Multi-Modal Learning for Real-Time Automotive Semantic Foggy Scene Understanding via Domain Adaptation

Naif Alshammari[1,2], Samet Akcay[1,3], and Toby P. Breckon[1,4]

***Abstract*— Robust semantic scene segmentation for automotive applications is a challenging problem in two key aspects: (1) labelling every individual scene pixel and (2) performing this task under unstable weather and illumination changes (e.g., foggy weather), which results in poor outdoor scene visibility. Such visibility limitations lead to non-optimal performance of generalised deep convolutional neural network-based semantic scene segmentation. In this paper, we propose an efficient end-to-end automotive semantic scene understanding approach that is robust to foggy weather conditions. As an end-to-end pipeline, our proposed approach provides: (1) the transformation of imagery from foggy to clear weather conditions using a domain transfer approach (correcting for poor visibility) and (2) semantically segmenting the scene using a competitive encoder-decoder architecture with low computational complexity (enabling real-time performance). Our approach incorporates RGB colour, depth and luminance images via distinct encoders with dense connectivity and features fusion to effectively exploit information from different inputs, which contributes to an optimal feature representation within the overall model. Using this architectural formulation with dense skip connections, our model achieves comparable performance to contemporary approaches at a fraction of the overall model complexity.**

I. Introduction

Semantic scene segmentation is an active research topic that targets robust pixel-level image classification. However, as the reported performance of many state-of-the-art scene understanding algorithms is limited to ideal weather conditions, extreme weather and illumination variation could lead to unexpectedly inaccurate scene classification and segmentation [2], [7], [10], [25], [28], [42]. To date, too little attention has been paid to address the issue of automotive scene understanding under extreme weather conditions (e.g., Foggy weather) [8], [36], as the multitude of proposed deep learning approaches are generally only evaluated on ideal weather conditions. To overcome this shortcoming, the present paper introduces an efficient algorithm that tackles the challenge of automotive scene understanding in extreme weather conditions using a novel multi-modal learning approach that translates foggy scene images to clear scenes and utilizes both depth and luminance information to achieve superior semantic segmentation performance.

Previous attempts to tackle the issue of scene understanding under non-ideal weather conditions for shadow removal and illumination reduction [2], [25], [28], [42], haze removal and scene defogging [16], [29], [43], and foggy scene understanding [8], [36] are mostly based on conventional image enhancement and dehazing methods. Despite the general trend of performance improvement within automotive scene understanding [4], [17], [27], [47], there is still significant room for improvement across the spectrum of non-ideal operating conditions. In parallel with using recent image segmentation techniques [15], [20], [21], [40], employing the concept of image-to-image translation to map one domain onto another [22], [48] is a useful step that enables accurate semantic segmentation performance under extreme weather conditions.

In this work, we propose an efficient end-to-end automotive semantic scene understanding capable of performing under foggy weather conditions. We employ domain adaptation within scene understanding as a method to correct for the degraded visibility present under foggy weather conditions. In addition, we use a lightweight semantic segmentation architecture that incorporates RBG colour, luminance and depth images via distinctive encoders contributing to a deeper extraction for the representation of different features, which leads to superior segmentation performance. As an

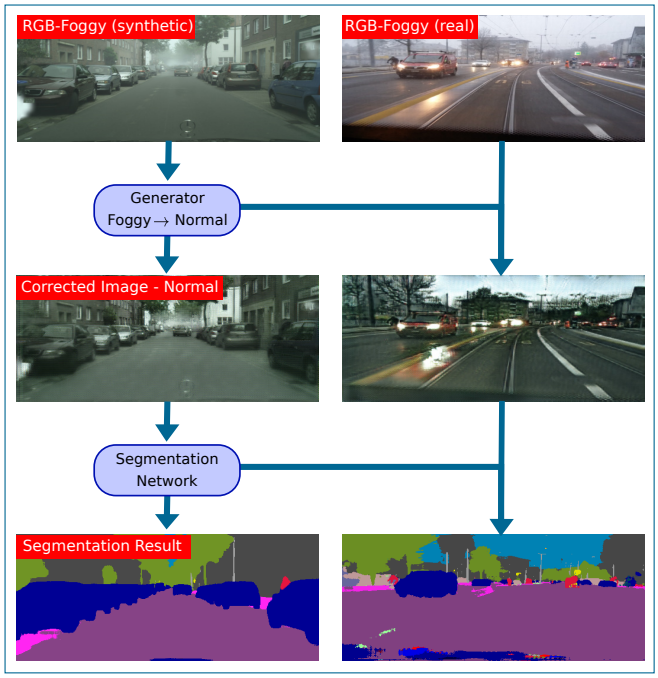

Fig. 1. An illustration of our semantic segmentation approach under foggy weather conditions, trained on *Foggy Cityscapes* [36] (partially synthetic data) and evaluated on *Foggy Driving* [36] (real foggy scenes). Degraded scenes visibility present under foggy weather conditions are corrected using domain adaptation [48] to serve a better semantic segmentation performance.

[1]Department of Computer Science, Durham University, UK
[2]Department of Natural and Applied Science, Community College, Majmaah University, Majmaah 11952, Saudi Arabia
[3]Intel, UK
[4]Department of Engineering, Durham University, UK

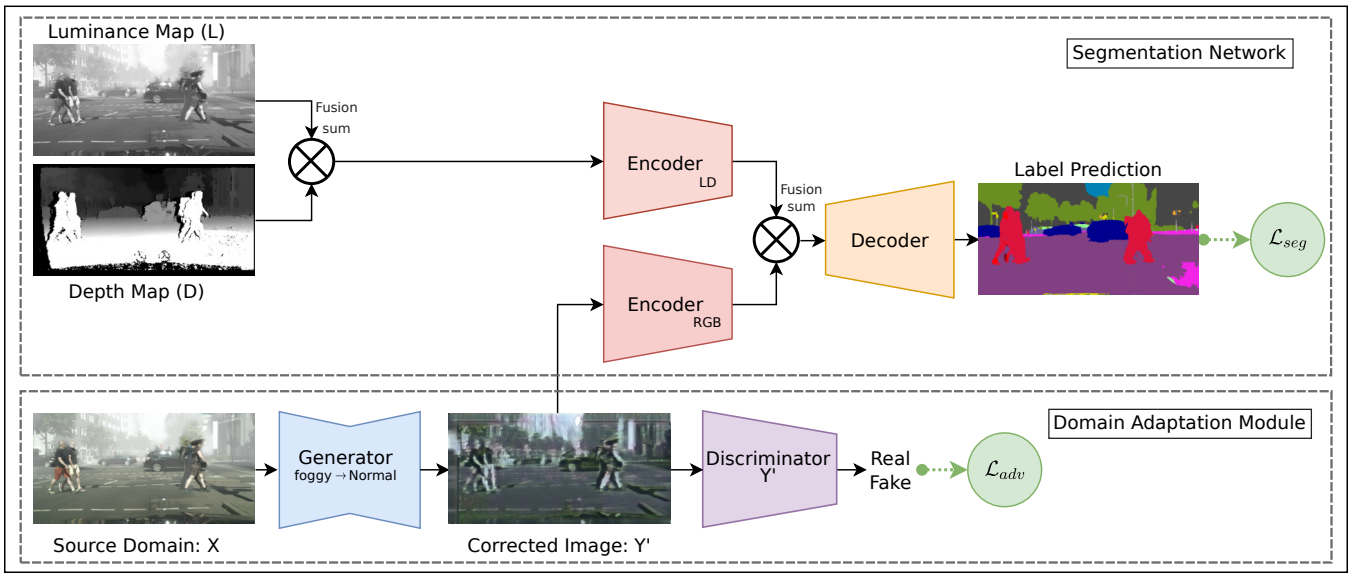

Fig. 2. Overview of our approach using [21], [48]. The source domain X (foggy scene) mapped to the target domain Y' (corrected image). Subsequently, the corrected image Y' is fed to the RGB encoder in the semantic segmentation network. The Depth (D) and luminance (L) images are incorporating RGB colour via the LD encoder. Finally, the output from the two encoders is passed to the semantic segmentation decoder.

integration methodology within encoders, we use a fusion-based connection. To avoid information loss and share high-resolution features in the latter reconstruction stages of CNN upsampling, we leverage skip-connections [30], [34], [40], [44]. Figure 1 provides an illustration of our foggy semantic scene understanding approach within domain adaptation.

II. Related Work

The related work is organized into two main categories: (i) semantic segmentation (Section II-A) and (ii) domain transfer (Section II-B).

A. Semantic Segmentation

Modern segmentation techniques utilize deep convolutional neural networks and outperform the traditional approaches by a large margin [3], [4], [17], [27], [47]. These contributions use a large dataset such as ImageNet [35] for pre-trained models. Recent segmentation techniques have distinct characteristics denoted by their design such as: (1) network topology: pooling indices [3], skip connection [34], multi-path refinement [27], pyramid pooling [47], fusion-based architecture [15] and dense connectivity [20], (2) varying input: colour RGB or RGB-D with depth [15], [19], depth and luminance [21], and illumination invariance [1], and (3) consideration of adverse-weather conditions [8], [14], [36]. As the main objective of this work is semantic scene segmentation under foggy weather conditions, recent studies in this specific domain are specifically presented in this section.

Different approaches have been proposed for tackling the issue of scene understanding under adverse weather conditions. To address illumination changes, an illumination-invariant colour space approach was proposed in [1], [23], [25] to minimize scene colour variations due to varying scene lighting conditions. Other approaches [8], [36], [38] addressed scene segmentation under foggy weather conditions using a semi-supervised approaches through generating synthetic foggy images from real-world data, and augmenting clear images with their synthetic fog images. By adapting segmentation models from day to night scenes, [37] addressed the issue of vision under nocturnal conditions.

Similarly, our model is trained on *foggy* scenes (synthetic images) adapted to *normal* using domain adaptation via style transfer (Section II-B) . Inspired by LDFNet [21], we employ the idea of incorporating luminance and depth alongside RGB, utilizing skip connections as well as fused features to perform semantic segmentation under foggy weather conditions.

B. Domain Transfer

Transferring an image from its real domain to another differing domain allows multiple uses of such images taken in complex environments or generated in different forms. Using recent advances in the field of image style transfer, [11], where target images are generated by capturing the style texture information of the input image by utilizing the Gram matrix, work by [26] shows that image style transfer (from the source domain to the target domain) is the fundamental process by which the differences between source and target distribution are minimized.

Recent methods [22], [39], [48] used Generative Adversarial Networks (GAN) [13] to learn mapping from source to target images. Based on training over a large dataset for specific image style, CycleGAN [48] shows an efficient approach to transfer image style from one image domain into another.

Within the context of semantic segmentation, we take advantage of GAN [13] to improve semantic segmentation by generating target scenes (clear-weather scenes) from the source domain (foggy scenes images) as source images

I_x mapped into a target domain I_y — hence significantly increasing our available image data training resources.

III. Proposed Approach

Our main objective is to train an end-to-end network that semantically labels every pixel in a scene that is invariant to both weather and illumination variations. We make use of *Foggy Cityscapes* dataset [36] for training. However, as the visibility is degraded due to fog, we attempt to reduce this sensing challenge using a model trained to transfer the style of *foggy* scenes to *normal*. *Foggy Driving* [36] and *Foggy Zurich* [8] are used as independent test sets comprising real world evaluation examples.

In general, our approach consists of two sub-components, namely domain transfer and semantic segmentation (each functioning as an integrated unit). These sub-components produce two separate outputs: a transformed clear-scene image (generated from a foggy domain) and semantic pixel labels. The pipeline of our approach is shown in Figure 2. In this section, we provide a detailed overview of these two sub-components: Domain Transfer (Section III-A) and Semantic Segmentation (Section III-B).

A. Domain Transfer

Our goal is to learn mapping $\mathcal{D}: X \rightarrow Y$ from the source domain X (foggy scenes) to the target domain Y (clear-weather) for which we assume that the scene visibility level in the constructed image is the optimal input to the subsequent Semantic Segmentation (Section III-B). We use GAN [13] to learn this domain transfer mapping function (shown in Figure 2, lower). A generator $G_{X \rightarrow Y}$ (generating clear scenes samples Y') and a discriminator D_Y (discriminating between Y and Y') are used to perform the mapping function from the source and target domains. The loss for each generator G coupled with a discriminator D is calculated as follows:

$$\mathcal{L}_{adv}(X \rightarrow Y) = \min_{G_{Y \rightarrow X}} \max_{D_Y} \mathbb{E}_{y \sim \mathbb{P}_d(y)}[\log(D)_{(y)}] + \mathbb{E}_{x \sim \mathbb{P}_d(x)}[\log(1 - D_Y(G_{X \rightarrow Y}(x)))], \quad (1)$$

where $\mathbb{P}_d$ is the data distribution, X the source domain with samples x and Y the target domain with the samples y.

B. Semantic Segmentation

As a subsequent component of the overall model, our pipeline performs the task of semantic segmentation on the corrected images Y' (mapped from *foggy* X to *normal* Y weather conditions via domain adaptation [13] as $G_{X \rightarrow Y}(X) = Y'$) incorporated with luminance L with or without depth D (shown in Figure 2, upper). Motivated by [21], we use an auto-encoder architecture for semantic segmentation, consisting of two distinctive encoders for image downsampling and features extraction: RGB encoder (E_{RGB}) and luminance with depth encoder (E_{LD}) or only luminance (E_L) if depth is not available (Figure 2, upper). Using explicit encoders for RGB colour, luminance and depth are considered to efficiently exploit information from

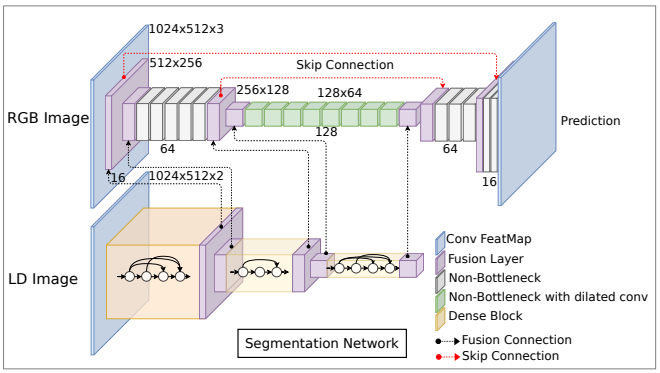

Fig. 3. Details of the segmentation network which consists of two encoders taking two types of inputs: **RGB Image** and **LD Image** (Luminance and Depth channels).

different inputs such as luminance and depth [15], [21]. As seen in Figure 3, we utilize dense-connections and features fusion to gain an optimal feature representation. To subsequently upsample the feature maps representation to the original input dimension, corresponding decoder is used (Figure 2, upper). Our encoder-decoder architecture is leveraging skip-connections to keep sharing high-level features which leads to a superior semantic segmentation performance (Figure 3).

RGB encoder: Designed to deal with a three-channel RGB input, the RGB encoder (E_{RGB}) (adopted from [33]) comprises of three downsampling stages ($\{16, 64, 128\}$) performed by three downsampler blocks consisting of convolutional and max pooling layers which are subsequently concatenated and followed by batch normalization and *ReLu()* activation function. The first downsampling stage only applies the downsamper block to extract the input features and reduce its dimensions. Subsequently, five non-bottleneck modules are implemented including the factorized convolutions (convolution kernel $n \times n$ factorized into $n \times 1$ and $1 \times n$ kernels), each followed by (*ReLu()*) and (batch normalization with *ReLu()*) respectively. With dilated and factorized convolutions, eight non-bottleneck modules were implemented as the last component of E_{RGB} (Figure 3).

LD encoder: Unlike the RGB encoder, the luminance and depth encoder (E_{LD}) (adopted from [21]) deals with luminance with or without depth images (concatenated as two-channel input). As a parallel functioning to E_{RGB}, E_{LD} is designed with a dense connectivity [20] technique to enhance the information flow from the earlier to the last layers. This design increases the effectiveness of the model by reinforcing information propagation when performance is degraded [15]. E_{LD} consists of three downsampling stages (the same as in E_{RGB}) design with three dense blocks; each has $\{4, 3, 4\}$ modules respectively. Similar to the first stage in E_{RGB}, the input is downsampled using the downsampler block discussed in E_{RGB}. In the second and third stages, a transition layer (1×1 convolution followed by batch normalization, *ReLu()* and 2×2 average pool layer). As some datasets do not contain depth maps, we used a luminance-only encoder (E_L) that is identical to (E_{LD}) except that it only takes a luminance channel. We make use of a distinct

Fig. 4. Sample image from *Cityscapes* [6] (top left) followed by (clockwise) foggy images (partially synthetic) with varying visibility (light to dense) from *Foggy Cityscapes* [8].

Fig. 5. Sample images from **Foggy Zurich** [8] and **Foggy Driving** [36] (real-world foggy datasets) along with their annotations.

encoder for luminance and depth to exploit deeper and better representation from the depth and luminance maps [15], [21] (Figure 3).

E_{RGB} and E_{LD} are linked by fusing output layers from blocks sharing the same number of channels among E_{RGB} and E_{DL}. Since it requires less computational cost, the fusion connectivity is implemented by summing the two layers such that for inputs x and y, the fused feature map is $E_{RGB}(x) + E_{LD}(y)$ or $E_L(y)$.

Decoder: After fusing the feature maps extracted from the last layer of E_{RGB} and either E_{DL} or E_L, a decoder upsamples the feature maps to the original resolution. The upsampling is implemented in three stages $\{64, 16, 19\}$. In the first two stages, transposed convolution, batch normalization, and *ReLu()* activation function, as well as two non-bottleneck modules, are employed. As the last component in the encoder, the transposed convolution layer maps the output to 19 class labels which we aim to predict (Figure 3).

Unlike LDFNet [21], we utilize skip connections for the fused features from the encoders into the decoder to avoid the loss of the high-level spatial features after being down-sampled (Figure 3). The fused feature maps {64, 16} passed from the encoders are concatenated with the corresponding upsampled feature maps in the decoder. As a semantic segmentation loss function, cross-entropy with pixel-wise *softmax()* is used summing over all pixels within a patch as follows:

$$P_k(x) = \frac{e^{ak(x)}}{\sum_{k'=1}^{K} e^{a_{k'}(x)}}, \tag{2}$$

$$\mathcal{L}_{seg} = -\log(P_l(S(x))), \tag{3}$$

where $S(x)$ denotes the output of the segmentation network, K is the number of classes, $P_k(x)$ is the approximated maximum function, and l is the ground truth label, $a_k(x)$ the feature activation for the channel k. As an overall loss, a joint overall loss function for our model is calculated as follows:

$$\mathcal{L} = \mathcal{L}_{adv}(I_y) + \mathcal{L}_{seg}(I_s). \tag{4}$$

The weighted sum of losses in the joint loss is dynamically updated using the homoscedastic uncertainty technique to weight and balance the two losses [5].

IV. Dataset

The availability of numerous well-annotated datasets [6], [9], [12], [35] has led to a proliferation of semantic segmentation studies. In this section, we will present the following datasets used in this paper: *Cityscapes dataset* [6] as the base dataset representing clear scenes, and *Foggy Cityscapes dataset* [36] as a partially synthetic data where fog is added into [6] (fine weather). As real-world datasets for *foggy* weather conditions, *Foggy Driving* [36] and *Foggy Zurich* [8] are used.

Cityscapes Dataset: We evaluate our approach on the *Cityscapes* [6] (Figure 4), a large dataset collected for urban-scene semantic segmentation. The dataset comprises of $2,975$ training and 500 testing image examples (resolution: 1024×2048) with 19 pixel classes: {*road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle and bicycle*}. In addition to semantic labels, *Cityscapes* provides disparity dataset labelled using Semi-Global Matching [18], used as a complementary information for semantic segmentation.

Foggy Cityscapes Dataset: *Foggy Cityscapes* [36] is a partially synthetic data generated from Cityscapes [6] by adding synthetic fog to the real images using fog simulation [36]. Three different versions of this dataset (shown in Figure 4) exist with varying fog density levels (controlled using attenuation coefficient $\beta \in \{0.005, 0.01, 0.02\}$ —from light to dense fog) and were used in the present study. This dataset inherits the annotations from *Cityscapes* [6] as labels for the synthetic foggy datasets as well as disparities. *Foggy Cityscapes* dataset consists of 8925 training and 1500 testing image examples (resolution: 1024×2048).

Foggy Driving Dataset: The *Foggy Driving* dataset [36] (Figure 5) is a real-world dataset collected in foggy-weather conditions, consisting of 101 images (resolution: 960×1280) with annotations for semantic segmentation and object detection tasks. Following *Cityscapes* [6] dataset, the *Foggy*

Methods				Foggy Zurich	Foggy Driving	Complexity of the Network	
Models	Network Architecture	Training	Fine-Tuning	Mean IoU	Mean IoU	Number of Parameters	FPS
CMDAda [8]	AdSegNet [41] - DeepLab-v2 [4]	C	—	25.0	29.7	44.0M	20
SFSU [36]	Dilated Conv. Net. (DCN) [46]	C	FC (498)	35.7	46.3	134M	-
CMAda2+ [38]	RefineNet-ResNet-101 [27]	C	FC (498)	43.4	49.9	118M	22
CMAda3+ [8]	RefineNet-ResNet-101 [27]	C	FC (498)	46.8	49.8	118M	22
Hanner *et al.* [14]	RefineNet-ResNet-101 [27]	C	FS (24,500)	40.3	48.4	118M	22
Hanner *et al.* [14]	RefineNet-ResNet-101 [27]	C	FS (498)	42.7	48.6	118M	22
Hanner *et al.* [14]	RefineNet-ResNet-101 [27]	C	FC+FS (498)	41.4	50.7	118M	22
Hanner *et al.* [14]	BiSeNet [45]	C	—	16.1	27.2	50.8M	-
Hanner *et al.* [14]	BiSeNet [45]	C	FC (498)	25.0	30.3	50.8M	-
Hanner *et al.* [14]	BiSeNet [45]	C	FS (24,500)	27.8	30.9	50.8M	-
Hanner *et al.* [14]	BiSeNet [45]	C	FS (498)	27.6	31.8	50.8M	-
Hanner *et al.* [14]	BiSeNet [27]	C	FC+FS (498)	35.2	30.9	118M	22
Ours w/o domain adaptation	—	C	FC (498)	8.7	17.6	**2.4M**	**42**
Ours w/ domain adaptation	—	C	FC (498)	21.4	29.4	**13.8M**	20

TABLE I

QUANTITATIVE COMPARISON OF SEMANTIC SEGMENTATION ON FOGGY ZURICH [8] AND FOGGY DRIVING [36] DATASETS OF OUR APPROACH AGAINST STATE-OF-THE-ART APPROACHES. **C**: CITYSCAPES [6]; **FC** FOGGY CITYSCAPES [36]; **FS**: FOGGY SYNSCAPES [14]. THE SPEED COMPARISON (FRAMES PER SECOND (FPS) IS BASED ON THE CITYSCAPES [6] TEST DATASET.

Driving dataset is labelled with 19 classes (33 images with fine annotations and 68 images coarsely annotated).

Foggy Zurich Dataset: The *Foggy Zurich* [8] (Figure 5) is a real-world foggy-scenes dataset consisting of 3808 images (resolution: 1920×1080) collected in Zurich. Following the approach of *Cityscapes* [6], *Foggy Zurich* provides pixel-level annotations for 40 scenes (finely annotated), including dense fog.

V. IMPLEMENTATION DETAILS

We implement our approach in PyTorch [31]. For optimization, we employ ADAM [24] with an initial learning rate of 5×10^{-3} and momentum of $\beta_1 = 0.5, \beta_2 = 0,999$. By following [32] and [21], we weight the classes of the dataset duo to imbalance number of pixels of each class in the dataset as follows:

$$\omega_{class} = \frac{1}{\ln(c + p_{class})}, \tag{5}$$

where c is an additional parameter empirically set to 1.10 to restrict the class weight and p_{class} is the probability of belonging to that class. We train the model for 100 epoch by using NVIDIA Titan X and GTX 1080Ti GPUs. We apply data augmentation in training using random horizontal flip for high resolution images (256×512). For semantic accuracy evaluation, we use the following evaluation measures: class average accuracy, the mean of the predictive accuracy over all classes, global accuracy, which measures overall scene pixel classification accuracy, and mean intersection over union (mIoU).

For the *Foggy Cityscapes* [36], *Foggy Driving* [36], and *Foggy Zurich* [8] datasets we train using the available information. For all datasets, as a complementary information source, we make use of the luminance transformation, which is a translated grayscale image L derived from $I_{RGB} \in \{I_R, I_G, I_B\}$ to both reduce the noise and improve feature extraction, defined as follows:

Methods \ Results	Global avg.	Class avg.	Mean IoU
Ours w/o domain adaptation	90.8	68.5	54.9
Ours w/ domain adaptation	**91.6**	**70.4**	**58.0**

TABLE II

QUANTITATIVE RESULTS OF SEMANTIC SEGMENTATION OVER THE *Foggy Cityscapes* [6] TEST DATASET (PARTIALLY SYNTHETIC DATA) OF OUR APPROACH WITH AND WITHOUT USING DOMAIN ADAPTATION.

$$L = 0.299(I_R) + 0.587(I_G) + 0.144(I_B). \tag{6}$$

VI. EVALUATION

We evaluate the performance of our proposed approach on the benchmark foggy weather conditions datasets: *Foggy Cityscapes* [36], *Foggy Driving* [36], and *Foggy Zurich* [8]. The evaluation was performed as follows:

1) We train the domain adaptation component (Section III-A) (employed later as a sub-component (Fig. 2) trained in step 3) on the *Cityscapes* dataset (*normal* weather) [6] and Foggy Cityscapes (*adverse* weather) [36] to map from *foggy* scenes to *normal*.
2) In the same manner, we train the semantic segmentation component (Section III-B) on the *Cityscapes* dataset [6] (*normal* weather).
3) Models obtained from steps (1, 2) are fine-tuned within a unified architecture using *refined Foggy Cityscapes* dataset [36] (a subset including 498 training and 52 testing better quality images).
4) The fine-tuned architecture in step 3 is evaluated on *Foggy Driving* [36] and *Foggy Zurich* [8].

With both qualitative and quantitative comparisons against the state-of-the-art approaches, we assess our approach on the aforementioned benchmark (foggy weather conditions

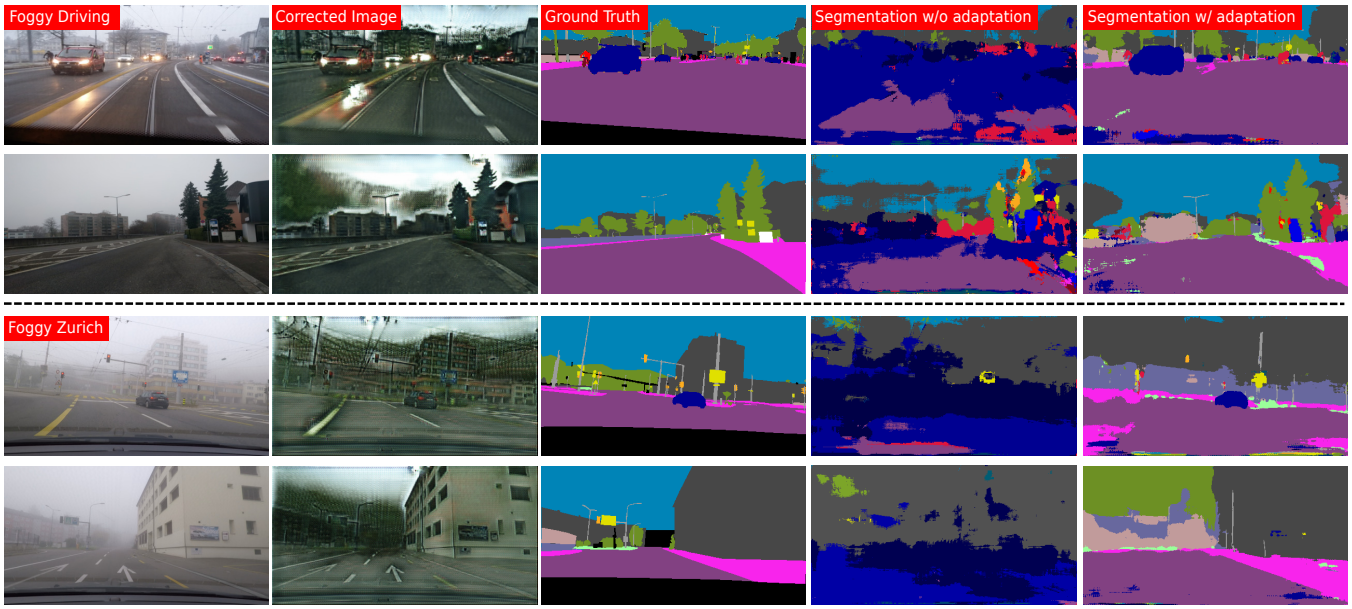

Fig. 6. Semantic segmentation predictions on **Foggy Driving** [36] and **Foggy Zurich** [8] for the proposed approach. The left column shows two scenarios of each dataset followed by **Corrected Image:** corrected image using *domain adaptation*; **Ground Truth:** ground truth segmentation; **Segmentation w/o adaptation:** segmentation results without using *domain adaptation*; **Segmentation w/ adaptation:** segmentation results with *domain adaptation*.

datasets). As an initial step, we evaluate the semantic segmentation performance directly on the foggy scenes without *domain adaptation* . Here, we deal with the semantic segmentation model (shown in Figure 3) as an independent model and isolated it from the entire pipeline (illustrated in Figure 2 and including the *domain adaptation* subcomponent) to investigate its performance on the foggy dataset. As seen in Table I and Figure 6, our model fails to produce any desirable improvements in terms of qualitative and quantitative results. However, our approach with *domain adaptation* (Figure 2, lower) provides the best results when compared without *domain adaptation* (Table I and Figure 6). Overall, we consider *domain adaptation* as a necessary step to correct foggy scenes before feeding them into the segmentation network.

As an initial evaluation applied to synthetic data, we test our approach on the *Foggy Cityscapes* [36]. In the evaluation, we consider the test set from the same (*Foggy Cityscapes*) used in training time, which leads to improved segmentation. As seen in the results shown in Table II, using *domain adaptation* contributes and increases to the mean intersection over union (IoU) by (3.1%) when compared with no domain adaptation. Figure 7 shows qualitative results on *Foggy Cityscapes* [36], for our approach with and without using domain adaptation.

Furthermore, we evaluate the performance of scene understanding and segmentation on the real-world datasets, *Foggy Driving* [36] and *Foggy Zurich* [8], with and without applying *domain adaptation* (III-A). This task is a more challenging as our model has not been training on the aforementioned datasets. Without any *domain adaptation*, our approach does not produce any qualitatively or quantitatively desirable results. However, with *domain adaptation* (Section III-A), our approach achieves superior results when compared to the absence of *domain adaptation*, in the mean intersection over union (IoU) scores of (29.4%) (*Foggy Driving*) and (21.4%) (*Foggy Zurich*) (see Table I). Figure 6 shows qualitative results on *Foggy Driving* [36] and *Foggy Zurich* [8] significant differences are evident between the two aforementioned methods.

As a comparison with the state-of-the-art semantic segmentation under foggy weather conditions, our approach with *domain adaptation* outperforms the work of [14] (see Table I). However, our proposed approach remains competitive with approaches such as [8], [14], [36], [38], as demonstrated in Table I. However, all comparators use off-the-shelf segmentation networks such as RefineNet [27], DeepLab [4], Dilated Conv Net [46], and BiSeNet [45], which offer higher segmentation accuracy due to their use of complex architectures at the expense of viable real-time performance. Using our purpose architecture requires less computational complexity and offers real-time inference performance, which represents an important aspect of our proposed approach. As shown in Table I, our approach with a significant number of parameters when compared to contemporary state-of-the-art architectures, enables a real-time inference speed of 20 – 42 fps with and without the use of *domain adaptation* respectively, enabling a real-time performance. Further evidence of the efficacy of our approach is being trained on less data (*Foggy* Cityscapes [36]), unlike [14] that have been trained on more datasets (*Foggy* Cityscapes [36] and Foggy *Synscapes* [14]), which contributes to higher accuracy but at the expense of higher computational complexity.

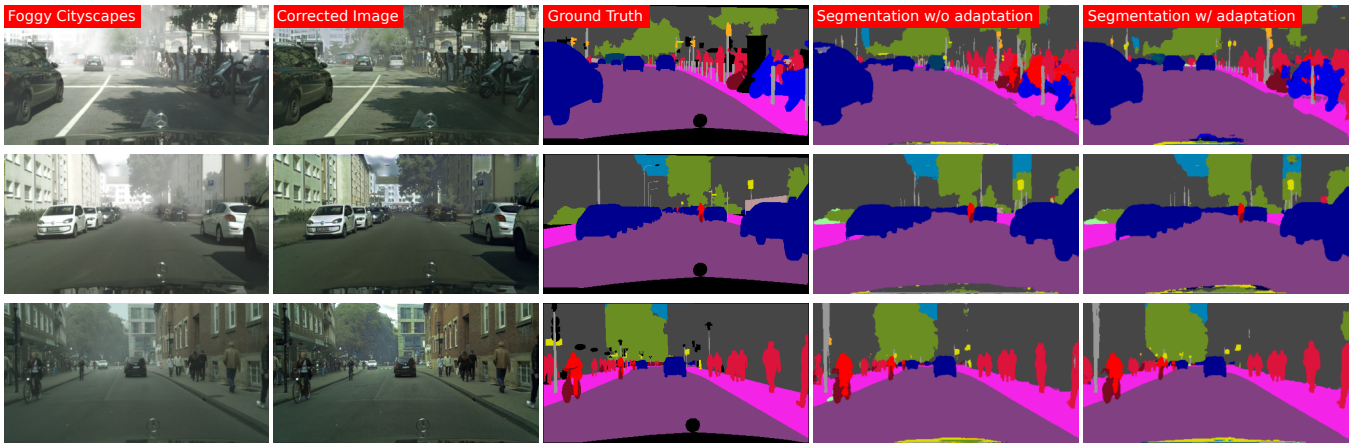

Fig. 7. Semantic segmentation predictions on the **Foggy Cityscapes** [36] for the proposed approach. The left column shows three scenarios of the dataset followed by **Corrected Image:** using *domain adaptation*; **Ground Truth:** ground truth segmentation; **Segmentation w/o adaptation:** segmentation results without using *domain adaptation*; **Segmentation w/ adaptation:** segmentation results with *domain adaptation*.

VII. Conclusion

This paper proposes a novel end-to-end automotive semantic segmentation within foggy scene understanding. Using a unified model, we make use of domain adaptation (GAN-based) [13] to adapt a scene taken in *foggy* weather conditions to *normal* thus increasing the scene visibility. Subsequently, the adapted images are fed to an effective semantic segmentation model for training. For real-time performance, our segmentation network is based on light-weight architecture that includes features fusion, dense connectivity and skip connections, making the approach real-time (20 – 42 fps with and without *domain adaptation* respectively). As a result, the performance of our approach has progressively improved and achieved significant performance over the state-of-the-art semantic segmentation under foggy weather conditions [8], [36], [38].

References

[1] N. Alshammari, S. Akcay, and T. P. Breckon, "On the impact of illumination-invariant image pre-transformation for contemporary automotive semantic scene understanding," in *Proc. Intelligent Vehicles Symposium*, 2018, pp. 1027–1032. 2

[2] J. Álvarez and A. Lopez, "Road detection based on illuminant invariance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 184–193, 2011. 1

[3] 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. 2

[4] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 40(4):834-848, 2018. 1, 2, 5, 6

[5] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491. 4

[6] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4, 5

[7] P. Corke, R. Paul, W. Churchill, and P. Newman, "Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation," in *the IEEE Int. Conf. on Intelligent Robots and Systems*, 2013, pp. 2085–2092. 1

[8] D. Dai, C. Sakaridis, S. Hecker, and L. Van Gool, "Curriculum Model Adaptation with Synthetic and Real Data for Semantic Foggy Scene Understanding," *In. arXiv e-prints*, Jan. 2019. 1, 2, 3, 4, 5, 6, 7

[9] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. Journal of Computer Vision (IJCV)*, vol. 88, no. 2, pp. 303–338, 2010. 4

[10] G. Finlayson, M. Drew, and C. Lu, "Entropy minimization for shadow removal," *International Journal of Computer Vision*, vol. 85, no. 1, pp. 35–57, 2009. 1

[11] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015. 2

[12] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. Conf. on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361. 4

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, pp. 2672–2680, 2014. 2, 3, 7

[14] M. Hahner, D. Dai, C. Sakaridis, J.-N. Zaech, and L. Van Gool, "Semantic understanding of foggy scenes with purely synthetic data," in *Intelligent Transportation Systems Conference*, 2019, pp. 3675–3681. 2, 5, 6

[15] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asian Conference on Computer Vision*, November 2016. 1, 2, 3, 4

[16] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, 2010. 1

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2

[18] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2008. 4

[19] C. J. Holder and T. P. Breckon, "Encoding stereoscopic depth features for scene understanding in off-road environments," in *International Conference Image Analysis and Recognition*. Springer, 2018, pp. 427–434. 2

[20] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. Conference on Computer Vision and Pattern Recognition*, July 2017. 1, 2, 3

[21] S. Hung, S. Lo, and H. Hang, "Incorporating luminance, depth and color information by a fusion-based network for semantic segmentation," in *Proc. Int. Conf. on Image Processing*, 2019, pp. 2374–2378. 1, 2, 3, 4, 5

[22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arxiv*, 2016. 1, 2

[23] T. Kim, Y. Tai, and S. Yoon, "Pca based computation of illumination-invariant space for road detection," in *Proc. Winter Conf. on Applications of Computer Vision*, 2017, pp. 632–640. 2

[24] P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *In Proc. Int. Conf. Learning Representations*, 2014. 5

[25] T. Krajnık, J. Blazıcek, and J. M. Santos, "Visual path following using intrinsic images," *European Conference on Mobile Robots*, pp. 1–6, 2015. 1, 2

[26] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," *arXiv preprint arXiv:1701.01036*, 2017. 2

[27] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 5, 6

[28] W. Maddern, A. Stewart, C. McManus, B. Upcroft, W. Churchill, and P. Newman, "Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles," in *Proc. Int. Conf. on Robotics and Automation*, vol. 2, 2014, p. 3. 1

[29] K. Nishino, L. Kratz, and S. Lombardi, "Bayesian defogging," *Int. Journal of Computer Vision*, vol. 98, no. 3, pp. 263–278, 2012. 1

[30] A. E. Orhan and X. Pitkow, "Skip connections eliminate singularities," *arXiv preprint arXiv:1701.09175*, 2017. 2

[31] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017. 5

[32] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016. 5

[33] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017. 3

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241. 2

[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *Int. Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015. 2, 4

[36] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *Int. Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, 2018. 1, 2, 3, 4, 5, 6, 7

[37] C. Sakaridis, D. Dai, and L. V. Gool, "Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," in *Proc. Int. Conference on Computer Vision*, 2019, pp. 7374–7383. 2

[38] C. Sakaridis, D. Dai, S. Hecker, and L. Van Gool, "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *Proc. European Conference on Computer Vision*, 2018, pp. 687–704. 2, 5, 6, 7

[39] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: Controlling deep image synthesis with sketch and color," in *Proc. Conference on Computer Vision and Pattern Recognition*, July 2017. 2

[40] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. Int. Conf. on Computer Vision*, 2017, pp. 4799–4807. 1, 2

[41] 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 *Proc. Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7472–7481. 5

[42] B. Upcroft, C. McManus, W. Churchill, W. Maddern, and P. Newman, "Lighting invariant urban street classification," in *Proc. Int. Conf. on Robotics and Automation*, 2014, pp. 1712–1718. 1

[43] Y.-K. Wang and C.-T. Fan, "Single image defogging by multiscale depth fusion," *IEEE Trans. on Image Processing*, vol. 23, no. 11, pp. 4826–4837, 2014. 1

[44] J. Yamanaka, S. Kuwashima, and T. Kurita, "Fast and accurate image super resolution by deep cnn with skip connection and network in network," in *Int. Conf. on Neural Information Processing*. Springer, 2017, pp. 217–225. 2

[45] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. European conference on computer vision*, 2018, pp. 325–341. 5, 6

[46] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015. 5, 6

[47] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. Conference on Computer Vision and Pattern Recognition*, July 2017. 1, 2

[48] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. Int. Conf on Computer Vision*, 2017, pp. 2223–2232. 1, 2