PanopticVis: Integrated Panoptic Segmentation for Visibility Estimation at Twilight and Night

Hidetomo Sakaino
Weathernews Inc.
sakain@wni.com

Abstract

Visibility affects traffic flow and control on city roads, highways, and runways. Visibility distance or level is an important measure for predicting the risk on the road. Particularly, it is known that traffic accidents can be raised at foggy twilight and night. Cameras monitor visual conditions like fog. However, only a few papers have tackled such nighttime vision with visibility estimation. This paper proposes a Panoptic Segmentation-based foggy night visibility estimation integrating multiple Deep Learning models: DeepReject/Depth/ Scene/Vis/Fog using single images. We call PanopticVis. DeepFog is trained for no-fog and heavy fog. DeepVis for medium fog is trained by annotated visibility physical scales in a regression manner. DeepDepth is improved to be robust to strong local illumination. DeepScene panoptic-segments scenes with stuff and things, booted by DeepDepth. DeepReject conducts adversarial visual conditions: strong illumination and darkness. Notably, the proposed multiple Deep Learning framework provides high efficiency in memory, cost, and easy-to-maintenance. Unlike previous synthetic test images, experimental results show the effectiveness of the proposed integrated multiple Deep Learning approaches for estimating visibility distances on real foggy night roads. The superiority of PanopticVis is demonstrated over state-of-the-art panoptic-based Deep Learning models in terms of stability, robustness, and accuracy.

1. Introduction

Camera-based scene understanding approaches have become attractive and wealthy to academia and industry due to the tremendous progress in Deep Learning models [21, 31, 32, 47, 49, 81] during the past few years. Many cameras in the cities, highways, on-board, and drone cameras monitor various objects such as vehicles, pedestrians, buildings, and vegetation. Different weather conditions like sunny, rainy, windy, snowy, and foggy events are shown periodically, which can cause unstable recognition and classification rate of such objects. In darker mornings, twilight, and nighttime, illumination from headlights, street lumps, traffic boards, reflected lenses, and raindrops on lenses can degrade image processing performance. 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 elements.

The representative metric is visibility levels or distances between a camera and a distant 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 lost in the nighttime road environment. Previously, computer vision-based visibility estimation methods with edge detection and geometrical coordinate have been proposed [3, 4, 30, 35, 57, 60, 61, 72, 87]. However, they are known to be vulnerable to illumination changes.

In Deep Learning (DL) models, semantic segmentation [49] and instance segmentation [2] have been reported and used for recognizing things or/and stuff [80]. Panoptic segmentation [8, 20, 28, 29, 39, 41, 78] 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 essential step towards scene understanding in autonomous vehicles since it provides object masks and attractive amorph regions like drivable road space or sidewalks. Although video-based panoptic segmentation models [1, 9, 24, 26, 28, 44, 46, 50, 53–55, 63, 75, 78, 84, 86, 89, 93, 99, 100] 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 [19, 58], and snowfall changes. Adversarial visual factors significantly degrade the accuracy of state-of-the-art (SOTA) DL-based segmentation. Raindrops [64, 91] are removed for better visibility. Defog or Dehaze [23, 32, 37, 40, 52, 85] is shown, with no visibility estimation. However, most papers have synthesized raindrops, rain streaks, and fog to obtain nearly perfect original daytime images under uniform illumination [36] SOTA

DL models are easy to fail in applications of natural foggy scenes due to the non-uniformity of fog and rainfall, ambient illumination, halo effect, and motion in depth [5, 31]. In dark backgrounds, night vision [22, 69, 79] is a challenging topic due to low light and less visible landmarks available. Although night-to-day translation by GAN [27, 101] 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 [33] is reported with no manual selection of complex scenes for multiple tasks with adversarial conditions and visibility estimation.

Evaluation image datasets are essential but very limited to scenes with clear, synthetic fog and real lighter fog [16, 66, 67, 92], where no or less adversarial conditions contain. Therefore, real heavier foggy night scenes with adversarial conditions have not been publicly available, as this paper uses. In particular, hundreds of meters of visibility distances are required on highways. Furthermore, although near and far objects can be landmarks for visibility estimation, foggy night causes extremely low contrast to recognize all of them. Therefore, overall daytime scenes are helpful to train image features in addition to segmented regions, as this paper proposes.

To this end, this paper proposes integrated panoptic segmentation-based visibility estimation on foggy night scenes under adversarial visual conditions using single images. Furthermore, multiple Deep Learning (DL) models with branched structures are integrated for efficiency in memory, training, and maintenance. Contributions of this paper are five folds as follows:

1. Transformer-based DeepScene segments images in which objects are partially or fully occluded by fog. Such segmented images and original images are trained by proposed DeepVis with the ground truth of image visibility distances using the distance chart in real road images in a regression manner. DeepFog takes charge of heavy or medium fog levels to classify.

2. DeepDepth is improved by alleviating intense illumination from a depth estimation method. DeepDepth can boost DeepScene panoptic segmentation performance like far tiny objects. DeepVis is integrated with DeepScene for a physical visibility distance estimation as well.

3. Since no reliable SOTA Defog and Derain DL models have been reported, DeepReject is proposed for images with adversarial conditions, which are newly trained for factors like intense illumination, raindrops, and darkness. This important preprocessing helps minimize recognition errors in DeepVis.

4. For our challenging, novel foggy day and night road images have been collected since publicly available image datasets, i.e., Cityscapes [16], Foggy Cityscape [66], and Foggy Zurich [67], are insufficient to train and test.

5. The proposed PanopticVis will enhance the camera-image-based visibility distance and level estimation accuracy for the surveillance monitoring, the safety of drivers, auto-driving, and rescue workers even under clear and adversarial conditions, i.e., extremely heavy fog. Moreover, the camera-based visibility estimation system will benefit from replacing the expensive visibility hardware sensors.

Using various road scenes from different cameras, experimental results show the superiority of the proposed method over SOTA DL models for visibility estimation under adversarial conditions in terms of accuracy, robustness, and stability.

2. Related works

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, and fog, i.e., 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 used as landmarks. Such objects may be obtained from segmentation. Dehaze [23, 32, 37, 40, 52, 85], denoise [45, 59, 76, 98], and derain/dedrops [64, 91] 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 [13, 17, 73, 95] 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 [52] has been alleviated by bridging the gap between clear and foggy images, i.e., city scenes. In [52], the intermediate gap/domain for the dual gap with style and fog is added in a pipeline into a unified framework

to disentangle the style and fog factors separately, and then the dual factor from images in different domains. However, since only daytime light foggy scenes have been experimented with, several objects were degraded by light fog. Issues in foggy and heavy snowfall at night remain unsolved only by this model [52], and no visibility estimation under fog requires further modelization, unlike the proposed approaches we show.

In image restoration [33], an all-in-one image restoration network (AirNet) for unknown corruption has been proposed. Single image restoration aims to generate a visually pleasant high-quality image from a given degraded correspondence, e.g., noise, rain, fog, or snow. 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. Therefore, the AirNet [33] enjoys two highly expected merits: an all-in-one solution to recover images with different corruption types and ratios and a single path network even with multiple corruption types, unlike previous multiple input and output heads. In [33], 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 [68]. Panoptic segmentation captures the full scene semantically and on an instance basis. Self-supervised monocular depth estimation uses geometric constraints derived from the camera measurement model to only measure depth from monocular video sequences. 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 [34]. HSS can exploit taxonomic semantic relations for structured scene parsing by slightly changing existing hierarchy-agnostic segmentation networks. By exploiting hierarchy properties as optimization criteria, hierarchical violation in the segmentation predictions can be explicitly penalized. Through hierarchy-induced margin separation, more effective pixel representations can be generated. 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. Single-image dehazing aims to restore the haze-free image from the hazy counterpart that suffers from the reduced contrast and dull colors caused by spatial variant haze densities. This task has been a longstanding and challenging problem with many applications.

The proposed method [20] combines the global modeling capability of the Transformer and the local representation capability of CNN with transmission-aware 3D position embedding. However, dehazing in [20] 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. Moreover, instead of 3D position embedding [20], panoptic segmentation is proposed to integrate for visibility distance estimation.

A unified framework for depth-aware panoptic segmentation (DPS) has been reported [29], 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. Monocular depth estimation is useful, but almost all trained Deep Learning models have been trained by a short distance range, i.e., several hundred meters. Basically, DL-based depth estimation is vulnerable to strong local illumination, which should be eliminated, as this paper proposes.

Thus, estimation of physical visibility or visibility level at foggy nights has not been reported under adversarial visual conditions. Although Cityscapes with 3000 images [16], Foggy Cityscape DBF with 500 synthetic foggy images [66], and Foggy Zurich with 3800 real light foggy images [67] are publicly available, they are almost all daytime and lighter fog data, whereas this paper uses heavier foggy nighttime images up to extremely heavy fog with adversarial conditions.

3. Proposed Method

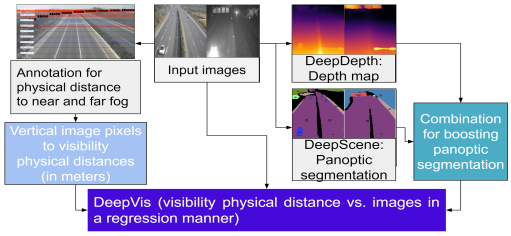
This section describes the proposed method to estimate visibility distance and level in various scenes, in particular, foggy nighttime scenes.

3.1. Overview of Proposed System

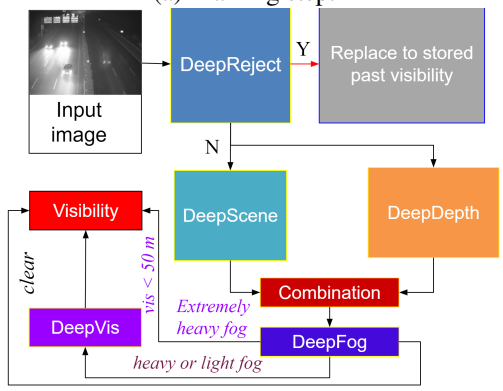
This section presents an overview of the proposed method for visibility estimation in road scenes under adversarial conditions. As shown in Figure 1, the framework consists of two steps: training and inference.

Training step: In Figure 1 (a), annotators are asked to classify and select various levels of real clear to foggy day and night images. All images have been manually selected. Thousands of real foggy images with different fog levels were collected from tens of camera locations, where many landmarks like poles and white lane lengths on road scenes were referred. Moreover, a fog synthetic physical equation was used to generate fog images as the ground truth data.

DeepScene is a transformer-based Deep Learning (DL) for segmentation, in which objects can deal with hundreds



(a) Training step.



(b) Inference step.

Figure 1. Overview of the proposed models.

of classes based on the COCO image database [43]. Fog occludes part of segmented objects. In fully covered heavy fog images, "sky" is obtained from overall images. Objects like "road" and "light" are detected in fog-free images.

Annotators are asked to provide physical scales in meters as the ground truth from known landmarks on the road scenes. Instead of inputting the original clear and foggy images like previous approaches, this paper proposes to input segmented images paired with physical visibility distances, i.e., 0 m – 500 m, > 500 m. Notably, segmented images are boosted by DL-based depth maps [65]. Training data is combined from day and night images to train the DeepVis model in a regression manner.

Inference step: In Figure 1 (b), DeepReject is adopted to eliminate low-quality images and increase the stability and accuracy of the night visibility estimation model. If the image is not rejected, the segmentation predicted by DeepScene will be fed into DeepVis to obtain visibility in meters. Actually, road operation based on visibility is translated into different fog levels. DeepFog takes charge of heavy or light fog levels to classify.

As shown in Table 1, an extremely heavy fog level is defined as a visibility distance < 50 m, where almost all objects are invisible, even illumination. Fewer white lines can be seen on the road. Light and heavy fog levels (50 m ∼ 500 m) are intermediate between extremely heavy fog and clear scenes, where the low and high risk for chain reaction accidents are assumed to happen. Far illumination at night is assumed to be seen when clear. It is assumed that such an extremely heavy or heavy fog may cause a higher risk for transportation.

Table 1: Definition of fog levels at night time.

Physical visibility range (m)		**Description**
From	*To*	
0	50	Extremely Heavy fog
50	200	Heavy fog
200	500	Light fog
500	+Inf	Clear

3.2. DeepReject (rejection to adversarial visual condition)

DeepReject is a classification-based DL model combined to classify input image quality as low or high. Few papers have applied the rejection approach. This is used to enhance the stability and accuracy of the other cascaded DLs. It is expected to make the system more robust than without DeepReject. Our original adversarial visual conditions are collected and used to train the DL model. When being rejected images, stored past image results are applied. DeepReject is an image classification model with an output of five different adversarial conditions: Normal: Normal visibility estimation allowed. Lens reflection: Effect of light refraction on the camera lens or affected by weather conditions with high humidity. Strong light: Strong artificial lighting such as road signs, headlights, and sunbeams. Low light: Insufficient light conditions. Raindrop: Raindrops on the camera lens. Data augmentation with flipping, rotating, cropping, and resizing images are utilized for training the DeepReject model.

3.3. DeepDepth (robust to local illumination)

This section describes DeepDepth improved from a monocular depth method [65, 71] using an RGB image. The algorithm [65] is based on a convolutional neural network architecture trained using a multi-objective optimization approach on a large dataset of paired RGB and depth images. The network extracts high-level features from the input image and then maps those to the corresponding depth information. However, the original depth estimation can be degraded due to local illumination. Therefore, this paper proposes to eliminate such factors through computer vision. The input image is first converted to the HSV color model and applied CLAHE with a clip limit of 2 and a tile grid size of $(8, 8)$. After that, a Gaussian filter is applied to the image to ensure its smoothness before implementing binary thresholding with a threshold value of 128. When the local illumination region is detected, inpainting is used with the Telea algorithm. However, DeepDepth is used only when DeepReject accepts input images.

3.4. DeepScene (segmentation model)

DeepScene is a Transformer-based panoptic segmentation DL model, i.e., Mask2Former [12], the backbone,

Swin-B [21]. In the road scenes, DeepScene can recognize objects such as "road", "signboard", "vehicle", or backgrounds as "sky" objects. At night, sky object contains the invisible region in the scene, and "light" are recognized. Note that such spotlighting is not assumed to be rejected by DeepReject. Majorly, DeepReject is assumed to reject glare and diffused illuminations. The segmentation results of Dscene represent the geometrical features of detected objects, such as the shape of the road and roadside objects, the position of vehicles, street lights, and the coverage of background objects. Like visual perception, spatial correlations between objects are used as an image feature to determine visibility by DeepVis.

4. Experiment and Discussion

This section conducts experiments to justify the proposed multiple Deep Learning approach for visibility estimation under foggy night scenes.

4.1. DeepReject evaluation

This subsection evaluates the performance of DeepReject. Our unique dataset consists of 4321 images with 5 different adversarial conditions types. Figure 2 illustrates rejected low-quality images by DeepReject due to non-weather factors such as large glare regions caused by (a) lens reflection, (b) strong headlight, (c) raindrops, and (d) low light. Single or mixed visual factors are contained in single images. As shown in Table 2, DeepReject achieves an average accuracy of 95.55%, which facilitates minimizing the visibility estimation error of the proposed method.

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

Figure 2. Example of rejected images: (a) Lens reflection. (b) Strong headlight. (c) Raindrop. (d) Low light.

Table 2: Evaluation of DeepReject on difficult scene classification under adversarial conditions at night.

	Image number	Correct recognition	Wrong recognition	Accuracy (%)
Normal	827	821	6	99.27
Lens reflection	627	600	27	95.69
Strong light	864	816	48	94.44
Low light	732	691	41	94.40
Raindrop	587	547	40	93.19
Total	3637	3475	162	**95.55**

4.2. DeepDepth under strong illumination

A monocular depth estimation [65] is assumed to boost the proposed DeepScene. Figure 3 shows fair-depth maps for normal scenes by using [65]. Figure 4 (a) shows two examples of foggy night images with strong headlights. Therefore, the original depth map [65] needs to be clarified. Strong illumination regions are eliminated using a method in Section 3.3. (b) shows improved depth maps. We call this DeepDepth.

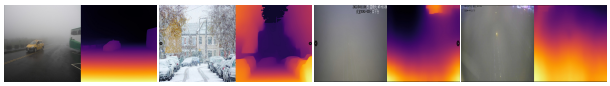
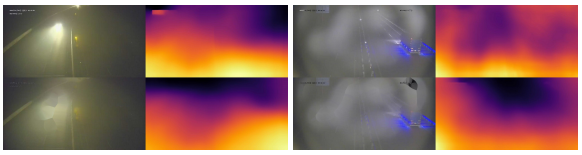


Figure 3. Results by original depth estimation [65].



(a) Original images. (b) Modified images.

Figure 4. Improved depth map from [65].

4.3. Segmentation by DeepScene + DeepDepth and SOTA at clear day, twilight, and night

In order to show the effectiveness of combining DeepScene and DeepDepth for scene segmentation, the six different scenes with clear, sunrise with glare, fog, snowfall, night snowfall, and night low light conditions have been selected. Two SOTA panoptic segmentation methods, DETR [20] and PanopticDepth [8], are compared with the proposed DeepScene and DeepDepth.

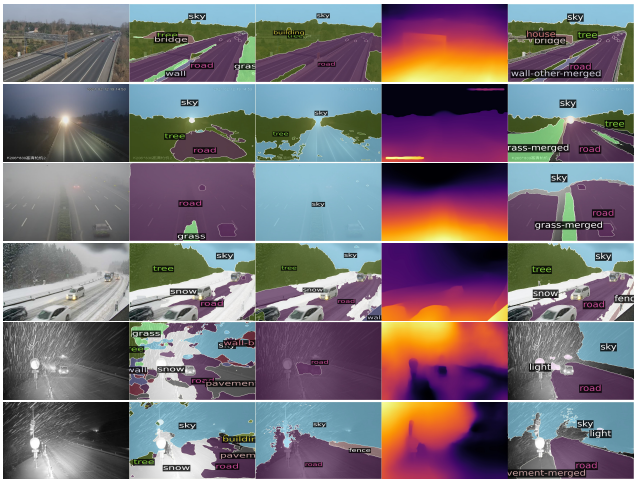
Figure 5 (a), (b), and (c) show original images, DETR [20], and PanopticDepth [8], respectively. Except for the clear image in 1st row, the remaining seven images (b), and (c) have become incomplete segmentation, e.g., no roads and mountains. On the other hand, the boosted DeepScene's images (e) by DeepDepth's images (d) have been segmented best among all, where stable visibility estimation will be allowed. Therefore, the proposed combination of DeepScene and DeepDepth has demonstrated the reliable segmentation.

4.4. Evaluation of PanopticVis

This section evaluates trained PanopticVis by 1174 foggy daytime real images. As shown in Table 1, visibility levels from clear to extremely heavy fog are defined. Figure 6 and Table 3 show that the overall accuracy is 87.82%. The F1 Score values show that the light and heavy classes are close to each other, often confused with the layers next to them, and these 2 classes are difficult to distinguish from each other, while the clear and extremely heavy fog is a good distinction.

4.5. PanopticVis and SOTAs at day and night scenes

This section devotes to simulating time-varying fog levels in actual highways in the day by comparing the proposed



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

Figure 5. Results of proposed DeepScene and SOTA: (a) original image. (b) DETR [8]. (c) PanopticDepth [20]. (d) Proposed DeepDepth. (e) Proposed DeepScene + DeepDepth.

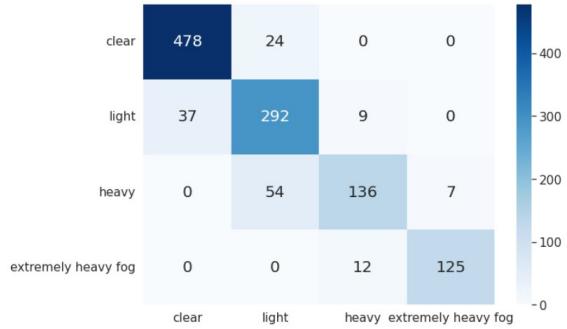


Figure 6. Evaluation of proposed PanopticVis in real fog visibility levels.

Table 3: Metrics for proposed PanopticVis evaluation [%].

Class	Precision	Recall	F1	Accuracy
Clear	95	93	94	94.80
Light	86	79	82	89.44
Heavy	69	87	77	93.02
Ex-heavy	91	95	93	98.38
Overall accuracy	85.25	88.5	86.5	92.66

DeepScene + DeepDepth and SOTA, FIFO [32]. Synthetic fog is generated by using a fog physical equation. Figures 7 and 8 illustrate the comparisons with different scenes at day and night respectively. Both figures show (1) no fog, and (2)-(3) light to heavy fog images with panoptic segmentation by the proposed model and FIFO [32]. As fog becomes heavier, the mountains and buildings located at different distances are disappeared.

Although FIFO [32] has been assumed to cope with any foggy scenes, the segmented results [32] have failed to segment full road surfaces as in clear day (1). Therefore, FIFO [32] could not show the fog-invariant on the roads and mountains, which cannot estimate visibility levels. Regions of the road have been expected to disappear from far to near according to light to heavier fog.

On the other hand, the proposed DeepScene + DeepDepth changes to the overall sky as fog changes heavier, which can lead to estimating visibility levels well. A further reconfirmation experiment will be conducted in the following sections.

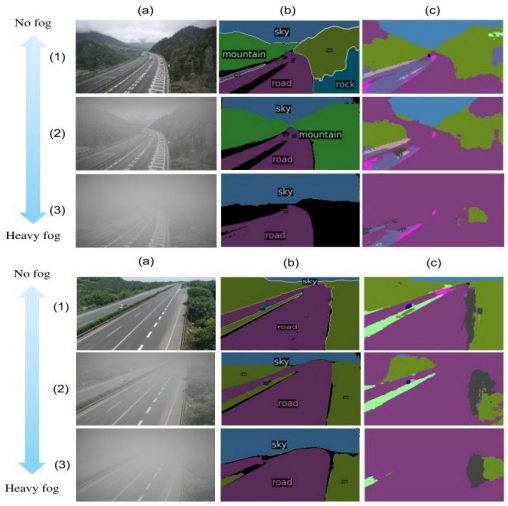


Figure 7. Comparative study of panoptic segmentation in different foggy levels at day: (1) Clear. (2) Light fog. (3) Extremely heavy fog. (a) Synthesized foggy images. (b) Segmented images by DeepScene + DeepDepth. (c) Segmented by SOTA, FIFO [32]. Purple: "road", green: "mountain", and blue: "sky".

5. Ablation study

To justify the proposed the proposed PanopticVis model, six experiments are conducted under adversarial conditions by comparing with many SOTA models.

5.1. Various foggy twilight and night scenes

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 [20] and Panoptic-DeepLab [11]. Figure 9 (a) compares highways and city roads. The proposed integrated model (b) has outperformed two SOTAs, (c) [20] and (d) [11], in terms of clearly segmented regions like roads, lights, vehicles, and trees. PanopticDepth [20] could not recognize several important stuff regions: roads and sky. Notably, the older method [11] presents more stable and better segmentation regions than [20]. Therefore, it has been proven that the proposed integrated model is robust and stable in adversarial visual conditions.

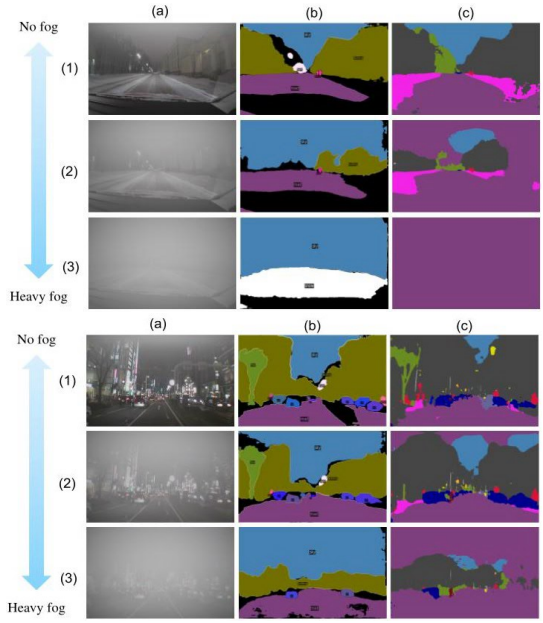
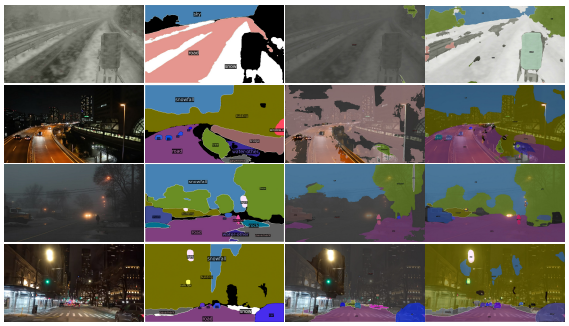


Figure 8. Comparative study of panoptic segmentation at different foggy levels at night: (1) Clear. (2) Light fog. (3) Extremely heavy fog. (a) Synthesized foggy images. (b) Segmented images by DeepScene + DeepDepth. (c) Segmented by SOTA, FIFO [32]. Purple: "road", green: "mountain", and blue: "sky".



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

Figure 9. Comparison of panoptic segmentation at twilight and night: (a) Original image. (b) Proposed method. (c) PanopticDepth [20]. (d) Panoptic-DeepLab [11].

5.2. Quantitative evaluation of SOTA panoptic segmentation models and the proposed PanopticVis using real foggy night scenes

This section quantitatively evaluates on real foggy night scene images using the proposed PanaoticVis and two SOTA panoptic segmentation models: PanopticDeep [11], PanopticDepth [20]. Mean Intersection over Union (mIoU), Average Precision (AP), and Panoptic Quality (PQ) are used for evaluation. As shown in Table 4, all three metrics by the proposed PanopticDepth show the highest scores between the two SOTA models. Therefore, based on the results of qualitative and quantitative evaluations, the robustness and stability of PanopticVis have been proven.

Table 4: Quantitative evaluations of SOTA panoptic segmentation models [11, 20] and the proposed PanopticVis. The best scores are in bold.

Metrics	Proposed PanopticVis (%)	PanopticDepth (%)	PanopticDeep (%)
mIoU	**66.0**	45.0	43.0
IoU	**47.3**	13.9	45.5
AP	**70.2**	70.1	63.8
PQ	**64.2**	24.4	62.5

5.3. Quantitative evaluation of the proposed method using real foggy night scenes

To evaluate the performance of the proposed method, the experiments are conducted on a set of real foggy night images labeled with physical visibility in meters. The test dataset consists of 1268 images collected under various adversarial conditions. Table 5 summarizes the accuracy by integrating different DL modules proposed in this paper. As their integration increases, accuracy becomes from 34.26% to 87.35%. Therefore, the effectiveness of the proposed PanopticVis combined with five Deep Learning models has been demonstrated.

Table 5: Evaluation results from only DeepVis to the full model of PanopticVis. Bold indicates the best accuracy.

Different deep learning module	Accuracy (%)
DeepVis	34.26
Combination of DeepVis, DeepFog, and DeepScene	65.13
Combination of DeepVis, DeepFog, DeepScene, and DeepDepth	84.90
PanopticVis: Combination with DeepVis, DeepFog, DeepScene, DeepDepth, and DeepReject	**87.35**

5.4. Visibility distance evaluation of the proposed PanopticVis using real foggy night scenes

In order to quantitatively evaluate the performance of the proposed PanopticVis in actual visibility distance and level, the experiment is conducted on real night images under various adversarial conditions. Errors between labeled and predicted visibility distances in meters are analyzed. Figure 10 (1) and (2) show eight different real foggy highways and segmented images, respectively. In each of the images, the actual and predicted visibility distances are contained as inset values in meters, e.g., 14.37 m - 1001.0 m. It is noted that the visibility distances in the actual fog are presented from no fog, i.e., clear, to extremely heavy fog.

Using 1268 images, the average error ratio between the labeled visibility and predicted values in meters is analyzed. Results in the mean error rates become 9.56% and 15.01% for extremely heavy and clear levels, respectively. Moreover, the average error ratios become 15.67% and 18.62%

for heavy and light fog levels, respectively. Moreover, as defined in Table 1, evaluation is added for four visibility levels. The average level error becomes 14.3%. In spite of images (Figure 10) under adversarial conditions like headlight, illumination, and low brightness, the proposed PanopticVis has proven to be useful for visibility distance and level estimation.

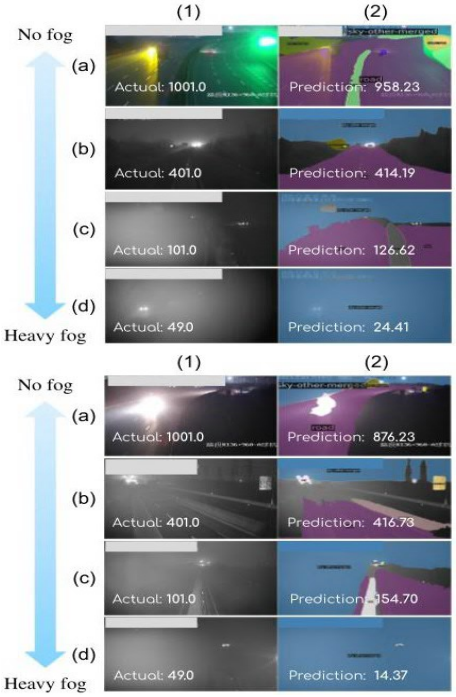


Figure 10. Prediction visibility distances in meters by PanopticVis: (a) Clear. (b) Light fog. (c) Heavy fog. (d) Extremely heavy fog. 8-paired examples of (1) Original image and (2) Segmentation by the proposed PanopticVis.

5.5. Limit of image enhancement with Night-to-Day models

This section conducts image enhancement by DL-based Night-to-Day models. CycleGAN [101] and pix2pix [27] are selected for image enhancement. In [27], image-to-image translation is a class of vision and graphics problems where the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs. Based on this property of CycleGAN and pix2pix, night images are translated into day images. Figure 11 shows (1) original foggy night images, (2) results of CycleGAN, and (3) pix2pix. Three road scenes (a)-(c) show different illumination and fog levels. Translated results of (2) and (3) could not present the daytime scenes, just brightening headlights and street lumps. In (b)-(2) and (c)-(3), the fog has become heavier than the original fog. Therefore, image enhancement cannot contribute to fog removal and estimation of visibility distances.

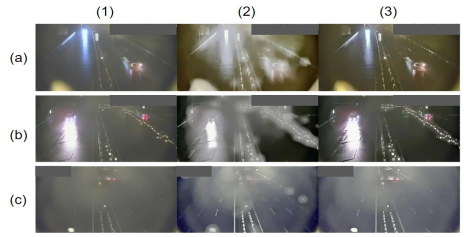


Figure 11. Night-to-Day translation: (a) Light fog. (b) Strong headlight. (c) Heavy fog. (1) Original image. (2) CycleGAN result. (3) pix2pix result.

5.6. Limit of SOTA image restoration model

In order to confirm another possibility for further processing images under adversarial conditions, image restoration by an all-in-one DL model [33] has been applied. Figure 12 shows results with (a) heavy snowfall, (b) raindrops on the lens, (c) light fog with sunbeam at dawn, and (d) a clear twilight scene. It is obvious that no image restoration has been achieved by SOTA DL [33]. Instead, false colors are generated in red and sky blue. Therefore, it is suggested that the proposed DeepReject plays an important role in avoiding visibility estimation in difficult images. This can stabilize overall system performance.

Figure 12. Limit of an all-in-one deep learning model [33] for adversarial weather conditions and clear scenes: (a) heavy snowfall. (b) raindrops on lenz. (c) light fog with a sunbeam. (d) clear twilight scene.

6. Conclusion

This paper has presented PanopticVis: an integrated panoptic segmentation with multiple different Deep Learning models, i.e., DeepScene/Vis/Fog/Reject/Depth, for visibility estimation at day, in particular, twilight and night. Moreover, many adversarial visual conditions have been considered to evaluate the proposed PanopticVis. Based on the Deep Learning model, it is the first time to estimate physical visibility distances under such adversarial visual conditions. For such difficult images, most SOTA Deep Learning models for image enhancement, image restoration, panoptic segmentation, and night-to-day translation show low performance under heavy fog and strong illumination. Since PanopticVis consists of changeable independent modules, it requires high efficiency in cost, retraining, memory, and extension. More, the proposed PanopticVis will be beneficial to the surveillance monitoring and the safety of drivers, auto-driving, and rescue workers. Finally, the camera-based visibility estimation system will also be replaceable with expensive visibility hardware sensors.

References

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

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] João Carreira, Viorica Patraucean, Laurent Mazaré, Andrew Zisserman, and Simon Osindero. Massively parallel video networks. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11208 of *Lecture Notes in Computer Science*, pages 680–697. Springer, 2018.

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

[11] 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.

[12] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 1280–1289. IEEE, 2022.

[13] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, June 2022.

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

[15] 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.

[16] 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.

[17] Xueqing Deng, Peng Wang, Xiaochen Lian, and Shawn Newsam. Nightlab: A dual-level architecture with hardness detection for segmentation at night. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16938–16948, June 2022.

[18] 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 recogni-

tion at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[19] 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.

[20] Naiyu Gao, Fei He, Jian Jia, Yanhu Shan, Haoyang Zhang, Xin Zhao, and Kaiqi Huang. Panopticdepth: A unified framework for depth-aware panoptic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 1622–1632. IEEE, 2022.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 603–612. IEEE, 2019.

[26] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 13352–13363, 2021.

[27] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. pages 1125–1134, 2017.

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

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

[30] 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.

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

[32] 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.

[33] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 17431–17441. IEEE, 2022.

[34] Liulei Li, Tianfei Zhou, Wenguan Wang, Jianwu Li, and Yi Yang. Deep hierarchical semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 1236–1247. IEEE, 2022.

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang,

Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 214–223. Computer Vision Foundation / IEEE, 2021.

[42] Yuanchu Liang, Saeed Anwar, and Yang Liu. Drt: A lightweight single image deraining recursive transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 589–598, June 2022.

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

[44] Dongfang Liu, Yiming Cui, Wenbo Tan, and Yingjie Victor Chen. Sg-net: Spatial granularity network for one-stage video instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9816–9825. Computer Vision Foundation / IEEE, 2021.

[45] Yang Liu, Zhenyue Qin, Saeed Anwar, Pan Ji, Dongwoo Kim, Sabrina Caldwell, and Tom Gedeon. Invertible denoising network: A light solution for real noise removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13365–13374, June 2021.

[46] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference. 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 X*, volume 12355 of *Lecture Notes in Computer Science*, pages 352–368. Springer, 2020.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

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

[53] Kevis-Kokitsi Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. Video object segmentation without temporal information. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(6):1515–1530, 2019.

[54] 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.

[55] Jiaxu Miao, Yunchao Wei, and Yi Yang. Memory aggregation networks for efficient interactive video object segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10363–10372. Computer Vision Foundation / IEEE, 2020.

[56] 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.

[57] 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.

[58] 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.

[59] Tongyao Pang, Huan Zheng, Yuhui Quan, and Hui Ji. Recorrupted-to-recorrupted: Unsupervised deep learning for image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2043–2052, June 2021.

[60] 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.

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

[62] Abhijith Punnappurath, Abdullah Abuolaim, Abdelrahman Abdelhamed, Alex Levinshtein, and Michael S. Brown. Day-to-night image synthesis for training nighttime neural isps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10769–10778, June 2022.

[63] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan L. Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3997–4008. Computer Vision Foundation / IEEE, 2021.

[64] 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.

[65] 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.

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

[67] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, volume 11217 of *Lecture Notes in Computer Science*, pages 707–724. Springer, 2018.

[68] Markus Schön, Michael Buchholz, and Klaus Dietmayer. Mgnet: Monocular geometric scene understanding for autonomous driving. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 15784–15795. IEEE, 2021.

[69] 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.

[70] 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.

[71] Jaime Spencer, C. Stella Qian, Chris Russell, Simon Hadfield, Erich Graf, Wendy Adams, Andrew J. Schofield, James H. Elder, Richard Bowden, Heng Cong, Stefano Mattoccia, Matteo Poggi, Zeeshan Khan Suri, Yang Tang, Fabio Tosi, Hao Wang, Youmin Zhang, Yusheng Zhang, and Chaoqiang Zhao. The monocular depth estimation challenge. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 623–632, January 2023.

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

[73] Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11563–11572, June 2022.

[74] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. 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*, pages 7283–7293. IEEE, 2021.

[75] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8741–8750. Computer Vision Foundation / IEEE, 2021.

[76] Zejin Wang, Jiazheng Liu, Guoqing Li, and Hua Han. Blind2unblind: Self-supervised image denoising with visible blind spots. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2027–2036, June 2022.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. pages 12077–12090, 2021.

[82] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and ef-

ficient design for semantic segmentation with transformers. pages 12077–12090, 2021.

[83] 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.

[84] 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.

[85] 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.

[86] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5187–5196. IEEE, 2019.

[87] 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.

[88] 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.

[89] 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. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8023–8032. IEEE, 2021.

[90] 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. *CoRR*, abs/1902.05093, 2019.

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

[92] 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.

[93] 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.

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

[95] Chang-Bin Zhang, Jia-Wen Xiao, Xialei Liu, Ying-Cong Chen, and Ming-Ming Cheng. Representation compensation networks for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7053–7064, June 2022.

[96] He Zhang and Vishal M Patel. Densely connected pyramid dehazing network. In *CVPR*, 2018.

[97] 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.

[98] Yi Zhang, Dasong Li, Ka Lung Law, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Idr: Self-supervised image denoising via iterative data refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2098–2107, June 2022.

[99] 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.

[100] 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.

[101] 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.