Night-time Scene Parsing with a Large Real Dataset

Xin Tan,　Ke Xu[†],　Ying Cao,　Yiheng Zhang,　Lizhuang Ma[†],　and　Rynson W.H. Lau[‡]

Abstract—**Although huge progress has been made on scene analysis in recent years, most existing works assume the input images to be in day-time with good lighting conditions. In this work, we aim to address the night-time scene parsing (NTSP) problem, which has two main challenges: 1) labeled night-time data are scarce, and 2) over- and under-exposures may co-occur in the input night-time images and are not explicitly modeled in existing pipelines. To tackle the scarcity of night-time data, we collect a novel labeled dataset, named *NightCity*, of 4,297 real night-time images with ground truth pixel-level semantic annotations. To our knowledge, NightCity is the largest dataset for NTSP. In addition, we also propose an exposure-aware framework to address the NTSP problem through augmenting the segmentation process with explicitly learned exposure features. Extensive experiments show that training on NightCity can significantly improve NTSP performances and that our exposure-aware model outperforms the state-of-the-art methods, yielding top performances on our dataset as well as existing datasets.**

Index Terms—**Autonomous Driving, Night-time Vision, Scene Analysis, Adverse Conditions.**

I. INTRODUCTION

SCENE Parsing is an important computer vision task for many applications, such as human parsing [1], image editing [2] and autonomous driving [3]. Although a lot of methods have been proposed, they mainly focus on day-time scenes. However, as night time and day time cover roughly about 50% of the time each (averaged over a year), it is equally important to build vision systems that perform well at night time, particularly for autonomous driving at night. In this paper, we address the night-time scene parsing (NTSP) problem.

When applied to night-time images, existing scene parsing methods designed for day-time images typically do not perform well. We observe that night-time scenes often contain both over-/under-exposures, which can seriously degrade the visual appearances and structures of the input images. For

X. Tan and K. Xu are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, and the Department of Computer Science, City University of Hong Kong, HKSAR, China. E-mails: tanxin2017@sjtu.edu.cn, kkangwing@gmail.com.

L. Ma are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, and the School of Computer Science and Technology, East China Normal University, Shanghai, China. E-mail: ma-lz@cs.sjtu.edu.cn

Ying Cao and Rynson W.H. Lau are with the Department of Computer Science, City University of Hong Kong, HKSAR, China. E-mails: caoying59@gmail.com, Rynson.Lau@cityu.edu.hk.

Y. Zhang is with the Department of Computer Science, Stanford University, CA, USA, and the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. E-mail: yihengz@stanford.edu.

† Corresponding Authors.

‡ Rynson W.H. Lau leads this project.

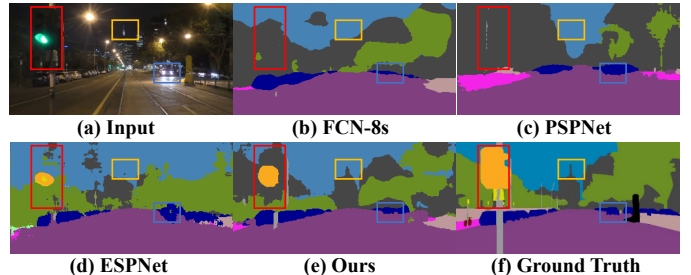Manuscript received November 24, 2020; revised xx xx, 2021.

Fig. 1. Night-time scene analysis. (a) shows an input night-time image, and parsing results from state-of-the-art methods: (b) FCN-8s [4], (c) PSPNet [5], and (d) ESPNet [6]. We also show the result by our model (e), and the ground truth (f). The yellow and blue boxes highlight under- and over-exposed regions, respectively. The red box highlights the region with a mixture of under- and over-exposures.

example, Figure 1(a) shows a night-time scene with both over-exposure (e.g., street lights and car headlights) and under-exposure (e.g., background and regions around the headlights). We can see from Figures 1(b–d) that state-of-the-art methods are not able to address this problem well. First, the building highlighted by the yellow box blends into the dark background due to under-exposure, causing it to be difficult to detect. Second, the texture and structure of the cars highlighted by the blue box are corrupted due to over-exposure, causing them to be difficult to segment correctly. Third, the traffic light highlighted by the red box is difficult to be detected or segmented correctly due to a mixture of over-/under-exposures.

There are two main challenges to the NTSP problem. First, large-scale *labeled* datasets of *night-time* scenes are not available. Existing large datasets for scene parsing mainly contain day-time images, with few or no night-time images [7], [8], [9]. Models trained on these datasets do not generalize well to the complexity of night-time scenes. Second, existing methods do not explicitly model over- and under-exposures, as they are primarily developed for day-time scenes. However, as demonstrated in our experiments, explicitly modeling the exposure is necessary in order to achieve robust performances.

To address the first challenge, we propose in this paper a new dataset, named *NightCity*, for the NTSP problem. It contains 4, 297 real night-time images of diverse complexity, with pixel-wise annotations. To our knowledge, NightCity is the largest labeled dataset for NTSP, and is an order of magnitude larger than existing datasets for NTSP [10]. As compared with Cityscapes [7], NightCity covers more diverse and challenging exposure conditions that are typical in night-time scenes. Our experiments show that NightCity can help significantly advance the NTSP performance. It can also serve as a benchmark for evaluating NTSP methods.

To address the second challenge, as we observe that the

drop in performance when applying existing scene parsing methods on night-time images is mainly due to the complicated exposure conditions of night-time scenes, we propose an Exposure-Guided Network (EGNet) to explicitly learn over- and under-exposure features to guide the NTSP process. Our model comprises two streams: the segmentation stream and the exposure stream. The segmentation stream learns to predict the semantic label map for the input image, while the exposure stream learns exposure-related features by explicitly predicting the exposure map and uses the learned features to augment the segmentation stream via an attention mechanism. Experimental results show that our model outperforms previous methods, achieving state-of-the-art performance on night-time images.

In summary, the main contributions of this paper include:

- We propose a large-scale labeled dataset of real night-time images for the NTSP problem. The dataset will be made publicly available, to help advance research along this direction.
- We present an end-to-end exposure-aware NTSP framework, which explicitly learns exposure features to address the NTSP problem.
- We have conducted extensive evaluations. Our results demonstrate that our proposed model outperforms the state-of-the-art methods on night-time scenes.

II. RELATED WORK

A. Scene Parsing.

There are a lot of works proposed for scene parsing. The performance has been improved significantly in recent years due to the development of Convolutional Neural Networks (CNNs). For example, in [11], prior location information at superpixel level was exploited with CNNs to strengthen object discriminativity in scene parsing. Multi-level-based methods [4], [12], [13], [14], [15] were widely used by learning multi-level features to extract the global context and to preserve the low-level details. Recently, attention-based methods [16], [17], [18] have shown promising performances. Fu et al. [16] adaptively integrated local features with their global dependencies using position attention and channel attention modules. Huang et al. [19] used a novel criss-cross attention module to model long-range contextual dependencies over local feature representations. Choi et al. [20] adopted the height attention to learn different categories that are distributed at different heights. Two-stream approaches were also proposed [21], [17]. For example, Takikawa et al. [17] introduced a two-stream network, with one of the streams explicitly wiring shape information for scene parsing. Feng et al. [22] transferred the knowledge of a teacher stream to a student stream via a pixel-wise similarity distillation module and a category-wise similarity distillation module, for balancing the accuracy and inference time.

Many methods are proposed to learn robust object representations for scene parsing. Yuan et al.[23] exploited object-contextual representations by exploring the features of corresponding object classes, to characterize a pixel. Zhu et al. [24] exploited video prediction models to predict the class labels. GPSNet [25] was designed to adaptively select receptive fields while maintaining a dense sampling capability for scene parsing. Huang et al. [26] proposed a Scale-Adaptive Network to tackle the varying scale problem of objects. To bridge the gaps between the domain of (limited) training data and that of test scenes, some domain adaptation-based methods [27], [28], [29], [30] were proposed to utilize useful knowledge of synthetic scenes to enrich the scene representations.

All the above methods achieved state-of-the-art results on day-time datasets. However, unlike these existing works, we focus our attention on night-time scenes with poor lighting conditions in this work.

B. Scene Parsing in Adverse Conditions.

Although most existing works focus on the "normal" scenarios with well-illuminated scenes, there are also some works that address the challenging scenarios. For rainy night scenes, [31] tried to solve the rainy night scene parsing problem via transferring day-time knowledge. It collected 226 images with eight categories. Tung et al. [32] also tried to handle this problem and proposed 95 annotated night-time images with three categories. Zheng et al. [33] presented a Fork-shaped Cyclic generative module, decoupling domain-invariant content and domain-specific style, to translate the scenes in adverse condition to images with better visualizations for parsing. For multiple and mixed adverse conditions, Valada et al. [34] proposed the convoluted mixture of deep experts fusion techniques to understand these adverse conditions, including rain, snow, sunset and night scenes among 13 categories. WildDash [35] and BDD100K [36] aimed to test segmentation performances by presenting many kinds of scenarios with 13 and 345 night-time images, respectively. Sakaridis et al. [10], [37] proposed guided curriculum model adaptation to solve the night-time semantic segmentation problem, with a small dataset of 151 night-time images. Recently, Wu et al. [38] proposed an unsupervised one-stage adaptation method for NTSP, by exploiting adversarial learning between labeled day-time dataset and unlabeled dataset of day-night aligned image pairs.

In contrast to the above works, we propose a large-scale dataset of real night-time images with semantic annotations (21 categories) and a novel exposure-aware framework to address the NTSP problem.

C. Image Enhancement/Correction.

One naive solution to our problem is to first apply image enhancement on the input night-time images and then perform scene parsing with an existing day-time method. Image enhancement methods, e.g., [39], [40], [41], [42], aim to remap the pixel values to improve the image visibility. Wang et al. [43] tried to address the under-exposure problem by estimating an illumination map. Cai et al. [44] proposed to learn a deep image contrast enhancer from multi-exposure images. In [45], an end-to-end network was proposed to convert an input LDR image first to HDR to recover the missing details due to under-/over-exposure, and then reproject it back to LDR as output while preserving the recovered details. Xu et al. [46] proposed a frequency decomposition framework to address the

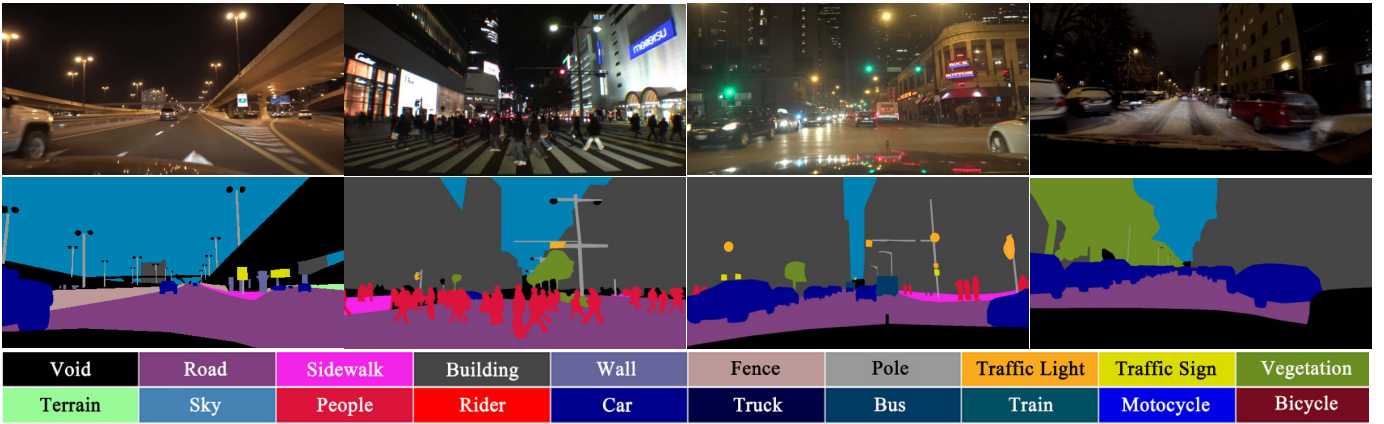| Void	Road	Sidewalk	Building	Wall	Fence	Pole	Traffic Light	Traffic Sign	Vegetation
Terrain	Sky	People	Rider	Car	Truck	Bus	Train	Motocycle	Bicycle

Fig. 2. Several example images from our NightCity dataset. Note that some regions around the car headlights are over-exposed and some background regions are under-exposed.

practical under-exposure problem with noise. However, night-time images often contain both under-/over-exposed regions (with pixel values very close to zero/one). The remapping process may not recover meaningful values. As demonstrated in our experiments in Section V-C, pre-processing the input images with a state-of-the-art image enhancement method before scene parsing cannot address our problem well.

In contrast, our method does not apply any pre-processing techniques. It achieves the state-of-the-art NTSP performances by (1) constructing the first large-scale NTSP dataset, and (2) exploiting exposure-aware features to enrich night-time scene representations via the proposed network.

III. THE NIGHTCITY DATASET

To construct our dataset, we first collect real night-time driving videos (which were captured using a Driving Recorder during car driving) over the Internet from various cities, e.g., Los Angeles, New York, Chicago, Hong Kong, London, Tokyo and Toronto. These videos cover urban street, highway and tunnel scenarios. We then manually select $4,297$ diverse images with no obvious motion blur from these videos for manual annotation, following the approach used to construct the Cityscapes dataset [7]. Table I shows the distribution of the videos over the considered cities.

Annotation. Like Cityscapes, we annotate semantic regions as layered polygons using LabelMe [47]. All our images are of resolution 1024×512. Given the difficulty of recognizing objects in the over-/under-exposed regions, we annotate our images by two separate annotators (A and B) and re-evaluate their results by a third one (C). Annotators may refer to the corresponding video if an image is difficult to see. For each image $I \in R^{H \times W \times 3}$, annotators A and B give the annotations $G_A \in R^{H \times W \times C}$ and $G_B \in R^{H \times W \times C}$, respectively. Annotator C compares the difference between G_A and G_B to produce a compromised map. There are two situations where Annotator C may correct a label. First, the two labels by the two annotators are different. Second, the two labels by the two annotators are the same, but Annotator C considers these labels as incorrect. In this case, C would discuss with the two annotators to come up with the final label.

The steps for Annotator C to follow are summarized as:

TABLE I
DISTRIBUTION OF THE VIDEOS OVER ALL CONSIDERED CITIES.

City	Number of images	Train	Test
Los Angeles	148	136	12
New York	375	261	114
Chicago	195	136	59
Toronto	132	92	40
Melbourne	183	128	55
London	74	52	22
Dubai	512	357	155
Helsinki	659	460	199
Hong Kong	367	256	111
Seoul	344	240	104
Nagoya	1025	715	310
Tokyo	181	126	55
Other cities in Japan	102	72	30

1) If C considers both G_A and G_B are correct, there will be no changes.
2) If G_A and G_B are different and C considers one is correct and the other is incorrect, C will simply select one of them as the final annotation (i.e., majority win).
3) If G_A and G_B are different but C considers both incorrect. C will discuss with Annotators A and B to make the final decision.

Statistically, there are in total 9.8% of the pixels that are labeled differently by the two annotators. Among these different annotations, 47.9% are finally corrected, i.e., (total number of different-labeled pixels finally corrected)/(total number of different-labeled pixels). The error proportions, i.e., (total number of error pixels per class)/(total number of pixels per class), of five object categories that are most difficult to annotate are shown in Table II. We find that poles have the highest annotation error proportion, as they can be easily ignored due to their small size. In addition, buildings and sky can be easily mistakable with each other.

Some regions that are too difficult to annotate even by humans are labeled as *invalid regions* so that they are ignored during training and evaluation. In GCMA [10], a pixel is
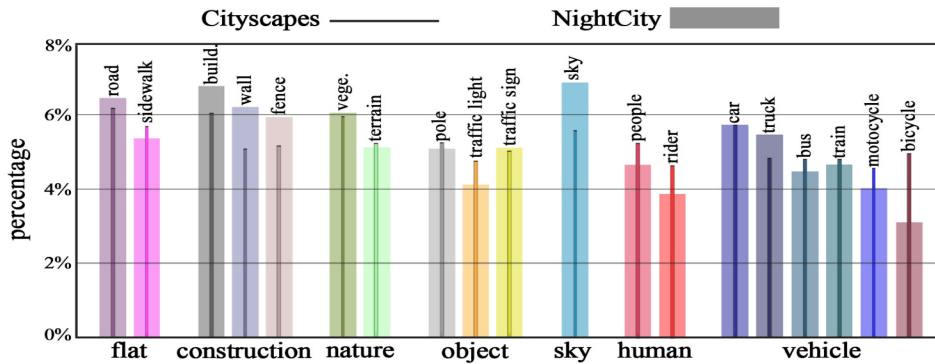
Fig. 3. Distributions of labeled pixels for each semantic class in Cityscapes and NightCity, shown in log scale.

TABLE II
PROPORTIONS OF ANNOTATION ERRORS OF FIVE MOST DIFFICULT OBJECT
CATEGORIES TO ANNOTATE.

Class	Pole	Building	Wall	Sky	Terrain
Error(%)	6.38	4.76	4.84	4.44	3.19

labeled as invalid if the annotator considers it as invalid. In our paper, we decide it based on the judgement of multiple annotators, i.e., if at least two of the three annotators consider a pixel as invalid (a majority-win strategy). The total proportion of invalid pixels in NightCity is 7.36%.

Figure 2 shows some example images from NightCity, demonstrating that our images are from complex scenes and their contents are difficult to recognize and segment even for humans, due to the under-/over-exposure problems caused by insufficient lighting, street lights and car headlights. Hence, our NightCity dataset represents a rather challenging training and evaluation dataset for NTSP.

Object Class Distribution. Figure 3 compares the distributions of labeled pixels in NightCity and Cityscapes. As some categories have significantly higher numbers of pixels (e.g., about 1×10^9 pixels belonging to "building") than some others (e.g., about 1×10^7 pixels belonging to "motorcycle"), we show the distributions in log scale. Overall, the pixel distributions of all classes in our night-time street scenes are similar to those of the day-time street scenes, except for Bicycle. This is reasonable because there are typically fewer bicycles on the street at night.

Exposure Distribution. To reveal the over-/under-exposure problems in our dataset, we also analyze the exposure conditions in our images by checking the pixel values. In photography, the exposure is determined by the shutter speed, lens aperture, scene luminance and ISO number. However, as this information is unknown to us, we use the V channel (i.e., intensity) of the images in the HSV color space to represent the exposure. In particular, we divide the exposure value equally into ten bins from 0 to 1 with an interval of 0.1. Figure 4 shows the average number of pixels per image that falls into each bin for NightCity and for Cityscapes. The [0, 0.1] bin stores the most under-exposed pixels, while the [0.9, 1.0] bin stores the most over-exposed pixels. From Figure 4, we can see that NightCity has significantly more under-
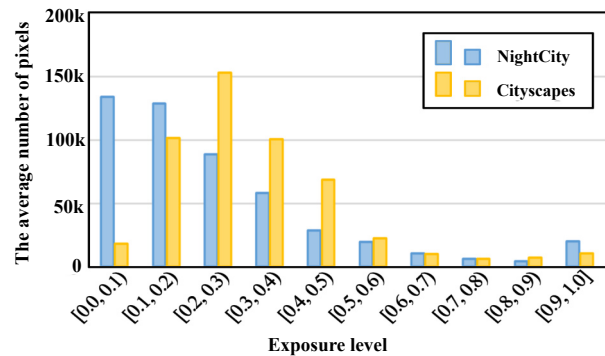

Fig. 4. The average number of pixels per image at each exposure level for Cityscapes and for NightCity, with image resolution of 1024×512.

exposed pixels than Cityscapes. Meanwhile, we observe that Cityscapes has a moderate exposure condition, with most of pixels falling between [0.2, 0.6]. Although the two datasets have a similar number of over-exposed pixels, we notice that the over-exposed regions in Cityscapes are mostly in the sky, which is easy to predict, while the over-exposed regions in NightCity can be produced by the street lights, car headlights, or traffic lights, which are very difficult to differentiate. This indicates that the NightCity dataset has various challenging exposure conditions, compared with Cityscapes.

Size and Splitting. Compared with existing datasets under adverse conditions, the number of images in NightCity is considerably higher. For example, *Foggy Driving* [48] proposes a real fog dataset with a total of 101 foggy images for testing (of which only 33 images are finely annotated and the remaining 68 images are only coarsely annotated). *The Dark Zurich Dataset* [10] contains only 151 night-time images with pixel-level annotations. BDD100K [36] contains 345 night-time images with some labeling errors. Hence, their ground truth is not very reliable. *Raincouver* [32] contains 95 coarsely annotated night-time images but only with 3 classes. *WildDash* [35] only has 13 finely annotated night-time images. Instead, all 4,297 images in our NightCity dataset are manually annotated with fine class labels. Thus, NightCity is the largest dataset of real night-time images with high-quality pixel-level annotations.

The NightCity dataset is split into training and test sets. We split them in such a way that they preserve similar distributions
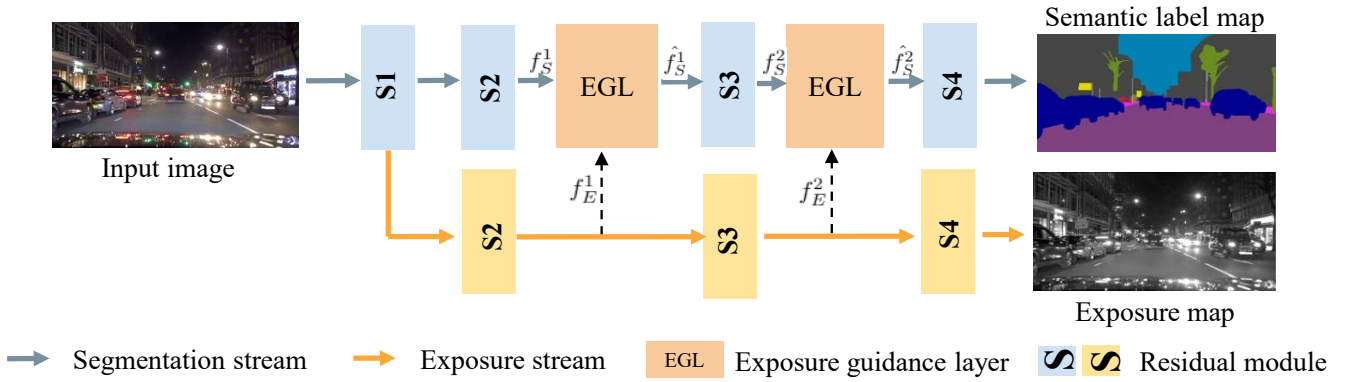
Fig. 5. The architecture of our exposure-guided network. It contains two streams, a segmentation stream in blue to predict a semantic label map, and an exposure stream in yellow to predict an exposure map. We introduce the exposure guidance layer (EGL) to augment segmentation features with exposure features.

to the whole dataset. In this way, our training and test splits include $2,998$ and $1,299$ images, respectively.

IV. EXPOSURE-GUIDED NETWORK

The core idea of our proposed Exposure-Guided Network (EGNet) is to explicitly learn exposure features and use them to augment the segmentation process. Hence, our network is designed to have two coupled streams: exposure stream and segmentation stream. The exposure stream learns to predict where exposure occurs and uses the predictions to guide the segmentation stream via the exposure guidance layers so that the segmentation stream can discriminate the under-/over-exposed regions more effectively.

A. Network Architecture

Figure 5 shows our network architecture. Given an input image $I \in R^{H \times W \times 3}$, we use a backbone encoder to transform it into convolutional features. The network is then split into the segmentation stream (top) and exposure stream (bottom). The segmentation stream predicts a class label map $M_S \in R^{H \times W \times C}$, where C is the number of classes. The exposure stream also outputs a pixel-wise exposure map $M_E \in [0,1]^{H \times W}$, indicating the magnitude of exposure at each pixel.

As discussed in Section III, we approximate the exposure map using the normalized V channel of the input image in the HSV color space.

Our network is mainly based on ResNet [49], like most existing scene parsing models [50], [51], [52], [5]. ResNet has four stages that extract hierarchical features at different scales, with earlier stages capturing low-level features and later stages capturing high-level semantics. In particular, we take stage1 (S1) of ResNet-101 as our backbone encoder. The segmentation and exposure streams have the same architecture, by combining stage2 (S2), stage3 (S3) and stage4 (S4) of ResNet, except for the last output layers. For the output layer, the segmentation stream uses a 19-channel convolutional layer to output a semantic label map, while the exposure stream uses another sigmoid nonlinearity function to output a soft binary exposure map. To train the network, we use a cross-entropy loss for the segmentation stream, and a ℓ_1 loss for the exposure stream. Our final loss is defined as $L = \alpha L_c + \beta L_e$, where α and β are weights of the two losses.

B. Exposure Guidance Layer

To extract exposure features from the exposure stream to guide the segmentation stream, we introduce the exposure guidance layer (EGL) to augment the intermediate features of the segmentation stream. Let f_S and f_E be the intermediate features of the segmentation and exposure streams. EGL updates f_S to obtain exposure-aware features $\hat{f}_S$ as:

$$\hat{f}_S = w_1 f_S + w_2 f_S \otimes W_r, \tag{1}$$

where $\otimes$ denotes element-wise multiplication. w_1 and w_2 are weight parameters. $W_r = \delta(W * f_E + b)$ is a soft spatial attention map as in the popular works [53], [54], where W and b are learnable parameters, and $\delta(\cdot)$ is a sigmoid function. The exposure-aware features $\hat{f}_S$ are then fed into the next stage of the network. As shown in Figure 5, both f_S and f_E are from the previous stages. All the operations in EGL are differentiable so that we can train the network end to end. In addition, EGL enables the gradients to be back-propagated from the output exposure map to the segmentation stream, thereby allowing the segmentation stream to exploit exposure information.

The intuition behind Eq. 1 is that our model needs to learn how to weight the features in the segmentation stream based on the exposure features, in order to generate more discriminative segmentation features particularly at under-/over-exposed regions. This formulation also forces our model to learn the exposure features that help predict the exposure maps and guide the scene parsing task towards an optimal performance.

C. Discussion: Is Noise Always Vital In Night-time Scenes?

Noise is inevitable in every imaging pipeline. The intuition is that noise is fatal in night-time scenes due to a lack of illumination, in which cases users may have to increase the

(a) Night-time Scenes   (b) Hist.Eq. of (a)   (c) Low-illuminated Scenes   (d) Hist.Eq. of (c)
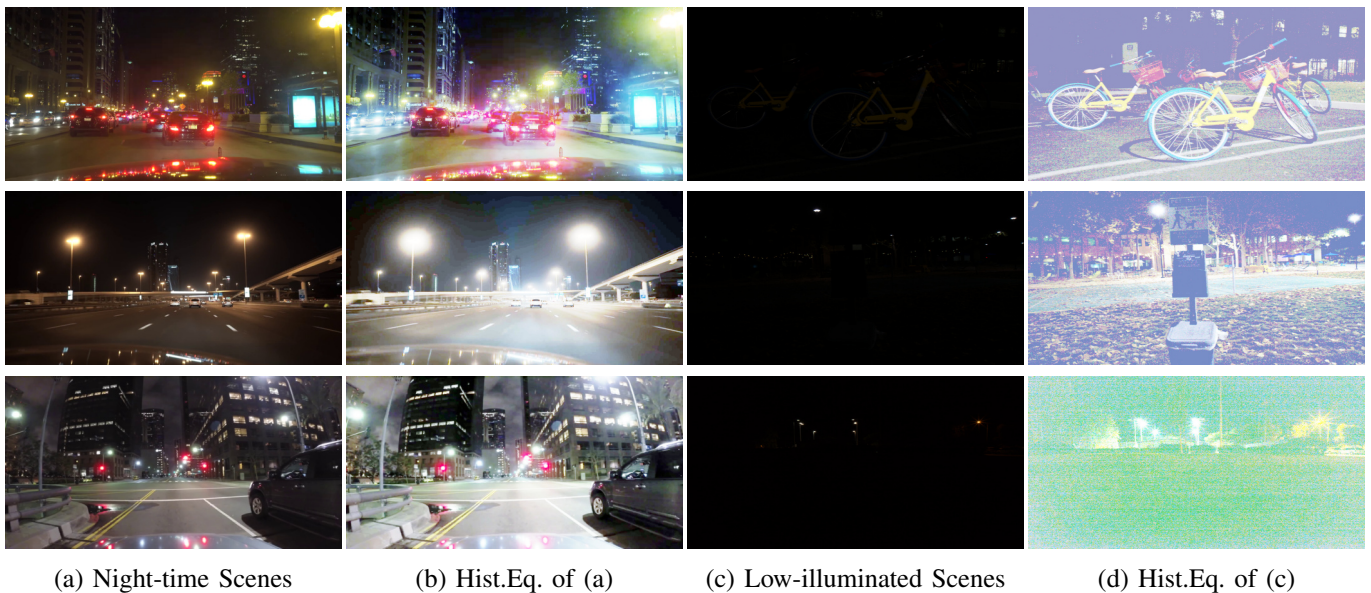
Fig. 6. Comparison of our night-time urban scenes (a) and low-illuminated scenes (c) from [46], and their enhanced ones using histogram equalization (see (b) and (d), respectively). Noise is negligible in the night-time urban scenes but non-negligible in the low-illuminated scenes.

ISO value to capture more photons. However, we find that in our night-time urban scenes, the noise issue could be ignored. This is because in night-time urban scenes, there are typically some man-made light sources, such as streetlights and headlights, appearing in the scene. The illuminance of the camera in this type of scenes is typically larger than 10 Lux. With a sufficient number of photons overall in the scene, noise could be ignored even at a low ISO value due to a relatively high signal-to-noise ratio.

For illustration, a group of visual comparisons between our night-time urban scenes and low-illuminated scenes from [46] are shown in Figure 6. We can see that due to the existences of man-made light sources, the illuminance in an auto-exposed camera is typically larger than 10 Lux, resulting in a negligible noise level (both in the original and histogram equalized night-time images). In contrast, noise can be a significant problem in the low-illuminated scenes if there is an insufficient number of photons captured by the camera due to a short exposure duration or a lack of light sources in the scene. In this situation (where the illuminance at the camera is typically less than 1 Lux), noise cannot be ignored due to the low signal-to-noise ratio. Hence, our method does not require any denoising mechanisms for the night-time scene parsing task.

V. EXPERIMENTS

In this section, we first introduce the exposure-aware Fi-score (EF1) as one of the evaluation metrics. We then compare our method to the mainstream day-time scene parsing methods, to reveal their limitations when applied to the night-time scene parsing task. Next, We justify the necessity of the proposed NightCity dataset. We further verify the effectiveness of the proposed model by comparing our method to ad hoc night-time scene parsing methods. Finally, we provide ablation studies to analyze different components of our proposed model.

Implementation Details. Our model is trained using the SGD optimizer with a batch size of 6 and an initial learning rate of 1e-5. We decrease the learning rate using a polynomial policy with a power of 0.9. Our model is trained for $40,000$ iterations, which takes about 12 hours on a PC with an i7-7700K CPU and two Nvidia 1080Ti GPUs. We set the image resolution to 500×500 for training and 900×900 for testing. All predictions are scaled to 1024×512, same resolution as the original image. We set α, β, w_1 and w_2 to 1, 0.01, 1, 0.3, respectively.

A. Evaluation Metrics

Following previous works [4], [50], [51], [13], we use mean IoU (mIoU) to evaluate the NTSP performance in our experiments. In addition, we propose an exposure-aware F1-score (EF1) to evaluate the segmentation performance under different exposure levels. EF1 takes both recall and precision into consideration. In particular, as in Figure 4, we divide night-time images into G groups according to the exposure levels. Hence, EF1 for group g is formulated as:

$$EF1_g = (1 + \beta^2) * \frac{(precision_g * recall_g)}{\beta^2 * precision_g + recall_g}, \quad (2)$$

where $precision_g$ and $recall_g$ are precision and recall for group g, respectively. We set β to 1 to balance the recall and precision values. To get a scalar-valued metric, we take the mean of EF1s of all groups as:

$$mEF1 = \frac{\sum_{g=1}^{G} EF1_g}{G}. \quad (3)$$

B. Comparison to Day-time Scene Parsing Methods

Compared Methods. We compare our model with several state-of-the-art scene parsing methods, including SegNet [55], FCN-8s [4], PSPNet [5], BiSeNet [56], PSANet [57], ESPNet [6], DFN [52], CCNet [19], VPLR [24] and HANet [20]. We use their released codes and train them using the hyper-parameters reported in their papers. Since all of these methods

TABLE III

COMPARISON OF OUR MODEL WITH STATE-OF-THE-ART DAY-TIME SCENE PARSING METHODS ON THE NIGHTCITY TEST SET. "C" COLUMNS: TRAINED ON CITYSCAPES. "C (IE)" COLUMNS: TRAINED ON CITYSCAPES, AND APPLYING DRHT [45] AS AN IMAGE ENHANCEMENT (IE) PRE-PROCESS DURING TESTING. "N" COLUMNS: TRAINED ON NIGHTCITY. "C + N" COLUMNS: TRAINED ON BOTH CITYSCAPES AND NIGHTCITY. THE BEST RESULTS ARE MARKED IN **BOLD**.

Methods	mIoU (%) ↑				mEF1 ↑			
	C	C (IE)	N	C + N	C	C (IE)	N	C + N
SegNet [55]	5.7	5.6	17.3	18.1	0.38	0.41	0.67	0.68
FCN-8s [4]	8.1	8.2	28.2	28.1	0.58	0.58	0.80	0.79
PSPNet [5]	12.6	11.1	46.3	46.5	0.61	0.54	0.87	0.86
BiSeNet [56]	6.1	8.8	50.0	46.2	0.42	0.53	0.87	0.85
PSANet [57]	7.8	8.2	28.8	28.5	0.53	0.57	0.70	0.70
ESPNet [6]	9.1	8.9	34.2	34.5	0.28	0.28	0.82	0.83
DFN [52]	9.6	9.6	50.1	52.3	0.33	0.34	0.87	**0.88**
CCNet [19]	11.1	10.9	49.2	49.8	0.42	0.46	0.85	0.86
VPLR [24]	12.5	11.2	50.4	52.4	0.65	0.60	0.87	**0.88**
HANet [20]	13.4	11.8	51.1	53.1	0.64	0.59	0.87	0.86
Ours	-	-	**51.8**	**53.9**	-	-	**0.88**	**0.88**

are developed for day-time domain, one naive solution for NTSP is to apply image enhancement to the input night-time images followed by an existing day-time method for scene parsing. Hence, we also evaluate this pre-processing approach using one of the latest enhancement methods DRHT [45] in our experiment.

Quantitative Results. Table III reports the experimental results. We have two observations here:

1) The existing day-time models trained on Cityscapes achieve poor performances ("C" columns), which are significantly worse than those of our model trained on NightCity, in terms of mIoU and mEF1. This shows that NTSP is rather challenging and the state-of-the-art scene parsing methods are not able to handle this problem well.

2) Adding image enhancement as a pre-process does not give obvious performance gains ("C (IE)" columns). To review the problems, we use DRHT [45] to enhance two example night images, as shown in Figure 7. We can see that while the under-exposed regions are enhanced, the over-exposure regions get worse. The enhanced images also have very different appearances from day-time ones, causing day-time scene parsing methods to fail.

Results of our model, when trained on our real night-time dataset, show superior performance over these existing methods trained on Cityscapes.

C. Evaluation of the Proposed NightCity Dataset

To investigate if our dataset can help improve the performance of existing methods, we have conducted three

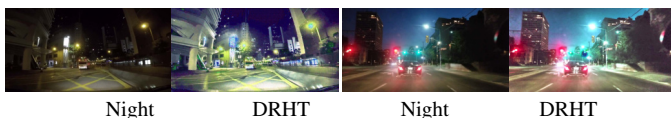Night   DRHT   Night   DRHT

Fig. 7. Examples of night images and the corresponding enhanced images using DRHT [45].

experiments. In the first experiment, we train the existing scene parsing models and our model on NightCity, instead of Cityscapes. In the second experiment, we train all models on Cityscapes and NightCity. In the third experiment, we compare the performances of existing methods trained on the existing night-time dataset BDD100K-night [36] and those trained on our NightCity.

Quantitative Results. We report the results of the first and second experiments on the NightCity test set in Table III. We have the following observations:

1) Cityscapes ("C" columns) vs. NightCity ("N" columns). We can see that after training on our NightCity, the performances of all existing methods have improved significantly, on both mIoU and mEF1. This suggests that our real night-time dataset is important to boosting the NTSP performance.

2) Although all models have significant performance gains after training on NightCity, our proposed model outperforms all existing methods. This demonstrates that our model, while simple, is very effective in handling night-time images, as compared with other methods. We also note that our model obtains the best mEF1 performance. To reveal how well our model performs under different exposure conditions, we also report the $EF1_g$ values for the 10 exposure groups (i.e., $g = \{1, \dots, 10\}$) in Table IV. We can see that our model outperforms the other models in all exposure groups, including [0, 0.2] (near under-exposure) and [0.8, 1] (near over-exposure).

3) NightCity ("N" columns) vs. Cityscapes + NightCity ("C + N" columns). We can see that after training on both datasets, some of the existing models have small performance gain, while others have small performance reduction. We believe that the performance gain of some methods may be due to the added context and object information from Cityscapes. On the other hand, the added day-time information may confuse the other methods, causing the performance reduction.

TABLE IV
COMPARISON OF OUR MODEL WITH STATE-OF-THE-ART METHODS ON EXPOSURE-AWARE F1 SCORE (EF1), WHICH QUANTIFIES THE SEGMENTATION
PERFORMANCE UNDER DIFFERENT EXPOSURE LEVELS. FROM LEFT TO RIGHT, EXPOSURE DEGREE INCREASES FROM UNDER-EXPOSURE TO
OVER-EXPOSURE. $EF1_g$ DENOTES DIFFERENT EXPOSURE LEVELS (OR BINS). ALL MODELS ARE TRAINED AND TESTED ON NIGHTCITY. THE BEST
RESULTS ARE HIGHLIGHTED IN **BOLD**.

Methods	$EF1_g$ ↑									
	[0, 0.1)	[0.1,0.2)	[0.2, 0.3)	[0.3, 0.4)	[0.4, 0.5)	[0.5, 0.6)	[0.6, 0.7)	[0.7, 0.8)	[0.8, 0.9)	[0.9, 1]
SegNet [55]	0.6335	0.6281	0.6945	0.7171	0.7132	0.6877	0.6625	0.6566	0.6507	0.6665
FCN-8s [4]	0.7717	0.7863	0.8069	0.8189	0.8209	0.8154	0.8027	0.8016	0.7952	0.8085
PSPNet [5]	0.8447	0.8634	0.8739	0.8750	0.8754	0.8739	0.8649	0.8624	0.8583	0.8882
BiSeNet [56]	0.8609	0.8674	0.8735	0.8735	0.8740	0.8736	0.8658	0.8637	0.8571	0.9073
PSANet [57]	0.7067	0.7087	0.7252	0.7228	0.7128	0.6964	0.6778	0.6744	0.6647	0.7252
ESPNet [6]	0.7860	0.8085	0.8325	0.8374	0.8399	0.8352	0.8257	0.8245	0.8178	0.8249
DFN [52]	0.8665	0.8697	0.8773	0.8763	0.8762	0.8762	0.8684	0.8671	0.8623	0.9119
CCNet [19]	0.8681	0.8623	0.8745	0.8737	0.8778	0.8765	0.8625	0.8678	0.8595	0.9109
VPLR [24]	0.8612	0.8655	0.8727	0.8725	0.8765	0.8711	0.8609	0.8613	0.8572	0.9021
HANet [20]	0.8656	0.8645	0.8788	0.8747	0.8812	0.8751	0.8703	0.8638	0.8632	0.9112
Ours	**0.8690**	**0.8751**	**0.8819**	**0.8819**	**0.8828**	**0.8814**	**0.8739**	**0.8738**	**0.8689**	**0.9155**

For the third experiment, we report the results in Table V. As there are currently no large-scale datasets with fine annotations for night-time scene parsing, we verify the effectiveness of the proposed dataset by comparing it with a subset of the BDD100K dataset (with annotations) [36] (denoted as BDD100K-night here), which contains 314 night-time images for training and 31 for test. We train three methods (i.e., PSPNet [5], CCNet [19], and HANet [20]) on the BDD100K-night training set and our training set, and test them on the BDD100K-night test set. The results in Table V show that our dataset can help improve the performances significantly, which demonstrates the effectiveness of our dataset.

Qualitative Results. Figure 8 qualitatively compares the results of our model with those of the best-performing six models (according to Table III) on some of the images in NightCity. These images have different degrees of under-/over-exposures, which render them difficult to recognize and segment. However, our model is able to handle them favorably. Particularly, in the first column, our model can produce more accurate and sharper boundaries on the building segmentation. In addition, it also gives more clear and complete sidewalks. In the second column, the traffic light with its pole is ignored by other models, but successfully recognized by ours. In the third column, our model gives more intact and clear shape of the pedestrian at the right end. In the fourth column, our model can detect a very small tree in an under-exposed region. Although BiseNet can also detect the tree, it generates a false positive segmentation of a person on the left side (the red segment marked by the yellow box). In the last column, while both our model and PSPNet are able to segment the buses near the camera well, PSPNet fails to give a correct segmentation of the distant bus (marked by the yellow box). These results once again demonstrate the superior performance of the proposed model on NTSP.

TABLE V
COMPARISON OF THE EXISTING METHODS TRAINED ON THE
BDD100K-NIGHT DATASET AND ON OUR NIGHTCITY DATASET.
METHODS ARE EVALUATED ON THE BDD100K-NIGHT TEST SET.

Methods	BDD100K-night [38]		NightCity-train	
	mIoU (%) ↑	mEF1 ↑	mIoU (%) ↑	mEF1 ↑
PSPNet [6]	36.8	0.69	51.2	0.76
CCNet [20]	41.4	0.72	54.6	0.81
HANet [21]	45.2	0.74	59.9	0.88

D. Comparison to Night-time Scene Parsing Methods

Quantitative Results. We compare our method to some ad hoc night-time scene parsing methods on one open challenge, i.e., the **Dark Zurich dataset** [10]. This dataset has 151 night-time images for evaluation. Since their labels are only available in their challenge, we submit our results to their challenge to obtain the results. Methods used for comparison include: GCMA [10], AdaptSegNet [58], ADVENT [60], BDL [61], DMAda [59], MGCDA [37], DANNet [38]. All these methods are domain-adaption-based methods, as there were no large-scale night-time scene parsing datasets with fine annotations for fully-supervised learning. The results are shown in Table VI, which verifies the superiority of our method.

Qualitative Results. We qualitatively compare our results with the results shown in GCMA [10]. As they only show results of eight images in their paper (including those in the supplemental), we therefore visually compare our results with their results of the eight images. Their results were from three NTSP methods, AdaptSegNet [58], DMAda [59], and GCMA [10]. These three compared methods are the same as those given in GCMA [10]. As shown in the top four rows of Figure 9, other methods mistakenly recognize a part of the sky as road (yellow boxes), while our model can produce the
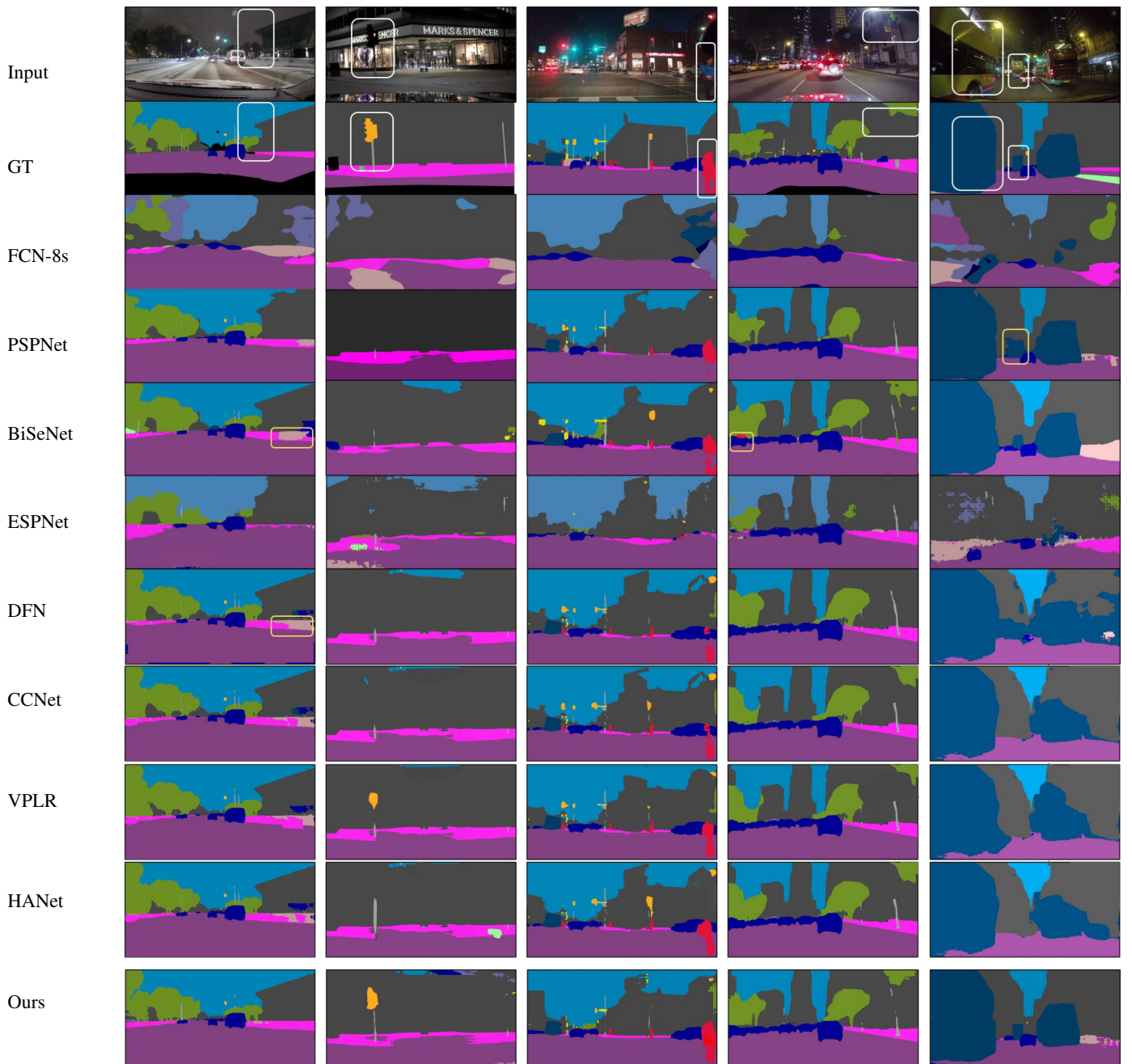
Fig. 8. Visual comparison of our results with those of the state-of-the-art methods. Our advantages are highlighted by white boxes. A few drawbacks of the other methods are marked by yellow boxes. All the methods are trained on NightCity.

very clear boundary of the tree (white boxes). In the fifth row, our model can detect a clear shape of the car. In the seventh row, our model produces a much cleaner building (white box), compared with another three methods. In the last row, only our model can identify the wall. These visual results show that our model can produce higher-quality segmentation maps under extreme lighting conditions.

E. Model Analysis

Model Robustness on Day-time Dataset. We evaluate our model on the existing day-time dataset Cityscapes to show its robustness. We provide the mIoU results of different methods trained on both Cityscapes and NightCity, and evaluated on the validation set of Cityscapes. The results are reported in Table VII. We can see that although our method is not specifically designed for day-time scene parsing, it can still achieve the second-best performance compared with the state-of-the-art methods on day-time images.

Ablation Study. To investigate the necessity of using the learned exposure features to guide the segmentation, we run an ablation study to compare our model against its four ablated alternatives: (1) remove the exposure guidance layer from our model ("$w/oEGL$"); (2) remove the entire exposure stream
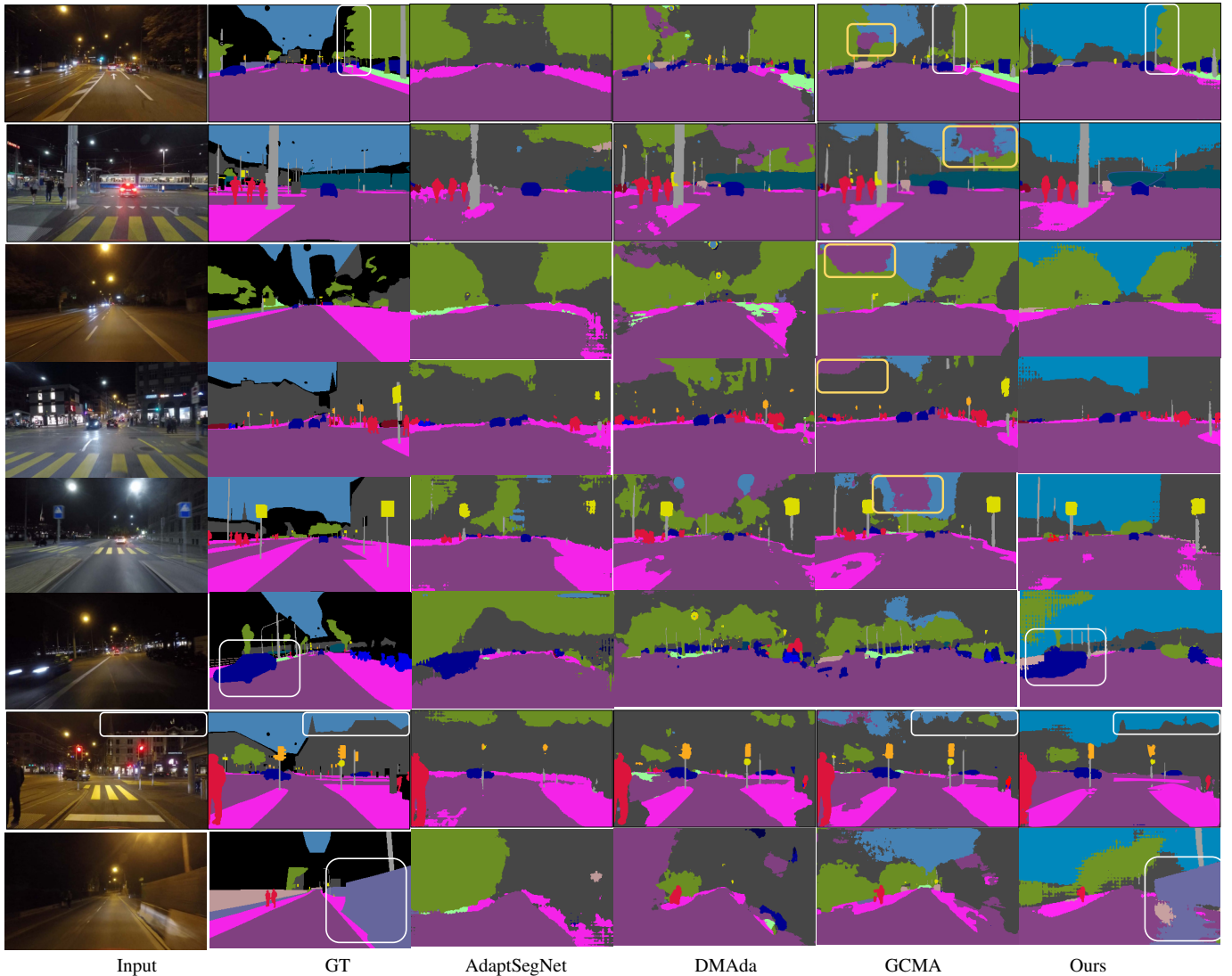
Fig. 9. Visual comparison of our results with those from AdaptSegNet [58], DMAda [59] and GCMA [10]. Our advantages are highlighted with white boxes and the limitations of existing methods are highlighted with yellow boxes. Our model can produce more accurate and robust segmentation.

TABLE VI
COMPARISON OF OUR MODEL WITH STATE-OF-THE-ART METHODS ON
DARK ZURICH-TEST [10]. THE BEST RESULT IS HIGHLIGHTED IN **BOLD**.

Methods	Venue&Year	mIoU (%) ↑
AdaptSegNet [58]	CVPR 2018	30.4
ADVENT [60]	CVPR 2019	29.7
BDL [61]	CVPR 2019	30.8
DMAda [59]	ITSC 2018	32.1
GCMA [10]	ICCV 2019	42.0
MGCDA [37]	TPAMI 2021	42.5
DANNet (DeepLab-v2) [38]	CVPR 2021	42.5
DANNet (RefineNet) [38]	CVPR 2021	44.3
DANNet (PSPNet) [38]	CVPR 2021	45.2
Ours	-	**45.4**

from our model ("w/oES"); (3) replace the attention operation in Eq. 1 with concatenation (Concat); and (4) with summation (Sum). We include alternatives (3) and (4) as they are straight-

forward ways of combining multiple features. Table VIII shows the results. We can see that when the exposure guidance layers are excluded, the performance of our model drops significantly, which confirms the importance of our exposure guidance. If we remove the exposure stream ("$w/o\ ES$"), the performance drops further, as there are no mechanisms for handling the over- and under-exposed regions. This suggests that learning to explicitly predict exposure information can help learn useful features for NTSP, Finally, we can see that us-ing concatenation and sum operations for fusing the exposure information can produce slightly better results. However, they still perform much worse than ours, which demonstrates the effectiveness of our attention operations in EGLs for exposure guidance, as the attention mechanism allows our method to aggregate different contextual information for segmentation according to the exposure-based attention map.

Strategies of Using the Exposure Map. To incorporate exposure information into a scene parsing model, one simple strategy is to directly use an exposure map as an additional input to the model or as an attention map to fuse the in-

TABLE VII

COMPARISON OF OUR MODEL WITH STATE-OF-THE-ART METHODS ON DAY-TIME IMAGES (THE VALIDATION SET OF CITYSCAPES) USING MIOU. THE BEST AND SECOND BEST RESULTS ARE MARKED IN **BOLD** AND BLUE, RESPECTIVELY.

Methods	Venue&Year	mIoU (%) ↑
FCN-8s [4]	CVPR 2015	62.2
SegNet [55]	TPAMI 2017	51.5
PSPNet [5]	CVPR 2017	75.3
BiSeNet [56]	ECCV 2018	39.6
PSANet [57]	ECCV 2018	41.2
ESPNet [6]	ECCV 2018	40.6
DFN [52]	CVPR 2018	75.8
CCNet [19]	ICCV 2019	**77.1**
CGNet [62]	TIP 2021	64.8
DSD [22]	TIP 2021	72.3
MagNet [63]	CVPR 2021	65.6
Ours	-	76.9

TABLE VIII

ABLATION STUDY. WE COMPARE OUR MODEL (OURS) WITH ITS ABLATED VERSIONS: WITHOUT EXPOSURE GUIDANCE LAYERS (W/O EGL), WITHOUT THE EXPOSURE STREAM (W/O ES), AND REPLACING THE ATTENTION OPERATIONS WITH CONCATENATION (CONCAT) AND SUM (SUM).

	w/o EGL	w/o ES	Concat	Sum	Ours
mIoU (%) ↑	41.6	40.1	42.5	43.1	51.8
mEF1 ↑	0.83	0.82	0.84	0.82	0.88

termediate features. To justify the advantage of our network design over these straightforward solutions, we compare our model with two baselines: (1) we use the exposure map as an additional input to our segmentation stream, by concatenating the exposure map with the RGB image, and remove the exposure stream (denoted as "as Extra Input"); (2) we replace the attention map W_r in Eq. 1 with the exposure map and remove the exposure stream (denoted as "as Attention Map").
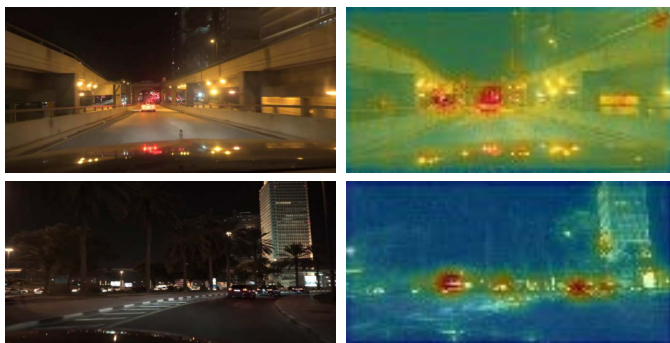

Fig. 10. Visualization of the attention maps learned by our guidance of exposure layer.

As shown in Table IX, our model outperforms the two ablated models by a large margin. If we feed the exposure map as input to the network, the network can only be supervised by the semantic labels. Since the exposure map does not contain much semantic information, its contribution to semantic segmentation prediction would not be modeled by

the network. Table IX shows that using the exposure map as input, instead of as the supervision signal, causes a significant performance drop from 51.8 to 35.6 (in terms of mIoU). In addition, by comparing Table IX to Table VIII, we can see that using the exposure map as input produces a worse performance (35.6 mIoU in Table IX) than the ablated model of removing the exposure stream "$w/o\ ES$" (40.1 mIoU in Table VIII). This suggests that the network may mistakenly use the exposure map as a weight map, which pays more attention to the over-exposed regions and less attention to the under-exposed regions. Table IX further shows that directly using the exposure map as an attention map produces a similar result (40.2 mIoU in Table IX) as the ablated model "$w/o\ ES$" (40.1 mIoU in Table VIII). This demonstrates that exposure information may not help in such an attentive manner.

Figure 10 shows examples of visualization of the attention maps learned by the exposure guidance layer. The first row shows that our model gives more attention to distant cars with lots of headlights (over-exposure) and the second row shows that our model pays more attention to the under-exposed regions. In contrast to the direct usage of exposure map as attention map that network would pay more attention to over-exposed regions, our model can learn to adaptively attend to both over- and under-exposed regions as all of them are crucial to segmentation performance of night-time images.

These experiments verify the effectiveness of our method of exploiting the exposure information, by first learning the exposure representation via the image-to-exposure translation step and then learning the fusion of semantic and exposure information via the exposure guidance layer.

TABLE IX

COMPARISON WITH TWO ABLATED MODELS OF USING THE EXPOSURE MAP: USING IT AS AN ADDITIONAL INPUT (AS EXTRA INPUT) AND AS AN ATTENTION MAP (AS ATTENTION MAP).

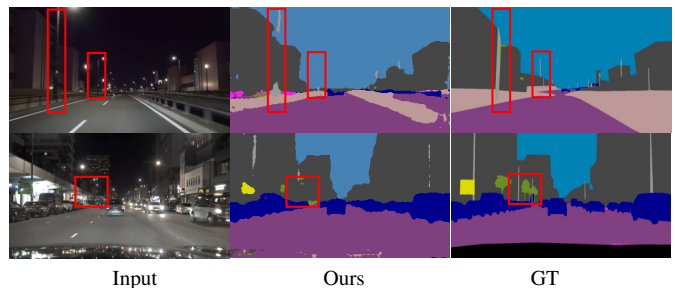|	as Extra Input	as Attention Map	Ours
mIoU (%) ↑	35.6	40.2	51.8
mEF1 ↑	0.73	0.76	0.88

Input   Ours   GT

Fig. 11. Failure cases. Our model may fail to detect objects that appear in large under-exposed regions, e.g., thin poles (top row) and trees (bottom row).

VI. CONCLUSION

In this paper, we have addressed the night-time scene parsing (NTSP) problem. To this end, we have proposed a large dataset of real night-time images with fine semantic annotations for training and benchmarking. We have also proposed a two-stream framework especially designed to address

the NTSP problem, which explicitly learns exposure features to augment the scene parsing process. Our results show that the proposed dataset can benefit existing scene parsing methods when applied to night-time scenes. We have also demonstrated that our proposed model trained on our dataset outperforms all existing methods, yielding state-of-the-art performance.

Although we have demonstrated the effectiveness of our model on night-time scenes, our model may fail in some extremely challenging situations, e.g., if an under-exposure region is large. Figure 11 shows two failure examples of our model, in which it fails to detect the thin poles (top row) and the trees (bottom row). All these objects are located in large under-exposed regions and are visually difficult to identify even for human. As a future work, we would like to consider using the raw data from the camera to address this problem.

REFERENCES

[1] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *IEEE CVPR*, 2017, pp. 6757 – 6765.

[2] X. Yang, K. Xu, S. Chen, S. He, B. Y. Yin, and R. Lau, "Active matting," in *NeurIPS*, 2018.

[3] L. Yang, X. Liang, T. Wang, and E. Xing, "Real-to-virtual domain unification for end-to-end autonomous driving," in *ECCV*, 2018, pp. 530–545.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE CVPR*, 2015, pp. 3431–3440.

[5] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE CVPR*, 2017.

[6] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *ECCV*, 2018.

[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE CVPR*, 2016.

[8] H. Alhaija, S. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, "Augmented reality meets deep learning for car instance segmentation in urban scenes," in *BMVC*, 2017.

[9] G. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[10] C. Sakaridis, D. Dai, , and L. Van Gool, "Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," in *IEEE ICCV*, 2019.

[11] Q. Wang, J. Gao, and Y. Yuan, "A joint convolutional neural networks and context transfer for street scenes labeling," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1457–1470, 2017.

[12] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "Exfuse: Enhancing feature fusion for semantic segmentation," in *ECCV*, 2018, pp. 269–284.

[13] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *IEEE CVPR*, 2017, pp. 1925–1934.

[14] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang, "Multi-scale context intertwining for semantic segmentation," in *ECCV*, 2018, pp. 603–619.

[15] B. Cheng, L.-C. Chen, Y. Wei, Y. Zhu, Z. Huang, J. Xiong, T. Huang, W.-M. Hwu, and H. Shi, "Spgnet: Semantic prediction guidance for scene parsing," in *IEEE ICCV*, 2019.

[16] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," *IEEE CVPR*, 2019.

[17] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-scnn: Gated shape cnns for semantic segmentation," in *IEEE ICCV*, 2019.

[18] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *IEEE ICCV*, 2019.

[19] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *IEEE ICCV*, 2019.

[20] S. Choi, J. Kim, and J. Choo, "Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks," in *IEEE CVPR*, 2020, pp. 9373–9383.

[21] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *IEEE CVPR*, 2017.

[22] Y. Feng, X. Sun, W. Diao, J. Li, and X. Gao, "Double similarity distillation for semantic image segmentation," *IEEE Trans. on Image Processing*, vol. 30, pp. 5363–5376, 2021.

[23] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *ECCV*, 2020.

[24] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. Newsam, A. Tao, and B. Catanzaro, "Improving semantic segmentation via video propagation and label relaxation," in *IEEE CVPR*, 2019.

[25] Q. Geng, H. Zhang, X. Qi, G. Huang, R. Yang, and Z. Zhou, "Gated path selection network for semantic segmentation," *IEEE Tran. on Image Processing*, vol. 30, pp. 2436–2449, 2021.

[26] Z. Huang, C. Wang, X. Wang, W. Liu, and J. Wang, "Semantic image segmentation by scale-adaptive networks," *IEEE Trans. on Image Processing*, vol. 29, pp. 2066–2077, 2020.

[27] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *ICCV*, 2017, pp. 2020–2030.

[28] Q. Wang, J. Gao, and X. Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," *IEEE Trans. on Image Processing*, vol. 28, no. 9, pp. 4376–4386, 2019.

[29] Y. Liu, W. Zhang, and J. Wang, "Source-free domain adaptation for semantic segmentation," in *IEEE CVPR*, 2021, pp. 1215–1224.

[30] F. Barbato, M. Toldo, U. Michieli, and P. Zanuttigh, "Latent space regularization for unsupervised domain adaptation in semantic segmentation," in *IEEE CVPR*, 2021, pp. 2835–2845.

[31] S. Di, Q. Feng, C.-G. Li, M. Zhang, H. Zhang, S. Elezovikj, C. C. Tan, and H. Ling, "Rainy night scene understanding with near scene semantic adaptation," *IEEE Trans. on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1594–1602, 2021.

[32] F. Tung, J. Chen, L. Meng, and J. Little, "The raincouver scene parsing benchmark for self-driving in adverse weather and at night," *IEEE Robotics and Automation Letters*, vol. 2, no. 4, pp. 2188–2193, 2017.

[33] Z. Zheng, Y. Wu, X. Han, and J. Shi, "Forkgan: Seeing into the rainy night," in *ECCV*, 2020.

[34] A. Valada, J. Vertens, A. Dhall, and W. Burgard, "Adapnet: Adaptive semantic segmentation in adverse environmental conditions," in *IEEE ICRA*, 2017, pp. 4644–4651.

[35] O. Zendel, K. Honauer, M. Murschitz, D. Steininger, and G. Fernandez Dominguez, "Wilddash-creating hazard-aware benchmarks," in *ECCV*, 2018, pp. 402–416.

[36] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *IEEE CVPR*, 2020, pp. 2636–2645.

[37] C. Sakaridis, D. Dai, and L. Van Gool, "Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2021.

[38] X. Wu, Z. Wu, H. Guo, L. Ju, and S. Wang, "Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation," in *CVPR*, 2021, pp. 15 769–15 778.

[39] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu, "Automatic photo adjustment using deep neural networks," *ACM TOG*, vol. 35, no. 2, pp. 1–15, 2016.

[40] S. Wang, J. Zheng, H. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Trans. on Image Processing*, vol. 22, no. 9, pp. 3538–3548, 2013.

[41] C. Li, J. Guo, F. Porikli, and Y. Pang, "Lightennet: a convolutional neural network for weakly illuminated image enhancement," *Pattern Recognition Letters*, vol. 104, pp. 15–22, 2017.

[42] Z. Ying, G. Li, Y. Ren, R. Wang, and W. Wang, "A new low-light image enhancement algorithm using camera response model," in *IEEE ICCV*, 2017.

[43] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia, "Underexposed photo enhancement using deep illumination estimation," in *IEEE CVPR*, 2019.

[44] J. Cai, S. Gu, L. Zhang, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. on Image Processing*, vol. 27, no. 4, pp. 2049 – 2062, 2018.

[45] X. Yang, K. Xu, Y. Song, Q. Zhang, X. Wei, and R. W. Lau, "Image correction via deep reciprocating hdr transformation," in *IEEE CVPR*, 2018.

[46] K. Xu, X. Yang, B. Yin, and R. W. Lau, "Learning to restore low-light images via decomposition-and-enhancement," in *IEEE CVPR*, 2020, pp. 2281–2290.

[47] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "Labelme: a database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[48] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, 2018.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv:1512.03385*, 2015.

[50] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *arXiv:1606.00915*, 2016.

[51] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters – improve semantic segmentation by global convolutional network," in *IEEE CVPR*, 2017, pp. 1743 – 1751.

[52] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *IEEE CVPR*, 2018.

[53] L. Chen, Y. Yang, J. Wang, W. Xu, and A. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *IEEE CVPR*, 2016, pp. 3640–3649.

[54] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *IEEE CVPR*, 2017, pp. 1–10.

[55] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for scene segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2017.

[56] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *ECCV*, 2018, pp. 325–341.

[57] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. Change Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in *ECCV*, 2018, pp. 267–283.

[58] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *IEEE CVPR*, 2018, pp. 7472–7481.

[59] D. Dai and L. Van Gool, "Dark model adaptation: Semantic image segmentation from daytime to nighttime," in *IEEE Intelligent Transportation Systems Conference*, 2018, pp. 3819–3824.

[60] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *CVPR*, 2019, pp. 2517–2526.

[61] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *CVPR*, 2019, pp. 6936–6945.

[62] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "Cgnet: A light-weight context guided network for semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 1169–1179, 2021.

[63] C. Huynh, A. T. Tran, K. Luu, and M. Hoai, "Progressive semantic segmentation," in *CVPR*, 2021, pp. 16 755–16 764.
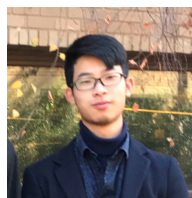
Xin Tan is now a Ph.D. candidate at the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. He is also a joint-Ph.D. student at the Department of Computer Science, City University of Hong Kong since 2019. He received his B.Eng. degree in Automation from the Chongqing University, China in 2017. His research interests lie in computer vision and deep learning, in particular, semantic segmentation and saliency detection. He serves as a program committee member/reviewer for CVPR 2021, ICCV 2021, AAAI 2021, IJCAI 2021 and International Journal of Computer Vision (IJCV).

Ke Xu is currently with the Department of Computer Science and Engineering at Shanghai Jiao Tong University. He obtains the dual Ph.D. degrees from Dalian University of Technology and City University of Hong Kong. He has served as a program committee member/reviewer for several CV and AI conferences and journals, including CVPR, ICCV, NeurIPS, ICLR, ICML, IJCV, TCSVT. His research interests include deep learning, image enhancement and editing.

Ying Cao received the Ph.D. degree in computer science from the City University of Hong Kong in 2014, and the M.Sc. and B.Eng. degrees in software engineering from Northeastern University, China, in 2010 and 2008, respectively. He was a Postdoctoral Fellow in the Department of Computer Science at the City University of Hong Kong from 2015 to 2016. His primary research interest lies in data-driven computational manga and graphic design.

Yiheng Zhang received his Honors Degree of Bachelor of Science in Computer Science at Shanghai Jiao Tong University in 2019 where he was a member of Zhiyuan Honors Program of Engineering. He is now a master of Science in Computer Science student at Stanford University, specialized in Real-World Computing.

Lizhuang Ma received his B.S. and Ph.D. degrees from the Zhejiang University, China in 1985 and 1991, respectively. He is now a Distinguished Professor, at the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China and the School of Computer Science and Technology, East China Normal University, China. He was a Visiting Professor at the Frounhofer IGD, Darmstadt, Germany in 1998, and a Visiting Professor at the Center for Advanced Media Technology, Nanyang Technological University, Singapore from 1999 to 2000. His research interests include computer vision, computer aided geometric design, computer graphics, scientific data visualization, computer animation, digital media technology, and theory and applications for computer graphics, CAD/CAM. He serves as the reviewer of IEEE TPAMI, IEEE TIP, IEEE TMM, CVPR, AAAI etc.

Rynson W.H. Lau received his Ph.D. degree from University of Cambridge. He was on the faculty of Durham University and is now with City University of Hong Kong.

Rynson serves on the Editorial Board of the International Journal of Computer Vision (IJCV) and Computer Graphics Forum. He has served as the Guest Editor of a number of journal special issues, including ACM Trans. on Internet Technology, IEEE Trans. on Multimedia, IEEE Trans. on Visualization and Computer Graphics, and IEEE Computer Graphics & Applications. He has also served in the committee of a number of conferences, including Program Co-chair of ACM VRST 2004, ACM MTDL 2009, IEEE U-Media 2010, and Conference Co-chair of CASA 2005, ACM VRST 2005, ACM MDI 2009, ACM VRST 2014. Rynson's research interests include computer graphics and computer vision.