Refined and Enriched Captions With Physical Scale For Dynamic Disaster Scene

Anonymous CVPR 2023 submission

Paper ID 2

Abstract

Vision-Language models (VLMs), i.e., image-text pairs of CLIP, have boosted image-based Deep Learning (DL). Unseen images by transferring semantic knowledge from seen classes can be dealt with the help of language models pre-trained only with texts. Two-dimensional spatial relationships and a higher semantic level have been performed. Moreover, Visual-Question-Answer (VQA) tools and open-vocabulary semantic segmentation provide us with more detailed scene descriptions, i.e., qualitative texts, in captions. However, the capability of VLMs presents still far from that of human perception. This paper proposes PanopticCAP for refined and enriched qualitative and quantitative captions to make them closer to what human recognizes by combining multiple DLs and VLMs. In particular, captions with physical scales and objects' surface properties are integrated by water level, counting, depth map, visibility distance, and road conditions. Fine-tuned VLM models are also used. An iteratively refined caption model with a new physics-based contrastive loss function is used. Experimental results using images with adversarial weather conditions, i.e., rain, snow, fog, landslide, flooding, and traffic events, i.e., accidents, outperform state-of-the-art DLs and VLMs. A higher semantic level in captions for real-world scene descriptions are shown.

1. Introduction

Segmentation has become an important task for real-world applications by Deep Learning (DL) [4, 5, 8–12, 15, 18–20, 26, 28, 33, 37, 39, 40, 50, 51, 56–58, 60, 61, 69–72, 74, 75, 79–81, 88, 93–95, 98, 106, 112, 115, 116, 120, 124, 126, 132–134, 152]. Segmentation has become diversified into semantic and instance segmentation. Although many variants of segmentation models are presented, issues with limits of training image datasets and robustness to illumination and noise remain unsolved. Only pretraining a finite number of images could not have enhanced segmentation accuracy in real-world scenes. Dynamic changes have been dealt with by rain drops [98, 134], defog, and dehaze [40, 61, 71, 74, 89, 132]; however, these methods fail to deal with heavy fog and snowfall events. Moreover, unpredicted disaster and traffic accident scenes require more pretraining image datasets; however, in spite of vital events, they are hard to collect sufficient images and videos due to rare chances. Therefore, segmentation to such conditions and events becomes degraded.

Recently, CV, DL, and NLP have been combined, i.e., Vision Language Model (VLM). It is known that unseen images that have not been pretrained have been recognized much better than only CV or DL models [3, 6, 14, 23, 25, 34, 41–43, 48, 54, 55, 64, 67, 83, 86, 91, 96, 97, 100, 108, 111, 114, 117, 121, 129, 139, 140, 142, 144, 146, 147, 151]. 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 [64, 91, 100] and fine-tuned on smaller datasets for specific tasks, allowing for efficient transfer learning [3, 6, 14, 23, 25, 34, 41–43, 48, 54, 55, 67, 83, 86, 96, 97, 108, 111, 114, 117, 121, 129, 139, 140, 142, 144, 146, 147, 151].

This can be useful in various applications, such as object detection [29, 30, 32, 38, 49, 59, 77, 82, 90, 92, 102, 119, 128, 138, 143, 149], segmentation [24, 36, 63, 76, 78, 84, 87, 110, 127, 131, 137, 145, 150], and classification. VLMs can save time and resources in various applications and improve semantic understanding by recognizing relationships between objects and concepts and developing a comprehensive understanding of visual content.

Image captioning is an important and challenging task in computer vision that involves generating natural language descriptions of complex visual scenes that include objects and their surrounding context. However, single VML is often weak for dynamic changes, i.e., disaster scenes [113]. Moreover, heavy rainfall and snowfall have been increasing, which may cause a chain reaction of natural disasters observed from the satellite images, i.e., landslides and flooding [17, 44, 125].

However, camera image-based post-disaster object recognition for dirt, water, and rocks remains unsolved on the road. Since domain adaptation segmentation DL models [46, 118] require manual selection of the optimal pre-

trained model, they are not useful for dynamic changes. However, few papers effectively integrate physical scale into the VLM model and combine multi-VLMs.

For images to be best captioned, they need to depict information most similar to Human perception. Human perception can simultaneously process different visual cues, i.e., texture, shape, and depth, to identify and label objects at varying distances. Therefore, a method for refining and enriching captions with different visual cues is needed.

To this end, this paper proposes PanopticCAP: a panoptic vision-language model under adversarial visual conditions using single images. This paper proposes refined and enriched captions for scene descriptions under adversarial conditions by the proposed PanopticCAP with multiple task-oriented DLs and VLMs.

PanopticCAP consists of eleven modules, i.e., Deep Visual Language Classification (Dvlc), Deep Visual Language Segmentation (Dvls), Visual-Query-Answer (VQA), Contrastive Language Physical scale Pretraining (CLPP), Deep Visibility estimation (Dvis), Deep Road conditions (Droad), Deep Depth (Ddepth), Deep anomaly (Danomal), Deep water-level (Dwater), Deep snowfall (Dsnow), and Deep Scene (DScene). The branched architecture allows us to efficiently maintain and upgrade each of the eleven modules. Contributions of this paper are fourfold:

1. Multiple vision language and Transformer-based Deep Learning (DL) models with branched structures for efficiency in light of memory, training, and maintenance. Danomal excludes difficult images, i.e., lenz reflection, to stabilize the overall system. Due to enormous datasets of VLMs, Dvls, and Dvlc are fine-tuned VLMs from SOTA models for segmentation and classification, respectively.

2. It is the first time to contain dynamic changes with physical scales, i.e., depth by Ddepth, fog visibility distance by Dvis, weather conditions by Dsnow, water level by Dwater, and road conditions by Droad. Unseen images like adversarial weather and disaster conditions can be dealt with. Moreover, more detailed scene descriptions of traffic accident events are shown. Captions with 3D-related adverbs, i.e., behind, rear, in front of, and far, enable to generate as SOTAs have used 2D-related adverbs, i.e., left and right.

3. More refined and enriched captions are generated based on fixed queries at VQA, CLPP, and the above multiple modules. A new contrastive loss function is proposed in CLPP to refine and enrich captions with the object's physical scale, i.e., size and position, under an iterative refinement process. API tools, i.e., Visual ChatGPT [123], may be hard to generate dynamic scene changes with physical scales as this paper presents.

4. Many experimental results show the superiority of the proposed PanopticCAP over SOTA DLs and VLMs. The proposed PanopticCAP 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 Computer Vision (CV), 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 [2] and Droad [1].

An all-in-one image restoration network for unknown corruption has been proposed [62]; however, this method can be degraded heavy fog and snowfall, as shown in Dvis [2] and Droad [1]. Dehazing in [33] is limited to closer views of daytime lighter foggy scenes, i.e., indoor and garden, unlike our proposed method [2] for distant scenes with heavy fog at night, i.e., highway. A unified framework for depth-aware panoptic segmentation has been reported [57] under clear weather conditions.

Although Cityscapes with 3000 images [21], Foggy Cityscape DBF with 500 synthetic foggy images [103], and Foggy Zurich with 3800 real light foggy images [104] are publicly available, they are almost all daytime and lighter fog data. Moreover, image datasets for road conditions have not been built, unlike Droad [1].

In recent years, the Vision Language model (VLM) field has undergone significant progress [141, 147, 148]. 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 [22, 73, 99], generative objectives [47, 122][108, 109], and alignment objectives [35, 52, 85, 101, 130, 136]. VLMs are transferred by Text-Prompt Tuning [3, 6, 23, 42, 43, 55, 83, 86, 108, 111, 129, 139, 142, 146, 147, 151].

Besides finetuning, knowledge distillation is a method to improve VMLs for downstream tasks, including object detection [29, 30, 32, 38, 49, 59, 77, 82, 90, 92, 102, 119, 128, 138, 143, 149] and semantic segmentation [24, 36, 63, 76, 78, 84, 87, 110, 127, 131, 137, 145, 150].

Unseen images that have not been pre-trained have become recognized by VLM frameworks [13, 31]. More diverse and out-of-distribution data for pre-training and evaluation are used [45]. Prompt learning to adapt VLMs to new tasks without fine-tuning is also shown [53]. Contents of captions have been enhanced for better descriptions of real-world objects [31].

Geometric reasoning or depth estimation to infer 3-D information from 2-D images [135, 141] 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 introduced [13, 31, 65, 68, 100, 105, 109, 153]. 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 [123] is that it can produce acceptable results on the general scene and unseen classes. However, since Visual ChatGPT [123] is trained on the limited data of the year 2021, it generates captions under older datasets. So far, Visual ChatGPT [123] is weak at generating dynamic scene descriptions like weather and road conditions. Moreover, the physical size of objects, fog visibility distance, and water level are contained.

Therefore, as aforementioned above, no SOTA VLM papers and API tools have challenged images with the physical scale in natural phenomena, i.e., fog visibility in 3D depth and water level. More refined and enriched captions will be provided by the proposed PanopticCAP in the following sections.

3. Proposed Method

This section describes the proposed PanopticCAP 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 and vision-language modules.

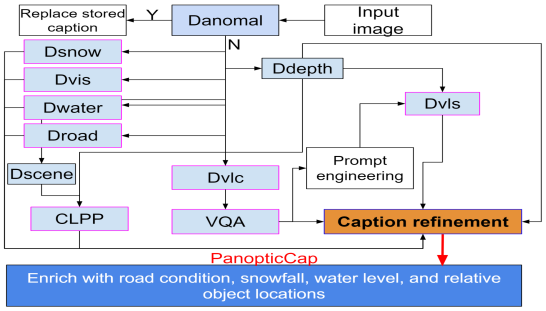


Figure 1: Overview of the proposed PanopticCAP model.

Figure 1 shows an overview of the proposed PanopticCAP. Since this paper deals with many challenging scenes with disasters and car accidents, adversarial conditions are taken into account. Further detailed explanations of the multiple modules will be given in Sub-sections 3.1 to 3.4.

3.1. Proposed Danomal

To identify and reject adversarial images, as shown in Figure 2, an algorithm to reject such images is proposed to avoid the degradation of the cascaded other recognition modules. Three major adversarial image patterns have been selected: a) lens reflection, b) strong headlight, and c) raindrops. These adversarial images were collected from over 2500 images and used to train by Swin Transformer [37] into 3 classes.

(a) (b) (c)

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

3.2. Proposed Dvlc and Dvls

Dvlc is a vision-language model 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, which is a significant accomplishment in Computer Vision.

Dvlc utilizes the input texts of five distinct disaster categories: *car crashes, flooding, fog, landslide, and rain*. Tailored textual 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 proposed to obtain semantic segmentation of these scenes. Dvls is finetuned from OvSeg [76] by adding a new physical constraint to the loss function. To obtain descriptions of disasters for the Dvlc, a classification task is performed using keywords corresponding to each disaster scene. These texts are used to generate text descriptions of the disasters that are fixed for each type of scene.

Therefore, since Dvlc and Dvls recognize texts and segmentation from a single image, this paper proposes to combine respective outputs.

3.3. Proposed CLPP

The proposed Contrastive Language Physical-Scale Pretraining (CLPP) is a VLM with inputs from a depth map, object location from image and text description pairs, and a modified contrastive loss function. Unlike SOTA VLMs with no physical models in contrastive loss functions, this paper proposes CLPP with additional physical constraints, as shown in Figure 3. The original contrastive loss function of CLIP [100] is defined by

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

where 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 is added. 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 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.

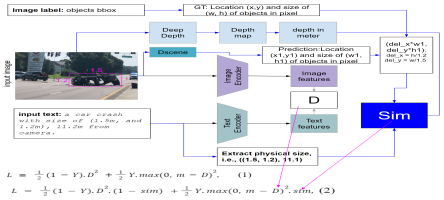


Figure 3: Proposed contrastive language for pre-training in physical scale.

3.4. Proposed Droad, Dsnow, and Dvis

This section discusses the proposed Droad, Dsnow, and Dvis as explained in Section 3. Unlike SOTA papers in DLs and VLMs, this paper aims to generate dynamic scene changes with the weather conditions, i.e., rain, snow, and fog, and road conditions, i.e., dry, wet, and snow. Dvis [2] and Droad [1] are applied for further detailed classes of segmented objects. Dscene [1, 2]is also applied to ensure snow conditions.

In Dvis [2], Droad [1], and Dscene [1, 2], Swinformer [126] is trained from over 7500 winter road images. It is noted that since publicly available annotation datasets are insufficient, various weather and road scenes from different countries under adversarial conditions have been collected and used to train.

3.5. Dwater

Dwater is a simple combination of Dscene and a transformer-based classifier, specifically the ViT [27] classifier. Dwater estimates the water level based on reference objects, such as cars, buses, humans, trees, poles, and traffic signs, through two steps: In Step 1, objects are extracted from the image using Dscene. In Step 2, the extracted objects are classified to their respective water levels, which are pre-defined for each object. Table 1 shows the physical height of the reference object for water levels.

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

4. Experiments and Discussion

4.1. Danomal for adversarial conditions

This section evaluates the performance of the proposed Danomal. The dataset comprises over 2500 images with 1 normal condition and 3 different adversarial conditions: lens reflection, strong light, and raindrops. In a comparative study, Droad, Dsnow, Dvis, and Dvlc are applied with and without Danomal. Evaluation is conducted using three classes of road condition, three classes of snowfall, four classes of visibility, and five classes of scene types.

Table 2 shows the results that Danomal can effectively reject images with adversarial conditions, where the accuracy using Danomal becomes better for each DL model without Danomal. No thresholding setting is required. Therefore, the proposed Danomal has been proven useful in rejecting such three adversarial factors in road images.

Table 2: Accuracy comparison with and without DAnomal under adversarial conditions for Droad, Dsnow, Dvis, and Dvlc.

	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 CLPP and Dvlc

To evaluate the performance of scene analysis of the proposed system, this section demonstrates the qualitative and quantitative accuracies of CLPP and Dvlc. The test dataset consists of six selected categories: a car crash in snow conditions, flooding with rain, low visibility with fog, landslide, wet road with rain, and traffic flow. Figure 4 (a)-(f) shows such images. In the test dataset, ground truth captions have been manually annotated and contain the following key criteria: the number of vehicles, scene category, and road conditions. Totally, 3500 images are collected. The experiments have been conducted to evaluate the performance of scene understanding using VLMs on dynamic scenes as well as the level of detail of the description based on the generated captions. BLIP [66] model has been chosen to

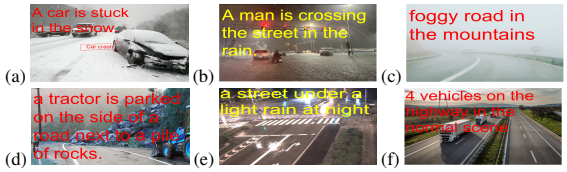
Table 3: Comparison of captions among ground truth, proposed CLPP, and BLIP [66]

	Ground truth	CLPP	BLIP [66]
(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

compare since the achievement of high capability of transfer flexibly for vision language understanding and captioning tasks.

Table 3 shows the captions containing ground truth and two captions generated by the proposed CLPP ($m = 4$ in eq. 2) and BLIP [66] on six images in Figure 4. CLPP and BLIP [66] can recognize the scene's context and key objects, i.e., traffic conditions and several objects. However, BLIP [66] is insufficient to capture detailed scene features compared with CLPP. In (a), (b), and (f), the counting number of vehicles by CLPP has been improved from one to many. In (b), no road conditions by BLIP [66] have been recognized, but CLPP recognizes the flooding highway. In (e) and (f), CLPP recognizes rain and sunny, respectively.

However, BLIP [66] shows no weather conditions. In (c), it is shown that the degree of fog is recognized as heavy fog by CLPP with time. In (d), CLPP recognizes that the road is occluded by the rock, but BLIP [66] shows a pile of rocks. Therefore, the proposed CLPP has demonstrated more refined and enriched captions over BLIP [66].



(a) (b) (c) (d) (e) (f)

Figure 4: Results of captions in images: (a) Car crash. (b) Flooding. (c) Fog. (d) Landslide. (e) Rain. (f) Traffic flow.

Table 3 shows the captions containing Ground truth and two captions generated by CLPP and BLIP [66] on five images in Figure 4. CLPP and BLIP [66] are able to recognize and count on things of the scenes; however, the captions generated by CLPP are semantically closer to the class of ground truths. Moreover, CLPP shows a higher level of detail in the caption since the reflection of the condition of the road surface is due to the effects of weather and disaster. In order to evaluate the classification capability of CLPP for dynamic scenes quantitatively, the similarity between the captions generated by CLPP and the annotated class of the images is evaluated based on the adjective or noun sets, i.e., frozen, flooded, damaged, and light as well as synonyms words occurring in the caption.

Table 4 demonstrates a comparison of the classification

Table 4: Accuracy comparison among ViT, Resnet, Vgg18, and Dvlc. The bold font indicates the best score.

Classes/model	Dvlc (%)	ViT (%)	Resnet101(%)	Vgg19(%)
Car crashes	**93.37**	92.41	91.12	87.67
Flooding	**90.69**	89.23	87.83	86.54
Fog	**92.98**	91.19	86.77	85.23
Landslide	**89.52**	87.63	87.19	84.89
Rain	**92.33**	87.58	88.92	83.11
Average	***91.78***	*89.61*	*88.37*	*85.49*

results among the proposed Dvlc, CLPP, and SOTA classifiers, i.e., ViT [27], VGG19, and Resnet101. Dvlc has classified best in the accuracy of 91.78% accuracy without retraining. Therefore, the effectiveness of Dvlc as prompt engineering facilitating to Dvls has been proven.

4.3. Refined Semantic Segmentation by Prompt Engineering

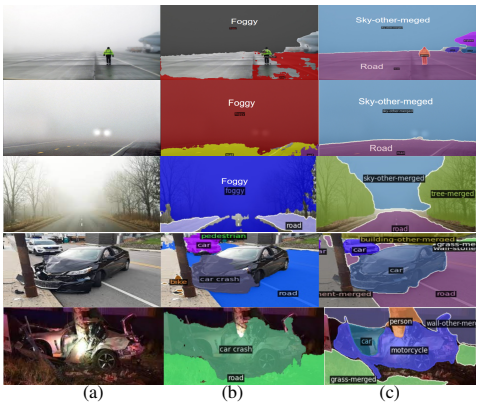
This section denotes the proposed Dvls and how to obtain the final refined captions using prompt engineering. The prompt for each scene is pre-defined as a list of words, i.e., (1) car crashes: ["pedestrian", "car", "car crash", "road", "bike", "tree"]; (2) flooding: ["water", "car", "person", "tree", "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 5 illustrates the effectiveness of our approach on images with 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 [19], 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 is helpful for detailing segmentation results under dynamic conditions.

4.4. Image Captioning by CLPP

This section challenges image captioning with obstacles on the road using the proposed Contrastive Language Physical-Scale Pretraining (CLPP) ($m = 4$ in eq. 2) in a 3D manner, unlike GraphVQA [1] with the 2D graph of on, left, or right. Figure 6 (a) shows post-disaster images suffered from the enormous typhoon, where many obstacles like dirt and rocks piled up on the road and other regions. In order to recognize whether obstacles are present on the road or not, the occluded road surfaces must first be identified. For this, DeepDepth (b) with vertical (Oy), horizontal (Ox), and depth (Oz) coordinates are used to recognize nearly flat



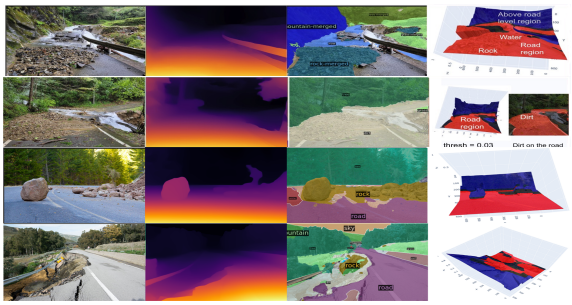
(a) (b) (c)

Figure 5: Results of segmentation by SOTA and proposed PanopticCAP: (a) Original image. (b) Proposed refined semantic segmentation. (c) Mask2Former [19]

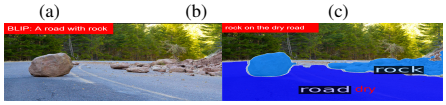
road surfaces that are assumed to be the normal road (xOz) below the obstacles.

On the other hand, Dscene (c) can provide segmented objects. Depth maps (b) and objects from (c) are fed into the image encoder of a VLM to make contrastive with description texts, respectively. The presentation of the depth map in 3D (d) shows the relative location of objects. The location of objects like tree and mountain classes are included in the caption for the image when encoding depth map with objects mask respectively with relative location by adverbs in texts, i.e., "on", "before", "above", "bottom", and "side" for 2D and adverbs for 3D are "behind," "in front of", and "rear".

The refined caption and segmentation (e) show that the proposed method can enrich the caption by adding detailed locations of objects and road conditions. SOTA single VLMs, i.e., BLIP [66], could not estimate the depth and road condition. It has been proven that CLPP can estimate physical factors on images, and CLPP caption result is more detailed than single VLM.



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



(e)

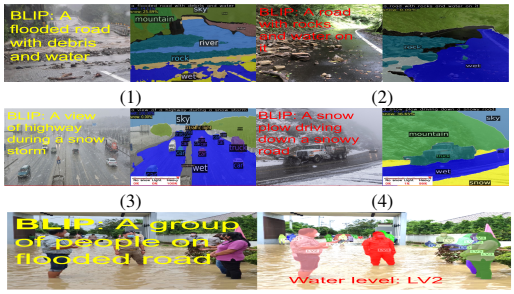
Figure 6: Proposed CLPP applied to post-disaster scenes to identify road regions with various obstacles (dirt, water, rock): (a) Input image. (b) Depth map. (c) Panoptic segmentation. (d) Refined road surface. (e) The left image is the input image with the BLIP caption result. The right image is the refined caption: "Rock on the dry road".

Table 5: Comparison of the refined captions with BLIP caption results.

	Proposed method	BLIP [66]
(1)	Rocks lay on the flooding road, within 938m in visibility	A flooded road in the rain
(2)	Rock debris lay on the wet road, within clear 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, 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
(6)	The highway under light snowfall with the snow on the side of the road, within 748m in visibility	Snow-covered road with a fence and a street light

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

To provide further complicated captions, this section conducts experiments on various weather conditions with traffic and disaster scenes. The proposed Dsnow, Droad, and Dvis are used by comparing a SOTA VL captioning model, BLIP [66]. Figure 7 shows six scenes. As a result, road conditions by Droad (1)-(6) are wet in blue and snow in yellow. Dsnow's indicators (3)-(6) present light to heavy snowfall. Dvlc recognizes overall scene objects like mountains, rivers, rocks, sky, and trees. Proposed Dvis [2] from images (1)-(6) can estimate a physical scale from weather phenomenon, i.e., foggy visibility distances in meters: 938, clear, 637, 512, 812, and 748 m, respectively. Therefore, the road condition and visibility distance have been included in the captions of Dvlc. Table 4 shows a comparison of the refined captions and a SOTA BLIP [66] result using six scenes of Figure 7. The comparison results show that a refined caption is detailed about the scene by adding road conditions, snowfall status, location of objects, and exact visibility in meters. Besides, the caption from BLIP lacks a description. The result has proven that the proposed method integrating Droad, Dsnow, and Dvis outperforms single VML, i.e., BLIP [66].



(1) (2) (3) (4) (5)

Figure 7: 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 water level, Lv2. (6) Light snowfall on the highway.

5. Ablation study

To justify the proposed PanopticCAP, many additional experiments are ablated below.

5.1. Caption Refinement by Dvls

To show the usefulness of refined Dvls, many unseen disaster scenes that have not been pre-trained are used to segment with classes. As shown in Figure 8 (a), images present disaster events. Two SOTAs of (c) MaskDINO [7]. (d) OVSeg [76] are compared. As a result, Table 6 summarizes classes of (b) proposed Dvls and (c), (d) two SOTAs. In (1), a track (c) or boat (d) has been annotated, whereas the proposed Dvls has refined to "car crash" over water (b). In (2)-(5), snow to water, landslide to rocks, pavement to rain, and tree to strong wind have been annotated by (b) the proposed Dvls, respectively.

Therefore, refined texts from SOTAs' texts could enhance original to higher semantic texts. In particular, (5) tree (c) is normal segmentation, but strong wind (b), (d) stands for intuitive weather conditions as humans may announce. When depth maps are added, combined with location prompts, Dvls can label segmented objects more semantically. Therefore, it has been proven that the proposed Dvls with texts will play an important role in messaging heavy disaster events more clearly than SOTAs' texts.

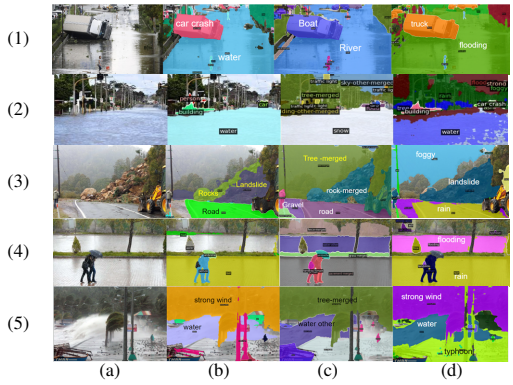


Figure 8: Comparison of the proposed method, MaskDINO [7], and OVSeg [76] (a) Input image. (b) Proposed Dvls. (c) MaskDINO [7] (d) OVSeg [76]

Table 6: Comparison of classes by SOTAs and proposed Dvls.

Image	SOTA	Proposed
(1)	boat, truck	car crash
(2)	snow, rain	water
(3)	rock-merged, rain	landslide
(4)	pavement-merged, rain	rain
(5)	tree-merged, typhoon	strong wind

5.2. Caption Enrichment Using Refined Segmentation and Query Loop

This section describes a combination of refined segmentation from the proposed Dvls with a query loop from VQA to further refine the caption for an image. A loop is necessary because the resulting segmentation of Dvls depends on the newly added class. By selecting the most appropriate word, the best segmentation result can be achieved during the looping. Figure 9 depicts a looping scheme between VQA and Dvls. The list of objects/events consists of words with similar meanings that describe an object/event. The process of selecting the best answer involves evaluating the cosine similarity between the answer and the image. The answer with the highest cosine similarity score is considered the best. To ensure the most accurate selection, all words with similar meanings are looped through until the best answer has been chosen. The effectiveness of these integrated models is compared to an online image captioning API, i.e., visual ChatGPT. The queries are designed to correspond to different types of events, such as "car crashes," "flooding," "fog," "landslide," "rain," and "snowfall." Predefined prompts for these scenes are used, including cars, water, snow, rocks, and debris flow. Figure 10 presents a

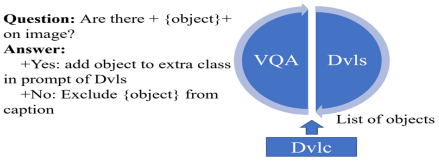


Figure 9: Query loop and segmentation combination.

comparison between two approaches: (a) visual ChatGPT [123] caption results and (b) finally refined segmentation using a query loop with consideration of the relative size and location of objects. The comparison reveals that visual ChatGPT [123] only provides an overview of scene descriptions with no physical scales. On the other hand, the proposed model presents more detailed physical scales of sizes and locations for accident events. Thus, the proposed method has proven capable of handling dynamic captions. More refined and enriched captions by the proposed model have been generated for such traffic accident scenes than visual ChatGPT [123].

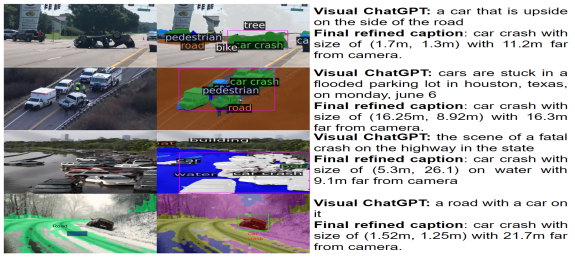
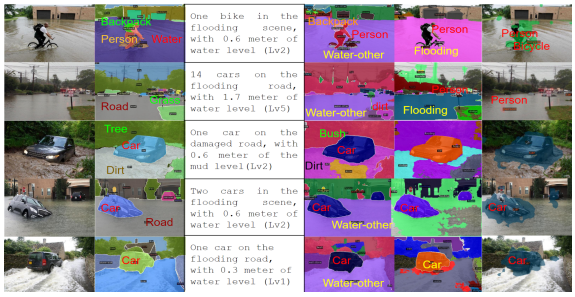


Figure 10: Comparison of vision-language models between visual ChatGPT and the proposed method: (a) Image captioning using visual ChatGPT. (b) Refined segmentation and captions with physical scales by the proposed CaptionCAP. (c) Captions from visual ChatGPT and the proposed CaptionCAP.

5.3. Compare the Proposed PanopticCAP with SOTA VMLs

To justify the performance of the proposed PanopticCAP, more complicated scenes under adversarial weather and disaster conditions are selected to provide captions, as shown in Figure 11 (a). VLM-based SOTAs, i.e., (d) ZegFormer [24], (e) OvSeg [76], and (f) ClipSeg [84], are compared. Dwater is used for water level estimation us-

ing pre-registered known standard objects' sizes, i.e., buildings, cars, poles, and human heights. Results by (f) ClipSeg [84] show the worst segmentation, particularly flooding regions. On the other hand, results by (d) ZegFormer [24] segment are better than results by (e) OvSeg [76]. However, all SOTAs are limited to generating captions with physical scales. Results by (b) the proposed PanopticCAP present segmented objects with bikes, cars, and water. In captions (c) from (b), water levels from 0.3 m to 1.56 m and several cars have been contained. These quantitative captions can describe more details than only qualitative texts, i.e., flooding. Therefore, it has been demonstrated that the proposed PanopticCAP provides dynamic physical scales that are helpful in understanding how cars and bikes are damaged. It is the first time to generate dynamic captions from such disaster scenes. When videos are available, spatio-temporal changes will also be captioned to monitor for rescue purposes.



(a) (b) (c) (d) (e) (f)

Figure 11: Results of SOTA VL-based segmentation: (a) Input image. (b) Proposed PanopticCAP. (c) Refined caption by the proposed PanopticCAP with scene descriptions and physical scale in water level. (d) ZegFormer [24] (e) OvSeg [76] (f) ClipSeg [84]

5.4. CLPP with Different Loss Function Parameters

This section presents an experimental comparison of various parameters for the contrastive loss function (L) used in the proposed CLPP. CLPP is applied in Sections 4.2 and 4.4 with $m = 0.4$ in equations 2 and 3. In this experiment, m is used from array list values $[0.2, 0.3, 0.4, 0.5]$. A comparison of equation (1) and the proposed equations (2) and (3) is carried out. Table 7 shows a comparison of the modified loss function and the original one for the physical scale-generated caption. The result shows that the performance of CLPP is the lowest in RMSE when m is set to 0.4. Therefore, it has been reconfirmed that the selected m is optimal.

Table 7: 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

5.5. Overall Evaluation PanopticCAP

This section presents an experiment that evaluates the final output of all eleven modules. The experiment measures performance using the BLEU score and is conducted on two datasets. The first dataset is publicly available and includes the COCO Caption dataset [16] and the Conceptual Captions dataset [107], 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 the Disaster dataset (1850 image-text pairs) and the Traffic accident dataset (2130 image-text pairs).

Table 8: Performance evaluation of proposed panopticCAP on public image-text datasets

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

Table 9: Performance evaluation of proposed panopticCAP on collected datasets.

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

Table 10: 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[66]	1.432	3214

According to the results in Table 8, PanopticCAP does not perform as well as Visual ChatGPT. This could be due to the fact that the text descriptions in the public image set do not include information about road conditions, water levels, snow conditions, or visibility, whereas PanopticCAP is capable of generating captions with these details. However, Table 9 presents contradictory results, where PanopticCAP outperforms Visual ChatGPT on datasets featuring disaster or traffic accident conditions. It has been proven that PanopticCAP can provide detailed semantics about the physical aspects of scenes. These can be highly useful for tasks such as traffic coordination and rescue operations.

In addition, we compared the computational cost and memory usage of the proposed system with that of SOTA methods on the same hardware device. Table 10 presents a comparison of the computational cost and memory usage for these methods.

6. Conclusion

This paper has proposed PanopticCAP with multiple DL and VLM models, which consist of 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., depth, fog visibility distance, weather conditions, water level, and road conditions. A physics-based loss function generates more refined and enriched captions at a contrastive loss. PanopticCAP will help notify detailed scene descriptions to drivers, auto-driving, and rescue workers from camera images.

References

[1] Anonymous. Panopticroad: Integrated panoptic road segmentation under adversarial conditions. *in CVPR Workshop*, 2023.

[2] Anonymous. Panopticvis: Integrated panoptic segmentation for visibility estimation at twilight and night. *in CVPR Workshop*, 2023.

[3] 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.

[4] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9736–9745. Computer Vision Foundation / IEEE, 2020.

[5] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. YOLACT: real-time instance segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9156–9165. IEEE, 2019.

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

[7] Christoph Busch and Eric Debes. Wavelet transform for analyzing fog visibility. *IEEE Intell. Syst.*, 13(6):66–71, 1998.

[8] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3981–3991. IEEE, 2022.

[9] Jiale Cao, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Sipmask: Spatial information preservation for fast image and video instance segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV*, volume 12359 of *Lecture Notes in Computer Science*, pages 1–18. Springer, 2020.

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020.

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020.

[12] João Carreira, Viorica Patraucean, Laurent Mazaré, Andrew Zisserman, and Simon Osindero. Massively parallel video networks. 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 IV*, volume 11208 of *Lecture Notes in Computer Science*, pages 680–697. Springer, 2018.

[13] 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.

[14] 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.

[15] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. Blendmask: Top-down meets bottom-up for instance segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8570–8578. Computer Vision Foundation / IEEE, 2020.

[16] 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.

[17] Y. H. Chen, P.J. Lee, and T.A. Buil. Multi-scales feature extraction model for water segmentation in the satellite image. *In Proc. IEEE Int. Conf. Consumer Electornics (ICCE)*, 2023.

[18] Bowen Cheng, Maxwell D. Collins, Yukun Zhu, Ting Liu, Thomas S. Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12472–12482. Computer Vision Foundation / IEEE, 2020.

[19] 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.

[20] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 17864–17875, 2021.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Dawei Du, Yuankai Qi, Hongyang Yu, Yi-Fan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. 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 X*, volume 11214 of *Lecture Notes in Computer Science*, pages 375–391. Springer, 2018.

[29] 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.

[30] 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.

[31] 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.

[32] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Open vocabulary object detection with pseudo bounding-box labels. 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, Pro-*

ceedings, Part X, volume 13670 of *Lecture Notes in Computer Science*, pages 266–282. Springer, 2022.

[33] 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.

[34] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *CoRR*, abs/2110.04544, 2021.

[35] 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.

[36] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. 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 540–557. Springer, 2022.

[37] Jiaqi Gu, Hyoukjun Kwon, Dilin Wang, Wei Ye, Meng Li, Yu-Hsin Chen, Liangzhen Lai, Vikas Chandra, and David Z. Pan. Multi-scale high-resolution vision transformer for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 12084–12093. IEEE, 2022.

[38] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[39] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 1777–1786. Computer Vision Foundation / IEEE, 2020.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] Daniel Hernández, José M. Cecilia, Juan-Carlos Cano, and Carlos T. Calafate. Flood detection using real-time image segmentation from unmanned aerial vehicles on edge-computing platform. *Remote. Sens.*, 14(1):223, 2022.

[45] 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.

[46] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. MIC: masked image consistency for context-enhanced domain adaptation. *CoRR*, abs/2212.01322, 2022.

[47] 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.

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

[49] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7010–7021. IEEE, 2022.

[50] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and

Alexei A Efros. Image-to-image translation with conditional adversarial networks. pages 1125–1134, 2017.

[51] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. *CoRR*, abs/2211.06220, 2022.

[52] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text 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 4904–4916. PMLR, 2021.

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

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

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

[56] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9856–9865. Computer Vision Foundation / IEEE, 2020.

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

[58] K. J. Kucharik and J. A. Norman. Effect of light and shade on visibility. *J. the Illuminating Engineering Society*, 13(3):117–125, 04 1984.

[59] 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.

[60] Hyunmin Lee and Jaesik Park. Instance-wise occlusion and depth orders in natural scenes. pages 21178–21189, 2022.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] Feng Li, Hao Zhang, Yi-Fan Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, PengChuan Zhang, and Lei Zhang. Vision-language intelligence: Tasks, representation learning, and large models. *CoRR*, abs/2203.01922, 2022.

[66] 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.

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

[68] 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.

[69] Qin Li, Yi Li, and Bin Xie. Single image-based scene visibility estimation. *IEEE Access*, 7:24430–24439, 2019.

[70] Ruoteng Li, Robby T. Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3172–3182. Computer Vision Foundation / IEEE, 2020.

[71] 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.

[72] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7026–7035. Computer Vision Foundation / IEEE, 2019.

[73] 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.

[74] Yu Li, Shaodi You, Michael S. Brown, and Robby T. Tan. Haze visibility enhancement: A survey and quantitative benchmarking. *Computer Vision and Image Understanding*, 165:1–16, 2017.

[75] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 214–223. Computer Vision Foundation / IEEE, 2021.

[76] 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.

[77] 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.

[78] 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.

[79] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021.

[80] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021.

[81] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3431–3440. IEEE Computer Society, 2015.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] 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.

[89] Xianzheng Ma, Zhixiang Wang, Yacheng Zhan, Yinqiang Zheng, Zheng Wang, Dengxin Dai, and Chia-Wen Lin. Both style and fog matter: Cumulative domain adaptation for semantic foggy scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18900–18909. IEEE, 2022.

[90] 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.

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

[92] 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.

[93] 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.

[94] Mihai Negru and Sergiu Nedevschi. Image based fog detection and visibility estimation for driving assistance systems. In *2013 IEEE 9th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 163–168, 2013.

[95] Lucas Prado Osco, José Marcato Junior, Ana Paula Marques Ramos, Lúcio André de Castro Jorge, Sarah Narges Fatholahi, Jonathan de Andrade Silva, Edson Takashi Matsubara, Hemerson Pistori, Wesley Nunes Gonçalves, and Jonathan Li. A review on deep learning in UAV remote sensing. *Int. J. Appl. Earth Obs. Geoinformation*, 102:102456, 2021.

[96] 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.

[97] 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.

[98] 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.

[99] 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.

[100] 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.

[101] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in vision-language models. *CoRR*, abs/2210.09996, 2022.

[102] 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.

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

[104] 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.

[105] 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.

[106] Aashish Sharma, Loong-Fah Cheong, Lionel Heng, and Robby T. Tan. Nighttime stereo depth estimation using joint translation-stereo learning: Light effects and uninformative regions. In *2020 International Conference on 3D Vision (3DV)*, pages 23–31, 2020.

[107] 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.

[108] 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.

[109] 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.

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

[111] 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.

[112] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[113] S. Sreelakshmi and S.S. V. Chandra. Machine learning for disaster management: insights from past research and future implications. *In Proc. IEEE Int. Conf. Comput. Communication, Security, and Intelligent Systems (IC3SIS)*, 2022.

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

[115] A. Tafreshi and M. Shahraeeni. Effect of atmospheric haze on visibility distance. *J. Environmental Science and Health*, 44(1):44–48, 04 2009.

[116] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019.

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

[118] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8495–8505. IEEE, 2021.

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

[120] Sanghyun Woo, Dahun Kim, Joon-Young Lee, and In So Kweon. Learning to associate every segment for video panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2705–2714. Computer Vision Foundation / IEEE, 2021.

[121] 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.

[122] 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.

[123] 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.

[124] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. *CoRR*, abs/2104.10834, 2021.

[125] Deqiang Xiang, Xin Zhang, Wei Wu, and Hongbin Liu. Denseppmunet-a: A robust deep learning network for segmenting water bodies from aerial images. *IEEE Trans. Geosci. Remote. Sens.*, 61:1–11, 2023.

[126] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12077–12090, 2021.

[127] 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.

[128] Johnathan Xie and Shuai Zheng. ZSD-YOLO: zero-shot YOLO detection using vision-language knowledgedistillation. *CoRR*, abs/2109.12066, 2021.

[129] 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.

[130] 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.

[131] 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.

[132] Wending Yan, Aashish Sharma, and Robby T. Tan. Optical flow in dense foggy scenes using semi-supervised learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13256–13265, 2020.

[133] Li Yang, Radu Muresan, Arafat Al-Dweik, and Leontios J. Hadjileontiadis. Image-based visibility estimation algorithm for intelligent transportation systems. *IEEE Access*, 6:76728–76740, 2018.

[134] Wenhan Yang, Robby T. Tan, Shiqi Wang, Yuming Fang, and Jiaying Liu. Single image deraining: From model-based to data-driven and beyond. *CoRR*, abs/1912.07150, 2019.

[135] 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.

[136] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *CoRR*, abs/2111.11432, 2021.

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

[138] 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.

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

[140] 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.

[141] 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.

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

[143] 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.

[144] 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.

[145] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In Shai Avi-

dan, 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.

[146] 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.

[147] 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.

[148] Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. VLUE: A multi-task benchmark for evaluating vision-language models. *CoRR*, abs/2205.15237, 2022.

[149] 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.

[150] 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.

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

[152] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251. IEEE Computer Society, 2017.

[153] 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.