DLCV

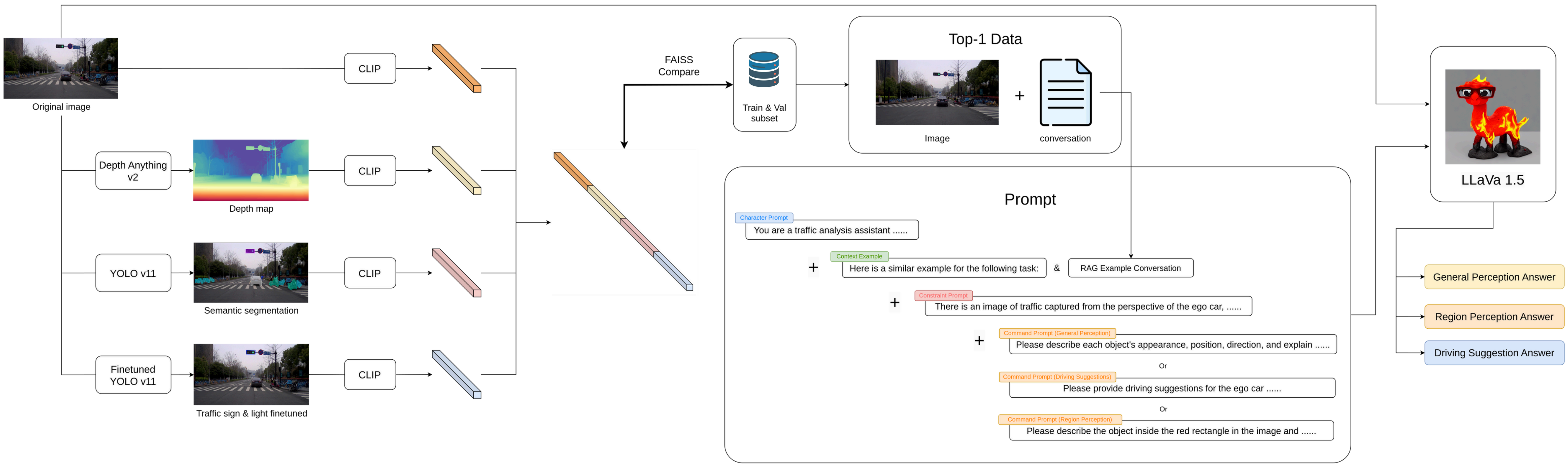
Don’t Learn Computer Vision

Authors:
NTU CSIE113 R13922A15 Xing-You, Chen
NTU CSIE113 R13922186 Yung-Chieh, Kao
NTU CSIE113 R13922193 Chu-Ching, Liang
NTU CSIE112 R12922205 Che-Wei, Lee

ABSTRACT

This study presents a comprehensive pipeline for multi-modal understanding and text generation, combining depth estimation, object detection, embedding-based retrieval, and fine-tuned visual-language models. The process begins with Depth Anything v2 generating a depth map while YOLOv11 performs object detection on the input image, identifying key objects such as vehicles and pedestrians, and creating semantic and bounding box representations. To enhance traffic sign detection, YOLOv11 is fine-tuned on a dedicated dataset, producing more accurate bounding boxes for traffic lights and speed limit signs. The original image, depth map, semantic image, and bounding box image are then encoded using CLIP into individual 1×512-dimensional embeddings, which are concatenated to form a unified 1×2048-dimensional representation. This embedding enables efficient retrieval of the most similar image and its associated conversation from a FAISS-trained dataset, providing contextual guidance. Finally, the fine-tuned LLava model processes the retrieved conversation and input image to generate descriptive text, demonstrating a novel integration of these components for enhanced image-text understanding and generation.

SYSTEM FLOW

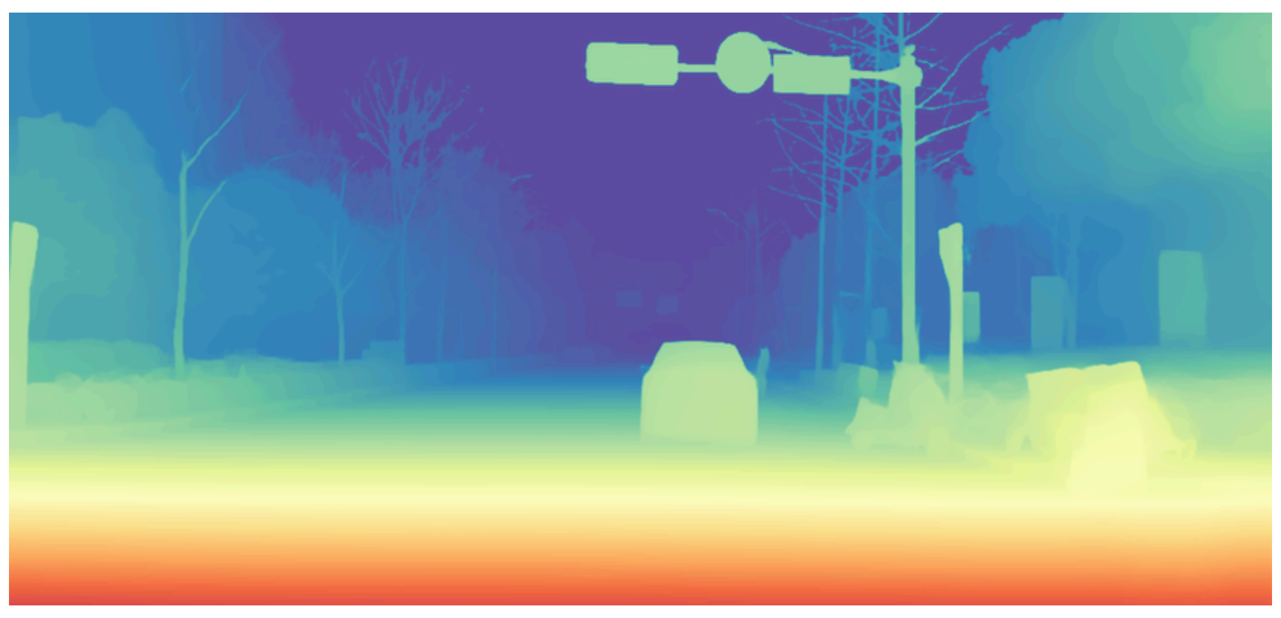

Original image
Depth Anything v2
Depth map
YOLO v11
Semantic segmentation
Finetuned YOLO v11
Traffic sign & light finetuned
CLIP
CLIP
CLIP
CLIP
FAISS Compare
Train & Val subset
Top-1 Data
Image
+
conversation
Prompt
Character Prompt
You are a traffic analysis assistant ......
+
Context Example
Here is a similar example for the following task:
&
RAG Example Conversation
+
Constraint Prompt
There is an image of traffic captured from the perspective of the ego car, ...
+
Command Prompt (General Perception)
Please describe each object's appearance, position, direction, and explain ......
Or
Command Prompt (Driving Suggestions)
Please provide driving suggestions for the ego car ......
Or
Command Prompt (Region Perception)
Please describe the object inside the red rectangle in the image and ......
LLaVa 1.5
General Perception Answer
Region Perception Answer
Driving Suggestion Answer

APPROACH

Our research proposes a robust pipeline for multi-modal understanding and text generation, integrating Depth Estimation, Semantic Segmentation, Retrieval-Augmented Generation (RAG), and LLaVa & Prompting. Each stage is meticulously designed to extract and combine spatial, semantic, and contextual information, enabling precise and enriched image-text generation.

Depth Estimation

We utilize Depth Anything v2 to generate depth maps from input images, capturing spatial relationships within the scene. Depth information plays a critical role in understanding the geometric structure of the environment, providing an additional dimension of context that complements visual features.



Semantic Segmentation

Object-level understanding is achieved through YOLOv11, which detects key objects such as cars, bicycles, and pedestrians, producing a semantic image with object masks. Additionally, YOLOv11 is fine-tuned on a dedicated traffic sign dataset to enhance its accuracy in identifying traffic-related elements, such as traffic lights and speed limit signs. This fine-tuning ensures detailed semantic segmentation and bounding box annotations for complex scenes.

Retrieval-Augmented Generation (RAG)

The outputs from the previous stages—original image, depth map, semantic image, and bounding box annotations—are encoded into 1×512-dimensional embeddings using CLIP. These embeddings are concatenated into a unified 1×2048-dimensional representation, which is then used to query a RAG dataset. The dataset contains pre-encoded embeddings according to each image. By retrieving the most similar image and its contextual conversation, this stage provides additional guidance for prompt generation.

LLaVa & Prompting

The retrieved conversation is incorporated into a dynamically generated prompt. This prompt, along with the input image, is processed by the fine-tuned LLaVa model, which generates descriptive text tailored to the visual and contextual information. The integration of RAG and LLaVa ensures that the generated text is not only accurate but also contextually enriched and semantically coherent.

EXPERIMENT

Our experiments aimed to evaluate the performance and adaptability of the proposed pipeline. We tested different LoRA ranks (rank 4, 8, and 16) to assess their impact on the model, analyzed the model’s ability to generate text when provided with detailed character descriptions, and conducted RAG testing by exploring various combinations of the four embedding vectors to determine the configuration that delivers the best results for multi-modal understanding and text generation.

Ep \ Rank	4	8	16
Epoch 2	3.891	3.972	3.755
Epoch 3	X	**3.973**	**4.118**

Table 1 ablation study on lora rank and epoch

Pt \ rank	4	8
With prompt tuning	2.636	2.089
W/o prompt tuning	**3.891**	**3.973**

Table 2 prompt tuning: “You are a traffic analysis assistant”

Embedding \ method	Few shot
Original	2.199
Original + yolo	**3.244**
Original+ yolo +depth map	3.492
Original+ yolo +depth map + traffic sign	2.242

Table 3 RAG: experiment on type of embedding

Pt	4
With threshold	**4.056**
W/o threshold	3.492

Table 4 Threshold on RAG retrieved data