Refined and Enriched Physics-based Captions For Unseen Dynamic Changes

Hidetomo Sakaino [1]

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 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 counting, 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 is shown.

1. Introduction

Segmentation has become an important task for real-world applications by Deep Learning (DL) (Long et al., 2015; Liu et al., 2021; Xie et al., 2021; Gu et al., 2022; Dosovitskiy et al., 2021; Tan & Le, 2019; Kim et al., 2020; Li et al., 2019; 2021). 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 (Quan et al., 2021; Yang et al., 2019), defog, and dehaze (Lee et al., 2022; Li et al., 2017; Yan et al., 2020; Guo et al., 2022; Ma et al., 2022); 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.

The single view metrology approach focuses on establishing correlations between low-level image features, such as vanishing points and lines, the 3D dimensions of objects, and their corresponding 2D positions and sizes in the scene. Several studies have utilized this approach to estimate object heights or camera height using camera parameters, image features, and annotated size information of reference objects (Lee et al., 2023). However, there is currently no research that addresses the inclusion of 2D physical scale, i.e., object location and size in meters, in image captions.

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. 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 (Radford et al., 2021; Miech et al., 2020; Li et al., 2022e) and fine-tuned on smaller datasets for specific tasks, such as object detection, segmentation, or 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

[1]Visual Recognition Group, Transportation Weather Lab., Weathernews Inc., Chiba, Japan. Correspondence to: Hidetomo Sakaino <sakain@wni.com>.

The 40^{th} International Conference on Machine Learning: The 2nd New Frontiers in Adversarial Machine Learning (ADVML FRONTIERS @ICML2023), Honolulu, Hawaii, USA. PMLR 202, 2023.

and their surrounding context. However, single VML is often weak for dynamic changes, i.e., disaster scenes (Sreelakshmi & Chandra, 2022) with 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 (Hernández et al., 2022; Xiang et al., 2023; Chen et al., 2023b). However, camera image-based post-disaster object recognition for dirt, water, and rocks remains unsolved on the road. Since domain adaptation segmentation DL models (Wang et al., 2021; Hoyer et al., 2022) require manual selection of the optimal pre-trained model, they are not useful for dynamic changes.

Unseen images that have not been pre-trained have become recognized by VLM frameworks (Chen et al., 2023a; Francis et al., 2021). More diverse and out-of-distribution data for pre-training and evaluation are used (Howard & Ruder, 2018). Prompt learning to adapt VLMs to new tasks without fine-tuning is also shown (Jiang et al., 2023). Contents of captions have been enhanced for better descriptions of real-world objects (Francis et al., 2021).

Geometric reasoning or depth estimation to infer 3-D information from 2-D images (Yu et al., 2022; Zhang et al., 2022) 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 (Chen et al., 2023a; Francis et al., 2021; Li et al., 2022b;d; Sanghi et al., 2022; Shi et al., 2022; Zhu et al., 2022; Radford et al., 2021). 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 (Wu et al., 2023) is that it can produce acceptable results on the general scene and unseen classes. However, since Visual ChatGPT (Wu et al., 2023) is trained on the limited data of the year 2021, it generates captions under older datasets. So far, Visual ChatGPT (Wu et al., 2023) is weak at generating dynamic scene descriptions like weather and road conditions. Moreover, the physical size of objects and fog visibility distance is contained. Besides, 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, and shape, 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 eight modules, i.e., Deep Visual Language Classification (Dvlc), Deep Visual Language Segmentation (Dvls), Contrastive Language Physical scale Pretraining (CLPP), Deep Road conditions (Droad), Deep anomaly (Danomal), Deep snowfall (Dsnow). 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. 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., weather conditions by Dsnow, 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.

3. More refined and enriched captions are generated based on fixed queries at 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 (Wu et al., 2023), 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. Proposed Method

This section describes the proposed PanopticCAP method/system for refinement and enrichment of captioning and classes from a single image input. In particular, this paper introduces a dynamic caption by a physical scale that cannot be pre-trained in a vision-language model.

To realize this, SOTAs in segmentation and vision-language models face their limits. Therefore, 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 shows an overview of the proposed PanopticCAP.

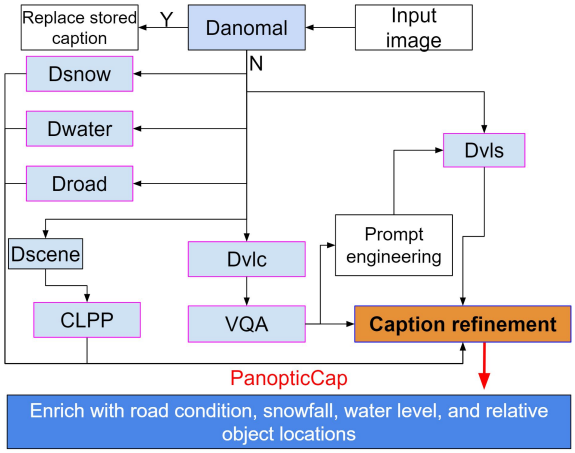

Figure 1. Overview of the proposed PanopticCAP model.

Since this paper deals with many challenging scenes with disasters and car accidents, adversarial conditions are considered. And a Danomal-like DeepReject in (Sakaino, 2023a;b; 2022) is proposed to avoid the degradation of the cascaded other recognition modules. Further detailed explanations of the multiple modules will be given in Sections 2.1 to 2.3.

2.1. 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-2 and -3. 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 such as pedestrian, airplane, debris flow, and eruption 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 (Liang et al., 2022) 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 segmented objects from a single image, this paper proposes to combine respective outputs.

2.2. Proposed CLPP

The proposed Contrastive Language Physical-Scale Pretraining (CLPP) is a VLM with inputs from object locations from pairs of images and text descriptions and a modified contrastive loss function. Unlike SOTA VLMs with no physical models in contrastive loss functions, this paper proposes CLPP with additional physical constraints, as shown in Figure 2. The original contrastive loss function of CLIP (Radford et al., 2021) is defined by

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

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

2.3. Proposed Droad, and Dsnow

This section discusses the proposed Droad, and Dsnow. 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. Droad (Sakaino, 2023b) is applied for further detailed classes of segmented objects. Dscene (Sakaino, 2023b;a) is also applied to ensure snow conditions.

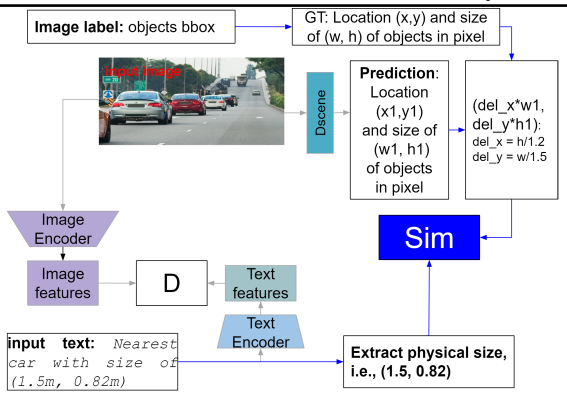

Figure 2. Proposed contrastive language for pre-training in physical scale.

In Droad (Sakaino, 2023b), and Dscene (Sakaino, 2023b;a;c), Swinformer (Liu et al., 2021) 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. Dsnow employs a transformer-based classifier trained on images captured during adverse weather conditions to estimate the level of snowfall. Our experts captured and labeled all the images used in the aforementioned DL models.

2.4. Caption refinement

The caption refinement process involves utilizing a large language model (LLM), namely GPT-4 (OpenAI, 2023), which incorporates the segmentation outcomes from Dvls, the physical scale from CLPP, and the captions generated by VQA. The output of Dvls comprises semantic segmentation along with corresponding locations and descriptions, expressed in a language-based segmentation format as a list of {object description: bounding box of the object in pixels}. The output of CLPP is a caption that includes details about the physical scale, i.e., object size, distance, water level, and visibility. VQA contributes additional descriptions that capture the overall dynamic conditions, including adverse weather conditions, to provide contextual information for the LLM. The final result of caption refinement is an enriched caption that encompasses information about road conditions, water levels, and relative object locations.

3. Experiments and Discussion

3.1. 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 for Dvls model are selected based on the classification results from Dvlc and the aforementioned pre-defined texts.

Figure 3 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 (Cheng et al., 2022), 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.

(a) (b) (c)

Figure 3. Results of segmentation by SOTA and proposed PanopticCAP: (a) Original image. (b) Proposed refined semantic segmentation. (c) Mask2Former (Cheng et al., 2022)

3.2. Dynamic Captions with Weather and Road Conditions by Proposed 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, and Droad are used by comparing a SOTA VL captioning model, BLIP (Li et al., 2022c).

Figure 4 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. Therefore, the road condition and visibility

distance have been included in the captions of Dvlc.

Table 1 shows a comparison of the refined captions and a SOTA BLIP (Li et al., 2022c) result using six scenes of Figure 4. 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, and Dsnow outperforms single VML, i.e., BLIP (Li et al., 2022c).

Table 1. Comparison of the refined captions with BLIP caption results.

	Proposed method	BLIP (Li et al., 2022c)
(1)	Rocks lay on the flooding road	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	A snowstorm on a highway
(4)	A truck on the wet highway, snow on the side of the highway, under heavy snowfall	A snow plow clears a road in the snow
(5)	12 people stand on a flooded road, 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	Snow-covered road with a fence and a street light

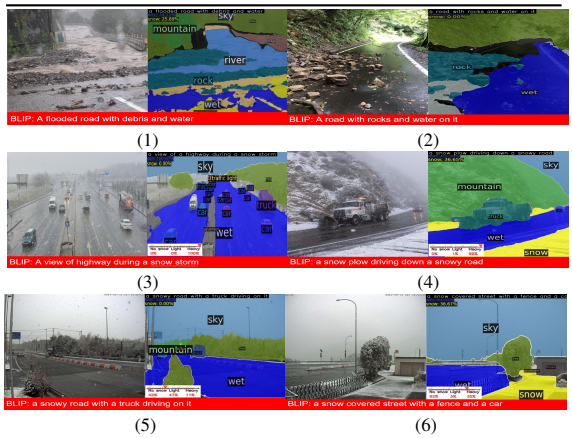

Figure 4. Results of proposed Dvls with refined and enriched captions in dynamic scenes: (1) Flooding road. (2) Landslide on the road. (3), (4) Heavy snowfall on the highway. (5) Flooded scene with water level, Lv2. (6) Light snowfall on the highway.

4. Ablation study

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

4.1. 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 5 (a), images present disaster events. Two SOTAs of (c) MaskDINO (Li et al., 2022a). (d) OVSeg (Liang et al., 2022) are compared.

As a result, Table 2 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 have 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 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.

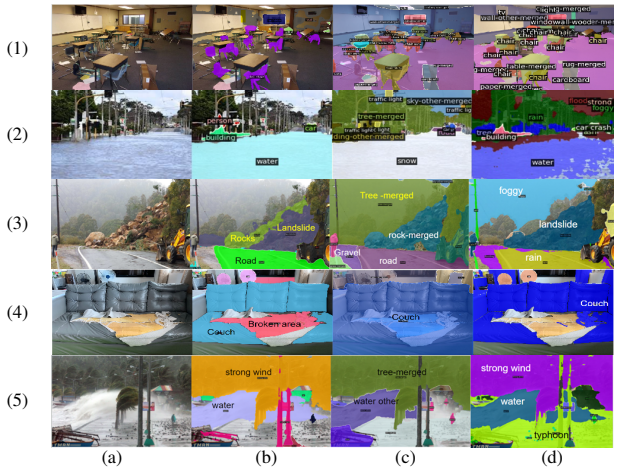

Figure 5. Comparison of the proposed method, MaskDINO (Li et al., 2022a), and OVSeg (Liang et al., 2022) (a) Input image. (b) Proposed Dvls. (c) MaskDINO (Li et al., 2022a) (d) OVSeg (Liang et al., 2022)

Table 2. Comparison of classes by SOTAs and proposed Dvls.

Image	SOTA	Proposed
(1)	table, chair	fell chairs
(2)	snow, rain	water
(3)	rock-merged, rain	landslide
(4)	couch	couch, broken area
(5)	tree-merged, typhoon	strong wind

4.2. 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. 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 3 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 be 0.4. Therefore, it has been reconfirmed that the selected m is optimal.

Table 3. 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

4.3. 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 (Chen et al., 2015) and the Conceptual Captions dataset (Sharma et al., 2018), 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).

According to the results in Table 4, 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 5 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.

Table 4. Performance evaluation of proposed panopticCAP on public image-text datasets

Dataset/Method	PanopticCAP	Visual ChatGPT
COCO Caption	0.4384	0.4415
Conceptual Caption	0.4319	0.4235

Table 5. Performance evaluation of proposed panopticCAP on collected datasets.

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

Furthermore, a comparison was made between the computational cost and memory usage of the proposed system and SOTA methods on the same hardware device. Table 6 presents a comparison of the computational cost and memory usage for these methods.

Table 6. 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(Li et al., 2022c)	1.432	3214

5. Conclusion

This paper introduces PanopticCAP, a novel framework that combines multiple DL and VLM models using efficient branched structures. It is the first approach to incorporate dynamic changes in 2D image captions, including physical scales such as object size and location, weather conditions, water level, and road conditions. By utilizing a 2D physics-based loss function, PanopticCAP generates refined and enriched captions surpassing those achieved by a contrastive loss. This framework has the potential to provide detailed scene descriptions for various applications, including drivers, autonomous systems, and rescue workers relying on camera images. However, future studies should focus on extending this understanding to 3D scenarios

References

Chen, F., Zhang, D., Han, M., Chen, X., Shi, J., Xu, S., and Xu, B. VLP: A survey on vision-language pre-training. *Int. J. Autom. Comput.*, 20(1):38–56, 2023a. doi: 10.1007/s11633-022-1369-5. URL https://doi.org/10.1007/s11633-022-1369-5.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

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

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. 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*, pp. 1280–1289. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00135. URL https://doi.org/10.1109/CVPR52688.2022.00135.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. 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. URL https://openreview.net/forum?id=YicbFdNTTy.

Francis, J., Kitamura, N., Labelle, F., Lu, X., Navarro, I., and Oh, J. Core challenges in embodied vision-language

planning. *CoRR*, abs/2106.13948, 2021. URL https://arxiv.org/abs/2106.13948.

Gu, J., Kwon, H., Wang, D., Ye, W., Li, M., Chen, Y., Lai, L., Chandra, V., and Pan, D. Z. 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*, pp. 12084–12093. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01178. URL https://doi.org/10.1109/CVPR52688.2022.01178.

Guo, C., Yan, Q., Anwar, S., Cong, R., Ren, W., and Li, C. Image dehazing transformer with transmission-aware 3d position embedding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5802–5810, 2022. doi: 10.1109/CVPR52688.2022.00572.

Hernández, D., Cecilia, J. M., Cano, J., and Calafate, C. T. Flood detection using real-time image segmentation from unmanned aerial vehicles on edge-computing platform. *Remote. Sens.*, 14(1):223, 2022. doi: 10.3390/rs14010223. URL https://doi.org/10.3390/rs14010223.

Howard, J. and Ruder, S. Universal language model fine-tuning for text classification. In Gurevych, I. and Miyao, Y. (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 328–339. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1031. URL https://aclanthology.org/P18-1031/.

Hoyer, L., Dai, D., Wang, H., and Gool, L. V. MIC: masked image consistency for context-enhanced domain adaptation. *CoRR*, abs/2212.01322, 2022. doi: 10.48550/arXiv.2212.01322. URL https://doi.org/10.48550/arXiv.2212.01322.

Jiang, J., Liu, Z., and Zheng, N. Correlation information bottleneck: Towards adapting pretrained multimodal models for robust visual question answering, 2023.

Kim, D., Woo, S., Lee, J., and Kweon, I. S. Video panoptic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9856–9865. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00988. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Kim_Video_Panoptic_Segmentation_CVPR_2020_paper.html.

Lee, B.-U., Zhang, J., Hold-Geoffroy, Y., and Kweon, I. S. Single view scene scale estimation using scale field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21435–21444, June 2023.

Lee, S., Son, T., and Kwak, S. 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*, pp. 18889–18899. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01834. URL https://doi.org/10.1109/CVPR52688.2022.01834.

Li, F., Zhang, H., xu, H., Liu, S., Zhang, L., Ni, L. M., and Shum, H.-Y. Mask dino: Towards a unified transformer-based framework for object detection and segmentation, 2022a.

Li, F., Zhang, H., Zhang, Y., Liu, S., Guo, J., Ni, L. M., Zhang, P., and Zhang, L. Vision-language intelligence: Tasks, representation learning, and large models. *CoRR*, abs/2203.01922, 2022b. doi: 10.48550/arXiv.2203.01922. URL https://doi.org/10.48550/arXiv.2203.01922.

Li, J., Li, D., Xiong, C., and Hoi, S. C. H. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022c. URL https://proceedings.mlr.press/v162/li22n.html.

Li, M., Xu, R., Wang, S., Zhou, L., Lin, X., Zhu, C., Zeng, M., Ji, H., and Chang, S.-F. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16420–16429, June 2022d.

Li, Y., You, S., Brown, M. S., and Tan, R. T. Haze visibility enhancement: A survey and quantitative benchmarking. *Computer Vision and Image Understanding*, 165:1–16, 2017. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2017.09.003. URL https://www.sciencedirect.com/science/article/pii/S1077314217301595.

Li, Y., Chen, X., Zhu, Z., Xie, L., Huang, G., Du, D., and Wang, X. 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*, pp. 7026–7035. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00719. URL http://openaccess.thecvf.com/content_

CVPR_2019/html/Li_Attention-Guided_Unified_Network_for_Panoptic_Segmentation_CVPR_2019_paper.html.

Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., and Jia, J. Fully convolutional networks for panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 214–223. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00028. URL https://openaccess.thecvf.com/content/CVPR2021/html/Li_Fully_Convolutional_Networks_for_Panoptic_Segmentation_CVPR_2021_paper.html.

Li, Y., Chang, Y., Gao, Y., Yu, C., and Yan, L. 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*, pp. 5831–5840. IEEE, 2022e. doi: 10.1109/CVPR52688.2022.00575. URL https://doi.org/10.1109/CVPR52688.2022.00575.

Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., and Marculescu, D. Open-vocabulary semantic segmentation with mask-adapted CLIP. *CoRR*, abs/2210.04150, 2022. doi: 10.48550/arXiv.2210.04150. URL https://doi.org/10.48550/arXiv.2210.04150.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. 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*, pp. 9992–10002. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00986. URL https://doi.org/10.1109/ICCV48922.2021.00986.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3431–3440. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298965. URL https://doi.org/10.1109/CVPR.2015.7298965.

Ma, X., Wang, Z., Zhan, Y., Zheng, Y., Wang, Z., Dai, D., and Lin, C. 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*, pp. 18900–18909. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01835. URL https://doi.org/10.1109/CVPR52688.2022.01835.

Miech, A., Alayrac, J., Laptev, I., Sivic, J., and Zisserman, A. Rareact: A video dataset of unusual interactions. *CoRR*, abs/2008.01018, 2020. URL https://arxiv.org/abs/2008.01018.

OpenAI. Gpt-4 technical report, 2023.

Quan, R., Yu, X., Liang, Y., and Yang, Y. Removing raindrops and rain streaks in one go. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 9147–9156. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00903.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. (eds.), *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*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Sakaino, H. Deepreject and deeproad: Road condition recognition and classification under adversarial conditions. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 382–389, 2022. doi: 10.1109/ICMLA55696.2022.00061.

Sakaino, H. Panopticvis: Integrated panoptic segmentation for visibility estimation at twilight and night. *in CVPR Workshop*, 2023a.

Sakaino, H. Panopticroad: Integrated panoptic road segmentation under adversarial conditions. *in CVPR Workshop*, 2023b.

Sakaino, H. Deepscene, deepvis, deepdist, and deepreject: Image-based visibility estimation system for uav. In *2023 IEEE Aerospace Conference*, pp. 1–11, 2023c. doi: 10.1109/AERO55745.2023.10115897.

Sanghi, A., Chu, H., Lambourne, J. G., Wang, Y., Cheng, C.-Y., Fumero, M., and Malekshan, K. R. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18603–18613, June 2022.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Gurevych, I. and Miyao, Y. (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018,*

Volume 1: Long Papers, pp. 2556–2565. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238/.

Shi, H., Hayat, M., Wu, Y., and Cai, J. 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)*, pp. 9611–9620, June 2022.

Sreelakshmi, S. and Chandra, S. V. 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.

Tan, M. and Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *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*, pp. 6105–6114. PMLR, 2019. URL http://proceedings.mlr.press/v97/tan19a.html.

Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., and Gool, L. V. Exploring cross-image pixel contrast for semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 7283–7293. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00721. URL https://doi.org/10.1109/ICCV48922.2021.00721.

Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., and Duan, N. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671, 2023. doi: 10.48550/arXiv.2303.04671. URL https://doi.org/10.48550/arXiv.2303.04671.

Xiang, D., Zhang, X., Wu, W., and Liu, H. Denseppmunet-a: A robust deep learning network for segmenting water bodies from aerial images. *IEEE Trans. Geosci. Remote. Sens.*, 61:1–11, 2023. doi: 10.1109/TGRS.2023.3251659. URL https://doi.org/10.1109/TGRS.2023.3251659.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. Segformer: Simple and efficient design for semantic segmentation with transformers. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 12077–12090, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html.

Yan, W., Sharma, A., and Tan, R. T. Optical flow in dense foggy scenes using semi-supervised learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13256–13265, 2020. doi: 10.1109/CVPR42600.2020.01327.

Yang, W., Tan, R. T., Wang, S., Fang, Y., and Liu, J. Single image deraining: From model-based to data-driven and beyond. *CoRR*, abs/1912.07150, 2019. URL http://arxiv.org/abs/1912.07150.

Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., and Lu, J. 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)*, pp. 19313–19322, June 2022.

Zhang, H., Zhang, P., Hu, X., Chen, Y., Li, L. H., Dai, X., Wang, L., Yuan, L., Hwang, J., and Gao, J. Glipv2: Unifying localization and vision-language understanding. *CoRR*, abs/2206.05836, 2022. doi: 10.48550/arXiv.2206.05836. URL https://doi.org/10.48550/arXiv.2206.05836.

Zhu, X., Zhang, R., He, B., Zeng, Z., Zhang, S., and Gao, P. Pointclip V2: adapting CLIP for powerful 3d open-world learning. *CoRR*, abs/2211.11682, 2022. doi: 10.48550/arXiv.2211.11682. URL https://doi.org/10.48550/arXiv.2211.11682.