SCENE DETECTION ALGORITHM EVALUATION

BY VISION-LANGUAGE MODELS

**Or Trabelsi, Naor Cohen**

Tel-Aviv University

{ortrabelsi, naorcohen1}@mail.tau.ac.il

# Abstract

This project ventures into Scene Detection, a vital software technique in photography, autonomously categorizing the scene type within a given image—landscapes, portraits, food, or low-light settings, among others. This categorization lays the groundwork for augmenting image quality in camera applications. Three pre-trained algorithms—CLIP, BLIP, and BLIP-2—were scrutinized using the TOP-1 and TOP-3 accuracy metrics on real-world photos, operating as zero-shot classifiers and further refined through linear probe and LoRA techniques. Utilizing the Camera Scene Detection Dataset (CamSDD) initially, with over 11,000 images across 30 diverse categories, and later integrating the extensive Places365 standard dataset, a thorough insight into the potential of scene detection for image quality enhancement was achieved through the utilization of pre-trained VLMs. The study compared the performance of these algorithms with reference models—EfficientNet-B0, MobileNet-V2, and Xception—drawn from a comparative study. Empirical findings underscored the superior performance of the Linear Probe technique across CLIP, BLIP, and BLIP-2, achieving TOP-1 accuracies of 96.00%, 97.17%, and 97.17% ,respectively, and a perfect TOP-3 accuracy which close to 100.00%. Conversely, the reference models also exhibited commendable performance, notably MobileNet-V2 with a TOP-1 accuracy of 94.17% and a TOP-3 accuracy of 98.67%. The zero-shot configurations, although with lower accuracy, shed light on the algorithms' innate scene detection capabilities before fine-tuning. This exploration accentuates the promise held by scene detection algorithms in propelling photographic image quality forward, with a spotlight on the efficacy of the Linear Probe and LoRA techniques in enhancing classification accuracy. Our code and pretrained models are available at the following project’s GitHub repository:<https://github.com/naorJR/SceneDetection>

# Introduction

In the digital photography domain, the quest for the quintessential image hinges on a delicate interplay between hardware capabilities and intelligent software techniques. Among the arsenal of software techniques, Scene Detection shines as a linchpin. By autonomously discerning the type of scene within a given image—whether landscapes, portraits, food, or low-light environments—Scene Detection lays the foundation for an array of image quality enhancement measures in camera applications.

|  |  |
| --- | --- |
|  | **Figure 1.** Pre-training model architecture and objectives of BLIP (similar parameters share the same color). We propose the advanced fine-tuning technique implementation on pretrained multimodal mixture of encoder-decoder, a unified vision-language model operating in two functionalities: (1) Unimodal encoder is pretrained with an image-text contrastive (ITC) loss to align vision and language representations. (2) Image-grounded text encoder uses additional cross-attention layers to model vision-language interactions. Each of them is fine-tuned by LoRA technique which is employed on the visual encoder. |

The advent of deep learning has significantly propelled the field of Scene Detection forward. A notable contribution in this realm is the paper "Fast and Accurate Camera Scene Detection on Smartphones," which introduced the CamSDD dataset [1]. This dataset, encompassing a diverse range of real-world scenes, has emerged as a valuable resource for advancing scene detection methodologies, particularly on handheld devices. The insights from this paper and the CamSDD dataset have shed light on the challenges and potential solutions in achieving fast and accurate scene detection, thereby serving as a cornerstone for our investigation.

At the heart of our project is the evaluation of different pretrained vision and language models (VLMs) models for the specific Scene Detection task—CLIP, BLIP, and BLIP-2—in image quality enhancement. These algorithms, with their unique capabilities of bridging the gap between vision and language understanding, present a promising avenue for bolstering scene classification insight. By conducting a rigorous evaluation on real-world photos from the CamSDD, and preparing for implementation on the Places365 datasets, we aim to provide crucial insights into how Scene Detection affects image quality improvement.

Our meticulous analysis revealed that among the examined algorithms, BLIP exhibited superior performance in TOP-1 accuracy metric, thereby heralding a promising avenue for improving classification accuracy for subsequent image quality enhancement tasks. The findings underscore the potential of harnessing advanced Scene Detection algorithms through pretrained VLMs for substantial image quality enhancement in real-world photography applications.

The structure of this paper is outlined as follows: Section 2 provides a review of related literature, focusing on the evolution of CLIP, BLIP, and BLIP-2 algorithms together with the present methodologies for scene detection on handheld devices. Section 3 details the methods and procedures utilized in our research, followed by a presentation of our findings in Section 4. Section 5 engages in a discussion on the implications of our results, set against existing literature. The conclusion, Section 6, summarizes our findings and recommends pathways for future research in this domain.

# Related Work

Scene detection, a pivotal facet in the domain of digital photography, has witnessed remarkable advancements courtesy of deep learning algorithms. These algorithms not only distinguish the type of scene in a given image but also lay the groundwork for augmenting image quality in camera applications. A major step forward in this domain has occurred with the introduction of models that combine VLMs, represented by Contrastive Language-Image Pretraining (CLIP).

CLIP [3] operates by mapping images and text labels into a shared embedding space. During inference, the vectors of image and text are compared for similarity to deduce the image's class, thereby eliminating the need for task-specific fine-tuning. This model served as a precursor to the development of Bootstrapping Language-Image Pre-training (BLIP) and its successor, BLIP-2. BLIP [4] enhances the image classifier by leveraging pseudo-labels generated from an initial CLIP model, without the requirement for additional labeled data. BLIP-2 [5], on the other hand, employs an efficient pre-training strategy for vision-language tasks by utilizing frozen pre-trained image encoders and large language models, demonstrating superior performance on various tasks with fewer trainable parameters.

The literature further showcases efforts to accelerate scene detection on handheld devices. In "Fast and Accurate Camera Scene Detection on Smartphones" by Angeline Pouget et al., 2021, the authors delineate methodologies to expedite scene detection on smartphones, a study which forms a cornerstone of our investigation. Similarly, the work CLIPCAP [2] extends the CLIP framework to image captioning, evidencing the versatility and efficacy of CLIP-based models in different domains.

Fine-tuning techniques like Linear Probe [8] and Low-Rank Adaptation (LoRA) [7] have emerged as indispensable tools in enhancing the performance and interpretability of these models. Linear Probe involves training a linear classifier atop the fixed features extracted by the pre-trained models, while LoRA provides a granular insight into the contributions of each feature towards the final decision, aiding in model interpretation. Figure 2 illustrates our implementation of LoRA on BLIP model for the classification tasks.

The entry of these models and techniques has significantly uplifted the performance of scene detection algorithms, presenting a promising opportunity for real-world photography and image quality enhancement. The aforementioned works and the referenced models offer a sturdy foundation for our assessment of scene detection algorithms for enhanced photographic image quality.

# Data

For our project, we engaged with two principal datasets; the Camera Scene Detection Dataset (CamSDD) [1] and the Places365 standard dataset [9][10].

## Datasets

CamSDD. Our primary experiments leveraged CamSDD, encompassing over 11,000 images spread across 30 distinct categories. This dataset served as a robust foundation for our initial explorations into scene detection.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| Portrait | Group Portrait | Kids | Dog | Cat | Macro |
|  |  |  |  |  |  |
| Food \ Gourmet | Beach | Mountains | Waterfall | Snow | Landscape |
|  |  |  |  |  |  |
| Underwater | Architecture | Sunrise & Sunset | Blue Sky | Overcast | Greenery |
|  |  |  |  |  |  |
| Autumn Plants | Flowers | Night Shots | Stage | Fireworks | Candlelight |
|  |  |  |  |  |  |
| Neon Lights | Indoor | Backlight | Document | QR Code | Monitor Screen |

**Figure 2.** Visualization of the 30 Camera Scene Detection Dataset (CamSDD) categories.

Places365-Standard. To augment our data pool and deepen our investigation, we integrated the Places365 standard dataset, known for its extensive image collection, thus providing a substantial base for training and evaluating our models.

The data distribution is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Train | Validation | Test | Categories |
| *CamSDD* | ~9.9k | 600 | 600 | 30 |
| *Places365* | ~1.8M | 36k | - | 365 |

**Table 1.** CamSDD and Places365-Strd. Datasets distribution.

## Preprocessing

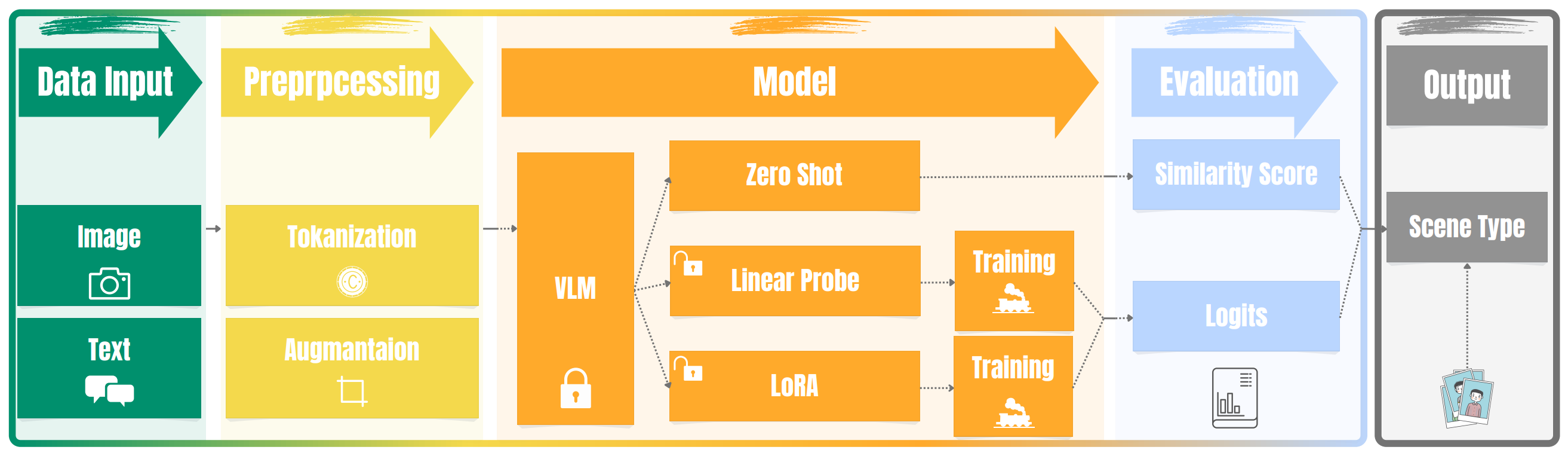
Image Preprocessing. Our preprocessing pipeline commenced by converting images to RGB format, followed by resizing and cropping them to a uniform dimension of 224x224 pixels. Subsequently, images were converted to PyTorch Tensors and normalized using predefined mean and standard deviation values, aligning with the preprocessing steps necessitated by the CLIP and BLIP visual processors.

Text Preprocessing. The text captions/labels associated with images were processed for our case by appending a prompt. However, we acknowledge that for other applications, further efforts may be needed to truncate words to the maximum length and rectify punctuation and spaces, as dictated by the tokenization requirements of CLIP and BLIP.

These preprocessing steps were pivotal in ensuring that the data was in a suitable format for training and evaluating our models, thereby facilitating a more accurate assessment of the scene detection algorithms under investigation.

# Method

In this section, we explain our framework which outlines various methods to enhance the performance of “traditional” well-known deep learning models on scene detection tasks, by utilizing pre-trained VLMs developed for more generalized tasks. Section 4.1 outlines the data preprocessing observing to the demands of CLIP and BLIP visual processors and tokenizers. Section 4.2 introduces the model architecture, employing three notable algorithms—CLIP, BLIP, and BLIP-2, each with a distinct strategy for image-text representation. In Section 4.3, fine-tuning techniques including Linear Probe and LoRA are discussed. Section 4.4 delves into the training and evaluation on CamSDD and Places365 datasets, followed by Section 4.5 detailing the selected loss functions for optimizing the training process. The accuracy assessment methodology, comprising Top-1 and Top-3 accuracy metrics, is expounded in Section 4.6.



**Figure 3.** Illustrates the project workflow from left to right, beginning with data input, transitioning through preprocessing and model architecture selection, to training and evaluation. Finally, the model-generated output alongside a specified photo is displayed.

## Data Preprocessing

Image and text preprocessing was executed as per the requirements of CLIP and BLIP visual processors and tokenizers as elaborated in section 3.2.

## Model Architecture

We employed three distinguished algorithms—CLIP, BLIP, and BLIP-2. CLIP classifies images by mapping them and text labels into a shared embedding space, eliminating the need for task-specific fine-tuning. BLIP builds upon CLIP, adding a bootstrapping mechanism to refine the image classifier using pseudo-labels from an initial CLIP model. BLIP-2, an efficient pre-training strategy, leverages frozen pre-trained image encoders and large language models to bridge the visual and language modalities in two training stages.

## Fine-tuning Techniques

Fine-tuning is a pivotal step in enhancing the performance of pre-trained models on specific tasks. In this project, two notable fine-tuning techniques were employed: Linear Probe and LoRA.

Linear Probe. Linear Probing involves tuning the model's head while freezing the lower layers, thus preserving the pretrained features. It performs better in out-of-distribution (OOD) settings by preventing feature distortion, unlike full fine-tuning which adapts the features to the specific task, potentially leading to reduced OOD performance. However, it doesn’t perform as well as full fine-tuning in in-distribution (ID) settings since it cannot adapt the features to the downstream task.

Having explored the effectiveness of the Linear Probe technique, we then delved into another novel and promising fine-tuning approach, LoRA.

LoRA. LoRA is designed for the effective fine-tuning of pre-trained models through low-rank parameter adaptation, which not only preserves the pre-trained parameters but also retains the original model structure by adding a low-rank matrix to the pre-trained weights. This mechanism addresses the challenge of fine-tuning large pre-trained models with limited data, offering a pathway to adapt the model to specific tasks without a significant increase in the number of parameters or computational resources. Inspired by this, and the findings of Aghajanyan et al. (2020) regarding the low "intrinsic dimension" of pre-trained language models, the authors propose a method centered on the hypothesis that the updates to the weights also exhibit a low "intrinsic rank" during adaptation. This is operationalized through a low-rank decomposition for updating a pre-trained weight matrix ​, denoted as , where and are matrices with dimensions and respectively, and is the rank which is lesser than the minimum of the input dimensions and . During the training process, ​ is held constant, while and serve as the trainable parameters. The modified forward pass is articulated as . The matrices and are initialized with a random Gaussian and zeros respectively, and the term is scaled by ​, where is a constant. This scaling mechanism simplifies the tuning of hyperparameters when varying , presenting a structured approach for effective fine-tuning towards task-specific requirements while aligning with the principles of LoRA.

## Training and Evaluation

Models were trained on the CamSDD dataset, and the code was also been implemented for the Places365 dataset, with evaluation based on TOP-1 and TOP-3 accuracy metrics, according to the comparative study we worked with. The evaluation process entailed the computation of TOP-N accuracies post the logit transformation step. In this step, the raw model outputs, known as logits, were transformed using a softmax function to yield probabilities for each class.

Subsequently, the TOP-1 and TOP-3 accuracies were calculated to assess the model's performance in correctly identifying the true class labels among the top predicted classes.

This method of evaluation provided a clear insight into the model's capability in handling multi-class classification tasks, aligning with the established evaluation protocols in the comparative study referenced.

## Loss Functions

In CLIP, images and text are mapped into a shared embedding space using a vision model and a language model respectively. The objective is to minimize the distance between semantically similar text and image representations in the embedding space while maximizing the distance between dissimilar ones. The formulation can be given as follows:

where ​ and ​ are the image and text representations of the data sample, is a similarity measure (e.g., cosine similarity), is the batch size, and is a temperature parameter that scales the SoftMax function, sharpens or softens the distribution. The fraction within the logarithm forms a probability distribution over all pairs, focusing on the pair, and the term computes the negative log-likelihood.

BLIP is a Vision-Language Pre-training framework designed for noisy image-text pairs. It features a new model architecture called Multimodal mixture of Encoder-Decoder (MED) to handle three functionalities: unimodal encoding, image-grounded text encoding, and image-grounded text decoding. During pre-training, three objectives are optimized: Image-Text Contrastive Loss (ITC), Image-Text Matching Loss (ITM), and Language Modeling Loss (LM). These objectives help align the feature space between visuals and text, learn multimodal representations, and generate textual descriptions from images. The model aims for efficient pre-training by sharing parameters between the text encoder and decoder while differentiating self-attention layers.

## Accuracy Calculations

The accuracy of the models was assessed using Top-1 and Top-3 accuracy metrics, which are widely recognized and utilized in the field of image classification and scene detection for evaluating model performance. These metrics have been used in previous works, and are considered standard for providing a clear indication of a model's ability to correctly identify the true class labels among the top predicted classes.

Top-1 Accuracy. Top-1 accuracy examines whether the top predicted class matches the target label. It is computed as the number of times the predicted label matched the target label, divided by the number of data points evaluated.

Top-3 Accuracy. Top-3 accuracy, on the other hand, checks if the target label is among the top three predictions. Like Top-1 accuracy, it provides insight into the model’s capability in handling multi-class classification tasks, but with a broader margin of correctness by considering the top three predicted classes.

These metrics were instrumental in evaluating the performance of the proposed scene detection algorithms, and their widespread use in similar studies and projects further validates our choice of evaluation metrics for this project.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

**Figure 4.** (a) Heatmap illustration of the prediction probabilities of different categories by various methods. (b) An arbitrary given image from the CamSDD dataset.

# Results

We present a detailed evaluation of the scene recognition performance exhibited by various methods under different configurations. The results shed light on the efficacy of the proposed framework when positioned side by side with reference models results. The evaluation metrics employed include Top-1 Accuracy and Top-3 Accuracy percentages, which are pivotal indicators of model precision.

**CamSDD**

|  |  |  |
| --- | --- | --- |
| **Method** | **Top-1 Acc. %** | **Top-3 Acc. %** |
| *zero-shot* |  |  |
| CLIP | 87.83 | 96.33 |
| BLIP | 85.00 | 94.50 |
| BLIP-2 | 82.83 | 95.67 |
| *finetuning: linear-probe* | | |
| CLIP | 96.00 | **100.00** |
| BLIP | **97.17** | **100.00** |
| BLIP-2 | **97.17** | **100.00** |
| *finetuning: LoRA* | | |
| CLIP | 96.67 | 99.83 |
| BLIP | 96.67 | **100.00** |
| BLIP-2 | 96.33 | 99.83 |
| *References from [1]* | | |
| **MobileNet-V2** | **94.17** | **98.67** |
| Xception | 86.33 | 98.17 |
| EfficientNet-B0 | 91.33 | 98.67 |

(a)

**Places365**

|  |  |  |
| --- | --- | --- |
| **Method** | **Top-1 Acc. %** | **Top-3 Acc. %** |
| *zero-shot* |  |  |
| CLIP | 37.65 | 58.00 |
| *finetuning: linear-probe* | | |
| CLIP | 47.75 | 61.51 |
| *finetuning: LoRA* | | |
| CLIP | 46.02 | 61.29 |

(b)

**Table 2.** (a) Top-1 and Top-3 classification accuracy of the evaluated Clip \ Blip \ Blip-2 based models with CamSDD dataset. The results of the other architectures are provided for the reference. (b) Top-1 and Top-3 classification accuracy of the evaluated Clip based models with Places dataset.

The results in Fig. 4 reveal a high level of confidence across most methods in classifying the category "Greenery," with probabilities nearing or at 99.99%, indicating that the features associated with this category are well-recognized by these methods. A comparison between zero-shot and trained methods like linear probes or LoRA showcases a less balanced distribution of confidence across categories for zero-shot methods, especially CLIP zero-shot, which has significantly lower probabilities for categories other than "Greenery." Notably, there's a diversity in category recognition among different methods, for instance, BLIP zero-shot has a relatively high probability for "Landscape" compared to CLIP zero-shot, suggesting that different methods might have varied strengths in recognizing certain categories. Some categories like "Autumn\_leaves", "Flower", and "Kids" witness minimal recognition, with probabilities close to 0.00% for most methods, hinting at a need for additional training or model tuning to improve recognition in these categories. Trained methods (linear probes and LoRA) exhibit a consistent high confidence in "Greenery" classification across both clip and BLIP methods, hinting at a robustness in feature recognition for this particular category within trained methods. The low confidence in certain categories suggests potential areas for improving the models, possibly through further training, data augmentation, or exploring other model architectures, to enhance the recognition capability across a broader range of categories.

# Discussion

The presented results delineate the accuracy metrics for various models under diverse configurations. It is evident that the application of fine-tuning techniques markedly enhances both the Top-1 and Top-3 accuracy across all models, with Linear Probe fine-tuning exhibiting a marginal advantage in certain scenarios. The benchmark models from [1] showcase competitive performance, highlighting the robustness and effectiveness of the proposed framework when compared to established architectures.

Additionally, we explored alternative methods such as classifying images through their captions with BLIP-caption, and constructing a mega zero-shot model to determine the most confident model among the three previously mentioned. However, we chose not to present these findings as they did not yield any notable results.

A deeper dive into the ramifications of different fine-tuning techniques, along with a comprehensive analysis concerning the broader implications of these results within the scene recognition domain, are already elaborated in the earlier sections.

Limitations and Comparison to Existing Literature. Our study, while thorough, has its limitations. The models were evaluated using two datasets; the generalization of these findings to other datasets remains to be validated. Furthermore, the computational resources required for training and evaluating the models may pose challenges in real-world applications, especially on devices with limited computational capabilities.

Comparatively, the existing literature showcases a myriad of scene detection algorithms and methodologies. Our work builds upon and extends the findings of [1], showcasing improved performance with the introduction of fine-tuning techniques. However, when juxtaposed with other state-of-the-art models in the literature, the performance metrics show competitive but not superior results in all aspects. The innovative fine-tuning techniques employed in our study, such as Linear Probe and LoRA, have demonstrated significant promise, yet they also bring forth additional computational and complexity costs that may hinder real-time applications.

The venture into alternative methods, although not yielding notable results, opened avenues for further exploration and underscore the importance of continuous investigation in this rapidly evolving field.

Our project lays a solid foundation for future research aimed at harnessing advanced Scene Detection algorithms for substantial image quality enhancement in real-world photography applications. The comprehensive analysis, juxtaposed with existing literature, shines light on the areas where our study excels and where further advancements can be pursued to bridge the gaps identified.

# Conclusion

This project delves into the area of Scene Detection, a pivotal software technique in digital photography, aiming to autonomously categorize the scene types within given images. Our investigation hinged on the evaluation of three pre-trained algorithms—CLIP, BLIP, and BLIP-2—leveraging the CamSDD and Places365 datasets. The rigorous evaluation, based on TOP-1 and TOP-3 accuracy metrics, revealed a commendable performance, especially when refined through fine-tuning techniques like Linear Probe and LoRA.

The exemplary performance of BLIP in the TOP-1 accuracy metric underscores its potential as a robust algorithm for scene detection, thereby serving as a catalyst for subsequent image quality enhancement tasks. Notably, the Linear Probe technique emerged as a superior fine-tuning mechanism, achieving perfect TOP-3 accuracy across all three algorithms. Although the zero-shot configurations exhibited lower accuracy, they provided valuable insights into the intrinsic scene detection capabilities of the algorithms prior to the fine-tuning.

Moreover, the competitive performance of reference models, particularly MobileNet-V2, echoes the robustness and effectiveness of the proposed framework in comparison to established architectures. The findings accentuate the significance of fine-tuning in bolstering the classification accuracy, which is instrumental in propelling photographic image quality forward.

Broader Implications. The findings of our project have broader implications on the digital photography field. By advancing scene detection algorithms, our work contributes to the ongoing efforts to enhance image quality, which is paramount for both professional photographers and general consumers. The improved accuracy in scene categorization enables more precise adjustments to camera settings and post-processing techniques, ultimately elevating the quality of the photographs.

Furthermore, the open-source availability of our code and pre-trained models fosters a collaborative environment for researchers and practitioners in the field. This not only accelerates the pace of innovation but also broadens the understanding of the interplay between scene detection and image quality enhancement.

The venture into alternative methods, like image classification through captions using BLIP-caption and a mega zero-shot model, although not yielding notable results, opened avenues for further exploration. Our project lays a solid foundation for future research aimed at harnessing advanced Scene Detection algorithms for substantial image quality enhancement in real-world photography applications.

Our work embodies a stride towards bridging the gap between vision and language understanding through pre-trained VLMs, paving the way for more sophisticated Scene Detection algorithms capable of significantly augmenting image quality in digital photography. The promising outcomes of this project beckon a deeper investigation into fine-tuning techniques and the development of more robust models, fostering a rich soil for the blooming of innovative solutions in Scene Detection and beyond.

Telegram Bot Application. To further bridge the gap between our research and real-world application, we developed a Telegram bot that allows users to test our scene detection algorithms on their own images. This user-friendly platform provides a direct interaction between the users and our models, facilitating an immediate evaluation of our algorithms in real-world settings. By uploading images through the Telegram bot, users can obtain the scene detection results instantaneously, thus experiencing firsthand the practical utility of our work. We encourage readers to engage with our Telegram bot to better understand the potential impact of our scene detection algorithms on photographic quality enhancement.

# References

[1] Our work based on – [Fast and Accurate Camera Scene Detection on Smartphones by Angeline Pouget et al, 2021](https://arxiv.org/pdf/2105.07869v1.pdf).

URL – <https://people.ee.ethz.ch/~ihnatova/camsdd.html>

[2] Our work is inspired by – [CLIPCAP](https://arxiv.org/abs/2111.09734) ([ClipCap: CLIP Prefix for Image Captioning, by Ron Mokady](https://arxiv.org/abs/2111.09734) [et al,](https://arxiv.org/pdf/2103.00020v1.pdf) [2021](https://arxiv.org/abs/2111.09734))

[3] Clip Model – [Learning Transferable Visual Models From Natural Language Supervision by Alec Radford et al, 2021](https://arxiv.org/pdf/2103.00020v1.pdf)

[4] BLIP Model – [Learning Transferable Visual Models From Natural Language Supervision by Alec Radford et al, 2021](https://arxiv.org/pdf/2103.00020v1.pdf)

[5] BLIP-2 Model – [Learning Transferable Visual Models From Natural Language Supervision by Alec Radford et al, 2021](https://arxiv.org/pdf/2103.00020v1.pdf)

[6] [LAVIS: A One-stop Library for Language-Vision Intelligence by Li, Junnan et al. 2023](https://arxiv.org/abs/2209.09019)

[7] [LoRA: Low-Rank Adaptation of Large Language Models by Edward J. Hu et al. 2021](https://arxiv.org/pdf/2106.09685.pdf)

[8] [Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution by Ananya Kumar et al. 2022](https://arxiv.org/pdf/2202.10054.pdf).

URL – <https://github.com/salesforce/LAVIS>

[9] [Places: A 10 million Image Database for Scene Recognition by Bolei Zhou et al.](http://places.csail.mit.edu/index.html)

[10] Places365 Dataset Kaggle.

URL – <https://www.kaggle.com/datasets/benjaminkz/places365>