EFFICIENT SCENE INTERPRETATION FRAMEWORK USING DYNAMIC DUAL-PROCESSING

Dr. T.T. Mirnalinee, Naveed Buhari and R. Nikhilesh,
SSN College Of Engineering, Chennai, India.

Task Description : To accurately detect and describe occluded objects using a hybrid detection model with inpainting and occlusion-aware image captioning module using BLIP with cross-attention fusion.

Objective :Design a dual-processing framework that fuses object detection and image captioning to interpret complex scenes with occlusions

Highlights of the Proposed Model

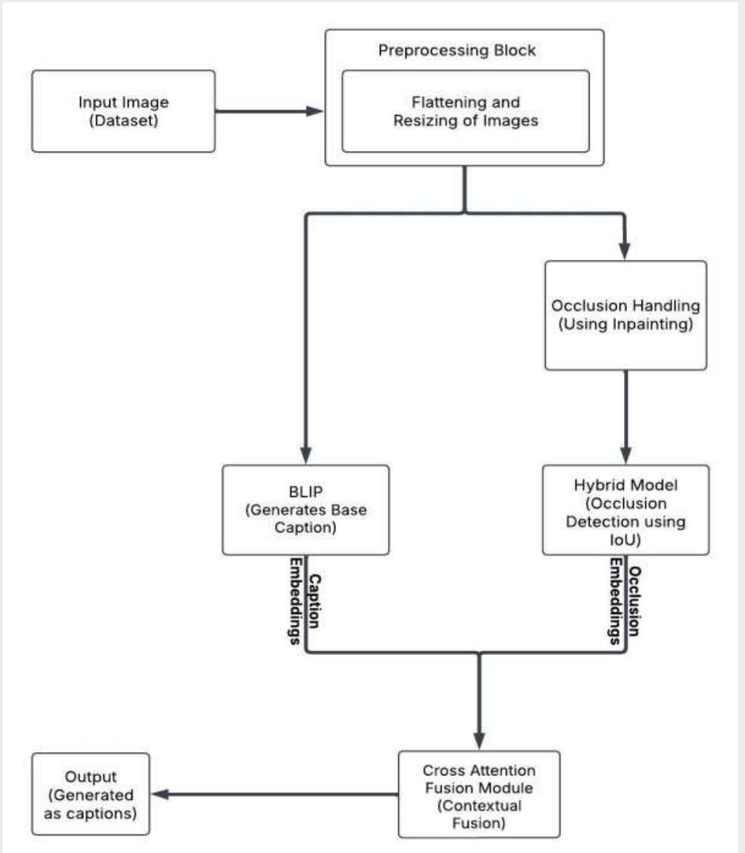
This framework highlights the importance of integrating object detection, occlusion recovery, and captioning to enable robust scene interpretation in complex visual environments. It addresses the limitations of single-method models and provides context-rich descriptions even when objects are partially hidden.

- Combines Faster R-CNN and DETR for hybrid object detection with both local and global context.
- Implements cross-attention fusion between detection outputs and BLIP captions for enhanced description accuracy.
- Generates coherent and context-aware image captions even in cluttered or partially visible scenes.
- Achieves improved precision (0.70) and recall (0.73) on the MS COCO dataset compared to standalone models.
- Delivers a METEOR score of 0.5916 and ROUGE-L of 0.6191 for captioning performance.

Dataset Description

- Utilized MS COCO for training object detection and captioning models with diverse, real-world scenes.
- Used OccludedPascal3D to evaluate model performance under various occlusion levels.
- MS COCO includes 200000 images across 80 object categories and OccludedPascal3D has10000 images with controlled occlusion scenarios across 12 categories.

Methodology



- Input images are preprocessed and passed through an inpainting block for basic occlusion handling.
- A hybrid detection model (Faster R-CNN + DETR) detects objects and generates occlusion summaries.
- BLIP generates base captions, which are refined using cross-attention fusion with the occlusion summaries.
- The framework effectively describes scenes with occluded or cluttered visual elements.

Performance

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

Table 1: Precision and Recall Comparison for Faster R-CNN, DETR, and Hybrid Model

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

Table 2: Evaluation of inpainting techniques

Metric	Average Score
BLEU	0.3173
METEOR	0.5916
ROUGE-L	0.6191

Table 3: Average Evaluation Metrics for the generated captions

Inferences

- Hybrid detection improves accuracy by leveraging both local and global object features.
- Cross-attention fusion enhances caption relevance in complex or occluded scenes.
- The framework generates coherent, scene descriptive captions even under visual clutter and partial occlusion.