PanopticRoad: Integrated Panoptic Road Segmentation Under Adversarial Conditions

Hidetomo Sakaino
Weathernews Inc.
sakain@wni.com

Abstract

Segmentation becomes one of the most important methods for scene understanding. Segmentation plays a central role in recognizing things and stuff in a scene. Among all things and stuff in a scene, the road guides vehicles in the cities and highways. Most segmentation models, i.e., semantic, instance, and panoptic segmentation, have focused on images with clear daytime weather conditions. Few papers have tackled nighttime vision under adversarial conditions, i.e., fog, rain, snow, strong illumination, and disaster events. Moreover, further segmentation of road conditions like dry, wet, and snow is still challenging under such invisible conditions. Weather impacts not only visibility but also roads and their surrounding environment, causing vital disasters with obstacles on the road, i.e., rocks and water. This paper proposes PanopticRoad with five Deep Learning-based modules for road condition segmentation under adversarial conditions: DeepReject/Scene/Snow/Depth/Road. Integration of them helps refine the failure of local road conditions where weather and physical constraints are applied. Using foggy and heavy snowfall nighttime road images and disaster images, the superiority of PanopticRoad is demonstrated over state-of-the-art panoptic-based and adaptive domain-based Deep Learning models in terms of stability, robustness, and accuracy.

1. Introduction

Semantic scene understanding under various weather conditions is important for monitoring and auto-driving. However, most semantic models have focused on clear daytime weather conditions in Computer Vision [89] and Deep Learning [1, 2, 3, 4, 5, 6, 13, 18, 19, 20]. In such weather conditions, city and highway scenes have been selected to recognize and evaluate various objects, such as buildings, traffic signals, vehicles, and pedestrians. Therefore, people's daily normal activities are monitored. On the other hand, camera images/videos inevitably must deal with weather changes like rain, snow, and fog. These weather phenomena can dramatically impact scene appearance changes over time. Moreover, sunbeams, rainfall, snowfall, and fog can degrade recognition and classification rates. More complicated scenes can happen by a mix of them and illuminations at twilight and night. For camera images, these factors are assumed to be adversarial visual conditions. In particular, road scene images are more complicated due to a mix of fog and adversarial factors.

The representative metric is visibility levels or distances between a camera and a far location. What is worse for visibility is darkness or low illumination at night. Therefore, the most important landmarks, as seen in the daytime, may be readily lost in the nighttime road environment. Previously, computer vision-based visibility estimation methods with edge detection and geometrical coordinate have been proposed [10, 11, 12, 15, 16, 17, 21, 22]. However, they are known to be vulnerable to illumination changes.

In Deep Learning (DL) models [20, 34, 63, 71], semantic segmentation [1, 39, 46, 49, 74] and instance segmentation [41, 42, 44] have been reported and used for recognizing things or/and stuff [65]. Panoptic segmentation [51, 52, 53, 64, 81, 82, 88] handles stuff and thing classes by fusing subregions by semantic and instance segmentation, providing a unique class label for each pixel in the image and instance IDs for countable objects. Panoptic segmentation is an important step towards scene understanding in autonomous vehicles since it provides object masks and interesting amorph regions like drivable road space or sidewalks [47]. Although video-based panoptic segmentation models [40, 43, 48, 50, 51, 54, 55, 56, 57, 58, 59, 60, 62, 64, 67, 68, 70, 72, 75, 76, 101, 73] have recently shown a new avenue to enhance accuracy, they require a temporally smooth change over time. Therefore, they are limited to applying to low frame rates, sudden changes of a moving camera [36, 37, 38], and snowfall changes.

Adversarial visual factors significantly degrade the accuracy of state-of-the-art (SOTA) DL-based segmentation. Raindrops [25, 32] are removed for better visibility. Defog and Dehaze [5, 23, 30, 31, 35, 77] are shown, but no visibility estimation. However, most papers have synthesized raindrops, rain streaks, and fog to obtain nearly perfect original daytime images under uniform illumination [8]. SOTA DL models are easy to fail in applications of real foggy scenes due to the non-uniformity of fog and rainfall, ambient illumination, halo effect, and motion in depth [6, 7]. In dark scenes, night vision [24, 28, 29] is a challenging topic due to low light and less visible landmarks available. Although night-to-day translation by GAN [26, 27] may enhance far landmarks to estimate visibility levels, real nighttime images are converted to false color images due to strong headlight, spotlighting, and fog image gradients. Therefore, as the visibility estimation task, DL models have not thoroughly explored images at foggy twilight and night. Moreover, few segmentation papers have explored visibility estimation. Physical distance and level of visibility by the DL model remain undone. An all-in-one image restoration model [78] is reported with no manual selection of difficult scenes for multiple tasks with adversarial conditions and visibility estimation.

Evaluation image datasets [84, 85, 86, 101] are important but very limited to scenes with clear, synthetic fog and real lighter fog [9, 84, 85, 86], where no or fewer adversarial conditions contain. Unlike rainfall and fog, snowfall [93] causes other difficult issues in visibility and road conditions. Snowfall can cover and accumulate on the road, causing icy barns and raising accident risks. Even a small amount of snowflakes significantly degrades visibility mixed with fog. Therefore, further segmentation is required for dry, wet, and snow road conditions [94].

Most SOTA DL-based segmentation models [79, 80, 81, 82, 83, 99, 100, 95] are limited to segmenting the road's details, i.e., conditions or statuses. In contrast, this paper proposes a pixel-based road condition segmentation method using DL models. Disaster scenes [92] with heavy rainfall and snowfall have been increasing, which may cause a chain reaction of natural disasters observed from the satellite images [95], i.e., landslides and flooding [91, 95, 96]. However, camera image-based post-disaster object recognition for dirt, water, and rocks remains unsolved on the road. Such stuff and objects may occlude the road surface, losing the normal road. Since domain adaptation segmentation DL-models [99, 100] require manual selection of the optimal pretrained model, they are not useful for unpredictable and sudden scene changes by disaster and weather conditions. Therefore, real heavier foggy night images with adversarial conditions and disaster images have not been fully publicly available, as this paper uses.

To this end, this paper proposes PanopticRoad: integrated panoptic road condition segmentation under adversarial visual conditions using single images. Multiple transformer-based Deep Learning (DL) models, i.e., DeepX, with branched structures are integrated for efficiency in light of memory, training, and maintenance. This paper's contributions are fourfold:

1. Multiple DL architecture with five independent DL modules is proposed for efficient model enhancement and maintenance. In order to stabilize the overall recognition system, DeepReject rejects difficult images with darkness and lenz reflection. SOTA DL, i.e., OneFormer [33], has not considered this concept yet. DeepSnow classifies snowfall among no-snow, light snowfall, and heavy snowfall. DeepScene is panoptic segmentation. DeepDepth estimates the depth map. DeepRoad recognizes road conditions.

2. Refinement to segmented regions is proposed. Integration of DeepScene, DeepDepth, and DeepRoad helps refine the failure of local road conditions due to incomplete segmentation by each DL module. Therefore, weather constraint is posed so that initial road conditions-based segmentation is refined by merging and changing, i.e., partial wet to fully covered snow (frozen) based on the surrounding snow. Moreover, identifying road locations is important to estimate for correct road condition decisions when a disaster causes occluded road images with many obstacles, i.e., dirt and rocks. Dirt and rocks may be on the slope (vertical) or road (on the ground). Therefore, DeepDepth and DeepScene are used to identify the location of pre-disaster normal roads under physical constraint, i.e., relative heights among roads, slopes, and cliffs. It is the first time to recognize the obstacles on the road by combining segmentation, depth map, and 3D cloud points under physical constraints, unlike SOTA without such constraints.

3. Novel foggy day and night road images, i.e., dry, wet, and snow, and post-disaster images have been collected since publicly available image datasets, i.e., Cityscapes [84], Foggy Cityscape [85], and Foggy Zurich [86] are insufficient to train and test.

4. Using foggy and heavy snowfall nighttime road images, the superiority of PanopticRoad for road conditions is demonstrated over SOTA panoptic-based and adaptive domain-based DL models in terms of stability, robustness, and accuracy. Moreover, obstacles on the road are recognized using 3D cloud points and 3D RANSAC [104].

2. Related work

This section briefly describes review methods and issues in scene understanding of camera images under various conditions. Visibility levels are one of the most important visual factors to estimate for monitoring and auto-driving. Weather conditions with sunbeams, rainfall, snowfall, fog, and haze impact visibility. Strong illumination like headlights, street lights, or darkness changes can also be added. A mix of these factors can lead to a worse visual condition. To estimate visibility, near and far objects can be landmarks. Such objects may be obtained from segmentation. Dehaze [5, 23, 30, 31, 35, 77], denoise, and derain [25, 32] may be useful to enhance such landmarks.

Although considerable progress has been made in semantic segmentation understanding under clear weather, it is still a tough problem under adversarial weather conditions, such as heavy fog and snowfall, due to the uncertainty caused by imperfect observations. SOTA segmentation models have become robust to the partial appearance of objects. However, they are stable mainly when opaque objects are occluded from each other. On the other hand, such natural phenomena pose a different challenge due to semi-transparent image features, i.e., stuff. This problem [77] has been alleviated by bridging the gap between clear and (light) foggy images, i.e., city scenes.

However, issues in foggy and heavy snowfall at night remain unsolved only by this model [77], and no visibility estimation under fog requires further modelization, unlike the proposed approaches we show. In image restoration [78], an all-in-one image restoration network (AirNet) for unknown corruption has been proposed. Almost all existing approaches could handle a specific degradation only, i.e., denoise, defog, deraining, and deblurring, where the user must know the correct corruption before applying a specific API. Since such degradations are rooted in natural phenomena, the degradation ratio can vary in space and time, letting the user retune manually. In [78], although AirNet experimentally shows superiority in three degradation factors with noise, rain, and haze (fog), at least only lighter fog has been used in the daytime scenes.

The monocular geometric scene understanding task combined with panoptic segmentation and self-supervised depth estimation has been reported as MGNet [79]. However, no adversarial weather conditions are shown, i.e., heavy fog. Moreover, the depth map may lose a lot of landmarks due to lower brightness at twilight and night. To enhance previous semantic segmentation problems, Deep hierarchical semantic segmentation (HSS) has been proposed in city scenes [80]. By exploiting hierarchy properties as optimization criteria, hierarchical violation in the segmentation predictions can be explicitly penalized. However, no physical scales of different semantic segmentation have been considered, like depth ordering from near to far objects along the road, i.e., multiple vehicles and pedestrians.

The proposed method [81] combines the global modeling capability of the Transformer and the local representation capability of CNN with transmission-aware 3D position embedding. However, dehazing in [81] is limited to closer views of daytime lighter foggy scenes, i.e., indoor and garden, unlike our proposed method for distant scenes with heavy fog at night, i.e., highway.

A unified framework for depth-aware panoptic segmentation (DPS) has been reported [82], aiming to reconstruct 3D scenes with instance-level semantics from one image. In contrast to previously predicting depth values for all pixels at a time, DPS manages to estimate depth for each thing/stuff instance, which also shares the way of generating instance masks. 3D cloud point images are generated. Domain adaptation segmentation [99, 100] is recently reported to refine locally insufficient segmentation. However, pretrained models are required to select manually based on target images. Therefore, they are hard to apply to images with unpredictable natural phenomenon changes.

This paper challenges dealing with road conditions even under adversarial nighttime snowfall conditions by the proposed PanopticRoad with multiple task-oriented Deep Learning models.

3. Proposed Methods

This section discusses the proposed PanopticRoad method/system for recognizing and classifying many road conditions under various adversarial conditions. Instead of recognization by a single-tasked DL, this paper integrates five proposed DL modules: DeepReject, DeepSnow, DeepScene, DeepDepth, and DeepRoad. As shown in Figure 1, a single image is an input with a city, highway, or mountain road. DeepReject may reject adversarial images. If rejected, past road status will be replaced. If heavy snow occurs, DeepSnow rejects and outputs the message. Images with light snowfall and no snowfall are used. If there is no rejection, an image goes into the three branches. DeepScene, DeepRoad, and DeepDepth are for panoptic segmentation, segmentation-based initial road condition, and depth map/3D cloud points, respectively.

In order to refine initial road conditions, such three modules are integrated, where weather and physical constraints are applied. These constraints may boost the incomplete segmentation of each DL module, like wet to snow condition, road class to snow condition, and giving classes to no segmentation region. Moreover, it may benefit from estimating whether obstacles of rocks and dirt are on the road or the slope due to the lower location in the 3D coordinates. Therefore, the refined road conditions will be estimated. The following explains each of the five Deep Learning modules: DeepReject, DeepSnow, DeepScene, DeepDepth, and DeepRoad further.

3.1. DeepReject

Roads at night pose several challenging factors, i.e., darkness. To identify and reject adversarial images, as shown in Figure 2, an algorithm to reject such images is proposed to avoid the degradation of the cascaded other recog-

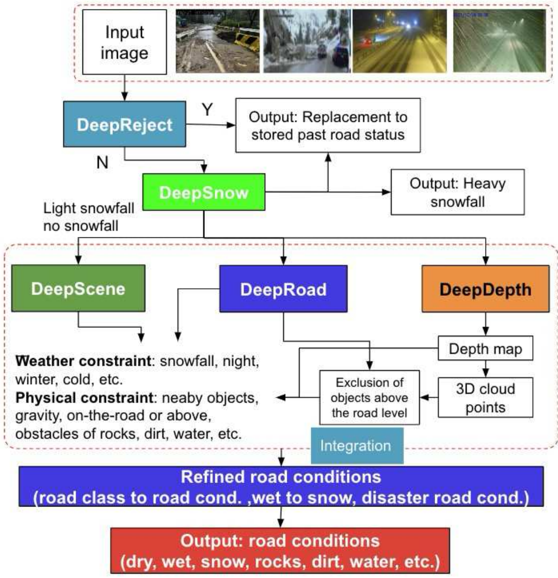
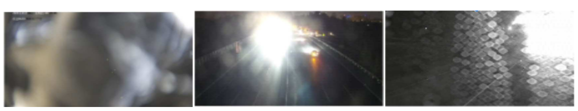


Figure 1. Proposed PanopticRoad model

nition modules. Such factors have been pre-analyzed using many city and highway images with different cameras. Therefore, three major adversarial image patterns have been selected: a) lens reflection, b) strong headlight, and c) raindrops. These adversarial images were collected from over 2500 images and used to train by Swin Transformer [4] into 2 classes: accept and reject.

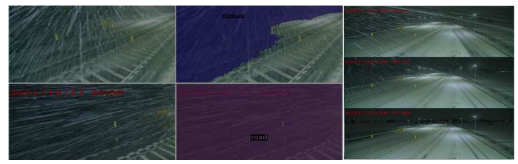


(a) (b) (c)

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

3.2. DeepSnow

Snowfall often appears in the scene of the winter season. Even with light snowfall, low visibility occurs in images. Such snowfall patterns can partially or overall occlude road surface view. Therefore, DeepScene can fail to recognize the road whenever heavy snowfall happens, as shown in Figure 3 (a). To detect snowfall, DeepSnow is proposed to apply, where no snowfall, light, and heavy snowfall are classified. Such images were collected from various countries' daytime and nighttime surveillance cameras. No video frames were used to detect snowfall. The snowfall classification model is based on pre-trained EfficientNet [14] with an input size of 512 × 512. Figure 3 (b) shows examples of heavy, light, and no snowfall. The global average is then applied to the output of the pre-trained EfficientNet followed by 512 units dense layer with ReLU activation [14]. To avoid the overfitting problem, this paper utilizes several augmentation techniques such as random brightness, random contrast, random translation, random horizontal flip, and random rotation. A dropout layer with a rate of 0.4 is also applied to enhance the model's representation. After 500 training epochs with a batch size of 32 and a learning rate of 0.001, the model is used to classify 3 snowfall levels in the image.



(a) (b)

Figure 3. Various snowfall images: (a) segmentation failure cases under heavy snowfall events. (b) Examples of heavy snowfall (upper), light snowfall (middle), and no snowfall (bottom).

3.3. DeepScene

This paper proposes globally and locally segmented images/objects to enhance the accuracy of road condition classification. DeepScene plays an important role in globally segmenting scene objects. On the other hand, DeepRoad is used for locally segmented road condition classification, as mentioned below. Segformer [3] is trained with COCO image datasets [90] to segment outdoor objects like mountains, fences, roads, rivers, and rocks, as shown in Figure 4. It is noted that DeepScene recognizes road conditions when there are snow, flooding, rocks, and other disaster classes. DeepScene recognizes "road" as "dry" or "wet" conditions.

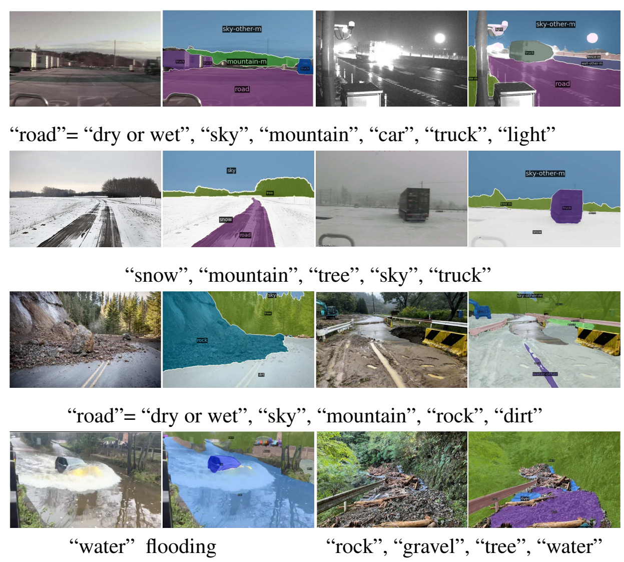


Figure 4. Various object detection segmented by DeepScene.

3.4. DeepRoad

To recognize road conditions, DeepRoad is proposed to apply. Particularly on winter roads, road conditions form from simple flat to complicated wheel-tread patterns, i.e., sherbet, frozen, pressed, and covered snow. In this paper, snow conditions are assumed to consist of a mix of sherbet, frozen, pressed, and covered snow. Wet is assumed to be from rainfall or melt of snow. DeepRoad recognizes three classes, dry, wet, and snow. It is noted that DeepScene also outputs “snow” by direct semantic segmentation. This paper integrates outputs from DeepRoad, DeepScene, and DeepDepth, as shown in the following sections. This integration will boost refinements of insufficient segmentation and road conditions under adversarial conditions. For example, a misrecognized wet class will be corrected to snow under weather constraints, and a far object will be better recognized. Swinformer [3] is trained from over 2500 winter road images. Since there are no publicly available annotation datasets, this paper created original 3-class road condition datasets from different country road images under adversarial weather conditions in different time zones.

3.5. DeepDepth

This section describes DeepDepth improved from a monocular depth method [87] using an RGB image. Such a depth map will be used to boost road conditions in several ways: road condition correction, road level estimation, and surface estimation. For example, in order to recognize road regions from a depth map, segmented objects by DeepScene are used to delete regions of the depth map corresponding to them. From this, an obstacle like a rock will be recognized as on the road under a physical constraint. A more detailed explanation of the experiments will be provided in Section 5.4.

4. Experiments and Discussion

4.1. Experimental result on DeepReject

This subsection evaluates the performance of DeepReject. The dataset comprises 3500 images with 3 different adversarial conditions: clear, lens reflection, strong light, and raindrops. In a comparative study, DeepRoad is applied to segment road conditions with and without DeepReject. Evaluation is conducted using all road condition classes. Table 1 shows that DeepReject can effectively reject images with adversarial conditions, where the accuracy using DeepReject becomes 86.0% better than 81.3% without DeepReject. No thresholding setting is required. Therefore, the proposed DeepReject has been proven useful in rejecting such three adversarial factors in images.

4.2. Experimental result on DeepSnow

Heavy snowfall can impede recognition of road surface. For this issue, this paper first considered the removal or rejection of snowfall. Raindrops and snowfall removal have recently been active research areas [25]. SOTA, Transweather [97], has been employed to compare its [97] performance using real images. Two heavy snowfall images

Table 1. Statistical analysis of DeepReject for adversarial conditions

	No DeepReject (%)	DeepReject (%)
Accuracy	81.3	86.0
Precision	72.9	78.2
Recall	75.5	80.6
F1 Score	74.0	79.3

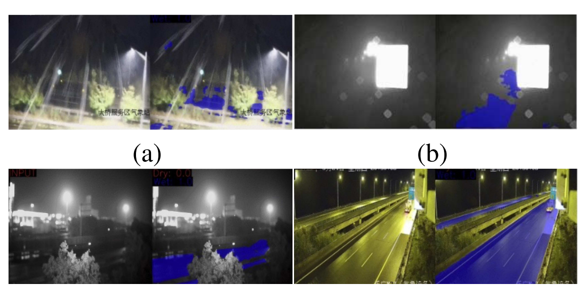
are shown in Figure 5 (a). (b) The proposed DeepSnow recognizes light and heavy snowfall. However, (c) SOTA [97] failed to remove overall snowfall patterns. Also, it cannot recognize light or heavy snowfall as the proposed DeepSnow. Therefore, this paper has applied DeepSnow to utilize the status of snowfall, where no snowfall and light snowfall images will be used for road conditions.

(a) (b) (c)

Figure 5. Comparison of snowfall detection in real images: (a) original image. (b) “Light snowfall” and “Heavy snowfall” by DeepSnow. (c) Failure cases by Transweather.

4.3. Experimental results on DeepRoad

This section conducts experiments of DeepRoad on adversarial night highway scenes. As shown in Figure 6, heavy rainfall, strong reflection from the traffic board, low lighting, and reflection from the road images are used for road conditions. Results show wet conditions in blue by DeepRoad. It is noted that such images with heavy rainfall, raindrops on lenz, and low illumination have not been rejected by DeepReject. Using 3080 images with day and night, 86.1% accuracy has been evaluated. Therefore, it has been proven that DeepRoad is useful for night-time road condition recognition.

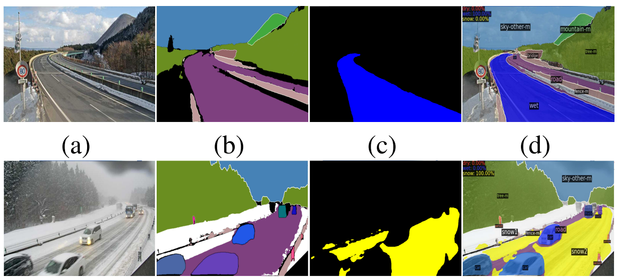


(a) (b) (c) (d)

Figure 6. Results of DeepRoad with wet road conditions in blue under adversarial images: heavy rainy and foggy night images.

4.4. Refinement on DeepScene by DeepRoad

This section conducts refinement experiments by the proposed multiple Deep Learning modules. Figure 7 shows (a)-(d) wet and (e)-(h) snowy road conditions. Results of (b)/(f) DeepScene and (c)/(g) DeepRoad are compared. Road regions are recognized in (b) and (f), but no road conditions are provided. Therefore, (c) wet and (g) snow conditions from DeepRoad are refined to (b) and (f), respectively. Final refined images (d)/(h) are generated. Therefore, road conditions and other objects like mountains and vehicles are shown in single images.



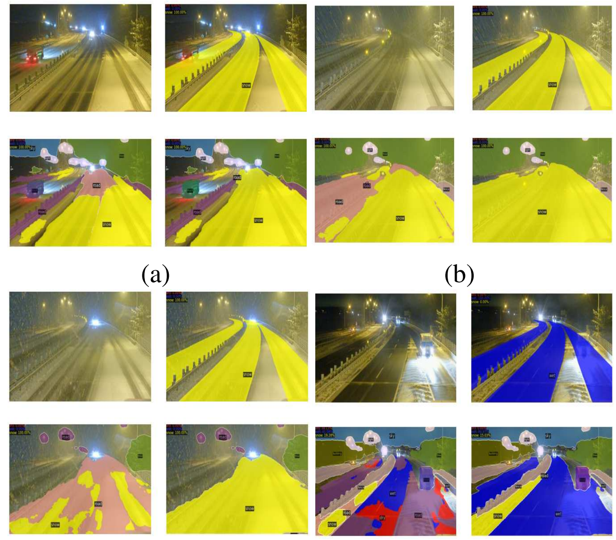
(a) (b) (c) (d)
(e) (f) (g) (h)

Figure 7. Refinement on DeepScene with DeepRoad: (a) Original image. (b) DeepScene. (c) DeepRoad only with road conditions. (d) Refined road conditions to wet. (e) Original image. (f) DeepScene. (g) DeepRoad. (h) Refined road conditions to snow.

4.5. Refinement on nighttime wet/snow conditions

In order to show the limit of DeepRoad, an improved method is proposed for road conditions. As shown in Figure 8, nighttime snowfall and foggy images are used in a 2 x 2 matrix. Ground truth images (upper right) of road conditions are annotated in the snow (yellow), wet (blue), or dry (red). Since only DeepRoad cannot recognize the road conditions, this paper first proposes integrating outputs of panoptic segmentation by DeepScene and road conditions by DeepRoad in (lower left). However, in (a)-(c), there was no output from DeepRoad, but only road and snow classes were segmented from DeepScene. Note that road class shows a possibility of dry or wet road conditions.

In order to enhance incomplete results of road conditions in images (lower left), this paper proposes to apply weather constraints on the road. Due to nighttime-covered snow on the roads, results in (a)-(c) with a mix of "road" and "snow" have been refined to all road regions with "snow" in yellow (lower right). In (d), the mix of dry, wet, and "road" have been refined to wet in main lanes and snow on the side road (lower right). SOTA adaptive domain semantic segmentation models [99, 100] have shown refinements of segmentation, but they are only applied to classes like incomplete road segmentation, not road conditions. Therefore, unlike the approaches in SOTA [99, 100], it is the first time to apply weather constraints to winter roads under snowfall and fog.



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

Figure 8. Integration of DeepScene and DeepRoad to enhance road condition accuracy: (upper left) Original image. (upper right) Ground truth. (lower left) DeepScene. (lower right) Refined road conditions to uniform snow or wet conditions. Yellow: snow, blue: wet, red: dry, purple: road, pink: light, green: sky.

5. Ablation study

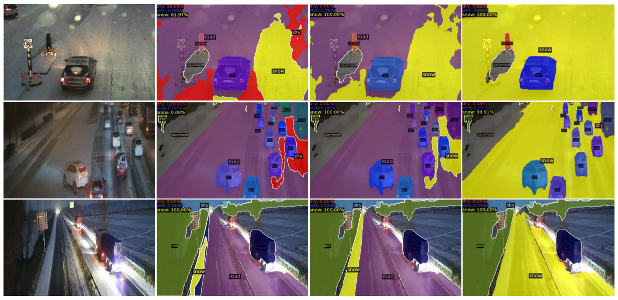
To justify the proposed PanopticRoad, many additional ablation studies are conducted below.

5.1. PanopticRoad for more complicated scenes

This section denotes the proposed PanopticRoad and how the final road conditions are refined. Figure 9 shows the results with (a) input images, (b) DeepScene + DeepRoad, (c) the Interaction over Unio (IoU) of (b), and (d) Refined weather constraints. Despite (a) the covered snow roads, DeepScene and DeepRoad have recognized the road and snow or dry due to low contrast, respectively. Next, the IoU of the two DL outputs is used. However, only local road regions have been refined to snow (c). Since the output of DeepScene with the road is suggested dry or wet conditions. Therefore, weather constraints are applied to the remaining road regions as wet to snow (d). It is noted that small side roads have been refined in [99, 100] but are mainly in dry conditions, unlike the original images of Figure 9 (a).

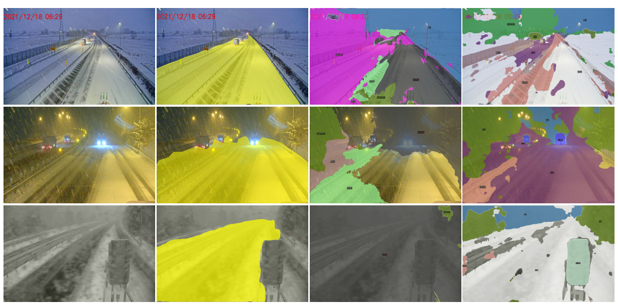
5.2. More comparison experiments of panoptic segmentation-based SOTA at foggy night

For further reconfirmation, various foggy twilight and night scenes are added to evaluate the performance of panoptic segmentation. Two SOTA panoptic segmentation methods are selected PanopticDepth [81] and Panoptic-DeepLab [45]. Figure 10 (a) compares highways and city roads. The proposed integrated model (b) has outperformed two SOTAs, (c) [81] and (d) [45], in terms of clear segmented regions like roads, light, vehicles, and trees. Panop-



(a) (b) (c) (d)
Figure 9. Proposed PanopticRoad: (a) Original image. (b) DeepScene + DeepRoad. (c) First refinement with the side road from (b). (d) Second refinement to full snow conditions from (c).

ticDepth could not recognize important stuff: roads and sky. Notably, the older method [45] presents more stable and better segmentation regions than the newer method in [81]. Therefore, it has been proven that the proposed integrated model is robust and stable in adversarial visual conditions.



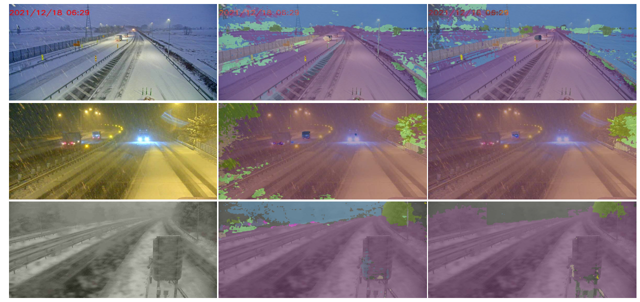
(a) (b) (c) (d)
Figure 10. Comparison of panoptic segmentation at twilight and night: (a) Original image. (b) Proposed method. (c) PanopticDepth [81]. (d) Panoptic-DeepLab [45].

5.3. Adaptive domain semantic segmentation-based SOTA at foggy night

In order to justify the proposed PanopticRoad, two SOTAs (DaFormer [99], MIC: CVPR2023 [100]) are used for nighttime snow roads. As in Figure 11, (a) original images are the same as those used in Figure 10. (b) DaFormer [99] and (c) MIC with the best selection of pretrained models [100] present similar results with segmented road classes. Therefore, further improvements by adding new pretrained models are desired for adversarial conditions. However, a manual selection of the optimal pretrained model may be required [99][100].

5.4. Post-disaster road conditions

This section challenges post-disaster road conditions to identify normal road surfaces and detect obstacles on the road using the proposed PanopticRoad. In addition to the aforementioned road conditions with dry, wet, and snow, road conditions can be dramatically changed by disaster events. Figure 12 (a) shows post-disaster images suffered



(a) (b) (c)
Figure 11. Adaptive domain semantic segmentation results from two SOTAs: (a) Original image. (b) DaFormer [99]. (c) MIC: CVPR2023 [100].

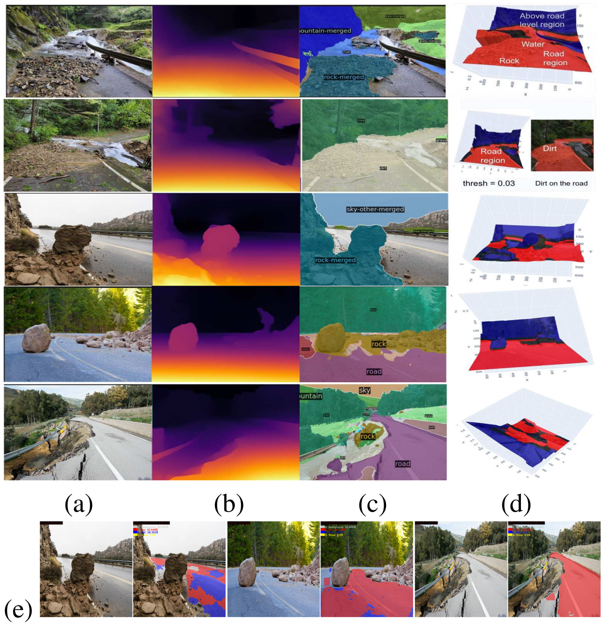
from the enormous typhoon, where many obstacles like dirt and rocks piled up on the road and other regions. In order to recognize whether obstacles are present on the road or not, first, the occluded road surfaces have to be identified. For this, DeepDepth (b) with horizontal (y), vertical (x), and depth (z) coordinates are used to recognize nearly flat road surfaces that are assumed to be the normal road (x-z) below the obstacles.

On the other hand, DeepScene (c) can provide segmented objects. From DeepDepth (b), vertical objects like tree and mountain classes can be removed from the depth maps from the physical viewpoint. However, non-road objects remain unremoved. Further removal to extract the road surface is needed. 3D cloud points that show a geometrical feature are converted from the depth maps. It is first assumed that the road is nearly flat. Based on this, the horizontal plane of a road is robustly estimated by Random Sample Consensus (RANSAC) 3D [104], which excludes outlier points with non-road points. Therefore, the points above this plane are used to remove objects above the road level.

Next, the refined depth maps and segmented regions are combined (d). (e) Road conditions with dry or wet are recognized as well. Finally, if the locations of dirt, water, or rocks are matched, the road conditions are assumed to be dirt, water, or rocks. It is the first time to utilize depth maps with physical constraints for road condition recognition. Although SOTA vision-language models [102, 103] suggest a promising framework related to this section, no depth maps have been utilized. Moreover, post-disaster image datasets are required to rebuild with depth maps.

5.5. Overall evaluation

To justify the performance of the proposed method, the experiment is conducted by comparing single-tasked DL models and combinations of various DL models. The test dataset is collected at various camera locations under different weather conditions. Evaluation results are calculated based on accuracy and mean IoU (mIoU) metrics. The accuracy metric is determined by the overall road conditions



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

(e)

Figure 12. Proposed PanopticRoad applied to post-disaster scenes to identify road regions with various obstacles (dirt, water, rock): (a) Input image. (b) Depth map. (c) Panoptic segmentation. (d) Refined road surface. (e) Road condition by DeepRoad.

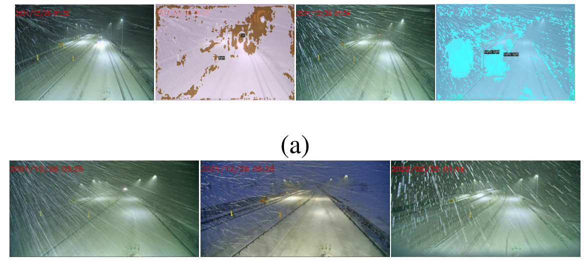
based on the class accounted for most of the coverage area, and mIoU is the mean of Intersection over union on three road conditions between ground truth and prediction mask. As shown in Table 2, the metrics become gradually better when combining the proposed five DL models: PanopticRoad.

Table 2. Comparison of PanopticRoad.

	Accuracy (%)	mIoU(%)
DeepRoad	91.75	55.15
Combination of DeepRoad and DeepScene	94.85	56.02
PanopticRoad	**96.15**	**58.83**

5.6. Evaluation of SOTAs under heavy nighttime snowfall

To understand the limitations of SOTAs, DETR [88] and DDSN [98], heavy nighttime snowfall events have been used. Figure 13 (a) shows two results by DETR [98], where the mix of heavy snowfall and strong illumination might have caused two failure cases: no segmentation regions with road and sky. It can be assumed that heavy snowfall seems to become foreground region. In (b), DDSN [98] failed to recognize and remove different snowfall events with light and heavy snowfall. It can be assumed that non-uniform streaks of snowfall in depth might not have been trained by DDSN [98]. Therefore, SOTAs could not demonstrate a satisfactory result in adversarial weather events.



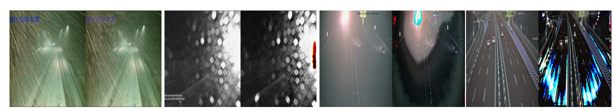
(a)

(b)

Figure 13. Failure cases (a) by SOTA, Transformer-based DETR [88] for panoptic segmentation, and (b) by SOTA, DDSN for snowfall removal [98].

5.7. Experiment on image restoration under adversarial weather conditions

To confirm another possibility for further processing under adversarial conditions, image restoration by an all-in-one DL model [78] has been applied. Figure 14 shows results with (a) heavy snowfall, (b) raindrops on the lens, (c) heavy fog with strong light, and (d) a clear scene. No image restoration has been achieved by SOTA DL [78]. However, heavy snowfall (a) and raindrops on lenz (b) could not be removed at all, unlike examples demonstrated in [78]. Moreover, false colors in (c) and (d) have been generated in red and sky blue. Therefore, the proposed DeepReject in this paper is important in avoiding visibility estimation in difficult images. This can stabilize overall system performance.



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

Figure 14. Limit of an all-in-one deep learning model [78] for adversarial weather conditions and clear scenes: (a) heavy snowfall. (b) raindrops on lenz. (c) heavy fog with strong headlight. (d) Clear scene.

6. Conclusion

This paper has proposed PanopticRoad with five Deep Learning-based modules, i.e., DeepReject/Scene/Snow/Depth/Road, for road condition segmentation under adversarial conditions, i.e., heavy nighttime snowfall and disaster. Integration of them helps refine the failure of local road conditions where weather and physical constraints are applied. On the other hand, most SOTA Deep Learning-based image enhancement and panoptic segmentation models show low performance. Recognizing obstacles, i.e., dirt, rocks, and flooding, on the road are novel road conditions. More complicated conditions will be considered for the deployment of auto-driving scenarios. The proposed PanopticRoad can be extended to video-based panoptic segmentation.

References

[1] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015.

[2] 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)*. IEEE, oct 2021.

[3] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *CoRR*, abs/2105.15203, 2021.

[4] Jiaqi Gu, Hyoukjun Kwon, Dilin Wang, Wei Ye, Meng Li, Yu-Hsin Chen, Liangzhen Lai, Vikas Chandra, and David Z. Pan. Multi-scale high-resolution vision transformer for semantic segmentation. *CoRR*, abs/2111.01236, 2021.

[5] S. Lee, T. Son, and S. Kwak. Fifo: Learning fog-invariant features for foggy scene segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18889–18899, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society.

[6] Hyunmin Lee and Jaesik Park. Instance-wise occlusion and depth orders in natural scenes. *CoRR*, abs/2111.14562, 2021.

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

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

[9] 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, Chubing 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 Transactions on Image Processing*, 29:5737–5752, 2020.

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

[11] D. Pomerleau. Visibility estimation from a moving vehicle using the ralph vision system. In *Proceedings of Conference on Intelligent Transportation Systems*, pages 906–911, 1997.

[12] 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.

[13] 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.

[14] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019.

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

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

[17] 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.

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

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

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

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

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

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

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

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

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

[27] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.

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

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

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

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

[32] Wenhan Yang, Robby T. Tan, Shiqi Wang, Yuming Fang, and Jiaying Liu. Single image deraining: From model-based to data-driven and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4059–4077, 2021.

[33] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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

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

[36] 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.

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

[38] Dawei Du, Yuankai Qi, Hongyang Yu, Yi-Fan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, volume 11214 of *Lecture Notes in Computer Science*, pages 375–391. Springer, 2018.

[39] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1280–1289, 2022.

[40] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9736–9745, 2020.

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

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

[43] João Carreira, Viorica Patraucean, Laurent Mazaré, Andrew Zisserman, and Simon Osindero. Massively parallel video networks. 11208:680–697, 2018.

[44] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. Blendmask: Top-down meets bottom-up for instance segmentation. pages 8570–8578, 2020.

[45] Bowen Cheng, Maxwell D. Collins, Yukun Zhu, Ting Liu, Thomas S. Huang, Hartwig Adam, and Liang-Chieh Chen.

Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12472–12482. Computer Vision Foundation / IEEE, 2020.

[46] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. pages 17864–17875, 2021.

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

[48] Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi. Temporally distributed networks for fast video semantic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8815–8824. Computer Vision Foundation / IEEE, 2020.

[49] Zilong Huang, Xinggang Wang, Yunchao Wei, Lichao Huang, Humphrey Shi, Wenyu Liu, and Thomas S. Huang. Ccnet: Criss-cross attention for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[50] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. pages 13352–13363, 2021.

[51] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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

[53] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 214–223, June 2021.

[54] Dongfang Liu, Yiming Cui, Wenbo Tan, and Yingjie Victor Chen. Sg-net: Spatial granularity network for one-stage video instance segmentation. pages 9816–9825, 2021.

[55] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference. 12355:352–368, 2020.

[56] Xiankai Lu, Wenguan Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. Video object segmentation with episodic graph memory networks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12348 of *Lecture Notes in Computer Science*, pages 661–679. Springer, 2020.

[57] K.-K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. Video object segmentation without temporal information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6):1515–1530, 2019.

[58] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. VSPW: A large-scale dataset for video scene parsing in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4133–4143. Computer Vision Foundation / IEEE, 2021.

[59] Jiaxu Miao, Yunchao Wei, and Yi Yang. Memory aggregation networks for efficient interactive video object segmentation. pages 10363–10372, 2020.

[60] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3997–4008, 2021.

[61] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. pages 7283–7293, 2021.

[62] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. pages 8741–8750, 2021.

[63] Mark Weber, Jun Xie, Maxwell D. Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, Aljosa Osep, Laura Leal-Taixé, and Liang-Chieh Chen. STEP: segmenting and tracking every pixel. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

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

[65] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, volume 11209 of *Lecture Notes in Computer Science*, pages 432–448. Springer, 2018.

[66] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8818–8826. Computer Vision Foundation / IEEE, 2019.

[67] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas S. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *CoRR*, abs/1809.03327, 2018.

[68] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. pages 5188–5197, 2019.

[69] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3684–3692. Computer Vision Foundation / IEEE Computer Society, 2018.

[70] Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Crossover learning for fast online video instance segmentation. *CoRR*, abs/2104.05970, 2021.

[71] Tien-Ju Yang, Maxwell D. Collins, Yukun Zhu, Jyh-Jing Hwang, Ting Liu, Xiao Zhang, Vivienne Sze, George Papandreou, and Liang-Chieh Chen. Deeperlab: Single-shot image parser. volume abs/1902.05093, 2019.

[72] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, volume 12350 of *Lecture Notes in Computer Science*, pages 332–348. Springer, 2020.

[73] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. pages 2491–2502, 2021.

[74] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 10326–10338, 2021.

[75] Tianfei Zhou, Jianwu Li, Shunzhou Wang, Ran Tao, and Jianbing Shen. Matnet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE Trans. Image Process.*, 29:8326–8338, 2020.

[76] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13066–13073. AAAI Press, 2020.

[77] Xianzheng Ma, Zhixiang Wang, Yacheng Zhan, Yinqiang Zheng, Zheng Wang, Dengxin Dai, and Chia-Wen Lin. Both style and fog matter: Cumulative domain adaptation for semantic foggy scene understanding. pages 18900–18909, 2022.

[78] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. pages 17431–17441, 2022.

[79] Markus Schön, Michael Buchholz, and Klaus Dietmayer. Mgnet: Monocular geometric scene understanding for autonomous driving. pages 15784–15795, 2021.

[80] Liulei Li, Tianfei Zhou, Wenguan Wang, Jianwu Li, and Yi Yang. Deep hierarchical semantic segmentation. pages 1236–1247, 2022.

[81] Naiyu Gao, Fei He, Jian Jia, Yanhu Shan, Haoyang Zhang, Xin Zhao, and Kaiqi Huang. Panopticdepth: A unified framework for depth-aware panoptic segmentation. pages 1622–1632, 2022.

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

[83] Yi Li, Yi Chang, Yan Gao, Changfeng Yu, and Luxin Yan. Physically disentangled intra- and inter-domain adaptation for varicolored haze removal. pages 5831–5840, 2022.

[84] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. pages 3213–3223, 2016.

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

[86] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. 11217:707–724, 2018.

[87] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(3):1623–1637, 2022.

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

[89] Emilio J. Almazan, Yiming Qian, and James H. Elder. Road segmentation for classification of road weather conditions. 9913:96–108, 2016.

[90] 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. 8693:740–755, 2014.

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

[92] S. Sreelakshmi and S. S. Vinod Chandra. Machine learning for disaster management: Insights from past research and future implications. In *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, pages 1–7, 2022.

[93] Martin Hahner, Christos Sakaridis, Mario Bijelic, Felix Heide, Fisher Yu, Dengxin Dai, and Luc Van Gool. Lidar snowfall simulation for robust 3d object detection. In

IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 16343–16353. IEEE, 2022.

[94] Zheng Shi, Ethan Tseng, Mario Bijelic, Werner Ritter, and Felix Heide. Zeroscatter: Domain transfer for long distance imaging and vision through scattering media. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3476–3486. Computer Vision Foundation / IEEE, 2021.

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

[96] Yu-Hsuan Chen, Pei-Jun Lee, and Trong-An Bui. Multi-scales feature extraction model for water segmentation in satellite image. In *IEEE International Conference on Consumer Electronics, ICCE 2023, Las Vegas, NV, USA, January 6-8, 2023*, pages 1–3. IEEE, 2023.

[97] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 2343–2353. IEEE, 2022.

[98] Kaihao Zhang, Rongqing Li, Yanjiang Yu, Wenhan Luo, and Changsheng Li. Deep dense multi-scale network for snow removal using semantic and depth priors. *IEEE Trans. Image Process.*, 30:7419–7431, 2021.

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

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

[101] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. pages 21001–21011, 2022.

[102] 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.

[103] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. Generalized decoding for pixel, image, and language. *CoRR*, abs/2212.11270, 2022.

[104] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4338–4364, 2021.