PanopticCAP: Refined and Enriched Physics-based Caption for Unseen Dynamic Changes

Hidetomo Sakaino sakain@wni.com
AI Image Group, Transportation Weather Lab., Weathernews Inc.

weathernews
Always WITH you!

ICML
International Conference On Machine Learning

Abstract

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

Low quality images to be rejected by proposed Danomal

Lens reflection Strong headlight Raindrops

Motivation

SOTA VLM failure cases

car crash Flooding Landslide
road flow Foggy strong light

SOTAs can't correctly caption under Dynamic Disaster Scene: flooding, fog, rain, landslide, and car crash.

Evaluation

Using BLEU score to compare on test dataset 2: two collected dataset, i.e., Disaster with 1850 image, and Traffic accident with 2130 images.

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

PanopticCAP has outperformed Visual ChatGPT.

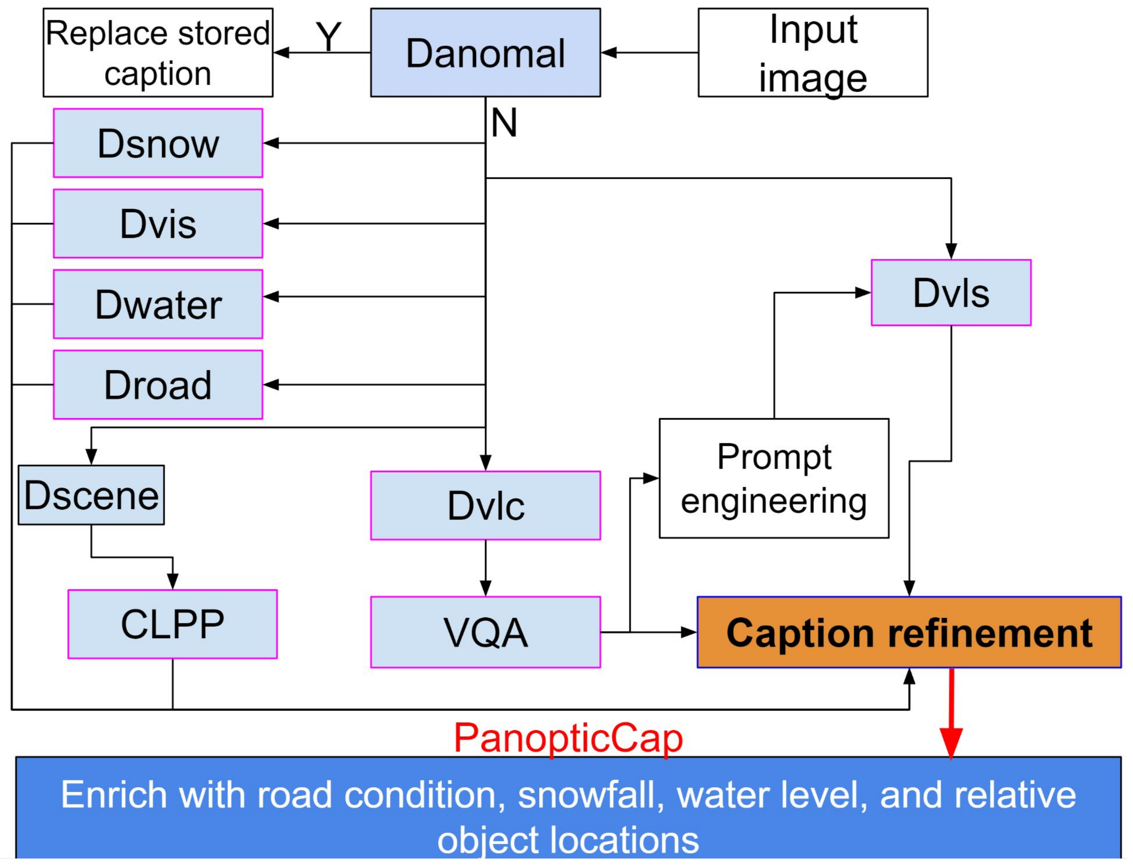
Contributions

1) PanopticCAP with multiple Deep Learning models
2) Combination of Deep Visual Lang. Seg. and Class.
3) The first time to contain dynamic changes with physical scales, i.e., depth, size, visibility, and water level.
4) Captions with 3D-related adverbs, i.e., behind, rear, in front of, and far, enable to generate as SOTAs have used 2D-related adverbs, i.e., left and right..
5) More quantitative texts for auto-driving and rescue workers from camera images

Conclusion

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

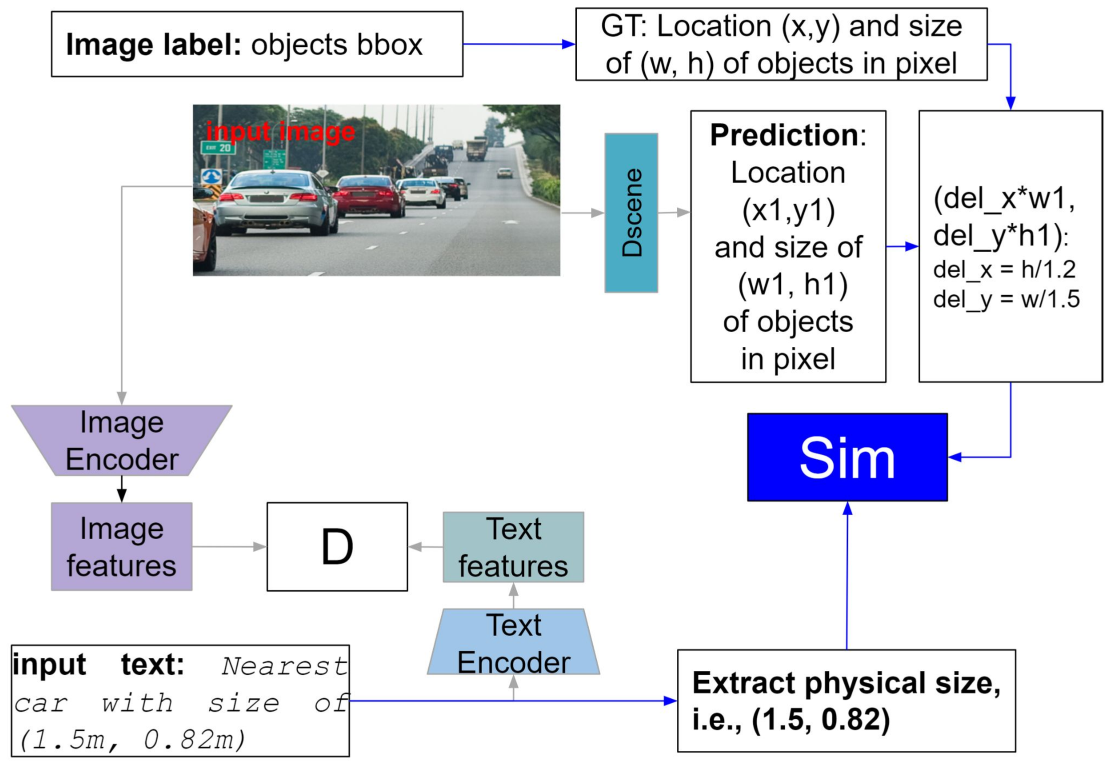
Proposed PanopticCAP



Caption refinement with multi-DL models, combine physical constraints:

- **Multi-DL models:**
  - DeepSnow (Dsnow): snow status detection
  - DeepVis (Dvis): visibility estimation
  - DeepRoad (Droad): road condition evaluation
  - Deep vision-language classification (Dvlc): v-l based classifier
  - Deep vision-language segmentation (Dvls): v-l based segmentator
  - DeepScene (Dscene): semantic segmentation.
- **Physical constraint: object sizes and locations.**

Contrastive Language Physical-Scale Pretraining (CLPP)

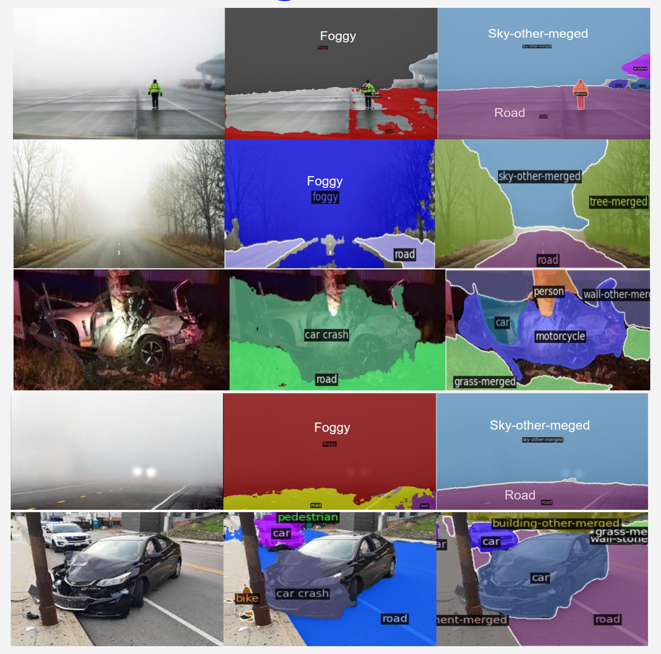


Custom loss function with new physical constraints

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

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

Refined segmentation and classification

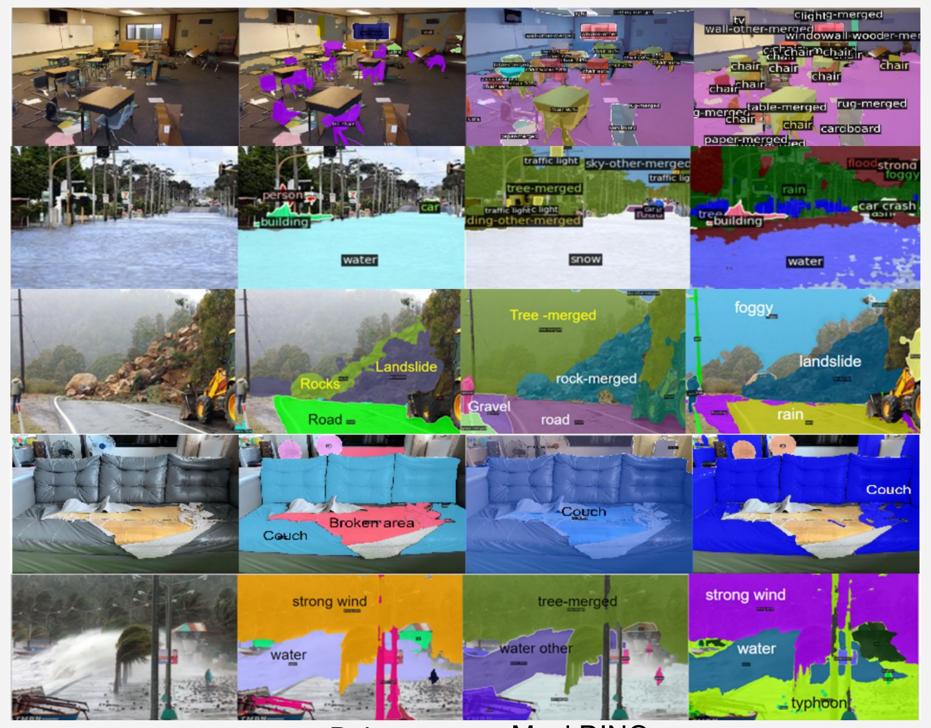


Input image PanopticCAP Mask2Former

Classes/ model	Dvlc (%)	CLPP (%)	ViT (%)	Resnet101 (%)	Vgg19 (%)
Car crashes	**93.37**	92.54	92.41	91.12	87.67
Flooding	**90.69**	90.02	89.23	87.83	86.54
Fog	**92.98**	89.56	91.19	86.77	85.23
Landslide	**89.52**	87.56	87.63	87.19	84.89
Rain	**92.33**	89.67	87.58	88.92	83.11
Normal	**94.67**	90.46	92.57	91.23	84.02
Average	***91.78***	*88.47*	*89.61*	*88.37*	*85.49*

Dvlc outperforms other models

Caption Refinement by Dvls

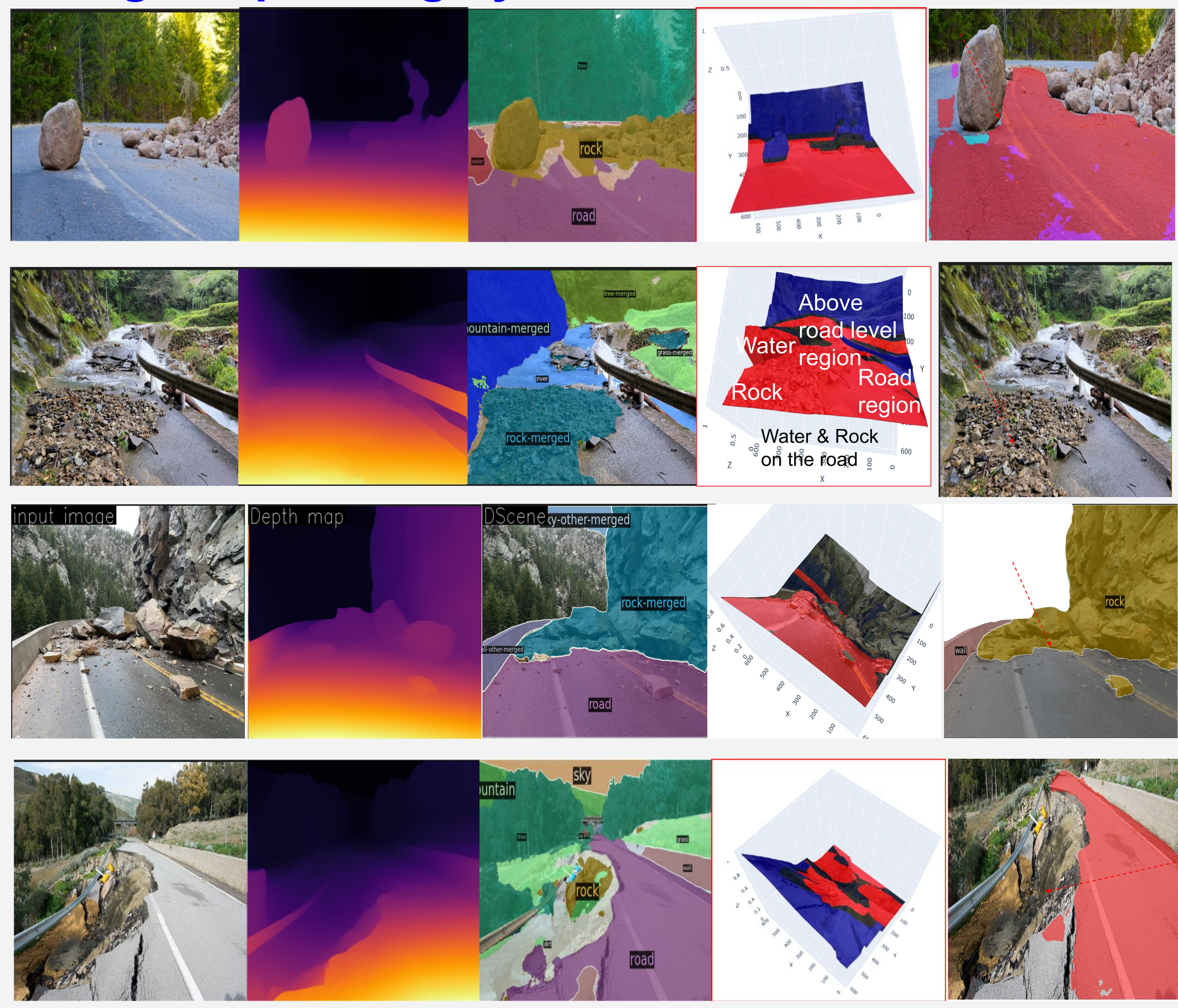


Input image Dvls MaskDINO OVSeg

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

Dvls can segment car crash, strong wind, landslide. SOTA only can segment general objects.

Image Captioning by CLPP



Rocks on the road

Rocks and water on the road

Rocks on the road

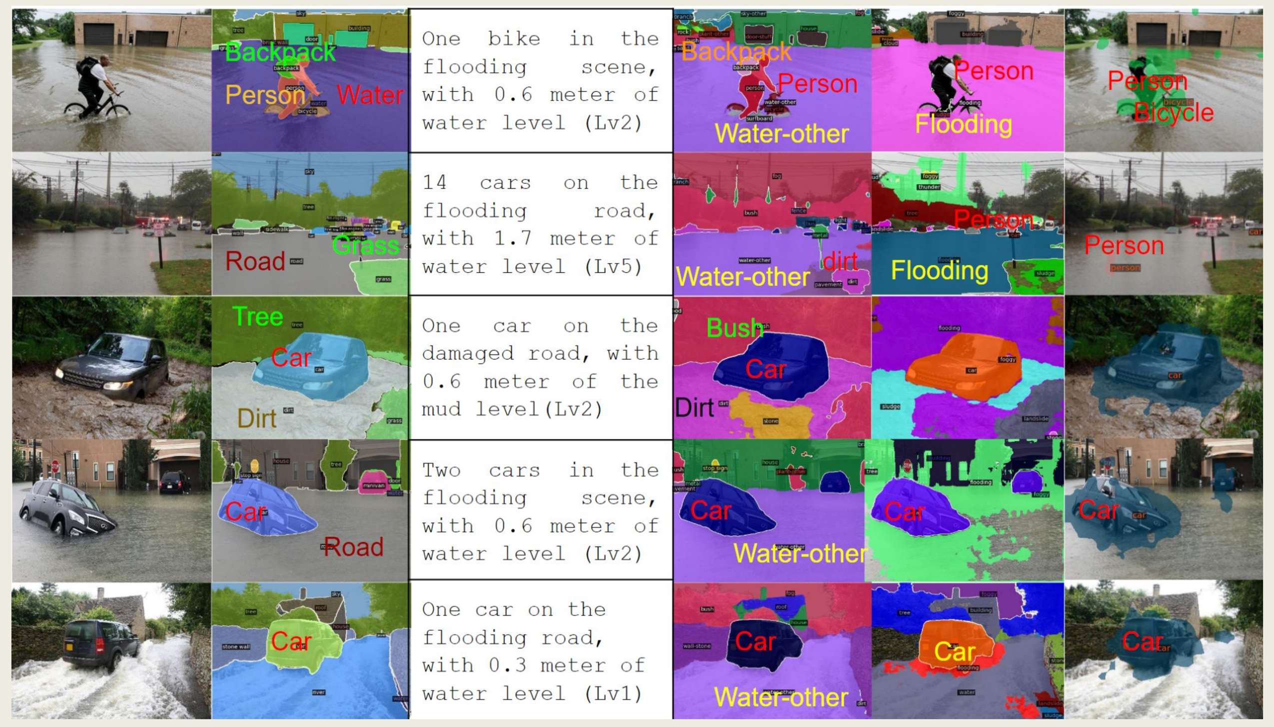
Collapsed road with a sudden narrowed road-width

Input image DeepDepth DeepScene 3D Depthmap separated road Physical constraints

References

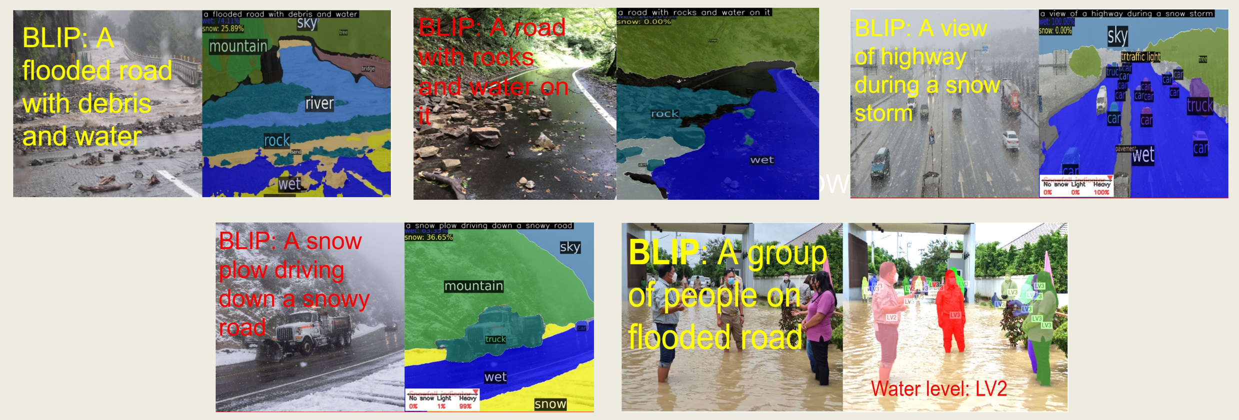
[1] H. Sakaino, 'PanopticVis: Integrated Panoptic Segmentation for Visibility Estimation at Twilight and Night', CVPR 2023. [2] F. Liang, et al., "OvSeg: Open-vocabulary semantic segmentation with mask-adapted clip," arXiv preprint arXiv:2210.04150, 2022. [3] B. Cheng, et al., "Masked-attention mask transformer for universal image segmentation," CVPR 2022. [4] Z. Zhou, et al., "Zegclip: Towards adapting clip for zero-shot semantic segmentation," arXiv preprint arXiv:2212.03588, 2022. [5] Y. H. Chen, et al, "Multi-scales feature extraction model for water segmentation in the satellite image," ICCE 2023. [6] F.Li, et al., "Towards A Unified Transformer-based Framework for Object Detection and Segmentation", CVPR 2022. [7] J. Li, et al., "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," CVPR 2022. [8] Xu, et al., "GroupViT: Semantic Segmentation Emerges from Text Supervision", CVPR 2022. [9] Z. Wang, et al., "Clip-driven referring image segmentation', CVPR 2022.

Proposed PanopticCAP vs. SOTA VMLs



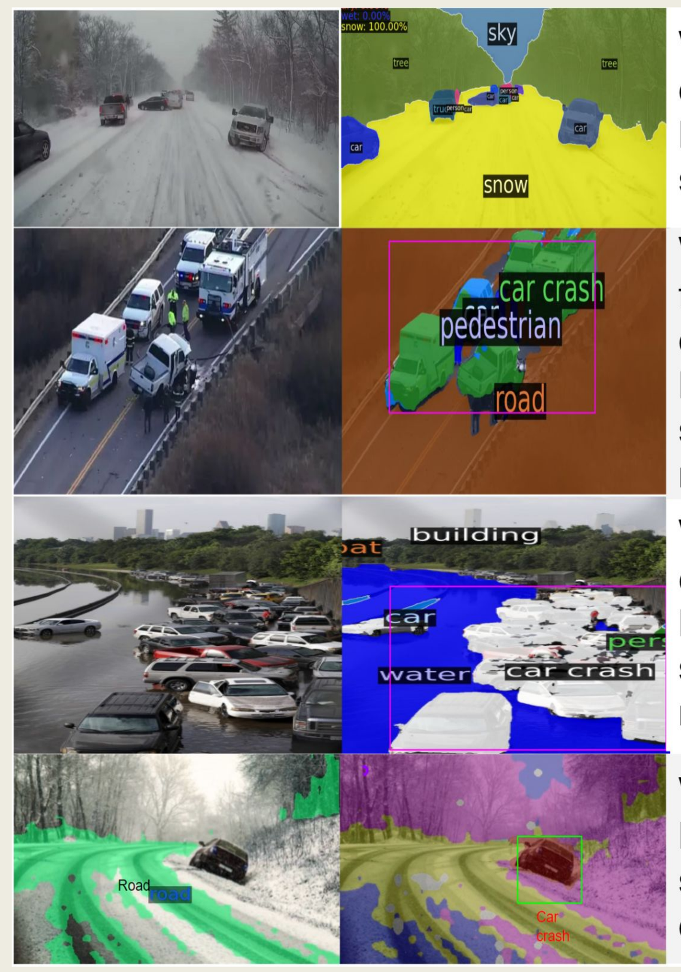
Input PanopticCAP ZegFomer OvSeg ClipSeg

The caption results of PanopticCAP contain physical scales, i.e., the number of vehicles, visibility and water level in meter.



	Proposed method	BLIP [216]
(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

Refined road conditions by proposed PanopticRoad



Visual ChatGPT: a car that is upside on the side of the road
Final refined caption: car crash with size of (1.7m, 1.3m) and is off road.

Visual ChatGPT: cars are stuck in a flooded parking lot in houston, texas, on monday, june 6
Final refined caption: car crash with size of (16.25m, 8.92m) and is on road

Visual ChatGPT: the scene of a fatal crash on the highway in the state
Final refined caption: car crash with size of (5.3m, 26.1) on water and side road.

Visual ChatGPT: a road with car on it
Final refined caption: car crash with size of (1.52m, 1.25m) is on the side of road.

The caption results of PanopticCAP contain physical scales, i.e., the number of vehicles, visibility and water level in meter.

PanopticCAP: Refined and Enriched Physics-based Caption for Unseen Dynamic Changes

Hidetomo Sakaino sakain@wni.com

AI Image Group, Transportation Weather Lab., Weathernews Inc.

weathernews Always WITH you!

ICML International Conference On Machine Learning

Abstract

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

Low quality images to be rejected by proposed Danomal

Lens reflection Strong headlight Raindrops

Motivation

SOTA VLM failure cases

car crash Flooding Landslide

road flow Foggy strong light

SOTAs can't correctly caption under Dynamic Disaster Scene: flooding, fog, rain, landslide, and car crash.

Evaluation

Using BLUE score to compare on test dataset 2: two collected dataset, i.e., Disaster with 1850 image, and Traffic accident with 2130 images.

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

PanopticCAP has outperformed Visual ChatGPT.

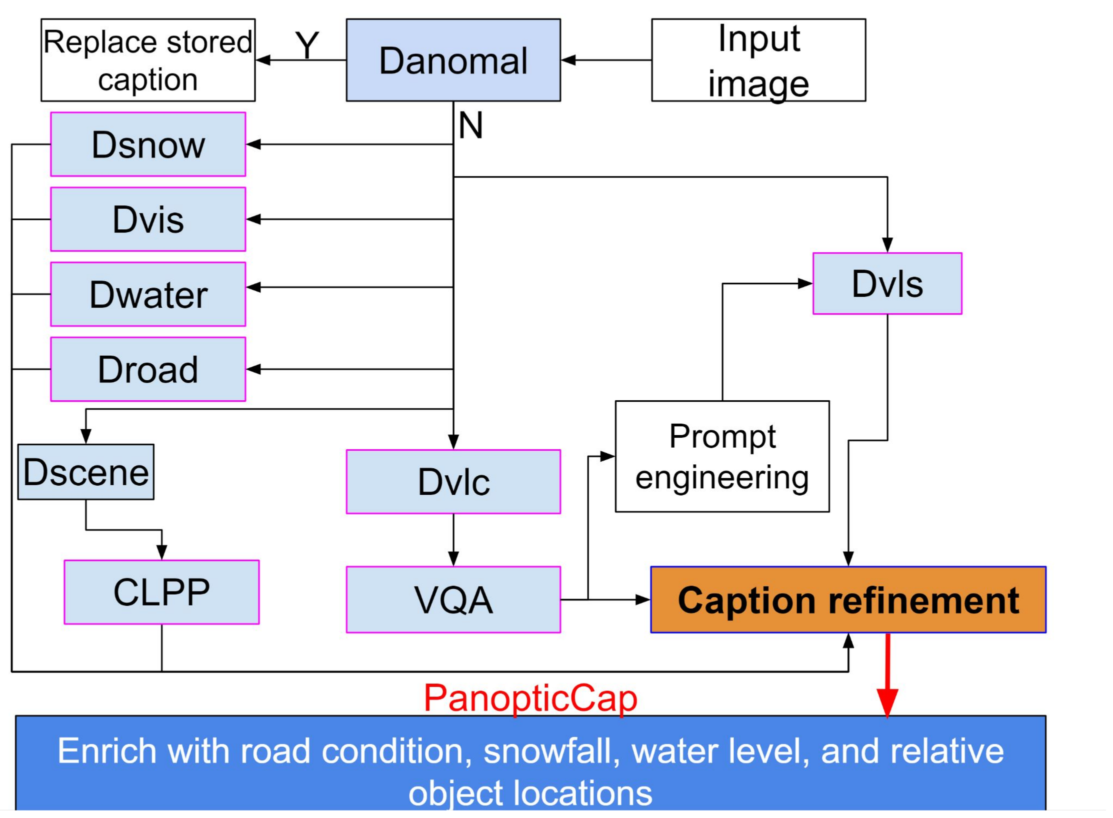
Contributions

1) PanopticCAP with multiple Deep Learning models
2) Combination of Deep Visual Lang. Seg. and Class.
3) The first time to contain dynamic changes with physical scales, i.e., depth, size, visibility, and water level.
4) Captions with 3D-related adverbs, i.e., behind, rear, in front of, and far, enable to generate as SOTAs have used 2D-related adverbs, i.e., left and right..
5) More quantitative texts for auto-driving and rescue workers from camera images

Conclusion

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

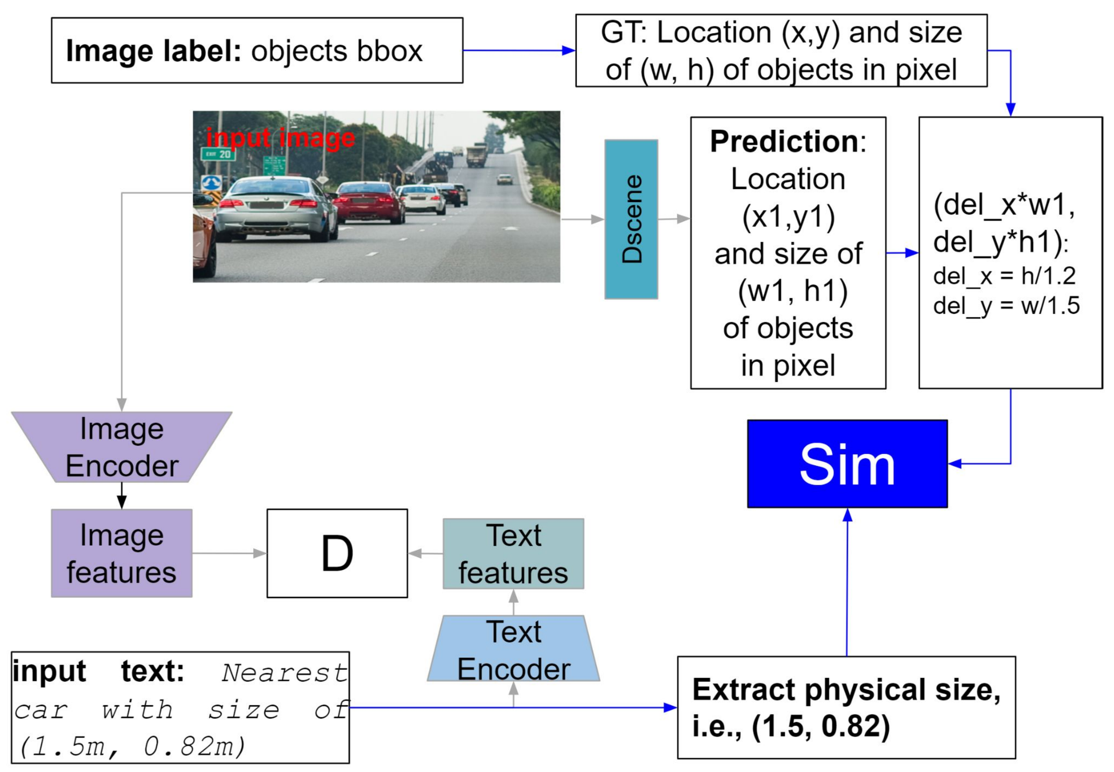
Proposed PanopticCAP



Caption refinement with multi-DL models, combine physical constraints:

- **Multi-DL models:**
  - DeepSnow (Dsnow): snow status detection
  - DeepVis (Dvis): visibility estimation
  - DeepRoad (Droad): road condition evaluation
  - Deep vision-language classification (Dvlc): v-l based classifier
  - Deep vision-language segmentation (Dvls): v-l based segmentator
  - DeepScene (Dscene): semantic segmentation.
- **Physical constraint: object sizes and locations.**

Contrastive Language Physical-Scale Pretraining (CLPP)

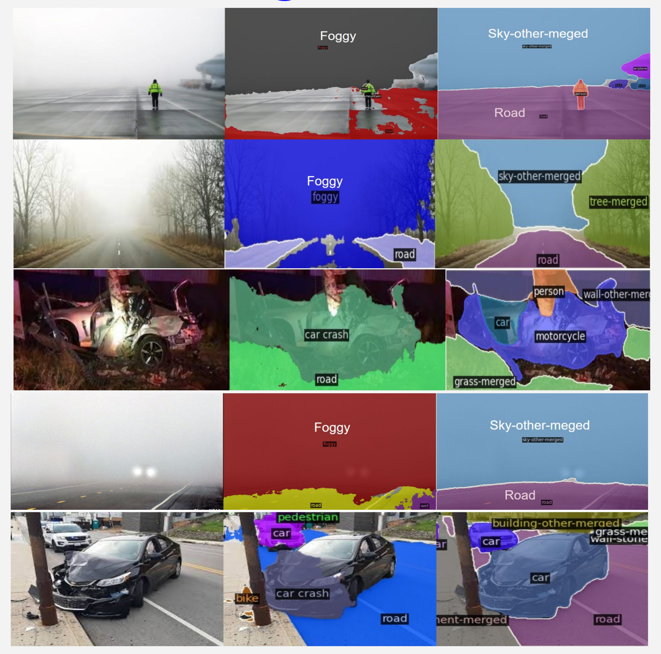


Custom loss function with new physical constraints

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

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

Refined segmentation and classification

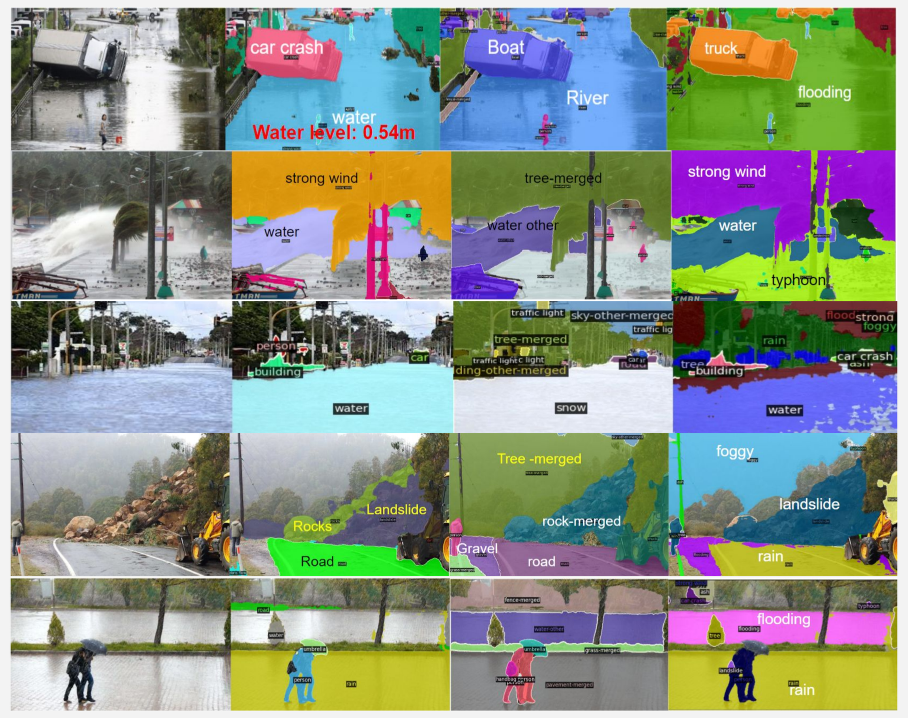


Input image PanopticCAP Mask2Former

Classes/ model	Dvlc (%)	CLPP (%)	ViT (%)	Resnet101 (%)	Vgg19 (%)
Car crashes	**93.37**	92.54	92.41	91.12	87.67
Flooding	**90.69**	90.02	89.23	87.83	86.54
Fog	**92.98**	89.56	91.19	86.77	85.23
Landslide	**89.52**	87.56	87.63	87.19	84.89
Rain	**92.33**	89.67	87.58	88.92	83.11
Normal	**94.67**	90.46	92.57	91.23	84.02
Average	***91.78***	*88.47*	*89.61*	*88.37*	*85.49*

Dvlc outperforms other models

Caption Refinement by Dvls

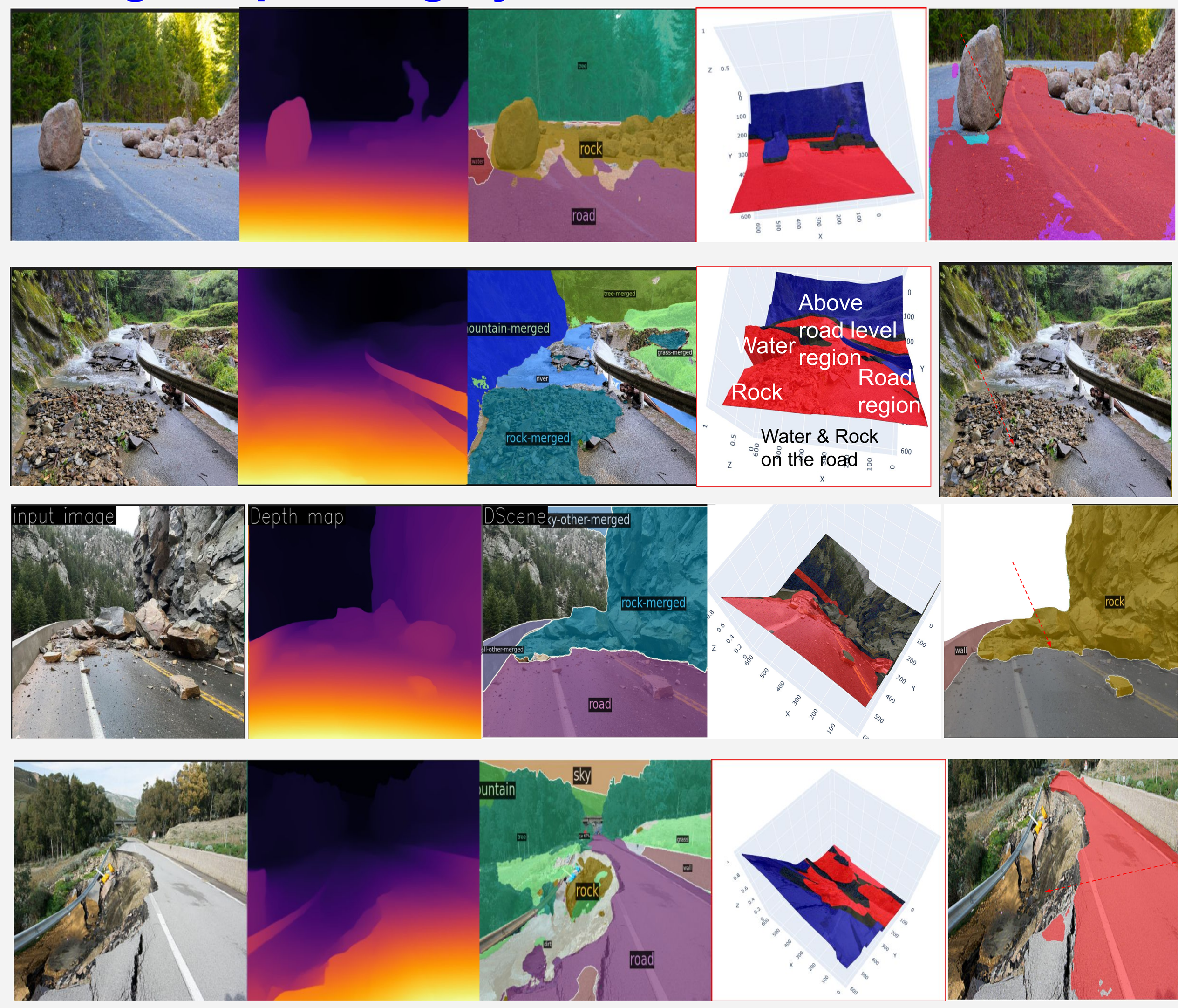


Input image Dvls MaskDINO OVSeg

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

Dvls can segment car crash, strong wind, landslide. SOTA only can segment general objects.

Image Captioning by CLPP



Rocks on the road

Rocks and water on the road

Rocks on the road

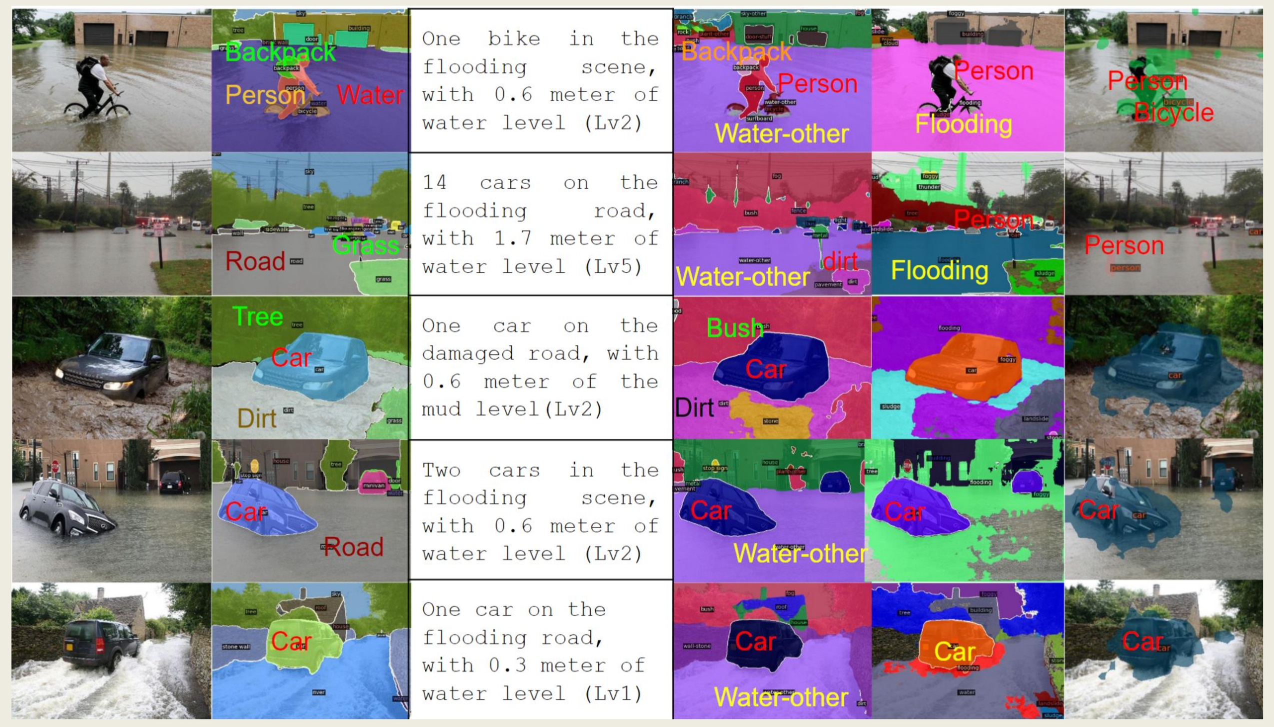
Collapsed road with a sudden narrowed road-width

Input image DeepDepth DeepScene 3D Depthmap separated road Physical constraints

References

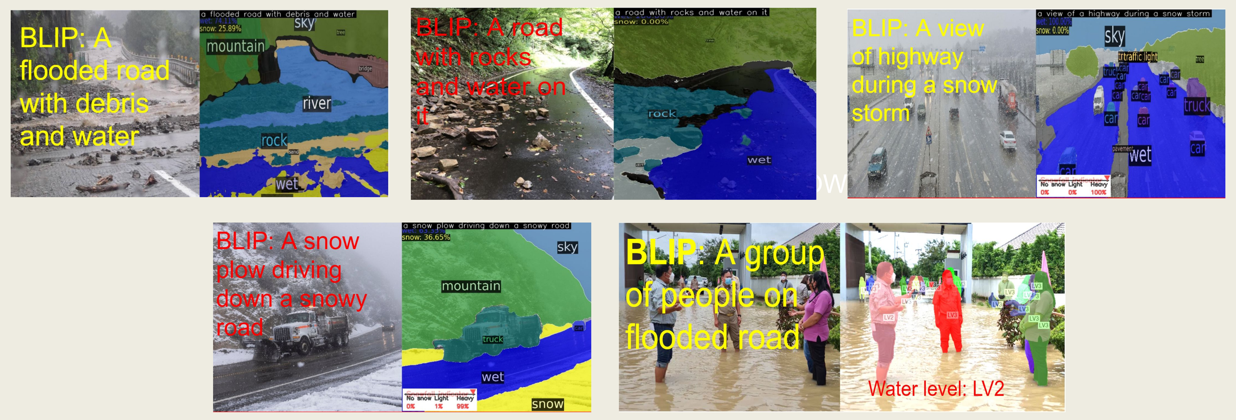
[1] H. Sakaino, 'PanopticVis: Integrated Panoptic Segmentation for Visibility Estimation at Twilight and Night', CVPR 2023. [2] F. Liang, et al., "OvSeg: Open-vocabulary semantic segmentation with mask-adapted clip," arXiv preprint arXiv:2210.04150, 2022. [3] B. Cheng, et al., "Masked-attention mask transformer for universal image segmentation," CVPR 2022. [4] Z. Zhou, et al., "Zegclip: Towards adapting clip for zero-shot semantic segmentation," arXiv preprint arXiv:2212.03588, 2022. [5] Y. H. Chen, et al, "Multi-scales feature extraction model for water segmentation in the satellite image," ICCE 2023. [6] F.Li, et al., "Towards A Unified Transformer-based Framework for Object Detection and Segmentation", CVPR 2022. [7] J. Li, et al., "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," CVPR 2022. [8] Xu, et al., "GroupViT: Semantic Segmentation Emerges from Text Supervision", CVPR 2022. [9] Z. Wang, et al., "Clip-driven referring image segmentation', CVPR 2022.

Proposed PanopticCAP vs. SOTA VMLs



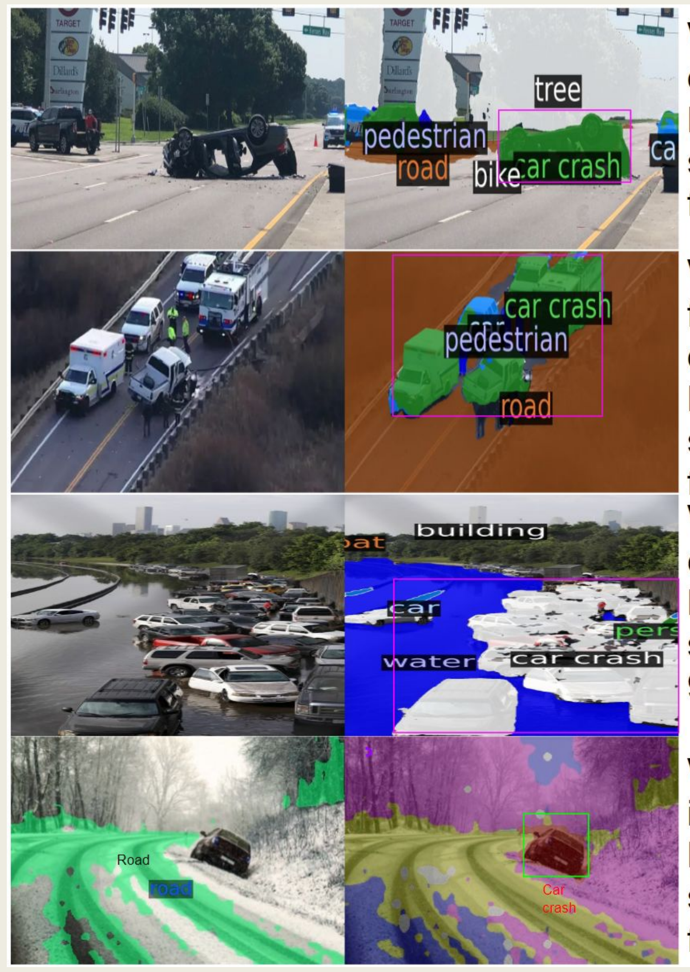
Input PanopticCAP ZegFomer OvSeg ClipSeg

The caption results of PanopticCAP contain physical scales, i.e., the number of vehicles, visibility and water level in meter.



	Proposed method	BLIP [216]
(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

Refined road conditions by proposed PanopticRoad



Visual ChatGPT: a car that is upside on the side of the road
Final refined caption: car crash with size of (1.7m, 1.3m) with 11.2m far from camera.

Visual ChatGPT: cars are stuck in a flooded parking lot in houston, texas, on monday, june 6
Final refined caption: car crash with size of (16.25m, 8.92m) with 16.3m far from camera.

Visual ChatGPT: the scene of a fatal crash on the highway in the state
Final refined caption: car crash with size of (5.3m, 26.1) on water with 9.1m far from camera

Visual ChatGPT: a road with a car on it
Final refined caption: car crash with size of (1.52m, 1.25m) with 21.7m far from camera.

The caption results of PanopticCAP contain physical scales, i.e., the number of vehicles, visibility and water level in meter.