YovaClip: A Multi-Model Approach to Forest Fire Detection

Zubair Salman Anwar
25100134@lums.edu.pk

Essa Jan
25100324@lums.edu.pk

***Abstract*—The increasing frequency and severity of forest fires presents significant threat to both human lives and natural ecosystems (flora and fauna). Given this problem, this project aims to provide a novel deep learning-based approach for the early detection of forest fires. The approach utilises the benefits of Large Multi-Models (LMM) and the You Only Look Once (YOLO) model for accurate and robust forest fire detection from surveilling cameras in forest areas. This methodology, leveraging the comprehensive feature extraction of LMMs and real time object detection of YOLO, aims to improve accuracy and efficiency of fire and smoke detection in these environments. The fusion of these 2 mechanisms allows them to compliment each other and filling in the gaps/weaknesses of both models. By detecting these incidents in their early stages, our solution aims to enable mitigating the risk of catastrophic wildfire outbreaks. Through rigorous experimentation and evaluation, we demonstrate the effectiveness and reliability of our proposed method, offering a promising tool for forest fire management and conservation efforts.**

I. Introduction

Forest fires, which leave a devastating impact on both human communities and natural habitats, are a critical concern worldwide. Such disasters have increased to 50,000-70,000 occurrences a year between 2017-2020 in the United States alone [1]. This increasing frequency and severity of these fires highlight the urgent need for proactive measures to detect and mitigate their outbreaks before they escalate into an uncontrollable disasters. Traditional methods of forest fire detection have often relied on manual surveillance and human intervention, leading to significant delays in response time and limited effectiveness in large and remote forested areas.

In recent years, advancements in computer vision and deep learning have provided promising tools for improving the early detection of forest fires through automated surveillance systems. However, despite the development of sophisticated models and techniques, the deployment of these solutions in real-world scenarios has been hampered by challenges such as high false positive rates and limited adaptability to diverse environmental conditions.

Motivated by these challenges, this paper proposes a novel deep learning-based approach for forest fire detection, aimed at addressing the limitations of existing methodologies. By leveraging a combination of state-of-the-art models, including LMM and the YOLO architecture, our approach seeks to enhance the accuracy and efficiency of fire and smoke detection in forest environments.

Central to our methodology is the integration of diverse data sources, including LUMS's footage from Pan-Tilt-Zoom (PTZ) cameras deployed in forested regions and publicly available datasets, such as the California forest fire dataset. Through model training and fine-tuning processes, we tailor our detection system to effectively differentiate between genuine fire and smoke signals and environmental phenomena such as clouds and haze, which often deter traditional detection methods.

Our approach uses a multi-stage pipeline, beginning with the classification of day and night images using Zero Shot CLIP Classification, followed by object detection utilising a custom-trained YOLO V9 model. Subsequently, detected bounding boxes are refined and analysed through Large Language and Vision Models (LLAVA) and CLIP, harnessing the combined power of visual and linguistic representations to augment fire detection accuracy. Such an approach allows us to take advantage of the benefits of each individual model.

Through extensive experimentation and evaluation, we demonstrate the efficacy and reliability of our proposed methodology in real-world deployment scenarios. By detecting forest fire incidents in their early stages, our solution holds the potential to significantly mitigate the risk of widespread wildfire outbreaks, thereby safeguarding both human lives and ecological balance in forested regions.

II. Background

In this section, we provide a detailed literature review on forest fire detection, highlighting the critical role of early detection in mitigating the impacts of wildfires on biodiversity and ecosystems. We discuss the utilization of advanced computer vision techniques, state-of-the-art models such as YOLO V9 and CLIP, and innovative solutions to challenges in smoke detection. Additionally, we explore the potential of CLIP for enhanced image analysis, underscoring its versatility in extending beyond traditional image classification tasks.

A. Forests and Forest Fires

Forests primarily serve as important ecosystems, hosting diverse flora and fauna, playing a crucial role in maintaining global biodiversity. However, the escalating threat of forest fires, propelled by changing weather patterns and rising temperatures due to global warming, poses a significant risk to these delicate ecosystems. Timely detection of forest fires is

critical to mitigate their impact and minimize damage to the environment.

B. State-of-the-Art Models in Forest Fire Detection

Recent improvements in Deep Learning and Computer Vision models have seen the emergence of state-of-the-art architectures, such as YOLO V9 and CLIP, as potential tools for early forest fire detection under varying environmental conditions. Notable studies, exemplified by Islam et al. [2] and Kim et al. [3], have showcased the efficacy of these models in achieving impressive accuracy rates and F1 scores in forest fire detection tasks.

Furthermore, smoke detection plays a crucial role in early forest fire detection, yet we found many algorithms suffer from high false positive rates, often mistaking clouds for smoke due to similarities. Innovative approaches, such as that proposed by Nikolay et al. [5], address this challenge by utilizing Gaussian filters to separate the sky from the image, thereby reducing false positive rates.

C. Image Detection Models

Studies, such as those conducted by Yu and Yao [6] and Li et al. [7], show the efficacy of ensemble CNNs and MobileNetV3 architecture in achieving commendable accuracy rates in forest fire detection tasks. Furthermore, research by Islam et al. [2] highlights the utilization of pre-trained EfficientnetB7 models augmented with attention mechanisms and Bayesian optimization to achieve remarkable accuracy and F1 scores.

The YOLO architecture, renowned for its efficiency in object detection, has garnered significant attention in the realm of forest fire detection. Pioneering works by Kim et al. [3] and Huo et al. [4] demonstrate the efficacy of convolutional attention mechanisms within YOLO variants in detecting forest fires and smoke with high accuracy rates and F1 scores.

Despite the recent development, YOLO still remains the most relied upon state of the art model for object detention. With frequent improvements, the YOLO V9 model improves its recall by 134% as compared to Yolo V8. Additionally, YOLO V9 has the largest AP (Average Precision) to number of parameters ratio [4].

D. CLIP

Contrastive Language-Image Pre-training (CLIP) provides a significant advancement in image classification through zero-shot learning. While traditionally employed for image classification, recent studies, including Radford and Kim [8], Lin and Gong [9], and Wang et al. [10], explore CLIP's potential for object detection and image segmentation. These studies demonstrate the versatility of CLIP in extending its applicability to various computer vision tasks without extensive local dataset training.

E. Other Large Multi-Models

In recent years, we have also seen the emergence of other LMM beyond CLIP, each offering unique capabilities and applications. Notable among these are models such as LlaVA [11] and BLIP, which have gained attention for their abilities in multimodal tasks.

LlaVA (Large Language and Vision Models), for instance, integrates textual and visual modalities to facilitate a wide range of tasks, from image classification to natural language understanding. Similarly, BLIP (Bi-Modal-PLarge) [12], a variant of Large Language Models (LLM), extends the capabilities of traditional LLMs by incorporating visual information for enhanced performance in multimodal tasks.

Despite their potential, these LMMs have yet to be explored in the domain of forest fire detection. While they offer promising pathways for analysis and feature extraction, their application in this context remains abscent in literature. Using the capabilities of these models could open new gates in forest fire detection, enabling more comprehensive and accurate analysis of visual data captured by surveillance systems.

III. Methodology

In this section, we outline the methodology used in our research, which includes dataset formation, baseline analyses, initial suggested improvements, and final suggested improvements. This section only covers the methodology leading up to the final pipeline; the final pipeline is discussed in the proceeding section.

A. Dataset

Our dataset comprises images sourced from two primary avenues: our own PTZ cameras positioned in the Himalayan region and the California fire dataset. The PTZ camera images, totaling around 1500, were meticulously curated to include approximately 500 images depicting fire, smoke, or no events. Additionally, to augment data variance, images from the California fire dataset were incorporated. Manual labeling and bounding box generation were performed, ensuring dataset accuracy and consistency. Labelled fire and smoke data examples are shown in Figure 1 and 2.

B. Baseline Analyses

Initial testing involved the evaluation of individual models, including CLIP and LLAVA, to gauge their performance in smoke and fire detection tasks. CLIP, employing a Vision Transformer architecture, demonstrated varying accuracies for smoke and fire classification. During zero-shot testing, CLIP achieved:

- Smoke Classification: An accuracy of 45%
- Fire Classification: An accuracy of 70%
- Haze Classification: A high accuracy rate of 96%

However, despite its strong performance in haze detection, CLIP displayed a significant issue of high false positives for smoke and haze, reaching approximately 86%. It's noteworthy that CLIP was tested on all classes in parallel, providing independent results for each class.

On the other hand, LLAVA exhibited distinct characteristics in its classification capabilities. LLAVA results showed:

- Smoke Detection: An accuracy rate of 92%

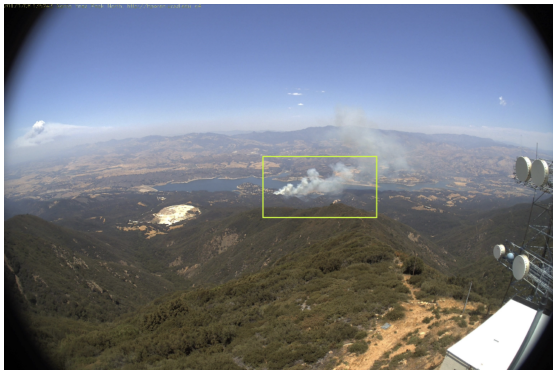

Fig. 1. Labelled Instance of Smoke

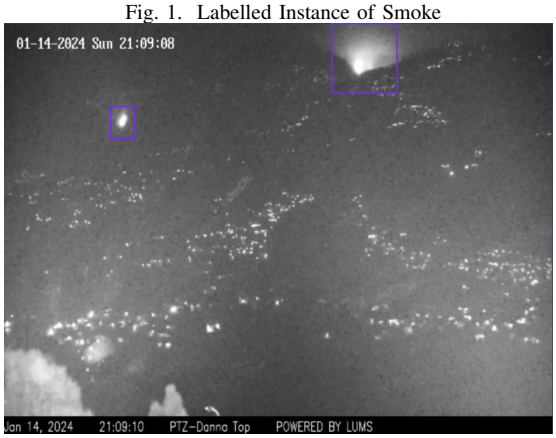

Fig. 2. Labelled Instance of Fire

- Fire Detection: However, a relatively low accuracy rate of 10%

Despite the difference in accuracy between smoke and fire detection, LLAVA showed a relatively low false positive rate across all classes compared to CLIP.

We also manually examined a stand-alone fine-tuned YOLOV9 model later to test its performance as well. The purpose of such testing was to gauge the individual performances of the models on the tasks. This provides a thorough understanding of their performances while also setting a baseline for comparison with our developed pipeline.

C. Initial Suggested Improvements

After observing the base performances of models and our dataset, we proposed utilising different strategies to improve the model performance. Initially we only aimed to make use of LMMs for fire and smoke detection; keeping the limitations of the models in mind, we devised two improvements. After baseline tests, we also decided to only make use of CLIP and LLAVA given our limited compute resources and the performance of these models.

For our first improvement, we conducted a detailed investigation into the influence of time of day on model performance. We noticed the different visual characteristics/features of daytime and nighttime images, we implemented separate pipelines to analyze their impact on fire and smoke detection. Initially, we divided the dataset into two subsets: day and night. This divide was achieved through gray-scale conversion and threshold-based segmentation, optimizing image categorization. This approach revealed a clear contrast in the occurrences of fire and smoke between daytime and nighttime images.

On implementing this strategy, we explored the improvements posed by this approach using qualitative analysis. For day and night classification, we studied the performance of both suggested approaches:

- Grayscale Thresholding: Achieved an accuracy of 96% in classifying images into day and night categories.
- Zero-Shot Classification with CLIP: Produced a remarkable accuracy of 100% in day-night classification.

Moving on to actual fire and smoke detection proceeding day-night classification, initial pipeline results mirrored the baseline performance:

- CLIP: Achieved 70% accuracy for fire and 45% for smoke detection.
- LLAVA: Demonstrated similar performance as before , indicating 10% accuracy for fire and 92% for smoke detection.

As based on the results, we noticed that the initial suggestion did not produce any significant improvements in detection.

Further analysis revealed crucial insights into the nature of fire and smoke events across different times of the day. We observed that all instances of fire occurred exclusively during nighttime, characterized by the distinct bright light emitted by flames. Conversely, smoke events were only observed during daytime, due to the visibility of smoke against the forest landscapes. Based on these findings, we implemented a night and day filter to refine the classification categories for our models. By restricting the LMM's classifications to either fire or smoke based on the time of day, we aimed to mitigate false positives and enhance model accuracy, by minimising false positives and eradicate model hallucinations. For example, at night time, we limited the model to only detecting fire; this completely removed false positives for smoke and we hoped to improve fire detection accuracy.

Moving on, for our second suggested improvement, we dived deeper into the spatial distribution of fire and smoke in the images to identify potential areas of improvement. Observing that the wide coverage of images often included smaller fire and smoke events near the image edges, we devised a novel approach. We partitioned images into a 3x3 grid, as shown in Figuare 3 allowing for local analysis of fire and smoke occurrences within each sub-image. By processing LMM independently on each grid segment, we aimed to improve model performance by limiting its field of view and reducing the influence of other irrelevant features present. We did test this mechanism manually on challenging images with small smoke or fire instances in a large scene, such as in Figure 3, and the strategy proved to enhance performance. Although no quantitative analysis was done on this approach.

Fig. 3. Example 3x3 Grid Visualisation on Sample Image

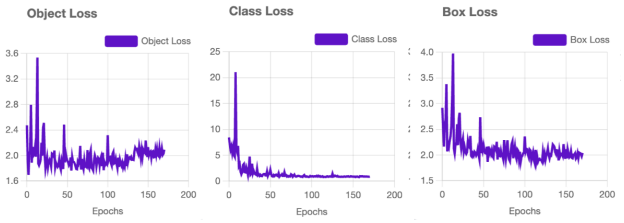

Fig. 4. Fine-Tuned YOLO V9 Object, Class, and Box Losses

D. Final Suggested Improvement

After acknowledging the limitations of the grid-based approach in our second initial improvement, we decided to take a more efficient and effective approach to identify regions of interest for fire and smoke detection. We turned to object detection models, choosing YOLO V9 as our primary choice due to its better recall and high precision-to-parameter ratio.

To use the strengths of YOLO V9 in detecting fire and smoke regions, we fine-tuned the model using our training dataset comprising 769 images resized to 640x640. Training YOLO V9 over 180 epochs for approximately 3 hours, we achieved a mean average precision (mAP) of 0.92 and box loss of 2.091 as shared in Figure 4. In addition to high mAP score our fine-tuned model achieves excellent Recall results of 93.3% and Precision results of 74.7% on Evaluation set as seen in Table I. These significant improvements came without any noticeable cost pertaining to inference speeds.

TABLE I
YOLO V9 FINE-TUNED EVALUATION

Class	Precision	Recall
Fire	0.74	1.00
Smoke	0.69	0.70

With YOLO V9 serving as our backbone, we adopted an approach to integrate object detection and LMMs seamlessly. Firstly, we used YOLO V9 to extract areas of interest (AOIs) within images, optimizing the detection of fire and smoke regions. By focusing on specific regions rather than using a brute-force grid-based segmentation, we significantly reduce computational inefficiencies and improved the precision of our analysis.

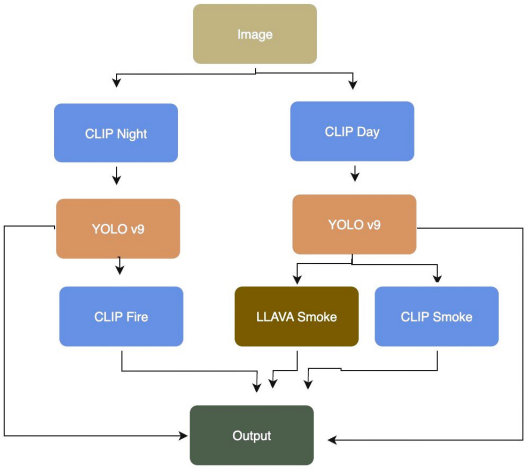

Fig. 5. Final Pipeline for Forest Fire Detection

Subsequently, we utilised these AOIs as inputs for our LMMs, including CLIP and LLAVA, to perform detailed classification and analysis. This integrated approach allowed us to make use of the strengths of both object detection and mutli-model analysis, maximizing the accuracy and efficiency of our forest fire detection system.

By combining object detection models with LMMs, we aimed to overcome the limitations of traditional detection approaches and enhance the robustness of our forest fire detection pipeline. This methodology not only improves the accuracy of fire and smoke detection but also ensures rapid inference and scalability for real-world applications.

IV. FINAL PIPELINE

Our final pipeline represents a fusion of three distinct models: YOLO V9, LLAVA, and CLIP, used in unison to optimize forest fire and smoke detection. The pipeline begins by routing incoming images through a CLIP instance, which exhibits 100% accuracy in classifying images as day or night.

Upon classification, the pipeline diverges based on whether the images are detected as nighttime or daytime scenes. In the case of nighttime images, YOLO V9 is employed to extract bounding boxes corresponding to potential fire regions. These bounding boxes are then cropped from the images and forwarded to another CLIP instance dedicated to fire detection. Conversely, for daytime images, the fine-tuned YOLO V9 model is utilized to identify bounding boxes for smoke regions.

To maximize the probability of capturing fire and smoke events, a low confidence threshold of 0.10 is applied to the YOLO bounding boxes. Thus, both daytime and nighttime images undergo parallel analysis through CLIP and LLAVA

models to detect smoke events. In the daytime scenario, if the confidence level of the YOLO model exceeds 0.25 or either of the LMMs detects smoke within any bounding box, the image is tagged as containing smoke. Similarly, for nighttime images, if the confidence level of the YOLO model surpasses 0.3 or CLIP detects fire within any bounding box, the image is classified as containing fire.

It is also important to note that the fine-tuned YOLO V9 model was not influenced by haze. Also, we lacked a dataset with both haze and smoke; therefore we remove the detection for haze as it is no longer relevant

Summarising this detailed explanation, we can see a visual representation of the final pipeline in Figure 5.

This final pipeline was tested, and our quantitative analysis clearly shows the enhancements introduced by these improvements. The pipeline was fed images from the test dataset; these images were not employed in fine-tuning YOLO V9 and therefore provide a good perspective on the pipeline's performance. This test set consists of 230 images with 130 depicting a clear scene with no fire or smoke, 29 images of fire, and 71 images of smoke instances.

Overall, the pipeline resulted in a macro average accuracy of 85% with perfect recall of 100% on Fire images and 75% precision on Smoke images. Other results are shown in Table II, where 'Empty' indicates images with clear scenes. Similarly, we can visually see the results in the form of a confusion matrix to further observe the distribution of true positives, false positives, and false negatives in Figure .

TABLE II
FINAL PIPELINE EVALUATION RESULTS ON TEST SET

	Precision	**Recall**	**F1-score**	**Instances**
Empty	0.84	0.85	0.85	130
Fire	0.94	1.00	0.97	29
Smoke	0.75	0.70	0.72	71
-	-	-	-	-
accuracy	-	-	0.83	230
Macro-Avg	0.84	0.85	0.85	230
Weighted-Avg	0.82	0.83	0.82	230

A. Analysing Final Pipeline Results

We can derive a comprehensive understanding of the pipeline's performance looking at the Table II and Figure 7.

For the "Empty" class the pipeline demonstrates a high precision of 0.84 and a recall of 0.85, indicating the its ability to accurately classify non-event instances. Additionally, the "Fire" class exhibits high precision and recall scores of 0.94 and 1.00, resulting in a high F1-score of 0.97. These metrics are evident of the pipeline's capability in correctly identifying fire instances with minimal false positives and false negatives.

However, the "Smoke" class shows slightly lower precision and recall scores of 0.75 and 0.70. While the precision score suggests a moderate level of accuracy in identifying smoke instances, the recall score indicates that there might be some instances of smoke events missed by the pipeline.

Fig. 6. Example of Smoke Instance (Difficult to Spot)

Overall, the pipeline achieves an accuracy of 0.83, indicating a better performance than our base model analysis in correctly classifying instances across all classes. The macro-average and weighted-average scores further confirm the pipeline's balanced performance across different classes, with precision, recall, and F1-score metrics averaging around 0.85.

It is important to note that the pipeline performs worst in case of smoke, however this also might be due to the nature of the images. Instances of smoke in the California forest fire dataset are often minute and hardly visible; such instances were difficult to label while preparing the dataset. These instances are near impossible to detect even for a human eye as they are usually in the image's periphery and in the distance making it difficult to spot. An example of such can be seen in Figure 6.

The empty cases falsely predicted as instances of smoke may also indicate poor labelling from our end; it is possible the model identified smoke where we failed to.

V. FUTURE WORK

Keeping our findings in mind, as well as obstacles and limitations we faced, we propose future direction in this domain to enhance forest fire detection.

One promising direction for further exploration involves the utilization of fine-tuning techniques to enhance the performance of both object detection models and large multimodal models (LMMs). We have seen that fine-tuning significantly improves accuracy for YOLO, suggesting that a similar approach could be beneficial for CLIP and LLAVA, as well as other LMMs in general. Fine-tuning these models on downstream tasks specific to forest fire and smoke detection could potentially lead to substantial improvements in accuracy and performance.

Moreover, while our pipeline currently employs YOLO V9, LLAVA, and CLIP, there exist other object detection models and LMMs that could be explored in future studies. Due to limited computational resources, we were unable to test larger

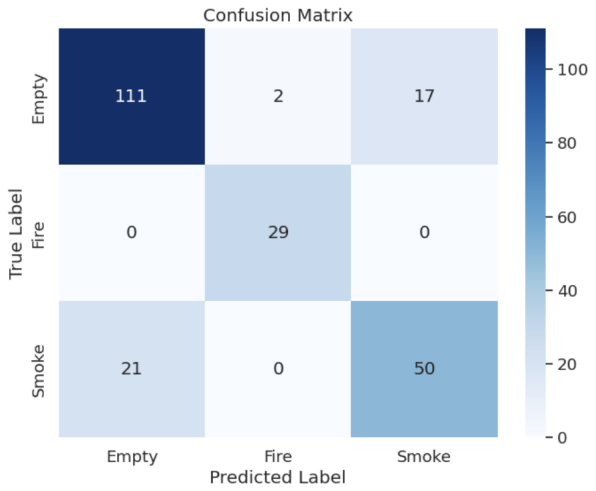

Fig. 7. Final Pipeline Confusion Matrix

models such as InstructBLIP. However, incorporating these larger models into the pipeline could provide enhanced capabilities and potentially improve detection accuracy. Exploring the performance of other state-of-the-art object detection models and LMMs could provide valuable insights and contribute to the development of more robust detection systems.

The current pipeline also primarily focuses on the detection of fire and smoke events. Future research could expand the scope to include the identification of other relevant phenomena, such as heat sources, vegetation changes, or abnormal weather patterns, which could provide valuable context for early fire detection and prevention efforts. The integration of additional data sources and modalities, such as satellite imagery, weather data, or historical fire incident records, could further improve the pipeline's capabilities and improve its accuracy in detecting and predicting forest fire events.

VI. Conclusion

In conclusion, this study presents a thorough study of early detection of forest fires and smoke using state-of-the-art computer vision and deep learning techniques. Through the development and evaluation of a novel pipeline including YOLO V9, LLAVA, and CLIP models, we have demonstrated promising results in accurately identifying fire and smoke events in forest environments. Our findings highlight the benefit of using a combination of object detection and large multi-models for improved detection capabilities.

Furthermore, our study identifies several areas for future research and improvement, including the exploration of fine-tuning techniques to enhance model performance, the investigation of alternative object detection models and LMMs, and the integration of additional data sources and modalities to enrich detection capabilities.

By continuing to refine and expand upon these proposed methodologies, we can possibly produce more effective early detection systems, ultimately aiding in the safety of our valuable natural ecosystems and the protection of lives and property from the impacts of wildfires.

References

[1] "Climate Change Indicators: Wildfires." EPA, United States Environmental Protection Agency, 4 Feb. 2024, www.epa.gov/climate-indicators/climate-change-indicators-wildfires.

[2] Islam, A.M.; Masud, F.B.; Ahmed, M.R.; Jafar, A.I.; Ullah, J.R.; Islam, S.; Shatabda, S.; Islam, A.K.M.M. An Attention-Guided Deep-Learning-Based Network with Bayesian Optimization for Forest Fire Classification and Localization. Forests 2023,

[3] Kim, S.-Y.; Muminov, A. Forest Fire Smoke Detection Based on Deep Learning Approaches and Unmanned Aerial Vehicle Images. Sensors 2023

[4] Huo, Y.; Zhang, Q.; Jia, Y.; Liu, D.; Guan, J.; Lin, G.; Zhang, Y. A Deep Separable Convolutional Neural Network for Multiscale Image-Based Smoke Detection. Fire Technol. 2022

[5] Abramov, Nikolay et al. Intelligent Methods for Forest Fire Detection Using Unmanned Aerial Vehicles. https://www.mdpi.com/2571-6255/7/3/89B14-fire-07-0008. 2024

[6] Yu, Y.; Yao, M. When Convolutional Neural Networks Meet Laser-Induced Breakdown Spectroscopy: End-to-End Quantitative Analysis Modeling of ChemCam Spectral Data for Major Elements Based on Ensemble Convolutional Neural Networks. Remote Sens. 2023

[7] Li, Y.; Zhang, W.; Liu, Y.; Jin, Y. A Visualized Fire Detection Method Based on Convolutional Neural Network beyond Anchor. Appl.

[8] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020. 2021

[9] Lin, J.; Gong, S. GridCLIP: One-Stage Object Detection by Grid-Level CLIP Representation Learning. arXiv preprint arXiv:2303.09252v1. 2023

[10] Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; Liu, T. CRIS: CLIP-Driven Referring Image Segmentation. arXiv preprint arXiv:2110.12009. 2021

[11] Islam, Ashhadul, et al. Pushing Boundaries: Exploring Zero Shot Object Classification with Large Multimodal Models.2023.

[12] Li, Junnan, et al. Blip: Bootstrapping Language-Image Pre-Training For Unified Vision-Language Understanding and Generation.2024.