Dynamic Texts From UAV Perspective Natural Images

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

Abstract

Drone-based image processing offers valuable capabilities for surveillance, detection, and tracking in vast areas, aiding in disaster search and rescue, and monitoring artificial events like traffic jams and outdoor activities under adversarial weather conditions. Nonetheless, this technology encounters numerous challenges, including handling variations in scales and perspectives and coping with environmental factors like sky interference and the presence of far and small objects. Besides, ensuring high visibility distance in 3D depth is crucial for safe flights in various settings, including airports, cities, and fields. However, local weather conditions can change rapidly during flights, leading to visibility issues caused by fog and clouds. Due to the cost of visibility measurement sensors, lower-cost methods using portable apparatus are desired for flight routines. Therefore, this paper proposes a camera-based visibility and weather condition estimation approach using complementary multiple Deep Learning (DL) and Vision Language Models (VLM) under adversarial conditions. Experimental results show the superiority of enhanced 2D/3D captions with physical scales over SOTA VLMs.

1. Introduction

Recent advances in manned and unmanned aerial vehicles (AVs and UAVs) have led to their widespread use in various applications [32, 2, 1]. These vehicles are equipped with cameras, enabling surveillance and recognition of objects and scenes using Image Processing, Computer Vision (CV), and Deep Learning (DL) techniques [26], [52]. AV and UAV cameras offer higher usability with 3D coverage and viewing angles than fixed cameras. However, challenges arise from adverse weather conditions, such as fog and heavy rainfall, causing low visibility that hinders safe flights. Expensive visibility sensors have been used at specific airports, while other locations rely on operators' visual observations. Robustness to such adversarial visual conditions is essential but requires augmenting training images, which can be time-consuming [32, 2, 1, 26, 52]. Over the past few years, there has been substantial advancement in the Vision Language Model (VLM) domain [55, 59]. Deep Learning (DL) and Vision Language model (VLM) are useful for object detection, segmentation, and classification tasks. VLMs can understand vision and text, enabling tasks like Vision-Question-Answer (VQA) and image captioning. Image reconstruction by DL models has been proven efficient for many tasks. However, almost all state-of-the-art (SOTA) DLs and VLMs present high performance in fair conditions, but the performance was degraded under adverse conditions. Besides, a single DL and VL model might not adapt, and segmentation limitations impact visibility, distance, and weather condition estimation results.

Thus, this paper presents an integrated system with multiple DL and VL modes to estimate visibility, distance, and weather condition from fixed and AV/UAV cameras.

The proposed system consists of seven modules, i.e., Deep Reject (Dreject), Deep Context (Dcontext), Deep Visual Language Segmentation (Dvls), Deep Visual Language Detection (Dvld), Deep Weather (Dweather), Deep Visibility (Dvis), and Deep Distance (Ddist). The branched architecture allows us to maintain and upgrade each of the eleven modules efficiently. 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. Dreject excludes difficult images, i.e., lens reflection, to stabilize the overall system. Due to images from adversarial weather conditions, Dvls, Dcontext, and Dvld have fine-tuned VLMs from SOTA models for segmentation, detection, and classification, respectively.

2. The system enables the realization of a low-cost and portable camera system by multiple DL models, i.e., Dreject, Dcontext, Dweather, Dvld, Dvls, Dvis, and Ddist. The proposed combination method enables the estimation of visibility levels for UAVs without the need for expensive systems.

3. More refined and enriched captions are generated based on multiple modules, i.e., Dweather, Dvis, and Ddist. Dweather estimates the weather condition of scenes. Ddist estimates the distance between far objects and a camera, and Dvis utilizes image-data regression for visibility estimation without the need for landmarks or geo-tagged scenes.

4. The proposed DL models and system have undergone experiments covering various lighting and weather conditions, confirming their stability, robustness, and accuracy. Additionally, the camera-based assistance for UAVs is useful for rescue during disaster events.

2. Related Work

This section describes related works on Computer Vision (CV), Deep Learning (DL), and Vision Language (VL) models, mainly in visibility estimation and segmentation using cameras.

2.1. Visibility Estimation

Visibility distance can also be considered one of the most important factors for a safe flight and accident reduction. At most airports, visibility is observed by experienced operators. Although visibility measurement can also be used, it is often restricted to the application due to the expensive and point-wise measurement tools. Various methods have been proposed for visibility estimation, such as using wavelet transform [5], spatially partial structure reconstruction [4], detecting road markings and calculating contrast [33], estimating the extinction coefficient from a single daylight photograph [20], and fog detection from onboard camera [31]. Additionally, a new approach using Retinex filtering for light intensity invariant images is presented in [43], and the Comprehensive Visibility Indicator (CVI) algorithm is proposed in [51] for more accurate visibility estimations in various driving situations. However, the previously mentioned visibility estimation methods on roads can only be effective in a short range of distance, i.e., 100-300 m. Recently, to estimate visibility, segmentation-based DL models have been reported and used by Dvis [37] and Droad [36]. However, landmarks such as white lanes cannot be used in UAV applications since a farther visibility distance, i.e., $500 - 3000$ m, is required. Therefore, this paper proposes a novel deep-learning-based method to recognize farther objects and regression-based methods to estimate the visibility distance.

2.2. Segmentation

Segmentation is an essential topic in Image Processing, Computer Vision (CV), and Deep Learning (DL) [26, 47, 17]. It can be categorized into three types: semantic segmentation, instance segmentation, and panoptic segmentation [16, 3, 24]. Semantic segmentation aims to classify each pixel into corresponding classes, while instance segmentation distinguishes different instances of the same class, making it more complex and computationally expensive. Panoptic segmentation integrates the advantages of both semantic and instance segmentation [24]. This paper uses panoptic segmentation to recognize objects (e.g., mountains) and non-objects (e.g., sky).

In DL, CNNs play a central role in solving various CV problems, including object semantic segmentation [26]. The fully convolutional network (FCN) with U-Net as the backbone has improved semantic segmentation accuracy [26]. Conditional Markov Random Fields are also used to enhance the boundaries of recognized objects [16].

Transformers were originally used in NLP [44] and later applied to 2D images. Vision Transformers (ViT) have significantly improved over CNN-based models by reducing dimensionality [8]. Swin Transformer [25] further enhances efficiency, making it more suitable for dense prediction tasks like semantic segmentation. Segformer [48] is the current state-of-the-art in semantic segmentation, utilizing encoder-decoder modules based on the Transformer architecture.

2.3. Image-Based Regression

Image-based regression has been used as one of the most convenient analysis and prediction methods pair-wise. This algorithm usually extracts features from given images and then regresses between image features and data. There have been a lot of variant algorithms like Haar features with boosting [60]. Haar features are effective only when detecting edges and lines. Neural networks [21] are used for image feature extraction, where the full-frame scene is fed into the network. In [21], data augmentation with both 2D and 3D is applied to avoid over-fitting. However, such local image feature-based regression would be unstable for visibility estimation due to partial occlusion by time-varying fog. Therefore, this paper considers more stable segmented regions from overall images. Moreover, image-based regression has majorly dealt with indoor scenes or fine weather scenes, whereas this paper challenges adversarial conditions like locally strong illumination that can degrade the original performance of image-based regression by ML and DL.

2.4. Vision Language Model

In recent years, the Vision Language model (VLM) field has undergone significant progress [55, 59]. 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 [34, 22, 7], generative objectives [45, 13], and alignment objectives [10, 50, 28].

VLMs are transferred by Text-Prompt Tuning [59, 58, 29]. Besides finetuning, knowledge distillation is a method to improve VMLs for downstream tasks, including object detection [56, 27] and semantic segmentation [57, 61, 30, 42].

Unseen images that have not been pre-trained have become recognized by VLM frameworks [6, 9]. More diverse and out-of-distribution data for pre-training and evaluation are used [12]. Prompt learning to adapt VLMs to new tasks without fine-tuning is also shown [14]. Contents of captions have been enhanced for better descriptions of real-world objects [9].

Geometric reasoning or depth estimation to infer 3-D information from 2-D images [53, 55] 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 [6, 19, 40, 41]. 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 [46] is that it can produce acceptable results on the general scene and unseen classes. However, since Visual ChatGPT [46] is trained on the limited data of the year 2021, it generates captions under older datasets. So far, Visual ChatGPT [46] is weak at generating dynamic scene descriptions like weather and road conditions. As mentioned in [39, 38], the physical factor was considered but did not orient to the scene's context.

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, distance, and weather condition.

3. Proposed Method

This section describes the proposed system with multiple DL models for visibility estimation in a fixed and AV/UAV camera applications.

3.1. Overview of Proposed System

This section overviews the proposed system for visibility estimation through cameras as shown in Figure 1. A single image is assumed to be monitored from a fixed or drone camera. In order to estimate visibility levels, the appearance of far objects is an important landmark for human eyes and image recognition. A location of a camera, drone, and objects can provide respective latitudes and longitudes from GPS and Google Maps data. Single or multiple mountains are assumed. The proposed Dvls is a Deep Learning-based segmentation in an image. Depending on the pretrained number of classes, tens or hundreds of objects will be recognized. Each object can set its latitude and longitude. Therefore, all major objects are arranged in order of depth. Fog, rainfall, and snowfall may cover such objects. Covered objects may not be recognized. Or partially occluded objects can happen. Thus, different visibility levels will be estimated through object recognition. However, more considerations are required in various scenes. A case of them is that far objects such as vehicles, pedestrians, and sky, may not provide GPS locators. Moreover, GPS information may not be useful when monitoring vertically downward to the ground from UAV. No objects may be also assumed on the ground. Therefore, this paper proposes Ddist, e.g., distance, and Dvis, e.g., visibility, to cope with this issue. Ddist is to utilize recognized far objects with a physical distance DL model. Such objects are recognized by Dvls. On the other hand, Dvis consists of Transformer and regression model to estimate visibility without object recognition when no object has been recognized on the ground. To stabilize Dvis under adversarial conditions like fog, Dreject is introduced to select images with good quality, where lens reflection, strong light, and heavy fog are presented. Therefore, the proposed system consists of four major DL models: Dreject, Dvls, Dvis, and Ddist. In the following, each of these models is mentioned in detail.

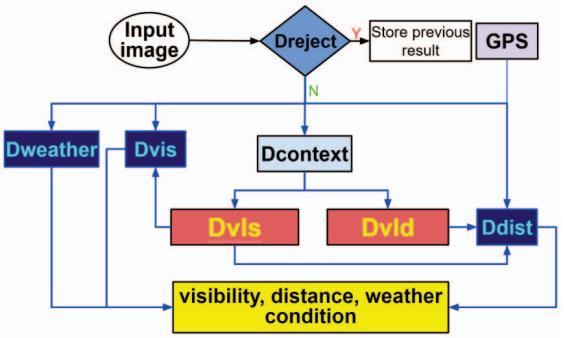


Figure 1. A system overview of the proposed visibility estimation with four proposed Deep Learning models: Dvls, Dreject, Ddist, and Dvis.

3.2. Dreject

Dreject is a classification-based DL model combined to classify input image quality with low or high. This is used to enhance the stability and accuracy of the other cascaded DLs. It is expected to the system more robust than no use of Dreject. When being rejected images, stored past image results are applied for continuous visibility estimation. Dreject is an image classification model with an output of 4 different adversarial conditions:

- Clear: Scenes with normal conditions and good visibility.
- Lens reflection: Scenes with the effect of light refraction on the camera lens or affected by weather conditions with high humidity and raindrops.
- Strong light: Scenes with strong artificial lighting such as beacons, road signs, and street rump.
- Heavy fog: Scenes with dense fog and very low visibility.

3.3. Ddist

This section describes Ddist for physical distance estimation as shown in Figure 2 Ddist is a DL-based regression model which estimates the physical distance based on distances between target objects and cameras. Target objects are recognized by Dvls, which is based on an image-data regression model. Data are physical distances between a camera and distant objects. Over 3000 images have been used to train. As shown in Figure 3, when a GPS signal is available, detected objects are used to estimate their distance. The predefined fog covers of each object serve as references in the order of depth to approximate the distances.

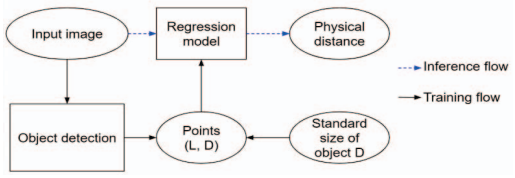


Figure 2. The proposed Ddist.

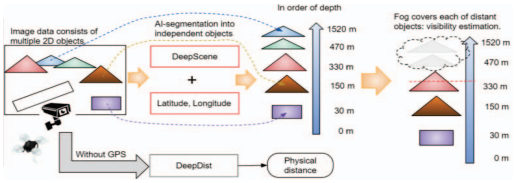


Figure 3. The proposed distance estimation method relies on the order of depth.

3.4. Dvis

In real scenes, whenever the fog appears, the recognition performance of Dvls can be degraded at far objects. For this issue, far objects cannot be used for physical distance estimation. Therefore, a no-object-based visibility estimation method is needed for a continuous flight. Dvis is trained from real and synthesized scenes with known physical distances and foggy images. Dvis is an image-data regression model as well as Ddist. Since overall image features have to be used for training, over 10000 images are needed to train. For the Dvis model, after preprocessing the raw image data, the image will be passed through the base models to extract the feature of that image to form a feature vector. then pass the feature vectors through the regression classes to find a unique value.

3.5. Proposed Dcontext

Dcontext is a VL 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. Dcontext 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. To reduce the computational cost and memory usage, BLIP2 [18] was used and fine-tuned on the collected dataset including 5000 images.

Dcontext utilizes the input texts of five distinct disaster categories: *the sky at the airport in the daytime, the sky at the airport at night time, the road scene captured by a drone camera, the city scene captured from a drone camera, outdoor activity scene captured from drone camera*. Tailored textual input descriptions are employed for each category to enhance natural language processing techniques in analyzing related data. These scenes are associated with domain-specific terms to improve the accuracy of automated disaster detection and classification.

3.6. Proposed Dvls, and Dvld

Dvls is suggested as a means to achieve semantic segmentation for these scenes. It is built upon the fine-tuned version of OvSeg [23], with the addition of a new physical constraint to the loss function. In order to obtain disaster descriptions for Dcontext, 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 vision language model with open-vocabulary object detection [11]. Unlike traditional object detection models that rely on fixed categories, Dvld can detect objects based on arbitrary text inputs from Dcontext. The model achieves this capability by leveraging the knowledge extracted from a pre-trained 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.

3.7. Dweather

Dweather is a combination of a DL-based classifier, i.e., transformer, and a VL model, i.e., BLIP-2. Figure 2 shows an overview of the proposed Dweather trained on 11237 images with rain, fog, and lightning classes. BLIP-2 is employed to generate image captions by combining the generated text from the input image. It uses the weather prompt *"How is the weather?"* to include additional weather conditions. The DL employs three classes: rain, fog, and lightning. The outputs of the VL model and classifier are summarized by DeepSummary (Dsum), i.e., GPT model. Dsum summarizes separated texts by inputting *"Summarize the weather condition without explanation: The weather is X. There is Y"*. In this input, X represents the output of the VL model, and Y represents the output of the weather classifier.

4. Experiments And Discussion

This section discusses experimental results and discussion on various road scenes that are recognized and seg-

mented under adversarial conditions.

4.1. Dreject Experiment

This subsection evaluates the performance of Dreject. The dataset consists of 3050 images with 4 different adversarial conditions types: Clear, Lens reflection, Strong light, and Heavy fog. Figure 4 shows drone-viewed images rejected by Dreject with strong light and lens reflection. Such images are from strong reflections from bright road surfaces in the daytime and the spotlight in the nighttime. As shown in Table 1, the average accuracy 96% shows under different adversarial conditions. Therefore, it has demonstrated the

Figure 4. Example of rejected drone images by Dreject.

Table 1. Evaluation of Dreject on difficult scene classification under adversarial conditions.

	Image number	Correct recognition	Wrong recognition	Accuracy (%)
Clear	827	821	6	99.27
Lens reflection	627	600	27	95.69
Strong light	864	816	48	94.44
Heavy fog	732	691	41	94.40
Total	3050	2928	122	**96.00**

effectiveness of Dreject to obtain acceptable drone images, which are useful for three DLs of Dvls, Dvis, and Ddist.

4.2. Segmentation By Dvls

This subsection presents the results of Dvls recognized from daytime and nighttime airport camera images. As shown in Figure 5, in each scene, a pair of two images show clear or cloudy weather conditions. Background objects are runways, buildings, and mountains. At night, many lighting spots are recognized. Note that such lighting spots are not assumed to be rejected by Dreject. Owing to panoptic segmentation results, wide sky regions and objects on the ground have been successfully separated. Experiments are conducted to obtain far horizons. However, due to the camera lens's distortion, it becomes difficult to obtain. In each of the scenes, different segmented objects are shown in the daytime and the nighttime. However, "sky" and "background objects" present a clear boundary. No threshold is applied for segmentation, unlike image processing-based edge detection with empirical threshold settings. Therefore, the robustness and stability of Dvls have been reconfirmed. On the other hand, for a comparison experiment, image processing to detect far horizons is conducted with thresholds. The results failed to detect such lines and curves (not shown here). Thus, Dvls can be used for obtaining the far horizon

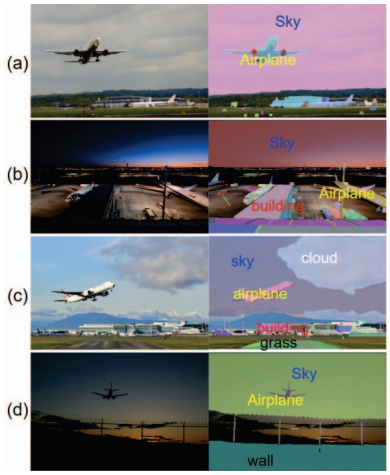


Figure 5. Panoptic segmentation by the proposed Dvls at airports: (a)(c) Daytime. (b)(d) Nighttime.

lines by the boundaries between the sky and bottom objects. It is assumed that the distance between such horizons can be preset by Google Maps geodata. For this, a physical visibility distance can be estimated.

4.3. Deconposition By Dvls

Dvls with querying objects are used to decomposite the input image into elements. Dvls is used for panoptic segmentation of an image. A number of objects in an image are obtained. Based on this, Figs. 6 show two decomposed images with mountains and no-mountain. 8 sub-images contain sky, buildings, trees, and roads. Each sub-image can be linked to a longitude and latitude from Google Maps. By following a flowchart in Figure 1, all recognized objects are arranged in order of depth from a camera's location.

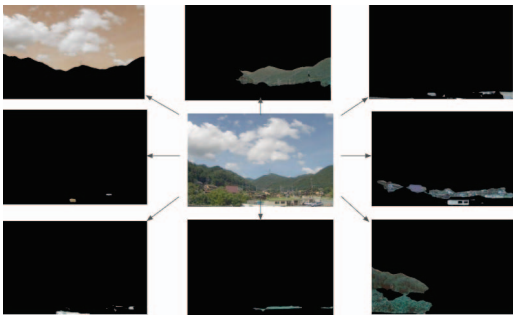


Figure 6. Decomposition of a scene with mountains by panoptic segmentation.

4.4. Dvis Experiments

In order to prove the performance of Dvis, further experiments are conducted. Dvis has been proposed for non-object recognition cases. The dataset includes 4424, 1109, and 1175 images on training, validation, and test sets, respectively. For the dataset used for training, the structure

of the datasets is similar to the datasets of the basic Regression problem. In order to train the Dvis model, the base model, SwinTransformerV2, combined with regression layers is used to obtain a single output. This paper defines visibility in five levels as follows:

- $Dvis > 5000$ m: Clear Image.
- 3000 m $< Dvis < 5000$ m: Fog level 1.
- 1500 m $< Dvis < 3000$ m: Fog level 2.
- 500 m $< Dvis < 1500$ m: Fog level 3.
- $Dvis < 500$ m: Fog level 4.

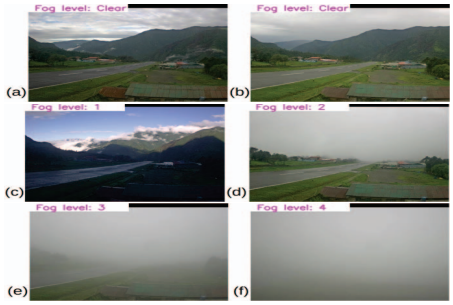


Figure 7. Results by the proposed Dvis for the airport cameras: (a) Clear Image. (b) Clear Image. (c) Fog level 1. (d) Fog level 2. (e) Fog level 3. (f) Fog level 4.

In order to justify the performance of Dvis we propose, different scenes with different fog levels are added to conduct an evaluation test. The accuracy scores of the Dvis model achieve 82.19% on 1175 images of the test dataset.

4.5. Dweather Results

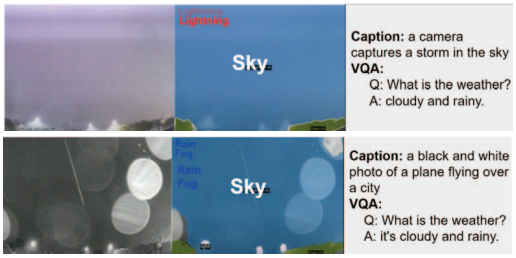
This section provides an evaluation of Dweather with a combination of a vision-language (VL) model and a transformer-based model. The weather classifier is evaluated on a test dataset consisting of 1411 images. Figure 8 presents results by DeepWeather. Sunny, rainy, and foggy scenes have been correctly recognized, where elaborate objects and background, i.e., trees and water, are present. For quantitative evaluation, the test dataset except sunny images is added, which includes five classes, i.e., fog, lightning, normal, rain, and snow, with 149, 67, 660, 223, and 312 images, respectively. The accuracy of the test data is 89.4%, and the loss is 0.397.

To validate the effectiveness of the proposed Dweather in PanopticBlue, camera images captured under adversarial weather conditions are utilized. Figure 9 shows the results of weather conditions: (1)-(b) lightning and (2)-(b) rain and fog from Dweather. (1)-(c) VL recognizes "cloudy and rainy", where "with lightning" has been added from the text at (1)-(b). In (2), "cloudy and rainy" + "rain and fog" has also been generated. Therefore, enriched auto-briefings have been demonstrated from image recognition and VL models.

4.6. Dvld Overall Evaluation

In this section, Dvld is estimated and compared with YoloV8 [35] for object detection on 1832 images of the VisDrone [62] dataset using pixel-wise accuracy and Intersection over Union (IoU). Table 2 shows that Dvld is better than YoloV8 [35].

Figure 8. Results on various weather conditions by DeepWeather: sunny, rainy, and foggy. Probability is shown in the parenthesis



(a) (b) (c)

Figure 9. Dweather results: (a) Input. (b) Weather classification. (c) VL model with weather prompt and refined auto-briefing

Table 2. Comparison of Dvld and YoloV8 [35].

	Dvld	YoloV8[35]
Accuracy	88.34	83.32
IoU	0.78	0.73

5. Ablation Study

This section discusses an ablation study in temporal fog variation and foggy disaster scenes.

5.1. Dvis With Different Models

This section devotes to justifying the performance of the proposed Dvis by comparing 7 DLs as baselines. Table 3 summarizes the results and accuracy of the different base models. It has proven that Swin Transformer V2 in Dvis presents the best performance among all previous DLs in terms of Macro F1, Weighted F1, RMSE, MAE, and Accuracy.

Model 1: VGG19 [32], accuracy 0.6689; Model 2: MobileNetV2 [38], accuracy 0.5577; Model 3: EfficientNetB7 [40], accuracy 0.57; Model 4: ResNet152V2 [41], accuracy 0.5229; Model 5: ViT [42], accuracy 0.6992; Model

6: Swin Transformer[43], accuracy 0.692; Model 7: Swin Transformer V2[43], accuracy 0.8219;

Table 3. Score and accuracy of the Dvis model with different base models.

	Macro F1	Weighted F1	RMSE	MAE	Accuracy
1	0.660	0.5026	250.16	154.64	0.6689
2	0.4099	0.5644	214.00	146.04	0.5577
3	0.3426	0.3986	245.18	171.08	0.570
4	0.3991	0.4523	310.58	222.48	0.6229
5	0.526	0.5375	232.97	147.89	0.6992
6	0.558	0.661	76.000	38.000	0.692
7	**0.6811**	**0.6976**	194.13	114.51	**0.8219**

5.2. Dcontext Evaluation

This section describes the Dcontext experiment on the collected dataset including 1500 images with five different contexts, i.e., the sky at the airport in the daytime, the sky at the airport at night time, the road scene captured by a drone camera, the city scene captured from a drone camera, outdoor activity scene captured from drone camera. The result of Dcontext is compared with different fine-tuned classifiers, i.e., ViT, ResNet, and VGG on the same dataset. Table 4 shows the comparison of accuracy between the proposed Dcontext and other classifier models. The comparison result has proven that Dcontext outperforms previous models.

Table 4. Performance Metrics for Different Contexts and Methods

Context/Method	Dcontext	ViT classifier	ResNet101 classifier	VGG19 classifier
daytime airport	86.45	81.26	80.24	79.32
nighttime airport	91.25	86.12	88.35	83.21
road by drone	86.32	81.87	82.65	81.14
city by drone	86.36	82.14	83.25	79.96
outdoor activity by drone	88.58	84.26	83.65	83.97
Overall	87.79	83.13	83.63	81.52

5.3. Visibility Based On Base-line Comparison With a Time-varying Fog

This section discusses visibility levels from images with a time-varying fog. Figure 10 shows temporal image changes from clear to fog changes. Visible several mountains (a) are between nothing (b) and reducing (c) while remaining most on-the-ground objects like a runway and green fields. In order to utilize temporal changes of the ridge curves, the initial ridge curve (a) in the sky is stored as shown in Figure 11. All objects, i.e., runway, house, and trees, below the mountains, have been merged into one class, i.e., mountain. All mountains are covered by heavy fog (b). Except for the sky region, all joint objects are merged. Fog as St cloud type begins to appear and cover part of mountains (c). In (b) and (c), the segmented mountain becomes disappeared. Therefore, the area between the

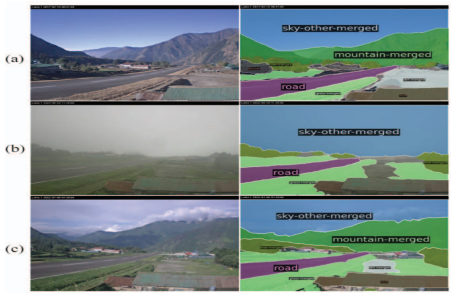


Figure 10. Dvls results with different fog levels: (a) Clear. (b) Heavy fog. (c) Light fog.

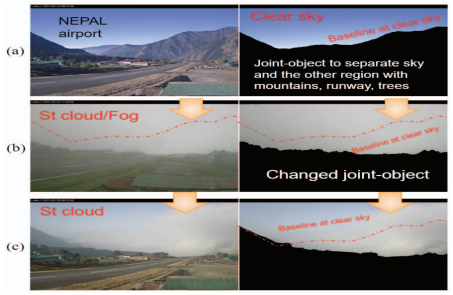


Figure 11. Ridge curve changes by merging several objects from Dvls.

initial and final ridge curves increases. Standard edge detection may be useful, but unstable edges are obtained due to sunbeam changes over time. However, the proposed join-object method by separating the sky region is robust to such sunbeam changes due to the object-based boundary detection method. Disappearing mountains have geo-data as shown in Figure 12. Results in Figure 12 can readily correspond to distances between 4.95 km and 1.22 km. Therefore, about 3 km has changed in visibility. For a better understanding, a time-series change of visibility is illustrated in Figure 13. In city scenes, Dvls is applied as shown in Figure 14. Many tiny objects, i.e., vehicles and ships, are recognized. Moreover, the far sky region is obtained. From these, visibility is assumed to be clear.

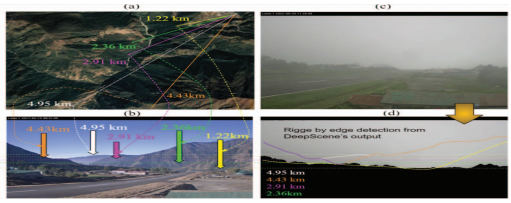


Figure 12. Visibility estimation from far objects with known distances in a few km.

5.4. Dvls Evaluation

In this section, Dvls is estimated and compared with Mask2former on the dataset using pixel-wise accuracy and Intersection over Union (IoU). Table 5 shows that Dvls is

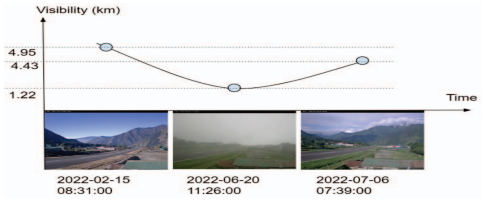


Figure 13. Visibility changes over time.

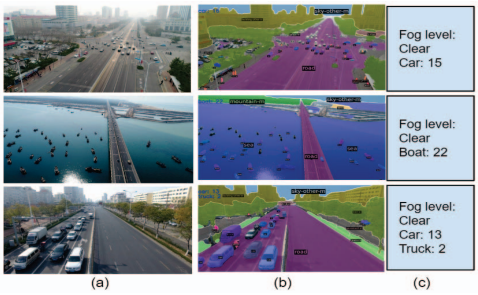


Figure 14. Results of far tiny objects recognition and sky by Dvls (a) input image (b) Dvls's output (c) fog level and objects in an input image.

better than Mask2former. In order to evaluate the robustness of Dvls in drone images and compare it with SOTA methods, i.e., ODISE [49], and SAM [15], experiments with various images in different altitudes, light conditions, and viewing angles are carried out. As shown in Figure 15, (c) ODISE [49], and (d) SAM [15] shows low performance in high altitude and glare from sunlight, i.e., at strong illumination conditions, where all objects on the ground are combined into one category. Moreover, in high altitude and frontal view, ODISE [49] and SAM [15] cannot detect tiny objects from a distant view. On the other hand, (b) Dvls performs better in recognition performance of major object categories, i.e., roads, pedestrians, buildings, and vehicles. It has been suggested that the proposed Dvls is robust to a wide range of far objects with sizes, orientations, and colors in spite of various camera angles. The results have proven that Dvls has shown the best performance with tiny and far objects.

Table 5. Comparison of Dvld and Mask2former

	Dvld	Mask2former
Accuracy	86.57	83.18
IoU	0.78	0.75

5.5. Images Generation

Data augmentation enhances the power of our deep learning model, while data labeling is labor-intensive. In this experiment, we generate images from drones with diverse weather conditions using linguistic descriptions. Consequently, the data is labeled in that language, eliminating the need for manual labeling. This section describes the results of image generation chosen by combining the vision language model and the stable diffusion model [54]. This section presents the results of image generation, achieved by combining a vision language model with a conditional stable diffusion model [54]. The conditional stable diffusion model [54] is responsible for generating images from prompts, while a vision language model, i.e., BLIP2 [18] is utilized to assess the similarity between the generated images and the input prompts, ultimately selecting the high-quality images. As shown in Figure 16, the image with the highest similarity, chosen by BLIP2, is selected from the 4-batch generated images by the diffusion model. Although the objects in the generated images do not appear realistic, the overall scene is similar to the text description.

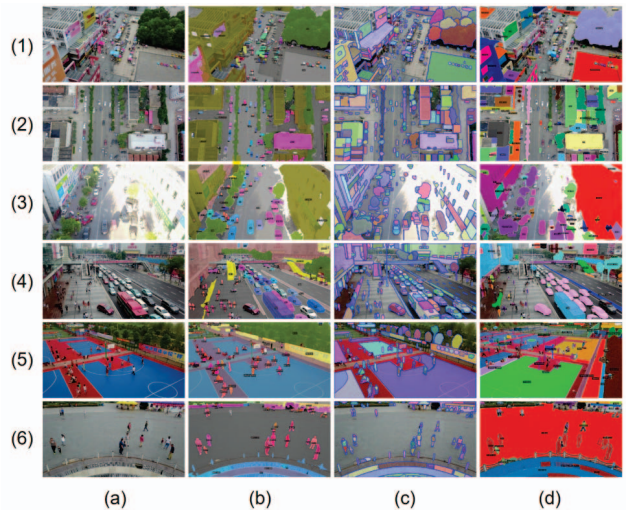


Figure 15. Comparison of recognition in (b) Dvls, (c) ODISE [49], and (d) SAM [15] in variant scenes

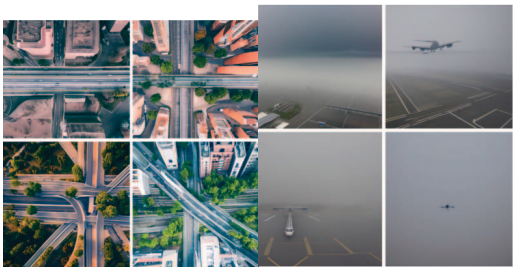


Figure 16. Generated images with various prompts selected by the VL model, i.e., BLIP2. (a) Aerial view of a foggy city road captured by a drone. (b) Foggy sky captured by an airport camera.

6. Conclusion

This paper has presented complementary multiple Deep Learning and Vision Language Models for enhancing segmentation, captions, and visibility under adverse conditions. Future UAVs will be required for disaster matters more than now due to the increments of extreme weather conditions all over the world.

References

[1] https://spacegrant.arizona.edu/sites/spacegrant.arizona.edu/files/spacegrant urep faq final.pdf. 1

[2] https://www.nasa.gov/sites/default/files/atoms/files/2019-rios-tm-220336-508.pdf. 1

[3] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. YOLACT++ better real-time instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(2):1108–1121, 2022. 2

[4] Clement Boussard, Nicolas Hautière, and Brigitte d'Andréa-Novel. Visibility distance estimation based on structure from motion. In *11th International Conference on Control, Automation, Robotics and Vision, ICARCV 2010, Singapore, 7-10 December 2010, Proceedings*, pages 1416–1421. IEEE, 2010. 2

[5] Christoph Busch and Eric Debes. Wavelet transform for analyzing fog visibility. *IEEE Intell. Syst.*, 13(6):66–71, 1998. 2

[6] 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. 3

[7] 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. 2

[8] 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. 2

[9] 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. 3

[10] 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. 2

[11] 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

[12] 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

[13] 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. 2

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

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 8

[16] Hyunmin Lee and Jaesik Park. Instance-wise occlusion and depth orders in natural scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 21178–21189. IEEE, 2022. 2

[17] 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. 2

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 4, 8

[19] 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. 3

[20] Qin Li, Yi Li, and Bin Xie. Single image-based scene visibility estimation. *IEEE Access*, 7:24430–24439, 2019. 2

[21] Xiaotian Li, Juha Ylioinas, and Juho Kannala. Full-frame scene coordinate regression for image-based localization. In Hadas Kress-Gazit, Siddhartha S. Srini-

vasa, Tom Howard, and Nikolay Atanasov, editors, *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, 2018. 2

[22] 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. 2

[23] 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. 4

[24] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. 2

[25] 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. 2

[26] 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. 1, 2

[27] 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. 3

[28] 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. 2

[29] 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. 3

[30] 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. 3

[31] M. Negru and S Nedevschi. Image-based fog detection and visibility estimation for driving assistance system. *in Proc. IEEE 9th Int. Conf. on Intelligent Computer Communications and Processing (ICCP)*, pages 163–168, 2013. 2

[32] 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. 1

[33] Pomerleau. Visibility estimation from a moving vehicle using the ralph vision system. *in Proc. IEEE Conference on Intelligent Transportation Systems*, pages 906–911, 1997. 2

[34] 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. 2

[35] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection with yolov8, 2023. 6

[36] Hidetomo Sakaino. Panopticroad: Integrated panoptic road segmentation under adversarial conditions. *in CVPR Workshop*, 2023. 2

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

[38] 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

[39] 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

[40] 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. 3

[41] 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. 3

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

[43] S. Varjo and J Hannuksela. Image bases visibility estimation during day and night. *in Proc. IEEE 9th Int. Conf. on Intelligent Computer Communications and Processing (ICCP)*, 2014. 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 2

[45] 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. 2

[46] 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

[47] 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. 2

[48] 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. 2

[49] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models, 2023. 8

[50] 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. 2

[51] 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. 2

[52] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J. Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, Yuqiang Zheng, Yanyun Qu, Yuhong Xie, Liang Chen, Zhonghao Li, Chen Hong, Hao Jiang, Siyuan Yang, Yan Liu, Xiaochao Qu, Pengfei Wan, Shuai Zheng, Minhui Zhong, Taiyi Su, Lingzhi He, Yandong Guo, Yao Zhao, Zhenfeng Zhu, Jinxiu Liang, Jingwen Wang, Tianyi Chen, Yuhui Quan, Yong Xu, Bo Liu, Xin Liu, Qi Sun, Tingyu Lin, Xiaochuan Li, Feng Lu, Lin Gu, Shengdi Zhou, Cong Cao, Shifeng Zhang, Cheng Chi, Chubin Zhuang, Zhen Lei, Stan Z. Li, Shizheng Wang, Ruizhe Liu, Dong Yi, Zheming Zuo, Jianning Chi, Huan Wang, Kai Wang, Yixiu Liu, Xingyu Gao, Zhenyu Chen, Chang Guo, Yongzhou Li, Huicai Zhong, Jing Huang, Heng Guo, Jianfei Yang, Wenjuan Liao, Jiangang Yang, Liguo Zhou, Mingyue Feng, and Likun Qin. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Trans. Image Process.*, 29:5737–5752, 2020. 1

[53] 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

[54] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 8

[55] 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. 1, 2, 3

[56] 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

[57] 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. 3

[58] 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. 3

[59] 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, 2, 3

[60] Shaohua Kevin Zhou, Bogdan Georgescu, Xiang Sean Zhou, and Dorin Comaniciu. Image based regression using boosting method. In *10th IEEE International Conference on Computer Vision (ICCV 2005), 17-20 October 2005, Beijing, China*, pages 541–548. IEEE Computer Society, 2005. 2

[61] 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

[62] Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge, 2018. 6