	0	1	0	0	0	0	0	0	0	0	0	5	0	0	0
barn	0	0	0	0	85%	0	0	2	0	1	1	1	0	0	0	1	0	0	0	0
beach	0	0	1	1	0	92%	0	1	0	0	0	0	0	0	0	0	0	0	0	0
bedroom	0	0	0	0	0	0	95%	0	0	0	1	0	0	0	0	0	1	0	0	0
bridge	0	2	0	0	0	1	0	65%	1	0	1	2	1	0	0	2	0	3	0	1
canyon	0	0	1	0	0	0	0	0	65%	0	0	0	2	0	0	11	0	0	0	0
castle	0	1	0	0	0	2	0	2	1	70%	4	2	0	0	0	0	0	0	0	0
church_outdoor	0	0	0	0	2	0	0	0	0	3	78%	0	0	0	0	0	1	2	0	1
forest_path	0	0	1	0	1	0	0	1	0	0	0	85%	2	0	0	1	0	0	0	0
highway	0	0	0	0	0	1	0	1	1	0	0	2	82%	0	0	0	0	2	0	0
kitchen	0	0	1	0	0	0	0	0	0	0	0	0	0	90%	1	0	2	0	0	0
library_indoor	2	0	0	0	0	0	0	0	0	0	0	0	0	1	80%	0	5	0	0	0
mountain	0	0	0	0	0	1	0	0	6	0	0	0	3	0	1	72%	0	0	0	0
restaurant	2	0	0	1	0	0	0	0	0	0	0	0	0	0	4	0	82%	0	0	0
skyscraper	0	0	0	0	0	0	0	2	0	0	1	0	1	0	0	0	0	90%	0	0
stadium_soccer	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	92%	0
swimming_pool_outdoor	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	95%

(a) ViT_B_32 model confusion matrix



	airport_terminal	amusement_park	aquarium	ballroom	barn	beach	bedroom	bridge	canyon	castle	church_outdoor	forest_path	highway	kitchen	library_indoor	mountain	restaurant	skyscraper	stadium_soccer	swimming_pool_outdoor
airport_terminal	80%	1	0	3	0	0	0	1	0	0	0	0	0	0	0	0	1	2	0	0
amusement_park	3	75%	0	0	0	0	0	3	0	1	1	0	0	0	0	0	1	1	0	0
aquarium	0	0	95%	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
ballroom	1	0	0	80%	0	0	1	0	0	0	0	0	0	0	0	0	6	0	0	0
barn	0	0	0	0	90%	0	0	1	0	1	1	0	0	0	0	1	0	0	0	0
beach	0	0	1	0	0	98%	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bedroom	0	0	0	0	0	0	95%	0	0	0	1	0	0	0	0	0	1	0	0	0
bridge	1	2	0	0	0	1	0	70%	0	0	0	2	1	0	0	3	0	2	0	0
canyon	0	0	1	0	0	0	0	0	72%	0	0	0	1	0	0	9	0	0	0	0
castle	0	0	0	0	1	2	0	2	0	62%	7	1	0	0	0	1	0	1	0	0
church_outdoor	0	0	0	0	0	0	0	0	0	7	80%	0	0	0	0	0	0	0	0	1
forest_path	0	0	3	0	0	0	0	1	0	0	0	85%	2	0	0	0	0	0	0	0
highway	0	0	0	0	1	0	0	1	1	0	0	2	82%	0	0	2	0	0	0	0
kitchen	0	0	0	0	0	0	0	0	0	0	0	0	0	95%	0	0	2	0	0	0
library_indoor	2	0	0	0	0	0	0	0	0	0	1	0	0	0	90%	0	1	0	0	0
mountain	0	2	0	0	0	0	0	0	5	0	0	1	2	0	0	75%	0	0	0	0
restaurant	0	0	0	1	0	0	0	0	0	0	0	0	0	0	3	0	90%	0	0	0
skyscraper	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	98%	0	0
stadium_soccer	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98%	0
swimming_pool_outdoor	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	1	0	88%

(b) ViT_B_16 model confusion matrix