OBJECT DETECTION AND RECOGNITION IN NON-IDEAL CONDITIONS COMBINED WITH SEMANTIC AUGMENTATION

Trương Lê Mỹ Thanh[1]

[1] Trường ĐH Công Nghệ Thông Tin

Introduce

A method to detect and recognize object (person, vehicles,…) in non-ideal conditions (rainy, night, foggy,...)

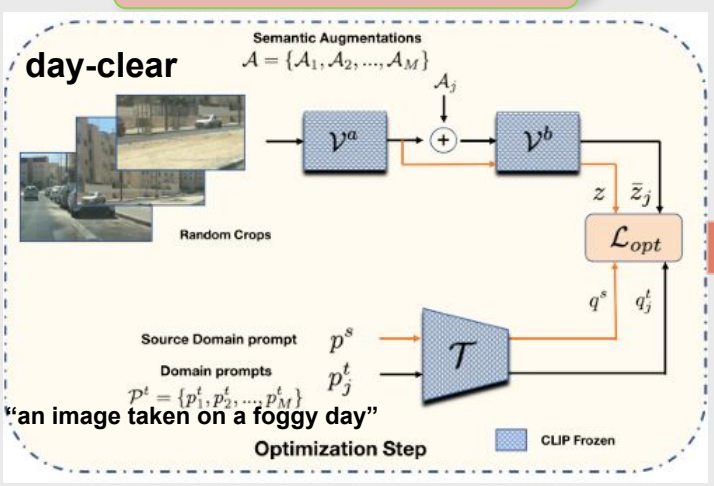
- Using backbone of CLIP: ResNet101, helps exploit pre-learned visual-linguistic features
- Use a list of text prompts describing each target domain condition and move the original image feature vector closer to the corresponding position of the new domain in the embedding space.

Target

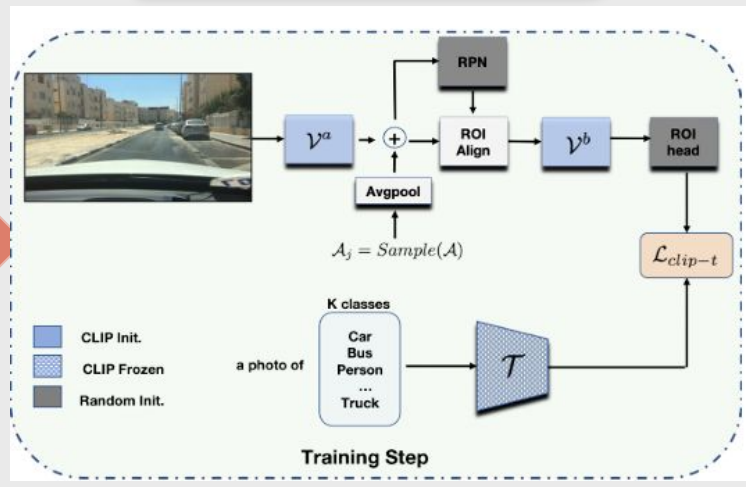
- Identifying and recognizing objects in non-ideal environmental conditions support the deployment of intelligent surveillance systems and more efficient rescue or reconnaissance robots, avoiding the use of humans in surveillance or dangerous tasks.
- Only need to use **one source data domain** but can **recognize unknown domains**

Overview

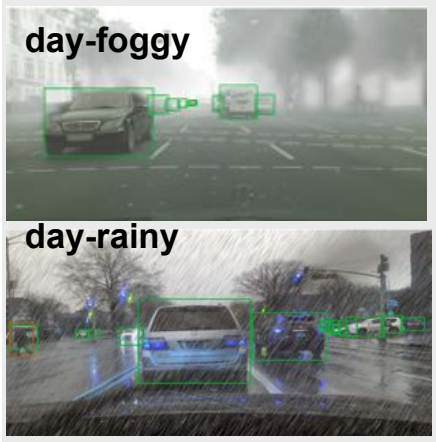
Optimization Stage



Training Stage



Testing



Description

1. Semantic Augmentation

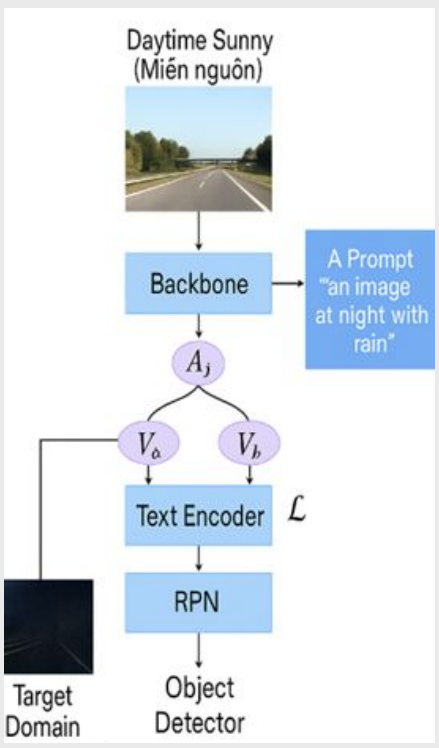


Figure 1. Apply Semantic Augmentation

"sunny day photo" is shifted in feature space to the new domain using text prompts
"rainy night photo"

NOT modify source image, enhances extracted features , help model learn domain variation without the target domain.

Use multiple prompts (prompt ensemble) 5-10 diverse prompts for each unknown domain.

Fog = {"a misty road", "low-visibility weather", "a fog-covered city", an image in poor visibility conditions, …}

2. Optimization Stage

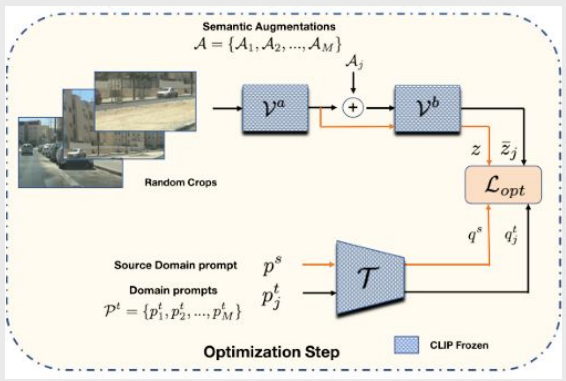


Figure 2. Optimization Stage for creating vectors Aj

- Use source domain images and text prompts describing weather conditions or time of day.
- Compute the characteristic displacement vector in the CLIP space.
- Optimize A□ so that the features of the enhanced image approximate those of the simulated target domain.

Goal: Generate feature enhancement vectors **Aj = {A₁, A₂,..., A□}** from Prompt to help the model learn domain changes (e.g. from daytime to rainy night).

3. Training and Testing Stage

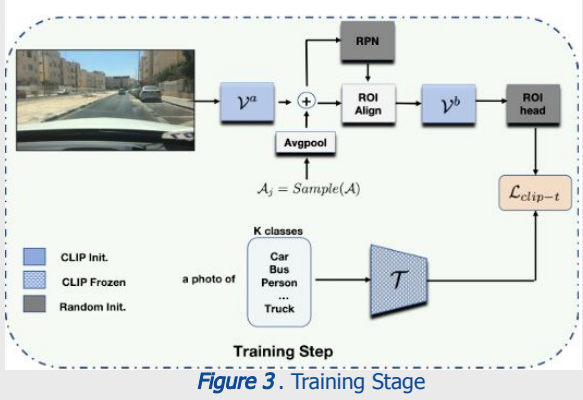


Figure 3. Training Stage

- The input image features are extracted and then added with vector **Aj** to simulate a new domain.
- Text prompts for each class (e.g., "a photo of a {class}", such as "car") are used to train classification via CLIP.
- RPN and ROI Align are combined to locate objects, while CLIP and text prompts are used to identify object classes in the image.

Goal: Use the optimized augmentation vectors **Aj** to train the CLIP model for object detection generalized to unseen domains

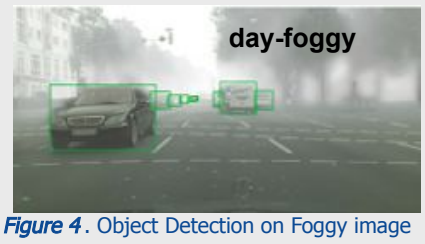


Figure 4. Object Detection on Foggy image

NII

Trương Lê Mỹ Thanh – Trường Đại học Công nghệ Thông tin. ĐHQG TP.HCM
TEL : 0876313757 Email : thanhtlm.19@grad.uit.edu.vn