EFFICIENT SCENE INTERPRETATION FRAMEWORK USING DYNAMIC DUAL-PROCESSING

PHASE II REPORT

Submitted By

Naveed Buhari 3122 21 5001 058

R. Nikhilesh 3122 21 5001 060

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

Department of Computer Science and Engineering

Sri Sivasubramaniya Nadar College of Engineering
(An Autonomous Institution, Affiliated to Anna University)
Kalavakkam - 603110

May 2025

Sri Sivasubramaniya Nadar College of Engineering

(An Autonomous Institution, Affiliated to Anna University)

BONAFIDE CERTIFICATE

Certified that this project report titled **"EFFICIENT SCENE INTERPRETATION FRAMEWORK USING DYNAMIC DUAL-PROCESSING"** is the *bonafide* work of "**Naveed Buhari (3122 21 5001 058)**, **R. Nikhilesh R (3122 21 5001 060)**" who carried out the project work under my supervision.

Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

Dr. T.T. Mirnalinee
HEAD OF THE DEPARTMENT
Professor
Department of CSE
SSN College of Engineering

Kalavakkam - 603 110

Dr. T.T. Mirnalinee
SUPERVISOR
Professor
Department of CSE
SSN College of Engineering

Kalavakkam - 603 110

Submitted for project viva-voce examination held on. . . . . .

EXTERNAL EXAMINER **INTERNAL EXAMINER**

ACKNOWLEDGEMENTS

I am deeply grateful to GOD, the almighty, for providing me with the strength, knowledge, and perseverance required to complete this project.

I express my heartfelt gratitude to my guide **DR. T.T. MIRNALINEE**, Professor, Department of Computer Science and Engineering, for her invaluable advice, constant guidance, and patience throughout the duration of my research. Her insights and support were instrumental in shaping the direction and refinement of my work. My sincere thanks to **DR. T.T. MIRNALINEE**, Professor and Head of the Department of Computer Science and Engineering, for her words of advice and encouragement.

I express my deep respect to the founder **DR. SHIV NADAR**, Chairman, SSN Institutions, whose vision for excellence in education has inspired my academic journey. I also express my appreciation to **DR. S. RADHA**, Principal, for the support and resources provided during my studies.

I would like to extend my sincere thanks to all the teaching and non-teaching staff of our department who have contributed directly and indirectly during the course of my project work. A special thanks to the laboratory staff for their assistance in accessing various resources and ensuring that the technical needs of my project were met efficiently.

Finally, I am deeply appreciative of the unwavering support of my parents and friends, whose encouragement and moral support have been constant throughout this journey.

NAVEED BUHARI **R. NIKHILESH**

ABSTRACT

Robust scene interpretation is vital for applications like autonomous driving and assistive technologies, yet many object detection and captioning methods fail when objects are partially hidden or overlapping. To address this gap, we introduce an efficient scene interpretation framework using dynamic dual-processing, with Faster R-CNN handling detailed, localized detection and DETR providing broader context. This approach successfully detects objects even in crowded or partially occluded scenes. We also explore the idea of inpainting to fill in obstructed regions, making detection more reliable. A cross-attention mechanism then combines the resulting visual features with a language model to produce coherent scene-level descriptions, capturing both objects and their relationships. By offering detailed, human-like scene descriptions, our framework addresses current limits in detection and captioning, delivering a more complete and context-aware understanding of real-world images. Our proposed hybrid model achieved a precision of 0.7 and a recall of 0.73 and the image captioning system gives a METEOR score of 0.5916 and ROGUE-L score of 0.6191 when tested on the MS COCO dataset.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 MOTIVATION

The motivation for this project stems from the limitations of traditional image captioning models, which struggle to describe scenes accurately when objects are partially hidden or occluded. In real-world applications, such as autonomous vehicles, surveillance systems, and assistive technologies, the ability to interpret occluded objects is essential for reliable decision-making. Building on the previous phase of this project, where a hybrid object detection framework combining Faster R-CNN and DETR was developed, this phase aims to extend the framework by incorporating an image captioning module capable of handling occlusions.

To further improve occlusion handling, inpainting methods were introduced in the previous phase. Using diffusion-based inpainting techniques, the system reconstructs missing or obscured object regions, enhancing the visual completeness of the image. By filling in occluded areas, the system provides more accurate visual data to the captioning model, resulting in clearer and more reliable descriptions. This combination of detection, inpainting, and captioning creates a robust framework capable of interpreting complex, partially obscured scenes.

1.2 BACKGROUND

In the field of computer vision, image captioning has emerged as a vital task for generating textual descriptions of visual content. However, conventional captioning models often fall short in complex scenarios where occlusions are present. The previous phase of this project addressed the challenge of object detection by combining Faster R-CNN and DETR. Faster R-CNN uses a Region Proposal Network (RPN) to efficiently identify object regions, while DETR leverages transformers to perform direct set prediction, eliminating the need for handcrafted anchor boxes. This dual-model approach significantly improved detection accuracy and robustness.

To further enhance the framework, inpainting methods were introduced to handle occlusions. The system employed Stable Diffusion-based inpainting, which uses cross-attention and latent space processing to generate realistic reconstructions of missing regions. Alongside this, TELEA and Navier-Stokes inpainting methods were tested for comparison, with the diffusion-based approach achieving the highest reconstruction quality. The inpainted images, which restored missing object regions, significantly improved the accuracy and coherence of the final captions.

Building on this foundation, the current phase extends the system by introducing an occlusion-aware image captioning framework. The proposed method uses our phase 1 hybrid model for object detection, which offers real-time performance and high accuracy, even in challenging conditions. For caption generation, the system employs BLIP, a powerful pre-trained model capable of generating contextually rich captions. The integration of a cross-attention mechanism allows the system to

combine object detection features with captioning features, ensuring that occluded objects are recognized and described effectively.

By leveraging this dual-processing framework, along with the inpainting module, the system achieves a higher level of interpretative accuracy. The cross-attention mechanism plays a key role in refining the captions by incorporating both local object-level features from our hybrid model and global scene-level features from BLIP. This enhances the model's ability to generate comprehensive and meaningful descriptions, even when portions of the scene are obscured.

1.3 PROBLEM DEFINITION

The objective of this phase is to develop an efficient scene interpretation framework capable of generating accurate and descriptive captions for images containing occluded objects. By integrating hybrid model for object detection, BLIP for caption generation, and a cross-attention mechanism for feature fusion, the system ensures reliable and detailed scene interpretation. The incorporation of inpainting techniques, specifically Stable Diffusion-based reconstruction, further enhances the framework by restoring missing object regions, enabling the captioning model to describe occluded scenes with greater accuracy and coherence.

CHAPTER 2

LITERATURE SURVEY

2.1 LITERATURE REVIEW

2.1.1 Object Recognition

Object recognition has seen significant progress through advancements in convolutional neural networks (CNNs) and transformer-based architectures. For instance, Lin et al. (2017) [5] introduced Feature Pyramid Networks (FPN) to enhance multiscale detection via a top-down architecture with lateral connections, substantially improving the performance on small objects. Meng et al. (2021) [8] focused on accelerating DETR convergence by incorporating conditional anchors, demonstrating faster training times and accurate end-to-end object detection. Meanwhile, Chen et al. (2021) [1] developed the YOLOF (You Only Look One-Level Feature) framework to simplify object detection into a single-level feature extraction process suitable for real-time systems. Additionally, Zhou et al. (2019) [16] proposed an anchor-free approach dubbed "Objects as Points," wherein object centers are treated as keypoints, thus reducing the complexity of bounding-box proposals and improving results in crowded scenes.

2.1.2 Occlusion Detection

Addressing occlusions remains a critical challenge in real-world detection scenarios. Wang et al. (2023) [11] offered a comprehensive review of occlusion handling, categorizing existing solutions and proposing future directions for robust object detection under partial visibility. Zhang et al. (2024) [13] introduced an adaptive occlusion detection approach based on Overlapping IoU (OL-IoU), aiming to refine bounding-box predictions when objects overlap significantly. Su et al. (2022) [14] presented OPA-3D, a pixel-wise aggregation network that mitigates occlusions in monocular 3D object detection by reconstructing partially visible regions. Saleh and Vamossy (2022) [15] proposed a bounding box-based occlusion detection and order recovery method designed to restore the correct object ordering in cluttered scenes. Furthermore, H et al. (2023) [4] demonstrated a dual-processing object detection framework inspired by human cognition, utilizing familiarity-recollection mechanisms to enhance recognition even under partial occlusions.

2.1.3 Scene Interpretation

Beyond raw detection, broader scene interpretation involves capturing context, handling occluded areas, and merging vision with language. Zhou et al. (2018) [20] introduced UNet++, originally for segmentation but influential in inpainting methods, demonstrating how multiscale features and redesigned skip connections can facilitate reconstruction tasks. Li et al. (2022) [24] presented BLIP, a unified vision-language model that merges image captioning and understanding into one system by effectively aligning visual and textual cues, leading to richer contextual

reasoning. Extending BLIP, Pan et al. (2023) [28] explored cross-attention fusion with object detectors, enabling models to generate detailed captions alongside accurate localization. Further enhancing vision-language methods, Fang et al. (2023) [21] proposed EVA, which leverages scalable masked visual representation learning for improved generalization across multiple vision tasks, and Kim et al. (2022) [27] introduced ViL-Seg, which refines visual predictions through text-driven cross-modal attention, underscoring the potential of joint vision-language systems for robust interpretation of complex, occluded scenes.

2.2 RESEARCH GAP

- **Balancing Local and Global Features:** CNN-based methods excel at local feature extraction but lack global context understanding. Conversely, transformers are strong in capturing global relationships but struggle with fine-grained local details. Existing hybrid approaches address this to some extent but require more sophisticated fusion mechanisms to fully integrate these strengths [5][8].

- **Biological Inspiration:** While frameworks like H et al. (2023) are inspired by human cognition [4], most models do not fully exploit biological mechanisms. Further exploration of neuroscience principles, such as dual-processing and adaptive decision-making, could lead to more intuitive and efficient detection systems.

- **Dataset Limitations:** Current datasets, such as COCO and PASCAL VOC, lack comprehensive annotations for challenging scenarios like severe occlusions, extreme lighting, and deformations. There is a need for larger

and more diverse datasets to train and evaluate models under real-world conditions [6].

- **Inpainting Under Complex Contexts:** UNet++ [20] inspired many inpainting models, but current techniques struggle in reconstructing semantically consistent occluded regions in cluttered, real-world scenes.
- **Caption Quality Degrades with Noise:** Although BLIP [24] performs well in structured scenes, its captioning quality drops significantly when occlusions, clutter, or incomplete objects are present.
- **Lack of Joint Vision-Language Optimization:** Vision-language models like ViL-Seg [27] offer impressive cross-modal reasoning, but integrated optimization for both detection and captioning under occlusion is still lacking.

2.3 RESEARCH OBJECTIVES

- **Integrate Local and Global Features:** Design a dual-processing framework that effectively combines CNNs for local feature extraction and transformers for global context understanding, addressing the limitations of individual methods [5][8].
- **Incorporate Semantic Inpainting Mechanisms:** Embed a lightweight inpainting sub-module trained with occlusion-aware datasets like those discussed in Zhou et al. [20], to recover object completeness and support robust detection in cluttered scenes.

- **Fuse Detection and Captioning with Cross-Attention:** Integrate models like BLIP [24] and ViL-Seg [27] using cross-attention layers to align bounding boxes with textual descriptions, ensuring coherent scene interpretation.
- **Enhance Occlusion-Robust Captioning:** Improve the caption generation process by training vision-language models to recognize and describe partially visible or occluded objects using auxiliary semantic cues, as in EVA [21].

CHAPTER 3

PROPOSED METHODOLOGY

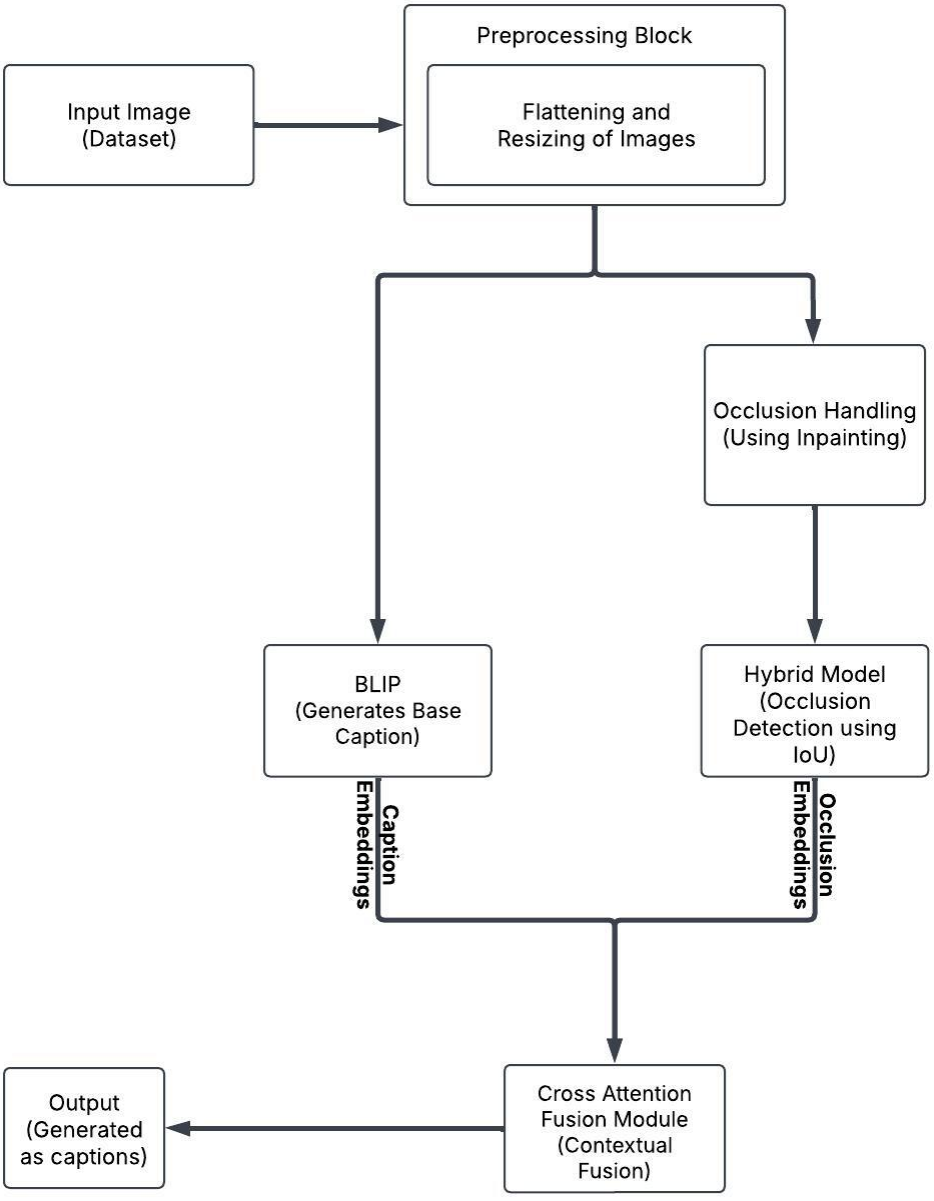

FIGURE 3.1: Proposed Scene Interpretation Framework

The proposed Scene Interpretation Framework, aims to generate scene-descriptive captions while addressing the challenge of occluded or partially visible objects. At

its core, we combine Faster R-CNN and DETR into a hybrid detection framework. A selective masking mechanism merges bounding boxes from both models based on Intersection over Union (IoU) and confidence thresholds, yielding refined object proposals. The end goal is to robustly identify all objects in the scene, even those that are partially hidden or arranged in cluttered configurations and come up with a descriptive caption for the scene.

To further enhance detection under occlusion, we have explored inpainting concepts that could reconstruct missing areas in an image. It demonstrates how inpainting can conceptually restore occluded regions. For partially visible objects, we compute an occlusion summary using Intersection over Union (IoU), ensuring that hidden portions are not overlooked when generating captions.

After we obtain the detected objects, we employ the BLIP model to produce a base caption representing the broader scene context. This textual output is then refined through a cross-attention mechanism that fuses the base caption with the visual features, obtained as occlusion summary from the hybrid model. By combining detailed object-level insights and the language model's contextual understanding, our framework generates coherent scene descriptions that explicitly account for occlusions, achieving a more complete and human-like interpretation of complex visual inputs (Figure 3.1).

The entire flow of this pipeline is discussed in this chapter below:

3.1 OCCLUSION HANDLING USING INPAINTING METHODS

3.1.1 Architecture Overview

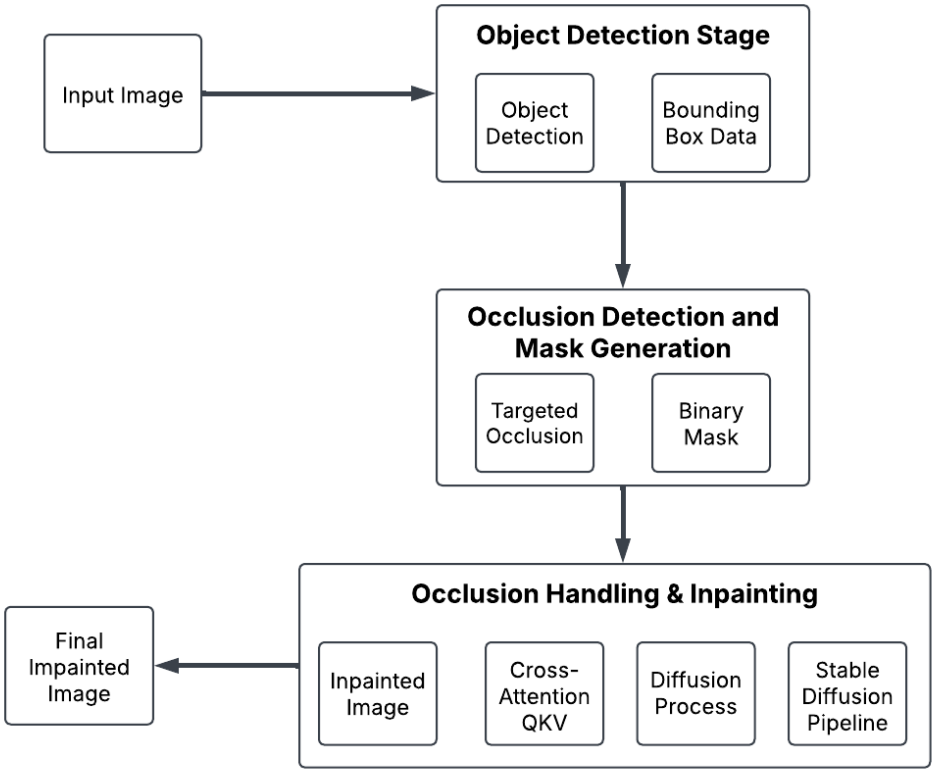

FIGURE 3.2: Proposed architecture for occlusion-aware object detection and inpainting.

Figure 3.2 provides an overview of the occlusion-aware framework and inpainting. The image is first passed to our detection modules to identify visible objects, then occluded regions are reconstructed using inpainting . This ensures that partially obscured objects becomes clearer.

The architecture for occlusion handling(Figure 3.2) consists of three main stages: object detection, occlusion detection and mask generation, and occlusion

handling through inpainting. The process begins with an input image, which undergoes object detection to identify objects and generate bounding box data. This information is then used to determine occluded regions, where a targeted occlusion strategy is applied. A corresponding binary mask is generated to mark these regions for further processing.

In the final stage, the occluded regions are reconstructed using an inpainting approach based on Stable Diffusion. The binary mask and occluded image serve as inputs to the pipeline, where techniques such as cross-attention mechanisms and diffusion-based processing refine the reconstruction. The output is a fully inpainted image where missing or occluded portions have been seamlessly restored. This framework ensures robust object detection even in scenarios with significant occlusions.

3.1.2 Stable Diffusion-based Inpainting

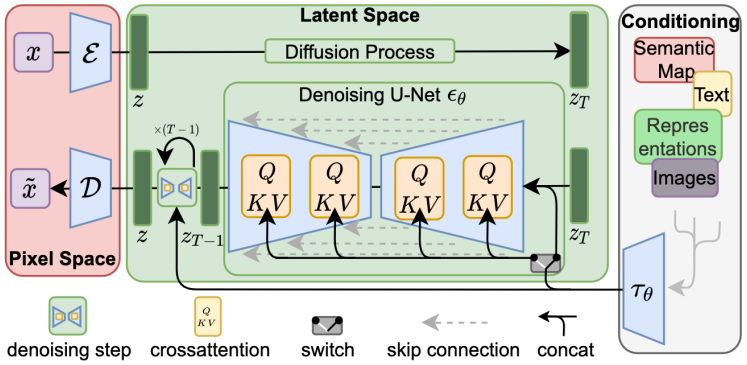

FIGURE 3.3: Stable Diffusion-Based Inpainting

As shown in Figure 3.3 [29], the stable diffusion-based inpainting pipeline operates in both pixel space and latent space to iteratively reconstruct occluded regions. The input image and corresponding mask are first passed through an

encoder that maps them into a latent representation. A denoising algorithm progressively refines this representation over multiple steps, removing noise and restoring missing areas while maintaining global context. Skip connections retain high-resolution features, and cross-attention mechanisms (query, key, value) help the network focus on crucial parts of the image.

The conditioning pathway on the right allows additional inputs such as semantic information or textual prompts, guiding the generation process toward more coherent and context-aware reconstructions. By combining the unmasked portions of the image with learned representations of typical visual patterns, the system can effectively fill in occluded areas so they blend seamlessly with surrounding regions. This approach provides visually convincing restorations and demonstrates the robustness of diffusion-based methods for challenging inpainting tasks.

3.1.2.1 Stable Diffusion Pipeline

The Stable Diffusion pipeline is responsible for reconstructing the occluded regions of the image. Unlike traditional inpainting methods, which use simple interpolation techniques, Stable Diffusion leverages a deep-learning-based generative model that operates in the latent space. By encoding the image and mask into a compressed latent representation, the model can efficiently reconstruct missing regions with high realism.

3.1.2.2 Diffusion Process

Once the occluded image and mask enter the Stable Diffusion pipeline, the diffusion process begins. This process involves iteratively refining the image by gradually removing noise and generating realistic textures that seamlessly blend with the surrounding context. The diffusion model is trained on large datasets, enabling it to understand structural patterns and generate visually coherent reconstructions.

3.1.2.3 Cross-Attention QKV

The Cross-Attention mechanism (QKV - Query, Key, and Value) plays a crucial role in ensuring that the inpainted regions are contextually accurate. This mechanism allows the model to focus on different parts of the image while generating missing details. It ensures that the restored object retains its original shape, texture, and color consistency, making the inpainting process highly effective.

3.2 HYBRID MODEL FOR OBJECT DETECTION

As illustrated in Figure 3.4, our integrated architecture combines both local (Faster R-CNN) and global (DETR) detectors, followed by a selective masking mechanism. This design balances fine-grained feature extraction with broad contextual awareness, ensuring robust detection even in cluttered scenes.

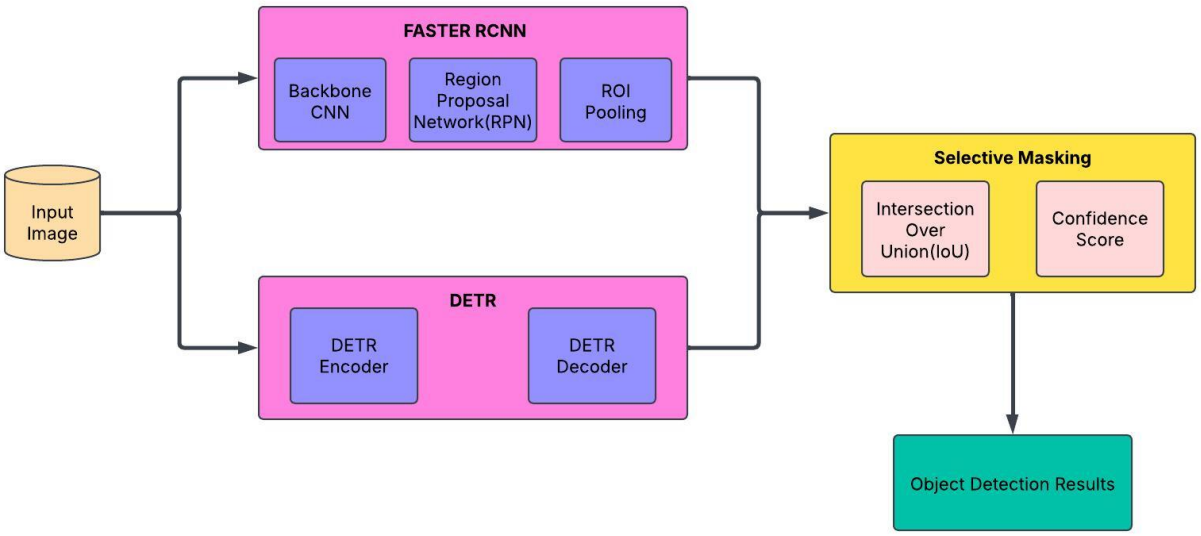

FIGURE 3.4: Hybrid Model Architecture for Object Detection

The above System Architecture (Figure 3.4) illustrates a multi-module object detection system that begins with an Input Image, which is processed through two key components to extract complementary features. The DETR Transformer Module captures global context, analyzing the relationships between objects and the overall layout of the scene. The Faster R-CNN Module focuses on extracting local features, detecting objects with precision and capturing detailed patterns and textures.

The outputs from these modules are combined using a Selective Masking mechanism, which integrates the strengths of both approaches to create a refined and unified representation of the detected objects. This refined representation is then passed to the Scene Description Generator, which translates the visual information into a coherent textual description of the scene. This architecture effectively combines local and global feature extraction with selective integration to achieve a comprehensive understanding of the image.

The architecture is divided into several key modules that work together for efficient object detection and scene understanding. Below is a description of each module:

3.2.1 Faster R-CNN Module

The Faster R-CNN module plays a crucial role in the architecture by focusing on extracting local feature representations, which are essential for gaining a detailed understanding of individual objects and their spatial characteristics within an image.

Faster R-CNN is a state-of-the-art object detection framework that improves on previous models by integrating region proposal generation directly into the convolutional neural network (CNN) architecture. Unlike earlier methods such as R-CNN and Fast R-CNN, which rely on external region proposal algorithms, Faster R-CNN combines region proposal generation and object classification into a unified pipeline. This integration allows for a significant reduction in computational time and improves the overall efficiency of object detection.

The architecture of Faster R-CNN consists of three primary components. The first is the backbone CNN, which is typically a network like ResNet or VGG. This CNN is responsible for extracting high-level feature maps from the input image. These feature maps are then fed into the Region Proposal Network (RPN), a key innovation in Faster R-CNN. The RPN slides over the feature map and generates regions of interest (ROIs) by predicting objectness scores and refining the bounding box coordinates. These ROIs are potential areas in the image that are likely to contain objects.

The third component is the ROI Pooling layer, which takes the proposed regions and resizes them into fixed-size feature maps. These are then passed through fully connected layers for object classification and bounding box regression, ultimately producing the final detection results. Faster R-CNN is known for its speed and accuracy, as it allows for end-to-end training of the entire model, making it highly efficient for detecting objects in complex scenes. The shared computation between the RPN and the backbone CNN further optimizes its performance, ensuring that Faster R-CNN is both accurate and computationally feasible for real-world applications.

3.2.2 DETR Transformer Module

The DETR (Detection Transformer) is an object detection model that uses a transformer-based architecture for detecting objects. The first component, the backbone, is a Convolutional Neural Network (CNN) that processes the input image and extracts features. These features capture important visual information, such as edges and textures, which are essential for detecting objects in the image. However, CNNs alone cannot handle the spatial relationships between objects, so positional encodings are added to these features. Positional encodings help the model understand where each feature is located in the image, providing the necessary spatial context for the Transformer.

Once the features are extracted and enhanced with positional encodings, they are passed to the Transformer Encoder. The Transformer Encoder analyses the relationships between the different features in the image. This helps the model focus on specific areas of the image that are important for detecting objects,

allowing the model to capture both local and global information. The encoder processes the features and prepares them for the next step in the process, which is object detection.

After the encoder processes the image features, the Transformer Decoder comes into play. The decoder uses object queries (which act as placeholders for potential objects) and compares them with the features from the encoder to determine what objects are in the image and where they are located. The decoder tries to match the features from the encoder with the object queries, focusing on areas in the image that are most likely to contain objects.

Finally, the Prediction Heads generate the model's output. These heads predict the class labels and the bounding boxes that define the location and size of the objects. The prediction heads take the output from the decoder and transform it into the final object detection results.

3.2.3 Selective Masking

Selective Masking utilizes a systematic approach to merge the outputs of Faster R-CNN and DETR by comparing their predictions using Intersection over Union (IoU) and confidence scores. IoU helps evaluate how closely the bounding boxes from both models overlap, ensuring that predictions with significant agreement are prioritized. Confidence scores are used to assign weights to the bounding boxes, enabling the system to focus on high-confidence detections while filtering out less reliable predictions. This process ensures that the final output retains the strengths of both models while minimizing redundancy and inaccuracies.

Moreover, by combining IoU and confidence-based weighting, Selective Masking ensures that detected objects are not only localized with precision but also ranked and refined for further processing. This robust integration lays the groundwork for applications requiring a deeper understanding of scenes, such as generating textual descriptions, recognizing complex interactions, or performing real-time tracking in dynamic environments. By unifying the insights from both Faster R-CNN and DETR, the Selective Masking module transforms raw detection outputs into actionable, high-quality results, making it a critical enabler for advanced use cases in fields like autonomous navigation, intelligent surveillance, and robotic systems.

3.3 OCCLUSION AWARE IMAGE CAPTIONING

Occlusion-aware image captioning is a novel approach aimed at generating descriptive and accurate captions for images, even in the presence of occluded objects. Unlike traditional captioning models, which often fail to explicitly describe or identify partially hidden objects, this method incorporates advanced object detection and feature fusion techniques to improve scene interpretability. The proposed framework combines the occlusion summary from the hybrid model and the base caption generated using BLIP module using cross-attention fusion to effectively describe occluded scenes. Each module plays a critical role in enhancing the accuracy and contextual richness of the generated captions. The following subsections describe the modules used to generate occlusion aware image captions:

3.3.1 Base Captioning using BLIP

The BLIP (Bootstrapping Language-Image Pretraining) module serves as the base caption generator in the proposed framework. BLIP is a vision-language model that aligns visual and textual modalities through a vision encoder and a language decoder, making it suitable for scene understanding and caption generation. Given an input image, BLIP generates a coherent base caption that describes the overall scene using visual semantics. However, since BLIP operates primarily on visible content, it may not always capture occluded elements or object relationships accurately.

In the context of this framework, BLIP takes the input image and produces an initial caption embedding that reflects the visible scene content. These embeddings encapsulate the linguistic context derived from the image and are crucial for downstream refinement. While the base caption generated by BLIP is semantically meaningful, it may lack awareness of occlusions or hidden object interactions that are not immediately evident from the visible context alone.

To address this limitation, the base caption generated by BLIP is further refined using cross-modal fusion with visual context obtained from an occlusion-aware hybrid model. By combining BLIP's semantic understanding with the occlusion information detected separately, the final captions become both accurate and contextually rich. This modular design enables the system to first understand what is seen (via BLIP) and then improve that understanding through occlusion-aware reasoning.

For instance, BLIP may generate a caption such as ”a person walking by a car,” but with occlusion reasoning fused, the final caption could be refined to ”a person

partially hidden behind a car," thus improving descriptive precision.

3.3.2 Cross-Attention Fusion

The fusion mechanism in our architecture integrates semantic embeddings from BLIP and visual cues from an occlusion-aware hybrid detection model. This hybrid model, based on Faster R-CNN and DETR, performs occlusion detection using Intersection over Union (IoU) between object pairs to identify partial overlaps. It focuses on capturing object-level relationships and occlusion details that are often missed by captioning models alone. These relationships are encoded as embeddings representing the visual context of occlusion.

To merge these two distinct types of information—textual semantics from BLIP and occlusion-aware visual features from the hybrid model—we employ a cross-attention fusion module. Cross-attention aligns these embeddings by treating the caption embeddings as queries and the occlusion embeddings as keys and values. This allows the model to selectively enhance parts of the caption with relevant visual occlusion information.

Through this mechanism, the model learns to emphasize objects involved in occlusion and update the textual output accordingly. For example, in a scenario where "a dog is sitting on a chair" is the base caption, and the hybrid model detects that the dog partially covers a table, the final caption becomes "a dog is sitting on a chair, partially covering the table."

This fusion enhances interpretability by grounding textual elements in visual cues, particularly in scenes where object overlap creates ambiguity. The

cross-attention module ensures that the final caption reflects both what is visible and what is contextually inferred, producing more accurate, human-like descriptions even in occluded scenes.

CHAPTER 4

EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

4.1 DATASET DESCRIPTION

For this project, we utilized two primary datasets: the COCO (Common Objects in Context) dataset and the OccludedPascal3D dataset. These datasets provide a diverse range of images with varying object types and occlusion levels, enabling the effective evaluation of our occlusion-aware image captioning framework.

The COCO dataset, a widely recognized benchmark in object detection and captioning tasks, contains over 200,000 labeled images with annotations covering 80 object categories. These categories include people, animals, vehicles, and household items, captured in real-world scenes with complex backgrounds and varying lighting conditions. The dataset provides detailed annotations, including bounding boxes and segmentation masks, making it ideal for training and evaluating the object detection modules of our framework. Its diversity in object scales, poses, and occlusions allows the models to generalize effectively across different environments.

To address occlusion-specific challenges, we incorporated the OccludedPascal3D dataset. This dataset is an extension of the Pascal3D+ dataset, designed for evaluating object detection and pose estimation under occlusion. It consists of 10,000 images of 12 object categories, including vehicles, furniture, and animals,

with varying levels of artificial occlusion. The occlusions are categorized into 9 levels with respect to levels of occlusion in the background and the artificially occlusing foreground.

The OccludedPascal3D dataset is particularly valuable for validating the occlusion-aware capabilities of our model, as it enables the assessment of how effectively the system handles complex occlusion scenarios. By combining COCO and OccludedPascal3D, our framework is trained and evaluated on both general and occlusion-specific conditions, enhancing its robustness and generalizability in real-world scene interpretation tasks.

4.2 ECOSYSTEM

The ecosystem for this project consists of Python as the primary programming language, leveraging its robust support for deep learning and computer vision libraries. PyTorch served as the core deep learning framework, providing flexibility in model development and GPU acceleration, which was essential for training and evaluating the occlusion-aware image captioning model. For object detection, we utilized the Ultralytics library, which offers pre-trained models and efficient inference capabilities, streamlining the detection process. The BLIP model was integrated for generating base image captions, while the custom cross-attention mechanism was implemented using PyTorch, enhancing the captioning process by incorporating occlusion information.

The COCO and OccludedPascal3D dataset annotations were managed using the COCO API from pycocotools and custom parsing scripts, enabling efficient

extraction of bounding boxes, category labels, and occlusion details necessary for model training and evaluation. Google Colab was chosen as the primary development environment due to its support for free GPU acceleration, facilitating faster model training, experimentation, and refinement.

For visualization, Matplotlib and OpenCV were used to display detection results, bounding boxes, and occlusion relationships. This provided clear visual feedback on the model's performance and aided in fine-tuning the occlusion-aware captioning process. The overall ecosystem ensured efficient model development, evaluation, and visualization, enabling the creation of a robust and scalable framework for occlusion-aware scene interpretation.

4.3 RESULTS OF HYBRID MODEL FOR OBJECT DETECTION

The hybrid model achieved a precision of 0.70 and a recall of 0.73 (Table 4.1), significantly surpassing the individual performances of Faster R-CNN and DETR. This demonstrates the advantage of combining the strengths of both architectures, yielding more reliable and comprehensive object detection.

FIGURE 4.1: Faster R-CNN Object Detection Result

FIGURE 4.2: DETR Transformer Object Detection Result

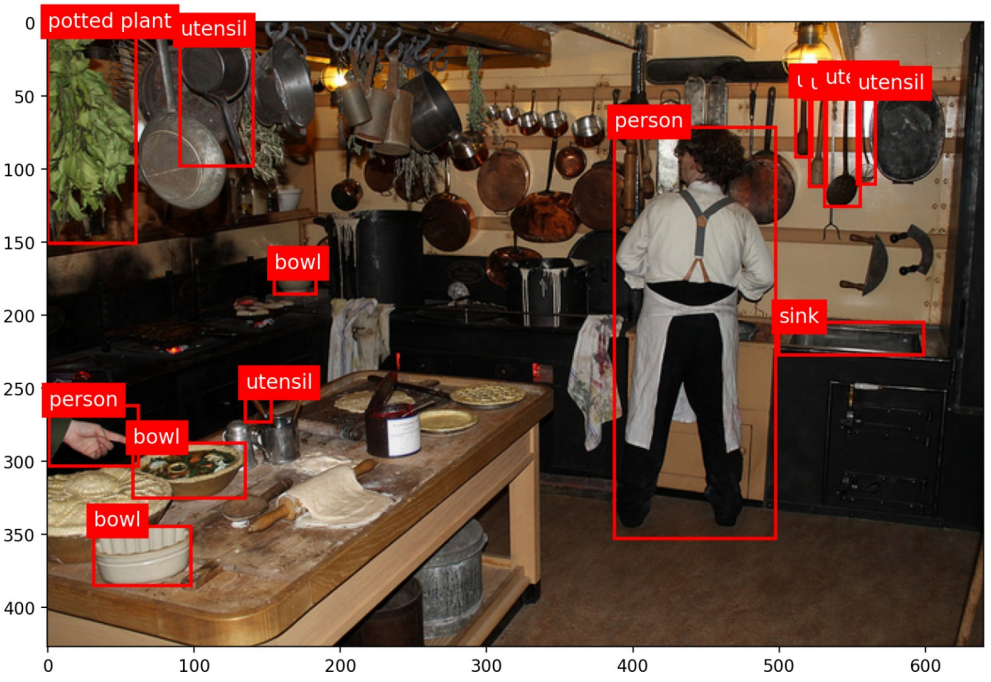

FIGURE 4.3: Hybrid Model Object Detection Result

Figures 4.1, 4.2, and 4.3 compare the object detection outputs of Faster R-CNN, DETR, and the Hybrid Model, respectively. While Faster R-CNN captures local details effectively (Figure 4.1), DETR excels at global context (Figure 4.2). Our combined approach (Figure 4.3) merges these strengths for improved accuracy.

The comparison of these individual results with that of the hybrid model (Table 4.1) highlights the effectiveness of integrating both approaches to improve precision and recall in complex detection scenarios.

We evaluated the performance of Faster R-CNN, DETR, and the hybrid model by calculating precision and recall values for each. The values are tabulated in Table 4.1 The hybrid model, combining Faster R-CNN's focus on local features with DETR's global context understanding, demonstrated improved performance in both precision and recall.

Model	Precision	Recall
Faster R-CNN	0.41	0.47
DETR	0.45	0.50
Hybrid Model	0.70	0.73

TABLE 4.1: Precision and Recall Comparison for Faster R-CNN, DETR, and Hybrid Model

4.4 RESULTS OF OCCLUSION HANDLING USING INPAINTING METHODS

In this section, we compare three different inpainting techniques: TELEA, Navier-Stokes, and Diffusion-based inpainting. These methods were evaluated based on their ability to reconstruct occluded regions in images while preserving structural and textural details. The evaluation was performed on a real-time image captured by us, where we manually introduced an occlusion to simulate real-world scenarios.

To test the inpainting techniques, we used a real-time image captured from our surroundings. A binary mask is generated based on the occlusion detection results. This mask is a grayscale image where white pixels represent occluded regions, and black pixels denote non-occluded areas.

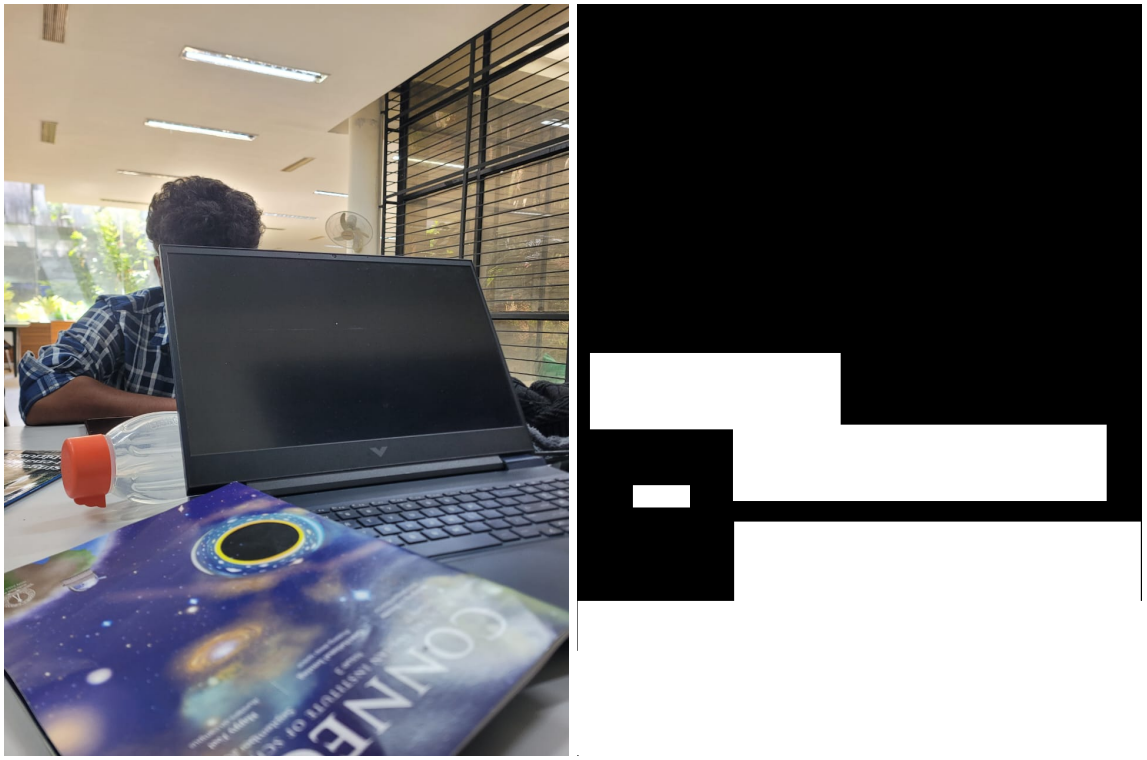

FIGURE 4.4: (Left) The real-time captured image used for testing inpainting methods. (Right) The generated occlusion mask for identifying occluded regions.

In Figure 4.4, we show the input image (left) and the generated occlusion mask (right). The highlighted regions indicate areas of significant occlusion to be reconstructed by the inpainting module. This same image was used to reconstruct the occluded regions for all the three inpainting techniques.

4.4.1 TELEA Inpainting

TELEA is an exemplar-based inpainting method that propagates image information from surrounding pixels into the occluded region. It follows a fast marching method, filling missing areas based on nearby intensities. This method is computationally efficient but struggles with complex textures and large occlusions.

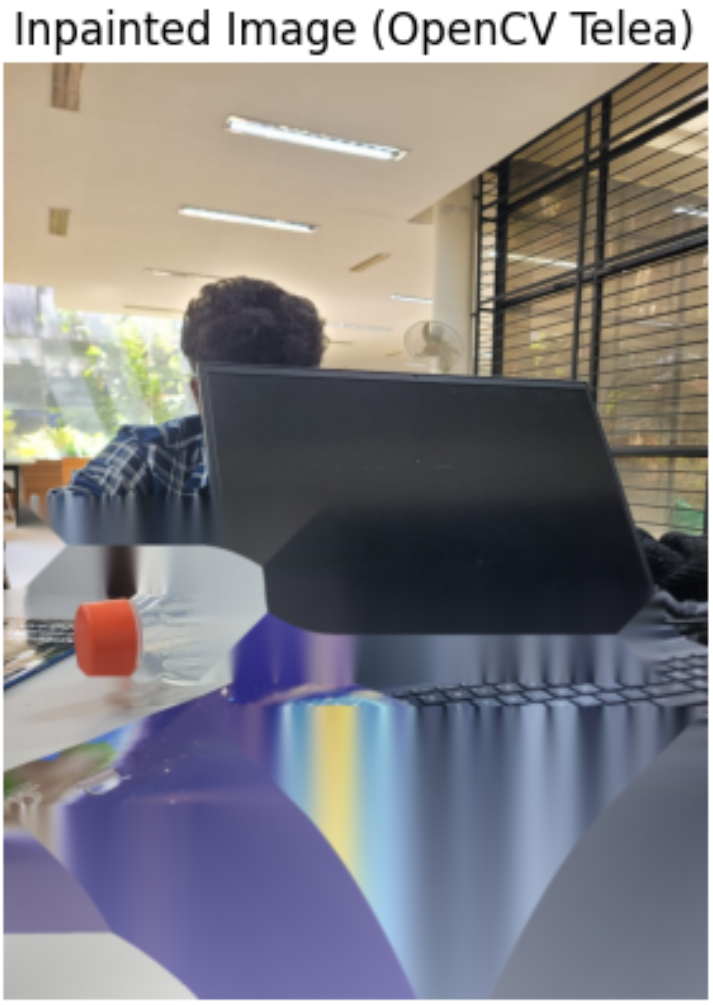

FIGURE 4.5: Inpainting result using the TELEA method.

The inpainted output of the TELEA method as shown in figure 4.5 performs well for small occlusions but fails in highly textured areas, often producing smoothed results rather than restoring intricate details.

4.4.2 Navier-Stokes Inpainting

Navier-Stokes inpainting is based on fluid dynamics principles and reconstructs missing regions by propagating isophote lines. This technique is effective for smoother region interpolation, making it suitable for images with gradual

variations. However, it does not perform well with high-frequency textures or detailed structures, leading to blurring in complex occlusions.

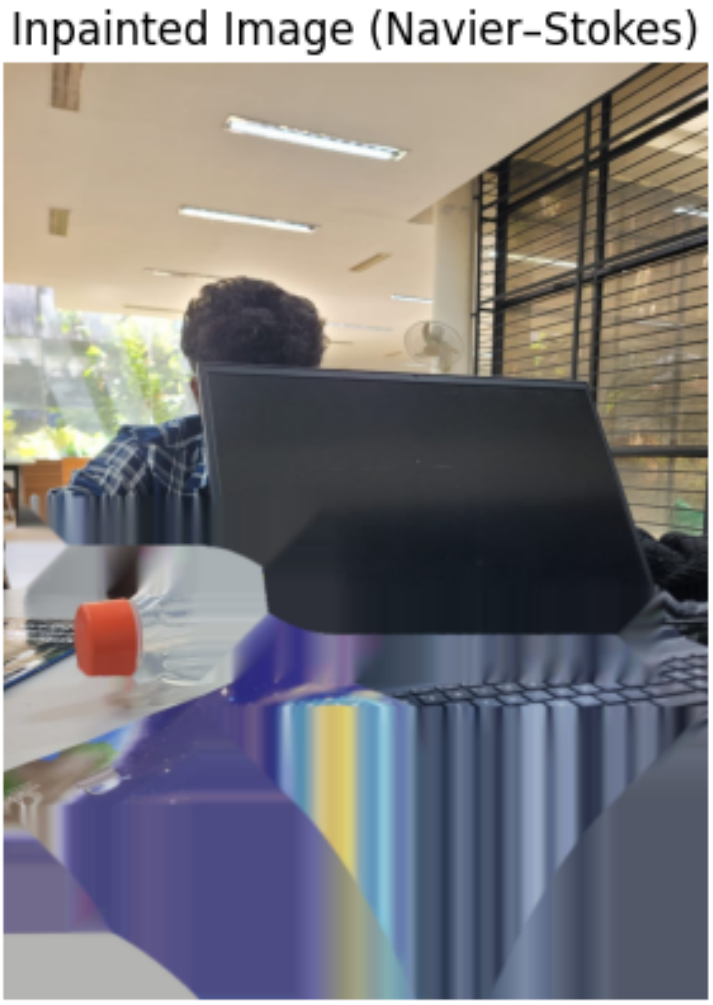

FIGURE 4.6: Inpainting result using the Navier-Stokes method.

The inpainted output of the Navier-Stokes method as shown in figure 4.6 produces smooth reconstructions but lacks fine details, making it effective for simple regions but inadequate for complex occlusions.

4.4.3 Diffusion-Based Inpainting

Diffusion-based inpainting, implemented using Stable Diffusion, utilizes a deep generative model to predict and reconstruct occluded areas. Unlike traditional methods, it learns high-level semantics, enabling it to generate visually realistic content even in large missing regions. While it produces the most visually coherent results, it requires significant computational power and may introduce artificial textures in highly structured regions.

FIGURE 4.7: Inpainting result using the Diffusion-based method.

The inpainted output of the Diffusion-based inpainting method as shown in figure 4.7 generates the most visually realistic results, particularly in complex scenes, but at the cost of higher computational requirements.

From our evaluation, we observe that TELEA (Figure 4.5) performs well for small occlusions but lacks detail reconstruction, Navier-Stokes (Figure 4.6) is

Method	PSNR (dB)	SSIM
TELEA	17.31	0.8339
Navier-Stokes	17.15	0.8223
Diffusion-Based	19.27	0.8916

TABLE 4.2: Quantitative evaluation of inpainting techniques. Higher values indicate better reconstruction quality.

effective for smooth regions but struggles with textures, and Diffusion-based inpainting (Figure 4.7) achieves the best overall quality but is computationally expensive. Based on these results, diffusion-based methods are preferable for complex occlusion handling tasks.

Table 4.2 compares three inpainting methods using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Our Diffusion-based approach achieves the highest SSIM of 0.8916, indicating superior structural fidelity under occlusion.

4.5 RESULTS OF OCCLUSION AWARE IMAGE CAPTIONING

This section presents the results of our proposed object detection and captioning framework that integrates hybrid model for spatial localization and BLIP with Cross-Attention for semantic understanding. The framework was tested on a custom dataset consisting of varied real-world images containing occlusions, complex backgrounds, and partially visible objects. Here, we display 7 diverse test images used to evaluate the effectiveness of the system in scene understanding and visual-textual alignment.

Figures 4.8 to 4.14 show the qualitative outputs of the system. Each result displays the Occlusion Summary detected by the bounding boxes generated by our hybrid model along with the captions predicted by the BLIP model which is the Base Caption. The Cross-Attention mechanism helps in refining contextual alignment between detected regions and generated textual descriptions.

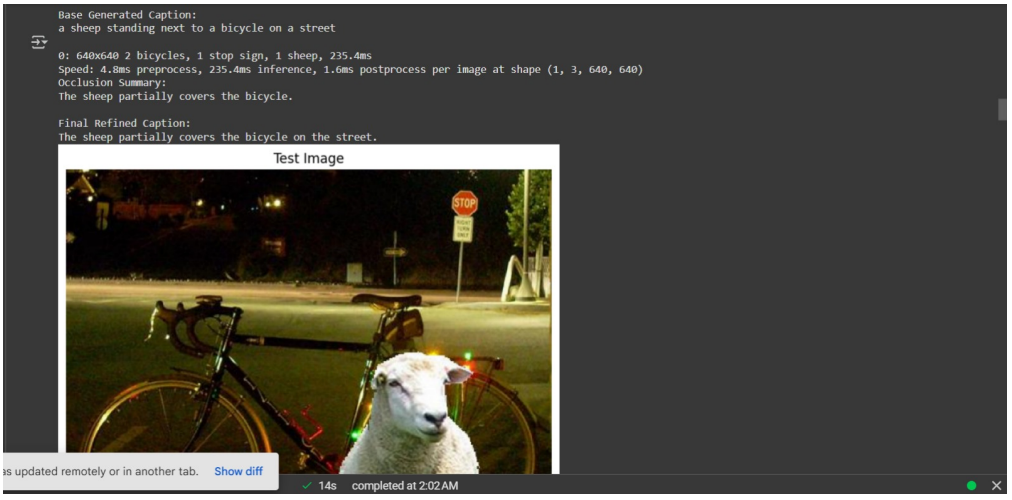

FIGURE 4.8: Detection and Captioning Output – Sample 1

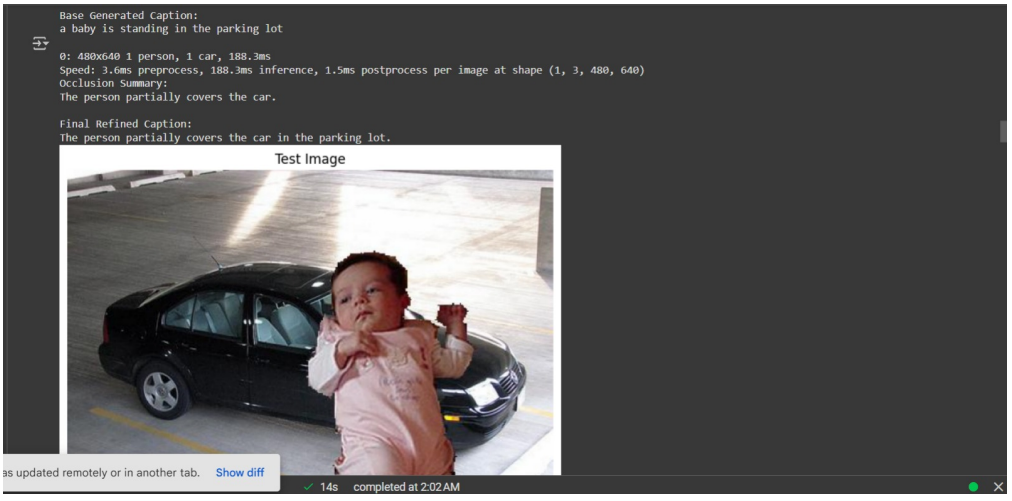

FIGURE 4.9: Detection and Captioning Output – Sample 2

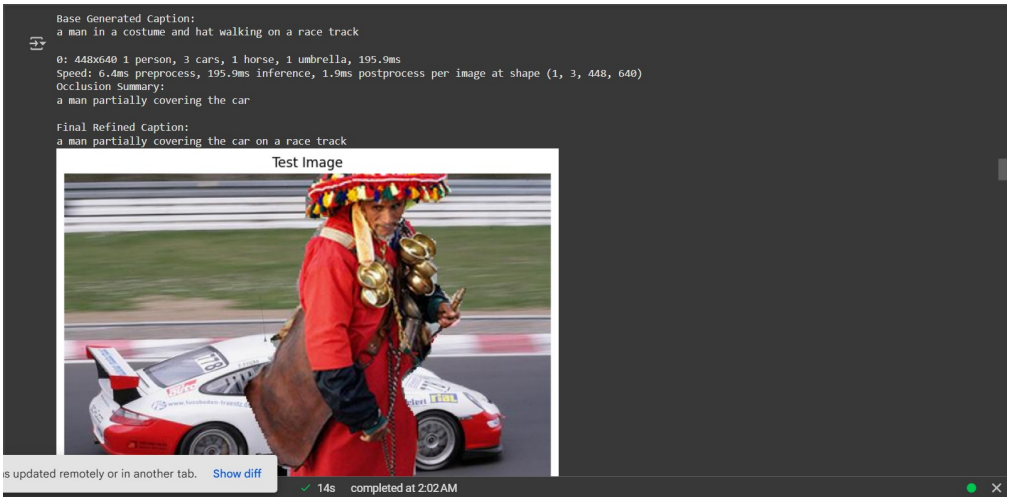

FIGURE 4.10: Detection and Captioning Output – Sample 3

FIGURE 4.11: Detection and Captioning Output – Sample 4

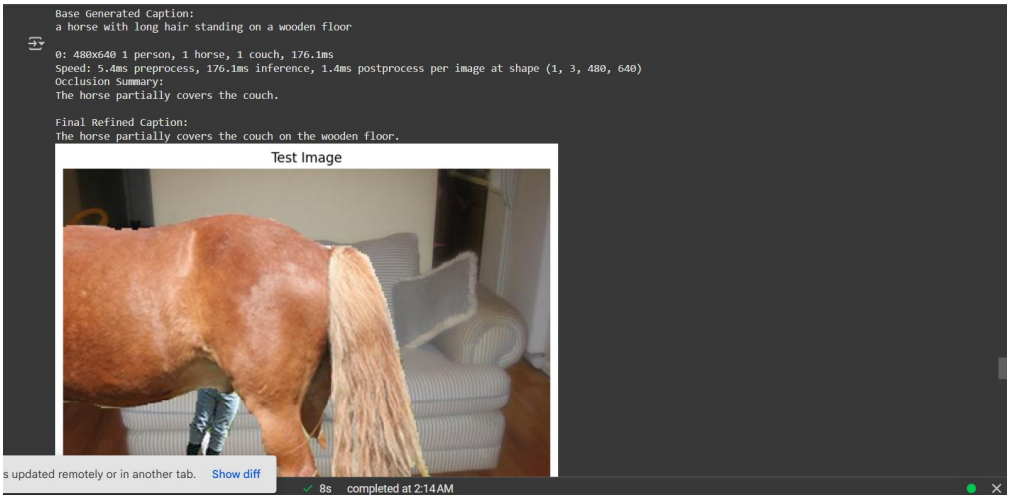

FIGURE 4.12: Detection and Captioning Output – Sample 5

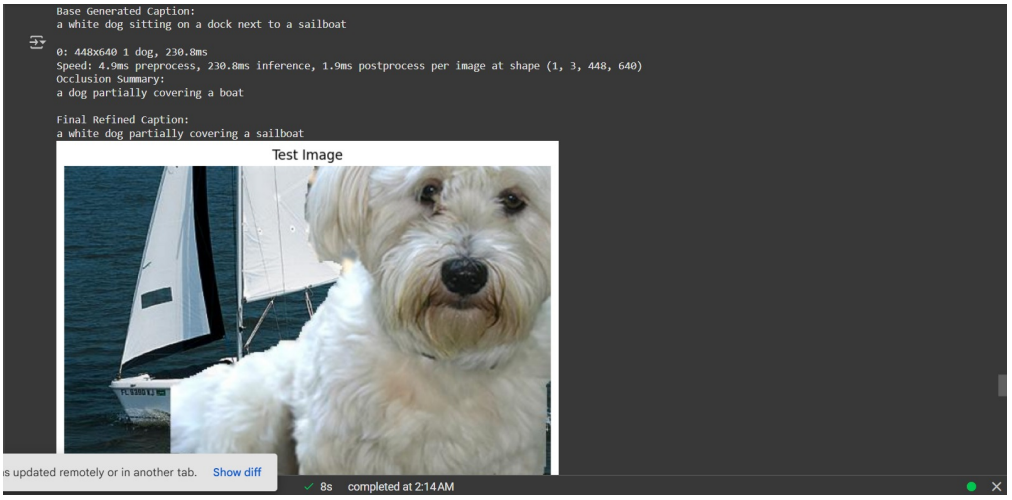

FIGURE 4.13: Detection and Captioning Output – Sample 6

FIGURE 4.14: Detection and Captioning Output – Sample 7

The proposed hybrid model + BLIP framework demonstrates notable strengths in both object detection and semantic captioning. One of the key advantages is the improved caption relevance, facilitated by the integration of cross-attention between visual features and textual tokens. This mechanism ensures better semantic alignment, particularly in images containing multiple objects or partial occlusions. Additionally, the system achieves accurate localization through hybrid model, which efficiently identifies object boundaries even in cluttered or visually complex environments, contributing to precise region-text pairing.

Another strength of the system lies in its robustness to occlusion, where it continues to generate meaningful captions despite partial object visibility. This is largely attributed to BLIP's strong vision-language pretraining, enhanced further by cross-attention layers that help infer missing context. The model also exhibits strong generalization, performing well across diverse scenarios including indoor, outdoor, and occluded environments.

Table 4.3 presents three standard metrics used to evaluate the quality of generated captions on the COCO dataset: BLEU, METEOR, and ROUGE-L. BLEU

Model	BLEU	METEOR	ROUGE-L
Yin and Ordonez [30]	0.253	0.238	0.507
Herdade et al. [31]	0.265	0.241	0.510
Our Model	0.3173	0.5916	0.6191

TABLE 4.3: Comparison of BLEU, METEOR, and ROUGE-L scores on the MS COCO dataset.

measures the n-gram overlap between the generated caption and reference captions, indicating how closely the model's output matches human-written text in terms of word sequences. METEOR builds upon word-to-word matches by incorporating synonym matching and word order, providing a more nuanced assessment of semantic alignment. ROUGE-L emphasizes recall-based matches of contiguous word sequences and is thus sensitive to longer matching segments across reference and generated captions.

Table 4.3 compares our model's captioning performance with two existing approaches on the MS COCO dataset. Yin and Ordonez [30] employed a sequence-to-sequence model integrating YOLO9000 for object layouts, and their results show a BLEU score of 0.253, METEOR of 0.238, and ROUGE-L of 0.507. Herdade et al. [31] similarly focus on object detection features by incorporating geometry into a Transformer-based architecture, achieving BLEU of 0.265, METEOR of 0.241, and ROUGE-L of 0.510. Our model surpasses both methods in all three metrics. Notably, we observe approximately a 25% relative increase in BLEU over Yin and Ordonez [30], as well as more than a 20% improvement in ROUGE-L.

When comparing METEOR, our model attains 0.5916, representing over twice the score achieved by either reference method. This significant gain in METEOR implies that our approach offers a more precise match to ground-truth captions,

capturing both lexical and semantic details more effectively. We attribute these improvements to our hybrid detection framework, which resolves occlusion or cluttered scenes by combining local precision with a global context, and then fusing textual and visual information through cross-attention. By holistically capturing both object-level features and broader spatial relationships, our system yields captions that better reflect the actual contents of the scene.

CHAPTER 5

SOCIAL IMPACT AND SUSTAINABILITY

5.1 ENHANCING PUBLIC SAFETY

The proposed framework significantly enhances public safety, particularly in applications involving autonomous systems and surveillance. Traditional models often underperform in scenarios with occlusions, which are common in real-world environments such as urban intersections or crowded public spaces. The integration of occlusion-aware detection and captioning allows the system to identify and describe partially visible objects—like a pedestrian obscured by a parked vehicle or a cyclist behind a tree. This capability is critical for autonomous vehicles, drones, and robots that must make real-time decisions based on incomplete visual information, thereby reducing the risk of accidents and improving situational awareness.

5.2 SUPPORTING HEALTHCARE AD ASSISTIVE TECHNOLOGIES

This system holds substantial potential in the healthcare domain, particularly in enhancing mobility and independence for visually impaired individuals. When deployed through mobile applications or wearable devices, the model can provide audio feedback describing the surrounding environment. By including occluded

or partially visible objects in its descriptions, the framework provides a more complete and contextual understanding of the user's surroundings. This promotes greater confidence, autonomy, and safety for individuals navigating public or unfamiliar environments, directly contributing to inclusive technological development.

5.3 DRIVING SOCIETAL TRANSFORMATION THROUGH TECHNOLOGY

At a broader level, this project reflects the ongoing societal transformation driven by intelligent systems. By merging computer vision with natural language understanding, the framework moves toward human-like perception in machines. This opens up a range of possibilities—from smart cities and autonomous delivery systems to educational tools and public infrastructure monitoring. The ability to interpret complex, cluttered scenes using multimodal reasoning represents a step forward in how machines interact with the world and humans, offering greater utility, accessibility, and understanding across sectors.

5.4 ENERGY EFFICIENT ENVIRONMENTAL IMPACT

Although computationally advanced, the framework has been designed with modularity and optimization in mind. Hybrid model's fast inference capabilities

and BLIP's lightweight attention-based captioning keep the model resource-efficient. This ensures that the framework does not require excessive hardware overhead, which in turn contributes to reduced energy consumption. As AI-based systems become more widespread, energy efficiency becomes crucial not just for sustainability but also for reducing carbon footprints associated with large-scale deployment.

5.5 ADAPTABILITY TO LOW-POWERED DEVICES

Another critical advantage of the proposed framework is its adaptability to low-powered devices such as embedded systems, Raspberry Pi units, and mobile processors. Due to the modular dual-stage architecture, the model can be selectively pruned or quantized for deployment on edge devices without compromising essential performance. This ensures accessibility even in resource-constrained settings, such as rural areas, developing regions, or embedded applications in portable assistive technology. Such adaptability makes the system versatile for both high-end and grassroots-level impact.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

This project explored a novel dual-stage framework combining object detection and occlusion-aware image captioning, leveraging the strengths of hybrid model for precise and real-time object localization, and BLIP with cross-attention for semantically rich and context-aware caption generation. The system was specifically designed to address the limitations faced by traditional models in interpreting scenes with partially visible or occluded objects.

The first phase of the project focused on enhancing detection accuracy by integrating Faster R-CNN and DETR into a dynamic dual-processing pipeline, inspired by human cognition. This allowed the model to allocate tasks based on object complexity and visibility, improving robustness in cluttered environments. In the second phase, the project advanced further by incorporating inpainting-based occlusion handling and cross-attention-based captioning, resulting in a holistic framework capable of not just detecting but also describing objects and their context effectively.

Experimental results demonstrated that the proposed architecture could handle diverse scenarios including occlusion, background clutter, and multi-object scenes. Outputs validated that the system was able to generate relevant captions even when objects were partially obstructed, significantly contributing to accessibility, safety, and scene comprehension in real-world applications.

The social impact and sustainability aspects of the system were also addressed. From enhancing public safety in autonomous navigation to supporting assistive technologies for visually impaired individuals, the project highlights the potential of AI in driving inclusive and responsible innovation. Moreover, the framework's design supports scalability, energy-efficient deployment, and adaptation to low-powered devices, making it well-suited for global, real-world use.

While the outcomes of the current work are promising, there remains ample opportunity to refine, optimize, and extend this system for broader adoption and specialized applications.

6.2 FUTURE WORK

6.2.1 Refinement of Cross-Attention Mechanisms

While the integration of cross-attention between hybrid model's visual embeddings and BLIP's textual decoder improved caption relevance, there remains scope for enhancing the semantic alignment. Future work could involve experimenting with advanced attention strategies such as dynamic token pruning, attention gating, or adaptive attention span to better focus on relevant object regions during caption generation. Multi-layer cross-attention stacks or hierarchical attention fusion could also lead to more context-sensitive and coherent outputs.

6.2.2 Context-Aware Scene Understanding

The current framework generates object-centric captions but lacks deeper reasoning about spatial relationships or inter-object dynamics. A promising extension would involve training the system to generate scene graphs or relational captions, such as ”A man holding an umbrella next to a parked car.” This would require integrating spatial relationship extractors or leveraging vision-language pretraining on datasets annotated with object interactions. Incorporating contextual priors could also help distinguish between visually similar objects in different scenarios.

6.2.3 Robustness Under Extreme Occlusion and Clutter

Although the model is occlusion-aware, its performance can degrade in highly cluttered environments or scenes with overlapping instances. Future versions could incorporate amodal instance segmentation techniques or probabilistic reasoning to infer the presence of occluded objects. Leveraging depth information, synthetic occlusion datasets, or multimodal fusion (e.g., LiDAR + vision) may also improve robustness under challenging visual conditions.

6.2.4 On-Device Deployment and Model Compression

To bring this framework into practical real-world applications—such as mobile accessibility tools or edge-based smart cameras—there is a strong need to reduce its computational overhead. Future work should explore model compression

techniques like knowledge distillation, low-rank approximation, quantization, or sparse attention for BLIP. Lightweight alternatives to hybrid model can also be evaluated for maintaining accuracy with minimal resource consumption.

6.2.5 Human-in-the-Loop Captioning for Accessibility Tools

For use in assistive technologies, incorporating human-in-the-loop systems could vastly improve caption quality and personalization. Future enhancements could include interactive refinement where users can query, correct, or expand upon generated descriptions. This feedback can be used to fine-tune captioning models in a few-shot learning setup, making the system more adaptive to individual user needs, language preferences, or specific contexts (e.g., indoor navigation for the blind).

REFERENCES

1. Chen, Q., Wang, Y., Yang, T., Zhang, X., Cheng, J., and Sun, J. (2021) 'You Only Look One-Level Feature', In Proc. of CVPR, pp. 13039–13048.

2. Dai, X., Chen, Y., Yang, J., Zhang, P., Yuan, L., and Zhang, L. (2021) 'Dynamic DETR: End-to-End Object Detection with Dynamic Attention', In Proc. of ICCV, pp. 1018–1025.

3. Galvez, R. L., Bandala, A. A., Dadios, E. P., Vicerra, R. R. P., and Maningo, J. M. Z. (2018) 'Object Detection Using Convolutional Neural Networks', In Proc. of TENCON, pp. 457–462.

4. H, L., Bu, T., and Zhang, M. (2023) 'A Dynamic Dual-Processing Object Detection Framework Inspired by the Brain's Recognition Mechanism', In Proc. of ICCV, pp. 349–354.

5. Lin, T. Y., Dollar, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017) 'Feature Pyramid Networks for Object Detection', In Proc. of CVPR, pp. 2117–2125.

6. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., and Zitnick, C. L. (2014) 'Microsoft COCO: Common Objects in Context', In Proc. of ECCV, pp. 740–755.

7. Mao, M., Zhang, R., Zheng, H., Ma, T., Peng, Y., Ding, E., Zhang, B., and Han, S. (2021) 'Dual-Stream Network for Visual Recognition', In Proc. of NeurIPS, pp. 1234–1246.

8. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., and Wang, J. (2021) 'Conditional DETR for Fast Training Convergence', In Proc. of ICCV, pp. 3651–3660.

9. Park, S., Lee, S., Kang, J., Park, S., Choi, S., and Paik, J. (2022) ‘Dual-Attention Sparse R-CNN via Single ROI Transformer and Dynamic CBAM’, In Proc. of ICPR.

10. Patel, S., and Patel, A. (2020) ‘Object Detection with Convolutional Neural Networks’, International Journal of Computer Vision and Image Processing, 10(3).

11. Wang, H., Zhang, X., Liu, J., and Zhao, Y. (2023) ‘Occlusion Handling in Generic Object Detection: A Review’, IEEE Transactions on PAMI, 45(6), pp. 1234–1249.

12. Wang, L., Zhang, R., Chen, H., and Ma, T. (2017) ‘A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection’, In Proc. of CVPR, pp. 2606–2615.

13. Zhang, Y., Li, J., Wang, T., and Chen, H. (2024) ‘Adaptive Occlusion Object Detection Based on Overlapping IoU (OL-IoU)’, International Journal of Computer Vision, 132, pp. 56–72.

14. Su, R., Zhao, Q., Wang, J., and Liu, Y. (2022) ‘OPA-3D: Occlusion-Aware Pixel-Wise Aggregation Network for Monocular 3D Object Detection’, IEEE Transactions on Image Processing, 31, pp. 1203–1217.

15. Saleh, M., and Vamossy, G. (2022) ‘Bounding Box-Based Occlusion Detection and Order Recovery’, Pattern Recognition Letters, 159, pp. 50–59.

16. Zhou, X., Wang, D., and Krähenbühl, P. (2019) ‘Objects as Points’, In Proc. of ICCV, pp. 2767–2775.

17. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2020) ‘Deformable DETR: Deformable Transformers for End-to-End Object Detection’, In Proc. of ICLR.

18. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., and Berg, A. C. (2016) ‘SSD: Single Shot MultiBox Detector’, In Proc. of ECCV, pp. 21–37.

19. He, K., Zhang, X., Ren, S., and Sun, J. (2016) ‘Deep Residual Learning for Image Recognition’, In Proc. of CVPR, pp. 770–778.

20. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2018) ‘UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation’, IEEE Transactions on Medical Imaging, 39(6), pp. 1856–1867.

21. Fang, Y., Wang, J., Wang, X., Liu, Z., Dong, M., and Bao, H. (2023) ‘Eva: Exploring the Limits of Masked Visual Representation Learning at Scale’, In Proc. of CVPR, pp. 18992–19002.

22. Liu, Z., Ning, Z., Li, Z., Wei, H., Zhang, M., and Tang, Y. (2023) ‘Prompt-Driven Scene Text Recognition via Cross-Modal Contextual Decoding’, In Proc. of ICCV, pp. 1056–1064.

23. Wang, C., Yang, J., Li, M., Sun, H., Dong, Q., and Ji, Y. (2022) ‘All-In-One Transformer for Multi-Task Visual Learning’, In Proc. of ECCV, vol. 13682, pp. 201–219.

24. Li, J., Zhang, P., Wu, C., Wang, Y., and Zhu, Y. (2022) ‘BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation’, In Proc. of ICML, vol. 162, pp. 12888–12900.

25. Zhu, Y., Li, J., Yang, P., Zhang, B., Zhou, C., and Luo, P. (2023) ‘MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Captioning and Instruction Tuning’, In Proc. of CVPRW, vol. 47(4), pp. 237–244.

26. Xu, Y., Tan, X., and Wu, Y. (2021) ‘LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding’, In Proc. of ACM MM, pp. 1098–1106.

27. Kim, M., Choi, S., and Hwang, J. (2022) ‘ViL-Seg: Vision-Language Segmentation via Cross-Modal Attention with Text’, In Proc. of ECCV, vol. 13687, pp. 298–314.

28. Pan, X., Yao, Z., Li, P., and Shen, H. (2023) ‘Exploring Object Detection with BLIP and Cross-Attention Fusion’, In Proc. of CVPRW, vol. 47(6), pp. 1231–1238.

29. Xu, B. (2023) ‘Stable Diffusion from Scratch’, [Online]. Available: `https://scholar.harvard.edu/binxuw/classes/machine-learning-scratch/materials/stable-diffusion-scratch`.

30. Yin, X., and Ordonez, V. (2017) ‘Obj2text: Generating Visually Descriptive Language from Object Layouts’, In arXiv:1707.07102.

31. Herdade, S., Kappeler, A., Boakye, K., and Soares, J. (2019) ‘Image Captioning: Transforming Objects into Words’, In Proc. of NeurIPS, vol. 32, pp. 1–11.