Semantically Enhanced Scene Captions with Physical and Weather Condition Changes

Hidetomo Sakaino
Visual Recognition Group, Weather Transportation Lab., Weathernews Inc.
sakain@wni.com

Abstract

Vision-Language models (VLMs), i.e., image-text pairs of CLIP, have boosted image-based Deep Learning (DL). Moreover, Visual-Question-Answer (VQA) tools and open-vocabulary semantic segmentation provide us with more detailed scene descriptions, i.e., qualitative texts, in captions. Images from surveillance, auto-drive, and mobile phone cameras have been used with segmentation and captions. However, unlike indoor scenes, outdoor scenes with uncontrolled illumination and noise can degrade the accuracy of segmented objects. Moreover, unpredictable events such as natural phenomena and accidents can cause dynamic and adverse scene changes over time. This greatly increases unseen objects due to sudden changes. Therefore, only a single state-of-the-art (SOTA) VLM and DL model cannot sufficiently generate and enhance captions. Even one time VQA is limited to generate a good answer. This paper proposes RoadCAP for refined and enriched qualitative and quantitative captions by DL models and VLMs with different tasks in a complementary manner. In particular, 2D-Contrastive Physical-Scale Pretraining (CPP) is also proposed for captions with physical scales. An iterative VQA model is proposed to further refine incomplete segmented images with the prompts. Experimental results outperform SOTA DL models and VLMs using images with adverse conditions. A higher semantic level in captions for real-world scene descriptions is shown as compared with SOTA VLMs.

1. Introduction

Deep Learning (DL) and Vision Language model (VLM) have become the most useful and effective for real-world applications, i.e., surveillance cameras in the cities and highways, auto-driving, and drone cameras [54, 6, 16, 38, 19, 50, 31]. Like DLs, VMLs are useful in various applications, such as object detection [13, 51, 62, 14, 77, 42, 30, 89, 99, 53, 44], segmentation [46, 10, 33, 88, 95, 49, 41, 86, 73, 82, 43], and classification. Various things and stuff such as roads, vehicles, pedestrians, buildings, nature, rivers, and waves have been used to train and infer by DLs and VMLs.

VLMs can understand vision and text, allowing them to perform tasks requiring multimodal understanding, i.e., Visual Question Answer (VQA), image captioning, or image retrieval. Moreover, VLMs can be pre-trained on large datasets [61, 52, 34] and fine-tuned on smaller datasets for specific tasks, allowing for efficient transfer learning [83, 21, 78, 36]. Notably, VLMs present higher performance than DLs and Computer Vision when unseen images that have not been pretrained have been recognized [98, 97, 48, 2, 45, 9, 101, 22, 4, 75, 20, 11, 25, 74, 1, 90, 71, 28, 83, 91, 56, 76, 27, 58, 93, 21, 78, 96, 36, 61, 34].

Most state-of-the-art (SOTA) papers in DLs and VLMs apply images with objects under stable and visible weather conditions, i.e., clear or cloudy. However, DLs and VLMs are still vulnerable to illumination changes, i.e., sunbeams, headlights, and reflection, and weather changes, i.e., rainfall, snowfall, and fog. In DLs, these have been tackled by Dynamic changes have been dealt with by De-raindrops [59], Defog, and Dehaze [31, 38, 19]. For heavier snowfall, rainfall, and fog at twilight and night, incomplete segmentation cannot be avoided. For this issue, SOTA papers [64, 63] show switching different pre-trained domain models, but they require a manual selection of domain models to different scenes. To solve this, SOTA papers [64, 63] have been proposed to auto-adapt such adverse scenes by switching different DLs.

On the other hand, the performance of VLMs with VQA (input of an image and text: output of an answer) and image captioning (input an image: output text) can be degraded by dynamic natural phenomena and dramatic scene changes, i.e., adverse conditions. Since a single VLM with a single VQA deals with limited tasks, more modifications to multiple elements of such adverse conditions are needed. Since a single question is predefined, an answer cannot be updated according to such scene changes. It is also noted that SOTA VLMs rely on segmentation by DLs, but incomplete segmentation is obtained. Therefore, incomplete im-

age2text results are shown accordingly. Moreover, since non-physical scale-based segmentation is made, resulting captions do not contain physics-based texts. SOTA VLM [3, 15, 37, 69, 72, 102, 61] paper presents indoor objects to describe captions with two-dimensional relationships between objects. 2D adverbs, i.e., left and next, have been used. However, three-dimensional relationships between objects, i.e., distance, will help capture outdoor objects more clearly using depths.

Adverse weather conditions impact environments and objects, where rainfall and fog cause low visibility to human drivers and auto-driving systems. Extremely disastrous events happen on roads with flooding and landslide. Such complicated visual conditions on roads may lead to severe traffic accidents. However, such physical scalings, i.e., meters and volumes of small or large objects, are important to estimate their depths, widths, and heights. In traffic scenes, road conditions, i.e., dry, wet, and snow, are the most impact on any automobile. During rush hours or disaster events, we can see heavy traffic jams. Therefore, counting of them on the road becomes helpful in understanding the scene. However, SOTA DLs and VLMs fail to count far and tiny objects, i.e., pedestrians, vehicles, birds, and fishes.

Data augmentation is the standard method for enhancing the performance of DLs and VLMs. However, it is difficult to collect a sufficient number of training images in all weather conditions and adverse conditions. On the other hand, adverse conditions like lens reflection, blur, and strong illumination may require several times larger image datasets than normal conditions when enhancing the generalization performance of DLs and VLMs. Therefore, it becomes more difficult to collect them. For this, rejection of such difficult images has been shown efficient and effective approach [64, 63]. This greatly eliminates to recognize errous objects.

In order to maximize the recognition capabilities of DLs and VLMs, tasks to them should be minimized. The simplest architecture is a cascaded model. However, all DLs and VLMs are necessarily used for a specified task. Moreover, the complementary modules between DLs and VLMs may boost overall performance. Therefore, the complementary combination of different DLs and VLMs with different tasks may be enhanced overall recognition accuracy in images and texts.

To this end, this paper proposes RoadCAP with multiple Deep Learning (DL) and Vision Language Models (VLMs) for the enrichment of captions from many scenes, particularly in adverse weather conditions. DLs and VLMs with different tasks are complementary in branched architecture so that each DL and VLM can work in a maximum mode.

Contributions of this paper are fourfold:

1. RoadCAP consists of thirteen modules, i.e., Deep Visual Language Classification (Dvlc), Deep Visual Language Segmentation (Dvls), Deep Visual Language Detection (Dvld), Visual-Query-Answer (VQA), Contrastive Physical scale Pretraining (2D-CPP), Deep Visibility estimation (Dvis), Deep Road conditions (Droad), Deep Depth (Ddepth), Deep anomaly (Danomal), Deep water-level (Dwater), Deep snowfall (Dsnow), and Deep Count (Dcount). Most of the modules are based on transformers. The branched architecture allows us to maintain and upgrade efficiently. Moreover, progressively processing each module at CPU/GPU is helpful to reduce excessive memory usage of overall modules at one time.

2. In DL modules, Dcount is proposed to count vehicles in heavy traffic jams, where the number of far tiny objects is predicted. Dwater is applied to estimate the physical depth of flooding.

3. In VLM modules, 2D-CPP is proposed with a 2D contrastive learning model. Iterative VQA is proposed for enhancing captions more than one-time VQA.

4. Many experimental results show the superiority of the proposed RoadCAP over SOTA DLs and VLMs. The proposed RoadCAP will help notify detailed scene descriptions, i.e., more quantitative texts, to drivers, auto-driving, and rescue workers from camera images.

2. Related Work

This section briefly describes Deep Learning (DL) and Vision Language Model (VLM) concerning methods and issues in scene understanding of camera images under various conditions. Visibility levels are one of the most important visual factors to estimate for monitoring and auto-driving. To estimate visibility, segmentation-based DL models have been reported and used by Dvis [64] and Droad [63].

An all-in-one image restoration network for unknown corruption has been proposed [32]; however, this method can be degraded heavy fog and snowfall, as shown in Dvis [64] and Improved Droad [63]. Dehazing in [16] is limited to closer views of daytime lighter foggy scenes, i.e., indoor and garden, unlike the proposed method [64] for far scenes with heavy fog at night, i.e., highway. A unified framework for depth-aware panoptic segmentation has been reported [29] under clear weather conditions.

Although Cityscapes with 3000 images [7], Foggy Cityscape DBF with 500 synthetic foggy images [67], and Foggy Zurich with 3800 real light foggy images [68] are publicly available, they are almost all daytime and lighter fog data. These datasets do not include specific physical values; instead, they provide relative position values. Moreover, image datasets for road conditions have not been built, unlike Droad [64, 63].

In recent years, the VLM field has experienced significant progress [92, 98]. But most of them are pre-trained with large-scale training datasets and fine-tuned with task-specific annotated training data. The pre-training of VLMs has been explored using three main approaches: contrastive objectives [60, 39, 8], generative objectives [79, 24], and alignment objectives [17, 85, 47]. VLMs are transferred by Text-Prompt Tuning [98, 97, 48]. Besides finetuning, knowledge distillation is a method to improve VMLs for downstream tasks, including object detection [94, 44] and semantic segmentation [95, 100, 49, 73].

Unseen images that have not been pre-trained have become recognized by VLM frameworks [3, 15]. More diverse and out-of-distribution data for pre-training and evaluation are used [23]. Prompt learning to adapt VLMs to new tasks without fine-tuning is also shown [26]. Contents of captions have been enhanced for better descriptions of real-world objects [15].

Geometric reasoning or depth estimation to infer 3-D information from 2-D images [87, 92] is shown using 3D point-cloud data and indoor scenes. Pretraining VLMs require over 100 million image-text paired datasets for high accuracy, more than DL models require. Therefore, many efficient models have been proposed [3, 37, 69, 72, 66, 65]. However, laborious and time-consuming tasks remain unsolved in pretraining VLMs.

Visual ChatGPT API tool has become famous as the image-text captioning tool. The advantage of Visual ChatGPT [81] is that it can produce acceptable results on the general scene and unseen classes. However, since Visual ChatGPT [81] is trained on the limited data of the year 2021, it generates captions under older datasets. So far, Visual ChatGPT [81] is weak at generating dynamic scene descriptions, i.e., natural phenomena and sudden events. Therefore, as aforementioned, no SOTA VLM papers and API tools have challenged images with the physical scale outside.

3. Proposed Method

This section describes the proposed RoadCAP method/system for refinement and enrichment of captioning and classes from a single image input. Instead of using only vision models or a single vision-language model, this paper proposes a new architecture that integrates multiple Deep Learning (DL) and Vision Language models (VLMs). Figure 1 shows an overview of the proposed RoadCAP. Since this paper deals with many challenging scenes with disasters and traffic accidents, adverse conditions are taken into account. Further detailed explanations of each module will be given in the following sub-sections 3.1 to 3.8.

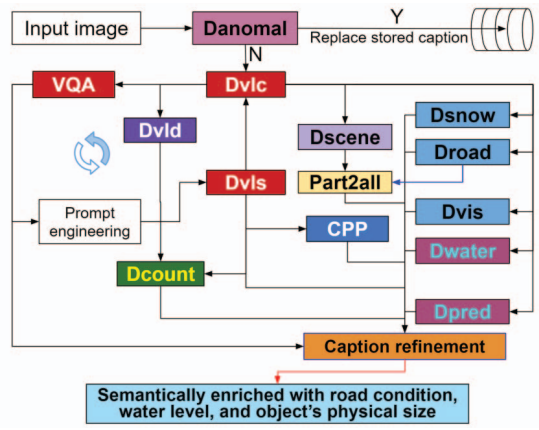


Figure 1. Overview of the proposed RoadCAP model.

3.1. Proposed Danomal

In real-world conditions, camera images suffer from unexpected illumination conditions, i.e., local strong sunbeams. When no-filtered images are input, detection and recognition accuracy become unstable, or targeted objects may interrupt to detect. The extra-illumination issue may be eliminated if the optic filter-like darkening can be provided at the camera lens. However, it is expensive to implement and the darker cloudy days and nighttime degrade overall object recognition performance.

Therefore, Danomal [64, 63] is designed to prefilter difficult input single images before applying the following DL and VLM modules. Such difficult images are called adverse image patterns, i.e., lens reflection, strong headlight, and raindrops, as shown in Figure 2. These are assumed to be major factors that degrade normal recognition processing. In this paper, fine-tuned Danomal using 2500 collected images is proposed to make it more robust to such adverse image patterns. In training, image experts define and classify two classes of rejection and no rejection. When an image is rejected, it will be pushed to the caption storage without undergoing processing, i.e., replacing the stored caption.

(a) (b) (c)

Figure 2. Examples of rejected images: (a) Lens reflection. (b) Strong headlight. (c) Raindrops.

3.2. Proposed Dvlc, Dvls, and Dvld

Dvlc is a VLM trained on image and text pairs that can predict the most relevant text given an image. It does not need to be directly optimized for this task and can perform "zero-shot" learning like GPT-3 and -4. Dvlc matches the performance of the original ResNet50 on ImageNet "zero-shot" without using any of the original 1.28M labeled examples.

Dvlc utilizes the input texts of five distinct disaster categories: *car crashes, flooding, fog, landslide, and rain*. Tailored text-input descriptions are employed for each disaster category to enhance Natural Language Processing techniques in analyzing disaster-related data. These scenes are associated with domain-specific terms to improve the accuracy of automated disaster detection and classification.

Dvls is suggested as a means to achieve semantic segmentation for these scenes. It is built upon the fine-tuned version of OvSeg [41], with the addition of a new physical constraint to the loss function. This proposed loss function will be explained in the next section. In order to obtain disaster descriptions for Dvlc, a classification task is undertaken, employing keywords that correspond to each disaster scene. These textual inputs are utilized to generate fixed text descriptions of the disasters, specific to each scene type.

Dvld is a VLM with open-vocabulary object detection [18]. Unlike traditional object detection models which rely on fixed categories, Dvld can detect objects based on arbitrary text inputs from Dvlc. The model achieves this capability by leveraging the knowledge extracted from a pretrained open-vocabulary image classification model. This knowledge is then utilized to create a two-stage detector, enabling Dvld to accurately identify and localize objects based on the textual descriptions provided.

Therefore, since Dvlc, Dvld, and Dvls recognize texts, objects, and segmentation from single imagea, all the outputs are integrated into captions. By this, more enriched captions are available than single VLMs.

3.3. Proposed 2D-CPP

Contrastive Physical-Scale Pretraining (CPP) is a variation of CLIP that incorporates inputs from a depth map, object location derived from image-text description pairs, and a modified contrastive loss function. Unlike SOTA VMLs that lack physical models in their contrastive loss functions, this paper introduces CLPP, which integrates additional physical constraints, as illustrated in Figure 3. The original contrastive loss function of CLIP [61] is defined by

$$L = \frac{1}{2}(1-Y) * D^2 + \frac{1}{2} Y * max(0, m-D)^2 \quad (1)$$

where * denotes a multiplication, Y is the binary label indicating whether the text and image are similar or dissimilar, D is the distance between the learned embeddings of the text and image, and m is the margin hyperparameter, i.e., 0.2. In order to incorporate the physical scale, including the size and location of objects, i.e., meters, a similarity metric.

The proposed 2D Contrastive Physical-Scale Pretraining (CPP) is a VLM with inputs from object locations from pairs of images and text descriptions and a modified contrastive loss function. Unlike SOTA VLMs with no physical models in contrastive loss functions, this paper proposes CPP with additional physical constraints, as shown in Figure 3.

The modified contrastive loss is then defined as

$$L = \frac{1}{2}(1-Y) * D^2 * (1-sim) + \frac{1}{2} Y * max(0, m-D)^2 * sim \quad (2)$$

where sim is the physical similarity between the text description and the image with object location. sim is computed as the Euclidean distance between the location of objects in the image and its description in the text. sim is defined by

$$sim = w_s * E(S_T, S_I) + w_l * E(R_T, R_I) \quad (3)$$

where: w_s is the weight of an object physical size, and w_l is the weight of object's physical location, normally, w_s and w_l are both set equal to 0.5. $E(S_T, S_I)$ is the Euclidean distance between the physical size in image S_I and in the text description S_T. $E(R_T, R_I)$ is *RMSE: Root Mean Square Error* between the physical object location in the image R_I and in text description R_T. The physical size of the object is determined based on the ratio between the object size in pixels and the object size in meters as labeled in the dataset.

When using cosine similarity as the distance metric D in the contrastive loss function, which ranges from -1 to 1, the margin hyperparameter is typically set to a small value, i.e., 0.2 to 0.5.

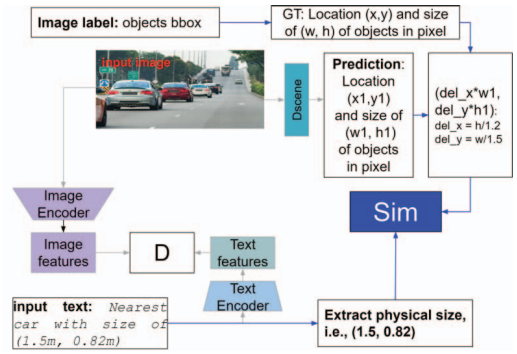


Figure 3. Proposed contrastive language for pre-training in physical scale.

3.4. Dscene, Dsnow and Dvis

This section introduces Dscene, Dsnow and Dvis [64, 63]. Dvis [64] is trained by known synthetic and real foggy images for visibility distances, i.e., 0 m − 1000m, in a regression manner. Dsnow classifies snowfall from light to heavy, which is trained by real snowfall images.

Dscene [64, 63] recognizes things and stuff in images within pretrained model, i.e., COCO or ADE20K. For example, things are car, pedestrians, mountain, road (dry/wet), building, and wall. Stuff contains snow, sky, and light. It is noted that road from Dscene cannot recognize road conditions of dry or wet. However, snow on roads, buildings, and cars is also recognition.

3.5. Proposed Part2All for Road Segmenation

Even with SOTA segmentation models, segmentation can be incomplete. In particular, illumination, shadow, noise, and weather impact segmentation quality. This section devotes to proposing the refinement of incomplete to complete segmented objects. Under dynamic scene changes, there are various factors impacting the performance of Droad. In such conditions, Droad is able to partially recognize the condition of the road only. Therefore, refinement to such incomplete road conditions cannot be ignored. Droad [63] is trained on labeled three road conditions by experts. In this paper, over 1500 images are added to train DL model. The original Droad has been improved, called improved Droad (iDroad). However, since iDroad is still weak in snow conditions, further refinement is required.

The proposal of refinement on road conditions consists of several steps. Overall Part2all module is shown in Figure 4. Firstly, Dscene is applied to a single image, where cars, roads, buildings, and pedestrians are recognized. Secondly, iDroad is used to segment road surfaces into three classes. The condition of the road surface is segmented as snow, however, this recognition is limited to the part region of the road that is in proximity to the camera, illuminated by the headlights. It is assumed that road conditions are homogeneous. Therefore, part of the snow region is overlapped onto the road surface of Dscene. Next, extrapolation from partial snow conditions in yellow is conducted to the remaining purple regions. Finally, all yellow regions over the road surface can be obtained by Part2All.

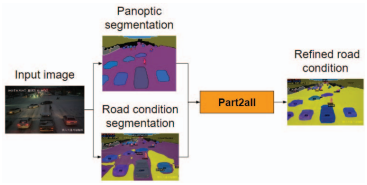


Figure 4. The proposed Part2all for refining road conditions.

3.6. Proposed Dwater

Dwater is a transformer-based classifier, i.e., the ViT [12] classifier. Through a two-step process, the water level is estimated based on the physical size, i.e., the height of recognized objects, i.e., cars, buses, humans, trees, poles, and traffic signs. In Step 1, Dvls is employed to extract objects from the image. In Step 2, the extracted objects are classified according to their pre-defined water levels. Table 1 displays the physical height of the reference objects for each water level.

Table 1. Physical height of reference objects.

Level/Objects	human (m)	car (m)	pole (m)
Lv1	0.3	0.3	0.5
Lv2	0.6	0.6	1
Lv3	0.9	0.9	2
Lv4	1.3	1.2	3
Lv5	1.7	1.5	4

3.7. Dcount

Road scenes are dynamically changing by traffic flow. The number of vehicles varies from 0 to over tens. In addition, there may be pedestrians and many other objects as well. Therefore, counting the number of objects is one of the most basic methods for the enrichment of captions. For this, this section proposes Dcount to count such objects. It is known that counting tasks become more difficult when the spatial density of objects becomes heavier, i.e., traffic jam. Therefore, a single counting model can fail to correctly detect objects. Based on this, Dcount consists of two different tasks: segmentation and object detection. Such two different modules are expected to be complementary due to their respective strengths.

As depicted in Figure 5, the inputs of Dcount consist of segmentation masks by Dvls and bounding boxes of objects by Dvld. The number of objects is estimated as the ratio between A and B of union classes of segmentation and detection. A represents the mask of each class in the number of pixels, and B represents the average of bounding boxes in the number of pixels.

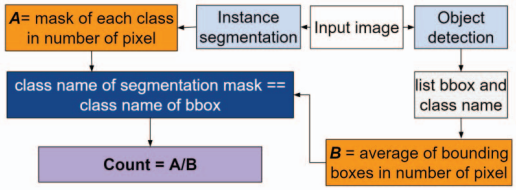


Figure 5. Flowchart of the object counting algorithm.

3.8. Caption Refinement

The caption refinement process involves utilizing a large language model (LLM), namely GPT-4 [55], which incorporates the segmentation outcomes from Dvls, the physical scale from 2D-CPP, and the captions generated by VQA. The output of Dvls comprises semantic segmentation along with corresponding locations and descriptions, expressed in a language-based segmentation format as a list of {object description: bounding box of the object in pixels}. The output of 2D-CPP is a caption that includes details about the physical scale, i.e., object size, distance, water level, and visibility.

VQA contributes additional descriptions that capture the overall dynamic conditions, including adverse weather conditions, to provide contextual information for the LLM. The final result of caption refinement is an enriched caption that encompasses information about road conditions, water levels, and relative object locations.

4. Experiments and Discussion

This section devotes many experiments and discussions on DL and VLM modules of the proposed RoadCAP by

comparing related SOTAs. Particularly, complementary enhancements by DLs and VLMs are confirmed.

4.1. Danomal for Adverse Conditions

This section evaluates the performance of the proposed Danomal. This Danomal is designed to prefilter difficult input single images before applying the following many main DL and VLM modules. For this, several main modules of Droad, Dsnow, Dvis, and Dvlc are selected. Droad, Dsnow, Dvis, and Dvlc are expected to apply for various road conditions, snowfall or not, visibility distance, and classification, respectively. These are used with and without Danomal as a prefiltering by inputting various test images.

The results in Table 2 show higher accuracy using Danomal (in bold) than without using Danomal. It is noted that any thresholding has been offered like Computer Vision applications. Thus, the proposed Danomal has demonstrated effectiveness for rejecting difficult input images leading stable image data processing, i.e., recognition and classification.

Table 2. Comparison of accuracy for Droad, Dsnow, Dvis, and Dvlc with and w/o Danomal using difficult images with adverse conditions.

	Without Danomal (%)	With Danomal (%)
Droad	81.31	**86.07**
Dsnow	72.90	**78.22**
Dvis	75.58	**80.86**
Dvlc	91.78	**92.65**

4.2. Scene Recognition Capability of 2D-CPP and Dvlc

This section evaluates the performance of the proposed 2D-CPP and Dvlc qualitatively and quantitatively. The test dataset comprises six carefully selected categories, including a car crash in snow conditions, flooding with rain, low visibility with fog, landslide, wet road with rain, and traffic flow. These images are shown in Figure 6 (a)-(f). For the test dataset, ground truth captions have been manually annotated, encompassing key criteria such as the number of vehicles, scene category, and road conditions. A total of 3500 images have been collected for this purpose. The experiments are evaluated the performance of captioning using VLMs on dynamic scenes. The BLIP [35] model has been chosen as a point of comparison due to its high transfer capability for both VL understanding and captioning tasks.

Table 3 presents the ground truth captions, captions generated by the proposed 2D-CPP (with $m = 0.4$ in equation 2), and BLIP [35], for the six images depicted in Figure 6. 2D-CPP and BLIP [35] exhibit the ability to recognize the contextual information and key objects in the scenes, such as traffic conditions and various objects. However, compared to 2D-CPP, BLIP [35] falls short in capturing detailed scene features.

In (a), (b), and (f), 2D-CPP improves the accuracy of vehicle counting by identifying multiple vehicles instead of just one. In (b), BLIP [35] fails to recognize the road conditions, whereas 2D-CPP accurately identifies the flooding highway. In (e) and (f), 2D-CPP successfully recognizes the presence of rain and sunny conditions, respectively. On the other hand, BLIP [35] does not provide any weather information. In (c), 2D-CPP accurately recognizes the degree of fog as heavy fog over time. In (d), 2D-CPP identifies that the road is obstructed by a rock, while BLIP [35] describes it as a pile of rocks. Consequently, the proposed 2D-CPP demonstrates more refined and enriched captions compared to BLIP [35].

Table 3. Comparison of captions among ground truth, proposed 2D-CPP, and BLIP [35].

	Ground truth	2D-CPP	BLIP [35]
(a)	7 cars under heavy snowfall in a crashed scene	7 cars on the frozen road, 1 severely damaged car	A car is stuck in the snow
(b)	4 vehicles under flooding scene	Cars and motorcycle on the flooded highway	A man is crossing the street in the rain
(c)	Empty highway under heavy foggy scene	Highway under heavy foggy at daytime	Foggy road in the mountains
(d)	One car on the road, in a landslide scene	One car on the damaged road occluded by the rock	A tractor is parked on the side of a road next to a pile of rocks
(e)	One car on the rainy road	A street under a light rain at night	The image is of a street intersection at night

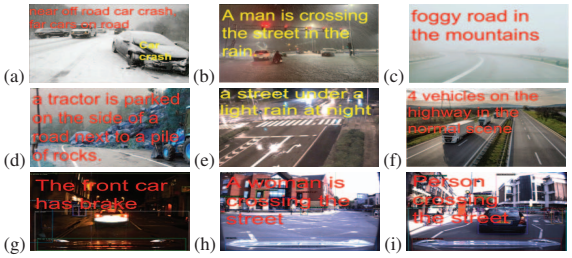


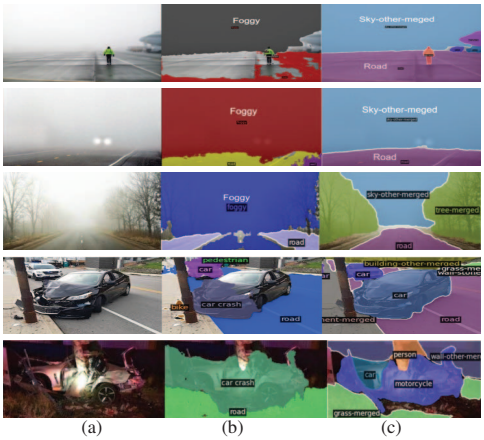
Figure 6. Results of captions in images: (a) Car crash. (b) Flooding. (c) Fog. (d) Landslide. (e) Rain. (f) Traffic flow. (g)(h)(i) Street.

4.3. Refined Semantic Segmentation by Prompt Engineering

This section denotes the proposed Dvls and how to obtain the final refined captions using prompt engineering. The prompt for each scene is pre-defined as a list of words, i.e., (1) car crashes: ["pedestrian", "car", "car crash", "road", "bike", "tree"]; (2) flooding: ["water", "car", "person", "tree", "sky"]; (3) fog: ["foggy", "mountain", "road", "car", "wet"]; (4) landslide: ["landslide", "debris flow", "rocks", "road", "dirt"]; (5)rain: ["water", "rain", "umbrella", "road", "person"]. Prompts are selected respectively by classification results from Dvlc.

Figure 7 illustrates the effectiveness of the proposed approach on images of foggy and traffic accident scenes. (a) shows the input images, while (c) displays the segmentation results generated by the transformer-based SOTA segmentation model, i.e., Mask2former [6], which shows generic

classes, i.e., ”sky-other-merged”, and ”car”. (b) presents improved segmentation results, and achieved prompt engineering, which provides more detailed semantic segmentation results, i.e., more detail from ”sky-other-merged” to ”foggy” for the foggy scene and from ”car” to “car crash” for the traffic accident scene. It has been demonstrated that prompt tuning for Dvlc helps detail segmentation results under dynamic conditions.



(a) (b) (c)

Figure 7. Results of segmentation by SOTA and proposed RoadCAP: (a) Original image. (b) Proposed refined semantic segmentation. (c) Mask2Former [6].

4.4. Dynamic Captions with Weather and Road Conditions by Improved Dvis, Dsnow, Droad, and Proposed Dwater

This section explores more complex captions by conducting experiments on traffic and disaster scenes under various weather conditions. The proposed models, Dsnow, Improved Droad, and Dvis, are compared with the SOTA VL captioning model, BLIP [35]. Figure 8 illustrates five scenes used in the experiments. The road conditions indicated by Improved Droad (1)-(5) are represented as wet (blue) and snowy (yellow). Dvlc recognizes objects in the overall scene, such as mountains, rivers, rocks, sky, and trees. The proposed Dvis [64], applied to images (1)-(5), estimates the physical scale based on weather phenomena, i.e., visibility in meters: 938, clear, 637, 512, and 812 meters for the respective images. Therefore, the captions generated by Dvlc include information about road conditions and visibility.

Table 4 summarizes the refined captions and the results obtained by SOTA BLIP [35] model, using the five scenes depicted in Figure 8. The comparison highlights that the refined captions provide detailed descriptions of the scenes, incorporating information about road conditions, snowfall status, object locations, and exact visibility distances in meters. In contrast, the captions generated by BLIP [35] lack descriptive elements. These results have demonstrated that the proposed method, integrating Improved Droad, Dsnow, and Dvis, outperforms the single VLM model, BLIP [35].

Table 4. Comparison of the refined captions with BLIP caption.

	Proposed method	**BLIP** [35]
(1)	Rocks lay on the flooding road, in front of river, within 938m in visibility	A flooded road in the rain
(2)	Rock debris lay on the wet road, withinclear visibility	A road in the rain with rocks and debris on the side
(3)	15 vehicles on the wet highway, under heavy snowfall and within 637m in visibility	A snowstorm on a highway
(4)	A truck on the wet highway with mountains in the rear, snow on the side of the highway, under heavy snowfall, and within 512m in visibility	A snow plow clears a road in the snow
(5)	12 people stand on a flooded road, within 812m in visibility, and 0.5m water level (Lv2)	A group of people on flooded road

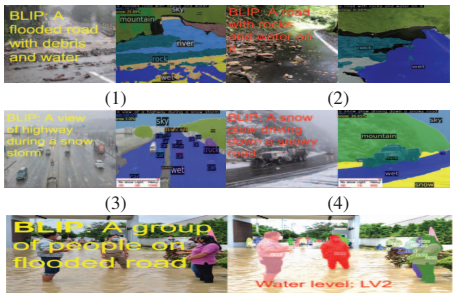


Figure 8. Results of proposed Dvls with refined and enriched captions in dynamic scenes: (1) Flooding road. (2) Landslide on the road. (3), (4) Heavy snowfall on the highway. (5) Flooded scene with the water level, Lv2.

4.5. Traffic Jams Detection by Dcount

To justify the proposed Dcount, SOTA objects counting methods, i.e., ZSC [84], DMDC [80], and CLTR[40], are compared. For robustness and stability evaluation, over one hundred images with adverse conditions, i.e., covered snow and fog, are selected.

Figure 9 shows the results of traffic jam recognition, where objects on the road tend to be occluded from each other, making it challenging to separate objects. CLTR[40] could not recognize vehicles resulting in counted zero. On the other hand, the proposed Dcound has successfully counted vehicles, i.e., truck: 2 or 15 and car: 11 or 2, where captions present ”crowded”. For the other SOTA results, Table 5 summarizes accuracy, where Dcount obtains best score, 86.54%.

Therefore, the propose Dcount has proven to enrich captions with traffic scenes even under adverse conditions over SOTAs, i.e., ZSC [84], DMDC [80], and CLTR[40].

Table 5. Counting objects in accuracy among the proposed Dcount, ZSC [84], DMDC [80], and CLTR [40].

Dataset/Method	**Dcount**	**CLTR [40]**	**ZSC [84]**	**DMDC [80]**
Accuracy	86.54	65.15	72.23	70.32

4.6. Overall Evaluation RoadCAP

This section presents an experiment that evaluates the final output of all thirteen modules. The experiment mea-

Figure 9. Object counting results in images with traffic jams. (1) Winter traffic jam images. (2) Proposed Dcount. (3) SOTA: CLTR[40].

sures performance using the BLEU score [57], a metric for evaluation of machine translation and is conducted on two datasets. The first publicly available dataset includes the COCO Caption dataset [5] and the Conceptual Captions dataset [70], both of which contain image-text pairs. The second dataset includes two images with accompanying text descriptions describing snowfall status, water level, and physical scale. These collections are disaster and traffic accident datasets, with 1850 and 2130 image-text pairs, respectively.

According to the results in Table 6, RoadCAP does not perform as well as Visual ChatGPT. This could be since the text descriptions in the public image set do not include information about road conditions, water levels, snow conditions, or visibility, whereas RoadCAP is capable of generating captions with these details. However, Table 7 presents different results, where RoadCAP outperforms Visual ChatGPT on datasets featuring disaster or traffic accident conditions.

It has been proven that RoadCAP can provide detailed semantics about the physical aspects of scenes. These can be highly useful for tasks such as traffic coordination and rescue operations. Moreover, the computational cost is analyzed and compared with that of SOTA methods on the same hardware device. Table 8 shows a comparison of the computational cost and memory usage for these methods.

Table 6. Performance of proposed RoadCAP on public datasets.

Dataset/Method	RoadCAP	Visual ChatGPT
COCO Caption	0.3854	0.4415
Conceptual Caption	0.3659	0.4235

Table 7. Performance of proposed RoadCAP on collected datasets.

Dataset/Method	RoadCAP	Visual ChatGPT
Disaster	0.4521	0.3124
Traffic accident	0.4315	0.3254

Table 8. Computational cost and memory usage comparisons.

Perform/Model	Computational cost (second)	Memory usage (Mb)
Proposed method	9.423	11231
Visual ChatGPT	8.123	6132
BLIP[35]	1.432	3214

5. Ablation Study

Various experiments are added to justify the usability, stability, and robustness of the proposed RoadCAP.

5.1. 2D-CPP with Different Loss Function Parameters

This section provides an experimental comparison of different parameters for the contrastive loss function (L) utilized in the proposed 2D-CPP. 2D-CPP is implemented in Sections 4.2 and 4.4 with m = 0.4 in equations 2 and 3. This experiment selects m from an array list of values [0.2, 0.3, 0.4, 0.5]. A comparison is conducted between equation (1) and the proposed equations (2) and (3). Table 5 compares the modified loss function and the original one for generating captions with physical scales. The results indicate that the performance of 2D-CPP achieves the lowest *RMSE* when m is set to 0.4. Therefore, this confirms the optimal selection of m.

Table 9. RMSE of different values m with/without sim.

m	Modified	Original
0.2	0.1985	0.2214
0.3	0.2043	0.2375
0.4	0.1894	0.2018
0.5	0.1964	0.2145

5.2. Limit of SOTA Image Restoration Model

In order to confirm another possibility for further processing images under adversarial conditions, image restoration by SOTA: all-in-one DL model [32] has been applied. Figure 10 shows results with (a) heavy snowfall, (b) raindrops on the lens, (c) light fog with a sunbeam at dawn, and (d) a clear twilight scene. It is obvious that no image restoration has been achieved by SOTA DL [32]. Instead, false colors are generated in red and sky blue.

Therefore, it is suggested that the proposed DeepReject plays an important role in avoiding visibility estimation in difficult images. This can stabilize overall system performance.

(a) (b) (c) (d)

Figure 10. Limit of an all-in-one deep learning model [32] for adverse weather conditions and clear scenes: (a) heavy snowfall. (b) raindrops on lenz. (c) light fog with a sunbeam. (d) clear twilight scene.

6. Conclusion

This paper has proposed RoadCAP with multiple DL and VLM models, which are complementary with branched structures for efficiency in light of memory, training, and maintenance. It is the first time to contain dynamic changes in captions with physical scales, i.e., fog visibility distance. A 2D physics-based loss function generates more refined and enriched captions at a contrastive loss. RoadCAP will help notify detailed scene descriptions to drivers, auto-driving, and rescue workers from camera images.

References

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 1(3):4, 2022. 1

[2] Adrian Bulat and Georgios Tzimiropoulos. Language-aware soft prompting for vision & language foundation models. *CoRR*, abs/2210.01115, 2022. 1

[3] Feilong Chen, Duzhen Zhang, Minglun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. VLP: A survey on vision-language pre-training. *Int. J. Autom. Comput.*, 20(1):38–56, 2023. 2, 3

[4] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. *CoRR*, abs/2210.01253, 2022. 1

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 8

[6] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 1280–1289. IEEE, 2022. 1, 6, 7

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3213–3223. IEEE Computer Society, 2016. 2

[8] Quan Cui, Boyan Zhou, Yu Guo, Weidong Yin, Hao Wu, Osamu Yoshie, and Yubo Chen. Contrastive vision-language pre-training with limited resources. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, pages 236–253. Springer, 2022. 3

[9] Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor Guilherme Turrisi da Costa, Cees G. M. Snoek, Georgios Tzimiropoulos, and Brais Martínez. Variational prompt tuning improves generalization of vision-language models. *CoRR*, abs/2210.02390, 2022. 1

[10] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11573–11582. IEEE, 2022. 1

[11] Kun Ding, Ying Wang, Pengzhang Liu, Qiang Yu, Haojian Zhang, Shiming Xiang, and Chunhong Pan. Prompt tuning with soft context sharing for vision-language models. *CoRR*, abs/2208.13474, 2022. 1

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 5

[13] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14064–14073. IEEE, 2022. 1

[14] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX*, volume 13669 of *Lecture Notes in Computer Science*, pages 701–717. Springer, 2022. 1

[15] Jonathan Francis, Nariaki Kitamura, Felix Labelle, Xiaopeng Lu, Ingrid Navarro, and Jean Oh. Core challenges in embodied vision-language planning. *CoRR*, abs/2106.13948, 2021. 2, 3

[16] Naiyu Gao, Fei He, Jian Jia, Yanhu Shan, Haoyang Zhang, Xin Zhao, and Kaiqi Huang. Panopticdepth: A unified framework for depth-aware panoptic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 1622–1632. IEEE, 2022. 1, 2

[17] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *CoRR*, abs/2204.14095, 2022. 3

[18] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 4

[19] Chunle Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5802–5810, 2022. 1

[20] Zixian Guo, Bowen Dong, Zhilong Ji, Jinfeng Bai, Yiwen Guo, and Wangmeng Zuo. Texts as images in prompt tuning for multi-label image recognition. *CoRR*, abs/2211.12739, 2022. 1

[21] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. CALIP: zero-shot enhancement of CLIP with parameter-free attention. *CoRR*, abs/2209.14169, 2022. 1

[22] Xuehai He, Diji Yang, Weixi Feng, Tsu-Jui Fu, Arjun R. Akula, Varun Jampani, Pradyumna Narayana, Sugato Basu, William Yang Wang, and Xin Wang. CPL: counterfactual

prompt learning for vision and language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3407–3418. Association for Computational Linguistics, 2022. 1

[23] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339. Association for Computational Linguistics, 2018. 3

[24] Runhui Huang, Yanxin Long, Jianhua Han, Hang Xu, Xiwen Liang, Chunjing Xu, and Xiaodan Liang. NLIP: noise-robust language-image pre-training. *CoRR*, abs/2212.07086, 2022. 3

[25] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *CoRR*, abs/2204.03649, 2022. 1

[26] Jingjing Jiang, Ziyi Liu, and Nanning Zheng. Correlation information bottleneck: Towards adapting pretrained multimodal models for robust visual question answering, 2023. 3

[27] Jonathan Kahana, Niv Cohen, and Yedid Hoshen. Improving zero-shot models with label distribution priors. *CoRR*, abs/2212.00784, 2022. 1

[28] Muhammad Uzair Khattak, Hanoona Abdul Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. *CoRR*, abs/2210.03117, 2022. 1

[29] Alexander Kirillov, Kaiming He, Ross B. Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. *CoRR*, abs/1801.00868, 2018. 2

[30] Weicheng Kuo, Yin Cui, Xiuye Gu, A. J. Piergiovanni, and Anelia Angelova. F-VLM: open-vocabulary object detection upon frozen vision and language models. *CoRR*, abs/2209.15639, 2022. 1

[31] Sohyun Lee, Taeyoung Son, and Suha Kwak. FIFO: learning fog-invariant features for foggy scene segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18889–18899. IEEE, 2022. 1

[32] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 17431–17441. IEEE, 2022. 2, 8

[33] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 1

[34] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Yong Jae Lee, Houdong Hu, Zicheng Liu, and Jianfeng Gao. ELEVATER: A benchmark and toolkit for evaluating language-augmented visual models. *CoRR*, abs/2204.08790, 2022. 1

[35] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 2022. 6, 7, 8

[36] Junnan Li, Silvio Savarese, and Steven C. H. Hoi. Masked unsupervised self-training for zero-shot image classification. *CoRR*, abs/2206.02967, 2022. 1

[37] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16420–16429, June 2022. 2, 3

[38] Yi Li, Yi Chang, Yan Gao, Changfeng Yu, and Luxin Yan. Physically disentangled intra- and inter-domain adaptation for varicolored haze removal. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5831–5840. IEEE, 2022. 1

[39] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3

[40] Dingkang Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. In *European Conference on Computer Vision*, pages 38–54. Springer, 2022. 7, 8

[41] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted CLIP. *CoRR*, abs/2210.04150, 2022. 1, 4

[42] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. *CoRR*, abs/2211.14843, 2022. 1

[43] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. CLIP is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. *CoRR*, abs/2212.09506, 2022. 1

[44] Yanxin Long, Jianhua Han, Runhui Huang, Xu Hang, Yi Zhu, Chunjing Xu, and Xiaodan Liang. P^3ovd: Fine-grained visual-text prompt-driven self-training for open-vocabulary object detection. *CoRR*, abs/2211.00849, 2022. 1, 3

[45] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In

IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 5196–5205. IEEE, 2022. 1

[46] Timo Lüddecke and Alexander S. Ecker. Image segmentation using text and image prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7076–7086. IEEE, 2022. 1

[47] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. *CoRR*, abs/2211.14813, 2022. 3

[48] Chengcheng Ma, Yang Liu, Jiankang Deng, Lingxi Xie, Weiming Dong, and Changsheng Xu. Understanding and mitigating overfitting in prompt tuning for vision-language models. *CoRR*, abs/2211.02219, 2022. 1, 3

[49] Chaofan Ma, Yuhuan Yang, Yan-Feng Wang, Ya Zhang, and Weidi Xie. Open-vocabulary semantic segmentation with frozen vision-language models. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, page 45. BMVA Press, 2022. 1, 3

[50] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5627–5636, 2022. 1

[51] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14054–14063. IEEE, 2022. 1

[52] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Rareact: A video dataset of unusual interactions. *CoRR*, abs/2008.01018, 2020. 1

[53] Matthias Minderer, Alexey A. Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. *CoRR*, abs/2205.06230, 2022. 1

[54] Yongguang Mo, Jianjun Huang, and Gongbin Qian. Deep learning approach to UAV detection and classification by using compressively sensed RF signal. *Sensors*, 22(8):3072, 2022. 1

[55] OpenAI. Gpt-4 technical report, 2023. 5

[56] Omiros Pantazis, Gabriel J. Brostow, Kate E. Jones, and Oisin Mac Aodha. Svl-adapter: Self-supervised adapter for vision-language pretrained models. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, page 580. BMVA Press, 2022. 1

[57] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. 8

[58] Fang Peng, Xiaoshan Yang, and Changsheng Xu. Sgva-clip: Semantic-guided visual adapting of vision-language models for few-shot image classification. *CoRR*, abs/2211.16191, 2022. 1

[59] Ruijie Quan, Xin Yu, Yuanzhi Liang, and Yi Yang. Removing raindrops and rain streaks in one go. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9147–9156. Computer Vision Foundation / IEEE, 2021. 1

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[61] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 1, 2, 4

[62] Hanoona Abdul Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman H. Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *CoRR*, abs/2207.03482, 2022. 1

[63] Hidetomo Sakaino. Panopticroad: Integrated panoptic road segmentation under adversarial conditions. *in CVPR Workshop*, 2023. 1, 2, 3, 4, 5

[64] Hidetomo Sakaino. Panopticvis: Integrated panoptic segmentation for visibility estimation at twilight and night. *in CVPR Workshop*, 2023. 1, 2, 3, 4, 7

[65] H. Sakaino. Physicscap: Dynamic captions for natural scene changes. In *ACM International Conf. Machine Learning (ICML), Workshop on Data-centric Machine Learning Research (DMLR)*, 2023. Nonarchival. 3

[66] H. Sakaino. Refined and enriched physics-based captions for unseen dynamic changes. In *ACM International Conf. Machine Learning (ICML), Workshop on the 2nd New Frontiers In Adversarial Machine Learning (ADVML FRONTIERS)*, 2023. Nonarchival. 3

[67] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *Int. J. Comput. Vis.*, 126(9):973–992, 2018. 2

[68] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, volume 11217 of *Lecture Notes in Computer Science*, pages 707–724. Springer, 2018. 2

[69] Aditya Sanghi, Hang Chu, Joseph G. Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18603–18613, June 2022. 2, 3

[70] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2556–2565. Association for Computational Linguistics, 2018. 8

[71] Sheng Shen, Shijia Yang, Tianjun Zhang, Bohan Zhai, Joseph E. Gonzalez, Kurt Keutzer, and Trevor Darrell. Multitask vision-language prompt tuning. *CoRR*, abs/2211.11720, 2022. 1

[72] Hengcan Shi, Munawar Hayat, Yicheng Wu, and Jianfei Cai. Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9611–9620, June 2022. 2, 3

[73] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. *CoRR*, abs/2206.07045, 2022. 1, 3

[74] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *CoRR*, abs/2209.07511, 2022. 1

[75] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *CoRR*, abs/2206.09541, 2022. 1

[76] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. *CoRR*, abs/2211.16198, 2022. 1

[77] Tao Wang and Nan Li. Learning to detect and segment for open vocabulary object detection. *CoRR*, abs/2212.12130, 2022. 1

[78] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7949–7961. IEEE, 2022. 1

[79] Bichen Wu, Ruizhe Cheng, Peizhao Zhang, Tianren Gao, Joseph E. Gonzalez, and Peter Vajda. Data efficient language-supervised zero-shot recognition with optimal transport distillation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3

[80] Chenshen Wu and Joost van de Weijer. Density map distillation for incremental object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2505–2514, June 2023. 7

[81] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671, 2023. 3

[82] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. CLIMS: cross language image matching for weakly supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 4473–4482. IEEE, 2022. 1

[83] Yinghui Xing, Qirui Wu, De Cheng, Shizhou Zhang, Guoqiang Liang, and Yanning Zhang. Class-aware visual prompt tuning for vision-language pre-trained model. *arXiv preprint arXiv:2208.08340*, 2022. 1

[84] Jingyi Xu, Hieu Le, Vu Nguyen, Viresh Ranjan, and Dimitris Samaras. Zero-shot object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15548–15557, June 2023. 7

[85] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas M. Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18113–18123. IEEE, 2022. 3

[86] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXIX*, volume 13689 of *Lecture Notes in Computer Science*, pages 736–753. Springer, 2022. 1

[87] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19313–19322, June 2022. 3

[88] Nir Zabari and Yedid Hoshen. Semantic segmentation in-the-wild without seeing any segmentation examples. *CoRR*, abs/2112.03185, 2021. 1

[89] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary DETR with conditional matching. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX*, volume 13669 of *Lecture Notes in Computer Science*, pages 106–122. Springer, 2022. 1

[90] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *CoRR*, abs/2210.07225, 2022. 1

[91] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *CoRR*, abs/2111.03930, 2021. 1

[92] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8552–8562, June 2022. 3

[93] Renrui Zhang, Longtian Qiu, Wei Zhang, and Ziyao Zeng. VT-CLIP: enhancing vision-language models with visual-guided texts. *CoRR*, abs/2112.02399, 2021. 1

[94] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, B. G. Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N. Metaxas. Exploiting unlabeled data with vision and language models for object detection. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX*, volume 13669 of *Lecture Notes in Computer Science*, pages 159–175. Springer, 2022. 3

[95] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, volume 13688 of *Lecture Notes in Computer Science*, pages 696–712. Springer, 2022. 1, 3

[96] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, volume 13688 of *Lecture Notes in Computer Science*, pages 696–712. Springer, 2022. 1

[97] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16795–16804. IEEE, 2022. 1, 3

[98] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022. 1, 3

[99] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX*, volume 13669 of *Lecture Notes in Computer Science*, pages 350–368. Springer, 2022. 1

[100] Ziqin Zhou, Bowen Zhang, Yinjie Lei, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting CLIP for zero-shot semantic segmentation. *CoRR*, abs/2212.03588, 2022. 3

[101] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *CoRR*, abs/2205.14865, 2022. 1

[102] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyao Zeng, Shanghang Zhang, and Peng Gao. Pointclip V2: adapting CLIP for powerful 3d open-world learning. *CoRR*, abs/2211.11682, 2022. 2