Competitive Simplicity for Multi-Task Learning for Real-Time Foggy Scene Understanding via Domain Adaptation

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

***Abstract*— Automotive scene understanding under adverse weather conditions raises a realistic and challenging problem attributable to poor outdoor scene visibility (e.g. foggy weather). However, because most contemporary scene understanding approaches are applied under ideal-weather conditions, such approaches may not provide genuinely optimal performance when compared to established *a priori* insights on extreme-weather understanding. In this paper, we propose a complex but competitive multi-task learning approach capable of performing in real-time semantic scene understanding and monocular depth estimation under *foggy* weather conditions by leveraging both recent advances in adversarial training and domain adaptation. As an end-to-end pipeline, our model provides a novel solution to surpass degraded visibility in *foggy* weather conditions by transferring scenes from *foggy* to *normal* using a GAN-based model. For optimal performance in semantic segmentation, our model generates depth to be used as complementary source information with RGB in the segmentation network. We provide a robust method for *foggy* scene understanding by training two models (*normal* and *foggy*) simultaneously with shared weights (each model is trained on each weather condition). Our model incorporates RGB colour, depth, and luminance images via distinct encoders with dense connectivity and features fusing, and leverages skip connections to produce consistent depth and segmentation predictions. Using this architectural formulation with light computational complexity at inference time, we are able to achieve comparable performance to contemporary approaches at a fraction of the overall model complexity. Evaluation over several foggy weather condition datasets including synthetic and real-world examples illustrates our approach competitive performance compared to other contemporary state-of-the-art approaches.**

I. Introduction

Semantic segmentation for automotive urban environments is a rapidly developing research topic illustrating successful state-of-the-art scene understanding approaches [4], [5], [22], [36]. Despite its successes, limited attention has been paid to the issue of automotive scene understanding under extreme weather conditions (*i.e.* foggy weather conditions) [7], [30], and by contrast we see deep learning approaches generally applicable to ideal weather conditions only. This paper proposes a robust solution to this challenge by taking advantage of domain adaptation for transferring knowledge from one domain to another – in this case, the between the domains of *normal* and *foggy* scene weather conditions.

Previous approaches for reducing adverse weather impact on automotive has seen differing methods proposed for reducing the illumination variance [1], [19], adapting scene understanding methods from day to night [28] or synthetic fog [7], [30]. Following recent advances in deep learning, scene understanding under such challenging conditions has also been addressed via domain adaptation [33] where a scene taken in foggy weather conditions is first pre-mapped onto a target domain (clear-weather), which is considered the optimal input for secondary scene understanding approaches.

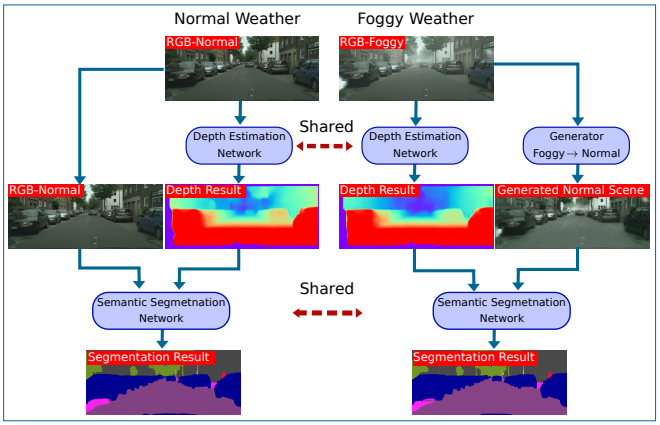

Fig. 1. A high-level illustration of our pipeline for semantic segmentation and depth estimation under foggy weather conditions.

In contrast to this pre-transformation approach, here we propose an end-to-end semantic scene understanding and monocular depth estimation framework using a novel multi-task approach specifically targeting the challenge of foggy weather operating conditions directly in the automotive environment. As the main objective of our work, we tackle the issue of semantic segmentation under *foggy* weather conditions in four steps. First, by employing the domain adaptation approach via image style transfer as proposed in [37] as a method to increase the level of visibility that suffers in *foggy* weather conditions. Second, by taking advantage of complementary depth information generated by a monocular depth estimator, which can be subsequently provided as an additional input to RGB colour into a semantic segmentor. These depth estimations and semantic segmentation components are trained via sub-models on both domains (*normal* and *foggy*), with shared weights to allow the implicit transfer of semantic and depth knowledge from one domain to another. Finally, our model is adversarially trained on output streams from depth estimation and semantic segmentation to force the multi-task model to produce predictions as close to the target outputs as possible. Figure 1 shows an illustration of our overall approach, including the steps mentioned above.

In summary, the main contributions of this paper are as

[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

follows:

- *Competitive low-complexity architecture* – enables semantic segmentation and depth prediction via multi-task learning and leveraging domain adaptation to correct images with degraded visibility in *foggy* weather conditions.
- *Optimal foggy scene understanding* – via adapting between two domains (*normal* and *foggy*) trained simultaneously with shared weights (each model is trained on one weather condition) and employing adversarial techniques on the output from each model.
- *Competitive performance* – outperforms the state-of-the-art foggy scene understanding [7], [12] on the benchmark datasets [7], [30], despite the fewer datasets our model is trained on.

II. Related Work

We review prior work in three key areas: semantic segmentation (Section II-A), monocular depth estimation (Section II-B), and domain adaptation (Section II-C).

A. Semantic Segmentation

Semantic segmentation is an essential task in scene understanding requiring robust per-pixels classification. Prior work has achieved promising results via deep convolutional networks [16], [22], [24], [26], [27], [36]. However, they differ by using different approaches for instance: pooling indices [4], skip connection [27], multi-path refinement [22], pyramid pooling [36], fusing-based [16] . As the basis for a number of semantic segmentation architectures, [24] leads the recent contributions by adopting [31] (an architecture designed for image classification) and subsequently decoding (mapping) low feature representations to pixel-wise output in an end-to-end model. Most prior work on semantic segmentation uses RGB and/or RGB-D data as an input [3], [4], [22]. As an incorporated technique, other studies have achieved some successes using luminance information [1], [16].

As a key challenge, several different approaches have been proposed to tackle the issue of scene understanding under adverse weather conditions. For instance, the issue of illumination changes is addressed in [1], [19] by minimising scene colour variations due to varying scene lighting conditions. Other approaches [7], [29], [30] address segmentation under foggy weather conditions using a semi-supervised approach through generating synthetic fog from real-world data and augmenting clear images to their synthetic fog images. By adapting segmentation models from day to night, [28] addressed the issue of poor scene visibility. Recently, domain adaptation as a technique (within the context of semantic segmentation) is employed to generate *normal* weather scenes from *adverse* ones [33] (this can be considered to be a defogging process) using [17], [37] (Section II-C). Subsequently, this generated input is fed into a semantic segmentor to perform pixel-wise segmentation [33].

Another method to achieve improved segmentation [32], propose a discriminator network using GAN [11] to encourage a segmentation model, with shared weights between two sub-models trained on different domains (real-world and synthetic images), to produce pixel-wise class labels.

Similarly, our semantic segmentation component is trained via two sub-models (each model on one weather condition) using real-world input representing *normal* weather conditions and synthetic *normal* inputs generated from a synthetic foggy dataset using domain adaptation [37] (discussed in Section II-C). However, inspired by [16], we employ the idea of incorporating luminance and depth alongside RGB colour via distinct encoders, utilising both skip connections and dense connectivity as well as fused features to gain better and deeper representation learning which leads to a superior semantic segmentation performance.

B. Monocular Depth Estimation

Although our main objective is semantic segmentation, using monocular depth estimation alongside semantic segmentation via multi-task learning may contribute to achieving better semantic segmentation performance [3]. Monocular depth estimation is a technique used to predict depth from a single image. In the literature, monocular depth estimation [2], [3], [9], [10] provides a solution for the shortcomings in depth estimation in terms of either the significant training data requirements or the final output (missing depth) of fundamental strategies [13].

Recent methods addressed monocular depth estimation using both supervised [2], [3], [8], [23] and unsupervised [9], [10] learning approaches. Employing a Generative Adversarial Network (GAN), [2] proposes monocular depth estimation using synthetic data transformed from real-world RGB colour. As a multi-task approach, [3] proposed temporally consistent depth prediction alongside semantic segmentation, which performed better than the single-task approach. Proposing an unsupervised depth estimation based on the ResNet-50 architecture, [10] uses the left image to generate depth for right-left images, and bilinear sampler and left-right disparity consistency loss to achieve significant improvement.

Motivated by [3], we estimate depth using a monocular depth estimation component that includes two sub-models with shared weights, each model trained using either the *normal* or *foggy* datasets.

C. Domain Adaptation

In the current literature, domain adaptation has been used to transfer an image from its real domain to different domain (image-to-image translation) [17], [37] allowing multiple uses of such images taken in complex environments or generated in different forms.

The idea behind this approach is that the generated images from the source domain can be transformed to be similar to the ones in the target domain through capturing the style texture information of the input by utilising the Gram matrix. Work in [21] shows that image style transfer (from the source domain to the target domain) is the process of minimising the differences between source and target distribution. Recent

methods [17], [37] use GAN [11] to learn the mapping from the source to the target images. Based on training over a large dataset for a specific image style, [37] shows an efficient approach to transferring image style from one image into another.

Another variation of domain adaptation has been performed within the same colour space of different domains (e.g. pixel-wise class labels for real-world and synthetic domains). In other words, the predictions derived from semantic segmentation components can be adapted by minimising the gap between them and the target ground truth [3], [32], [33].

In this work, we employ the idea of [37] to map between *normal* and *foggy* weather conditions as a method to increase the degraded visibility level due to *foggy* weather conditions. As a result, our model semantically segments a scene (taken in *foggy* weather conditions) based on a synthetic *normal* input (generated from *foggy* scenes), which are considered as optimal inputs to the subsequent scene understanding process. As an additional step to achieving better segmentation performance, we use the technique proposed in [3], [33] to jointly constrain depth estimation and semantic segmentation prediction close to the target domain (ground truth).

III. Proposed Approach

In simple terms, our main objective is to train an end-to-end network that semantically labels every pixel in a scene, and estimates the depth at each pixel from the monocular image, with both tasks occurring under foggy weather conditions. To achieve semantic segmentation under foggy weather conditions (the primary focus of our approach), we make use of knowledge adaptation [32] between models operating under *normal* and *foggy* weather conditions by simultaneously training two sub-models; each model is trained on one weather condition. Since scene visibility suffers due to foggy weather conditions, we make use of domain adaptation (Section III-A) as a method to increase the scene visibility level in the *foggy* weather datasets, for semantic segmentation task.

As an initial step towards improved semantic segmentation, monocular depth estimation is trained on both *normal* and *foggy* domains to produce depth maps for both domains. This step serves the semantic segmentation task by incorporating depth as a complementary information source with RGB colour [14], [16]. In addition, we consider using a multi-task approach as a feedback network [3], in which the output from a previous task serves as the input for the subsequent task, and the network recursively back propagates and updates its weights. Ultimately, the semantic segmentation is trained via two sub-models using *normal* and synthetic *normal* images (generated using the domain adaptation component in Section III-A).

In general, our approach consists of three sub-components: (i) Domain Adaptation (Section III-A), (ii) Semantic Segmentation (Section III-C), and (iii) Monocular Depth Estimation (Section III-D) (each functioning as an integrated unit). Our overall model produces three separate outputs: synthetic *normal* images (generated from foggy weather condition), pixel-wise class labels, and depth. Figure 2 shows our proposed approach. In the remainder of this section, we discuss the details of the aforementioned three primary sub-components.

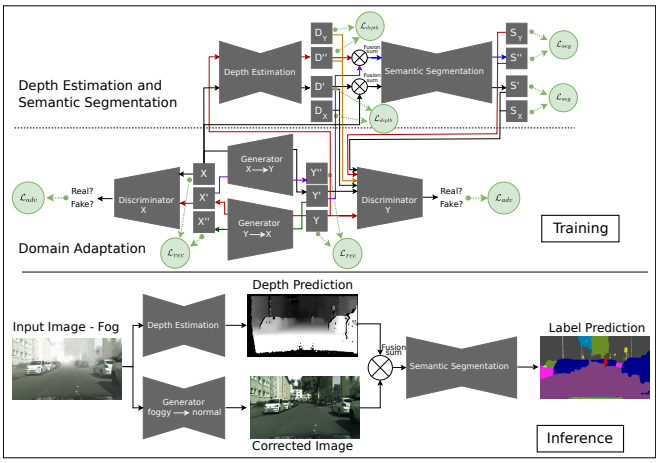

Fig. 2. A conceptual overview of our scene understanding approach via domain adaptation using [37] (**Inference**) and the detailed outline of the entire pipeline (**Training**). Our overall model consists of two main components: domain adaptation using [37] and an encoder-decoder submodule for semantic segmentation and depth estimation. Foggy scenes from (domain X) are transformed to fine scenes (domain Y) and vice versa, resulting in Y', X' (the desired domains), and cyclically mapping them back to their original domains, producing X'', Y''; D_X, D_Y: ground truth depths and D', D'': depth predictions; S_X, S_Y semantic labels and S', S'': semantic segmentation predictions.

A. Domain Adaptation

Our goal of employing domain adaptation [37] (Figure 2 DA) in the context of semantic segmentation and monocular depth estimation is to increase the level of visibility under *foggy* weather conditions via learning to map $\mathcal{D}: X \rightarrow Y$ from source domain X (*foggy* weather) to the target domain Y (*normal* weather) for which we assume such visibility corrected image is the optimal input to Semantic Segmentation (Section II-A). We use GAN [11] with the cycle consistency method of [37] for mapping between *foggy* and *normal* weather conditions (Figure 2). Two different generators $G_{X\rightarrow Y}$ (generating Y'), $G_{Y\rightarrow X}$ (generating X') and two discriminators D_X (to discriminate between X and X') , G_Y (to discriminate between Y and Y') are used to perform the mapping function from the source and target domains. The loss for each generator G with associated discriminator D is as follows:

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

$$\mathcal{L}_{adv}(Y \rightarrow X) = \min_{G_{X\rightarrow Y}} \max_{D_X} \mathbb{E}_{y\sim\mathbb{P}_d(x)}[\log(D)_{(x)}] + \mathbb{E}_{x\sim\mathbb{P}_d(x)}[\log(1 - D_Y(G_{Y\rightarrow X}(y)))], \quad (2)$$

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

In addition to the adversarial loss $\mathcal{L}_{adv}$, a cycle-consistency loss $\mathcal{L}_{cyc}$ is used to map the transferred image

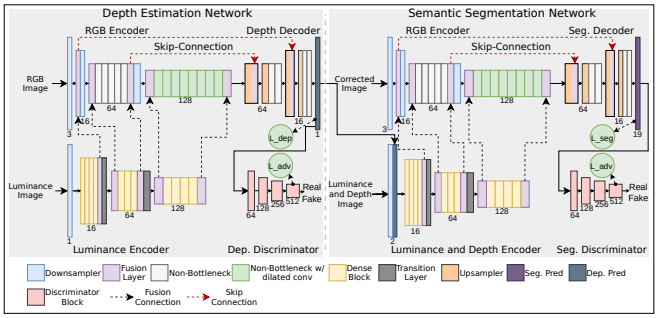

Fig. 3. A detailed outline of the encoder-decoder architecture for semantic segmentation and depth estimation. Each network consists of two sub-encoders taking two types of inputs: (i) **RGB** and (ii) luminance **L** with or without depth **D** images (depending on the task); as well as their respective decoders and discriminators.

(Y') back to the source domain (X). The cycle-consistency loss is implemented as follows:

$$\mathcal{L}_{cyc} = \|G_{Y\to X}(G_{X\to Y}(x)) - x\|_1 + \|G_{X\to Y}(G_{Y\to X}(y)) - y\|_1 \quad (3)$$

Subsequently, the joint loss function for the domain adaptation component is as follows:

$$\mathcal{L}_{domain-adapt} = \mathcal{L}_{adv}(X \to Y) + \mathcal{L}_{adv}(Y \to X) + \mathcal{L}_{cyc} \quad (4)$$

B. Overall Segmentation and Depth Estimation Architecture

As a subsequent task to domain adaptation in Section III-A, the semantic segmentation and depth estimation components are trained on two domains (normal and foggy scene weather conditions). First, we train the monocular depth estimation network on two sets of scenes: (i) real-world images (the original Cityscapes) Y for *normal* weather conditions, and (ii) partially synthetic images (Foggy Cityscapes) X for *foggy* weather conditions. This produces the depth maps in addition to the corrected images of the foggy scenes previously produced (Section III-A). Subsequently, we train the semantic segmentation network on two sets in the previous step, although here we use the transferred images $Y' = G_{X\to Y}(X)$ that represent the *foggy* weather conditions, as well as the generated depth as a complementary information source with RGB colour for improved segmentation performance.

As seen in Figure 3, the overall segmentation and depth estimation architecture is designed with an auto-encoder which includes two distinct encoders: (i) RGB encoder (E_{RGB}) and (ii) Luminance Encoder (E_L) or Luminance with Depth Encoder (E_{LD}) for depth estimation or semantic segmentation, respectively (Figure 3). The two encoders are incorporated within the features encoder stage and linked by fusing output layers from the corresponding blocks. The fusion connectivity is simply 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 $E_L(y)$.

Following the encoders, two decoders: the semantic segmentation decoder (D_{Seg}) and the depth estimation decoder (D_{Depth}) are designed to upsample the feature maps to the original input dimension for the two tasks of our model: pixel-wise segmentation with 19 class labels and depth images (Figure 3). Below we present in detail the encoders and decoders of the semantic segmentation and depth estimation components.

RGB encoder: Designed to deal with a three-channel RGB input, the RGB encoder (E_{RGB}) (adopted from [16]) comprises three downsampler blocks with convolutional and max pooling layers followed by batch normalization and *ReLu()* activation function ($\{16,\ 64,\ 128\}$ respectively). The first downsampling downsamper block is to extract the input features and reduce its dimensions. Five non-bottleneck modules are implemented in the second downsampler block including the factorized convolutions (convolution kernel $n \times n$ factorized into $n \times 1$ and $1 \times n$), each followed by batch normalization and *ReLu()* with residual connections. With dilated and factorized convolutions in the third downsampler block, eight non-bottleneck modules with residual connections were utilised as a last component of E_{RGB} to increase the RGB encoder efficiency.

Luminance encoder: Unlike the RGB encoder, the Luminance Encoder (E_L) for depth estimation (adopted from [16]) deals with the luminance image. We make use of a distinct encoder for luminance to exploit better learning and representation from the luminance maps that may not be possible when stacking with RGB colour [14], [16]. As a parallel function to E_{RGB}, the E_L encoder is designed using a dense connectivity technique for information flow enhancement from earlier to the final layers. Specifically, E_L consists of a downsampler (as in E_{RGB}) followed by three dense blocks; each has $\{4,\ 3,\ 4\}$ modules, respectively (E_L has the same number of channels as E_{RGB}). Each dense block is followed by a transition layer designed with 1×1 convolution layer and followed by 2×2 average pool layer.

Luminance and depth encoder: To achieve our goal in semantic segmentation, we use the luminance and depth maps (concatenated as a two-channel input) in the Luminance and Depth encoder (E_{LD}) which is identical to (E_L) except than it takes luminance and depth maps as an input.

Decoder: After fusing the last extracted feature maps from E_{RGB} and either the E_{LD} (for semantic segmentation) or E_L (for monocular depth estimation), the depth decoder (D_{Depth}) and Semantic Segmentation decoder (D_{Seg}) perform upsampling upon the feature maps to the original resolution. This upsampling is implemented in three stages. In the first two stages $\{64,\ 16\}$, convolutional transpose, batch normalisation and *ReLu()* activation function, as well as two non-bottleneck modules, are employed. To this end, D_{Seg} and D_{Depth} perform the same process. As the last component in the D_{Seg}, a convolutional transpose layer maps the generated output from the previous layer to the 19 class labels we aim to predict. For monocular depth prediction, pyramid depth predictions are produced via D_{Depth} at two scales to gain consistent representation following [10]. Specifically, a

Methods				Mean IoU			Complexity of the Network			
Models	Network Architecture	Training	Fine-Tuning	Fog Zurich	Fog Driving	Fog Cityscapes	Multi-Task	Real-Time	Number of Parameters	FPS
CMDAda [7]	AdSegNet [32] w/ DeepLab-v2 [5]	C	—	25.0	29.7	—	—	—	44.0M	20
SFSU [30]	Dilated Conv. Net. (DCN) [35]	C	FC (498)	35.7	46.3	—	—	—	134M	-
CMAda2+ [29]	RefineNet [22]	C	FC (498)	43.4	49.9	—	—	—	118M	22
CMAda3+ [29]	RefineNet [22]	C	FC (498)	46.8	49.8	—	—	—	118M	22
Hanner *et al.* [12]	RefineNet [22]	C	FS (24,500)	40.3	48.4	—	—	—	118M	22
Hanner *et al.* [12]	RefineNet [22]	C	FS (498)	42.7	48.6	—	—	—	118M	22
Hanner *et al.* [12]	RefineNet [22]	C	FC+FS (498)	41.4	50.7	—	—	—	118M	22
Hanner *et al.* [12]	BiSeNet [34]	C	FC (498)	25.0	30.3	—	—	—	50.8M	-
Hanner *et al.* [12]	BiSeNet [34]	C	FS (24,500)	27.8	30.9	—	—	—	50.8M	-
Hanner *et al.* [12]	BiSeNet [34]	C	FS (498)	27.6	31.8	—	—	—	50.8M	-
Hanner *et al.* [12]	RefineNet [22]	C	FC+FS (498)	35.2	30.9	—	—	—	118M	22
Ours w/o domain adaptation	—	C	FC (498)	13.9	17.6	59.4	✓	✓	**4.8M**	**31**
Ours w/ domain adaptation	—	C	FC (498)	26.1	31.6	**60.3**	✓	✓	**16.2M**	16

TABLE I

QUANTITATIVE COMPARISON OF SEMANTIC SEGMENTATION ON FOGGY ZURICH [7], FOGGY DRIVING [30] AND FOGGY CITYSCAPES [30] DATASETS OF OUR APPROACH AGAINST STATE-OF-THE-ART APPROACHES. **C**: CITYSCAPES [6]; **FC** FOGGY CITYSCAPES [30]; **FS**: FOGGY SYNSCAPES [12]. THE SPEED COMPARISON (FRAMES PER SECOND (FPS) IS BASED ON THE CITYSCAPES [6] TEST DATASET.

convolutional transpose layer maps the predicted depth to match the original input dimension followed by a sigmoid activation function. Besides, the previous stage (64 channels) are also mapped as in the final stage but at half the size of the original input dimension. As the final component in the semantic segmentation and depth estimation networks, a discriminator adapted from [17] is trained to adapt the predictions to the target labels.

C. Semantic Segmentation

Our semantic segmentation model provides semantic predictions for two different scenes: (1) *normal weather conditions*; and (2) *foggy weather conditions*. To be more specific, two semantic segmentation sub-models with shared weights are trained on each weather conditions (*normal* and *foggy* weather conditions). We assume that sharing weights within the sub-models will allow transferring knowledge between *normal* and *foggy* domains and may contribute to improved segmentation in the later domain. As the scene visibility is very poor in the *foggy* weather conditions, we use the corrected images Y' (mapped from *foggy* to *normal* via domain adaptation [37] (Section III-A)) as an alternative to *foggy* scenes, assuming they are the optimal inputs to semantic segmentation.

As a complementary information source, depth images are incorporated with RGB colour contributing to improved semantic segmentation performance [14], [16]. As an initial step serving semantic segmentation task, our model provides complementary depth images via monocular depth estimation (Section III-D), which allows benefits from using depth with RGB colour to improve the performance of semantic segmentation. Serving the same goal, our semantic segmentation model uses the luminance input image, which is a translated grayscale image employed in [16]. As a semantic segmentation loss function ($\mathcal{L}_{seg}$), cross-entropy is used.

To force our semantic segmentation sub-model (*foggy* weather conditions) to generate better segmentation labels close to performance under *normal weather conditions*, we use an adversarial training approach [11] that is used in the literature [2], [3], [32], [37] to produce similar segmentation distributions in *foggy* to *normal* weather conditions. Specifically, we feed the predicted semantic labels from the segmentation sub-model (*foggy* scenes) along with the corresponding ground truth labels into a discriminator (D) adapted from [17] (Figure 3) to adapt output predictions by distinguishing predicted labels $G(x) = \tilde{y}$ from ground truth y. The adversarial loss ($\mathcal{L}_{adv}$) is used for our semantic segmentation and described in Eq. 1. As an overall loss for the segmentation task, a joint segmentation loss defined as follows:

$$\mathcal{L}_{joint-seg} = \mathcal{L}_{seg} + \mathcal{L}_{adv}. \tag{5}$$

D. Monocular Depth Estimation

Although monocular depth estimation is not the main objective of this paper, it has been used alongside semantic segmentation (our main objective) to improve the latter. Unlike when semantic segmentation performs individually, multi-modality allows us to gain deeper representation features in the overall model [3] and perform inference in real-time [18]. In a similar vein to the earlier semantic segmentation component (Section III-C), our model performs depth prediction via two sub-models over two scenes (*normal* and *foggy* weather conditions), each model on each weather condition. However, the depth estimation architecture deals only with RGB and luminance information as inputs. The loss function has been for depth estimation ($\mathcal{L}_{depth}$) is $\mathcal{L}_1$. We employ adversarial training to minimize the gap between the predicted depth on *foggy* weather conditions against *normal* weather conditions using a discriminator (D) [17] takes predicted depth from the depth estimation sub-model (*foggy* scenes) along with the corresponding ground truth to distinguish the predicted depth $G(x) = \tilde{y}$ from ground truth y. The adversarial loss ($\mathcal{L}_{adv}$) which described in Eq. 1 is used for depth estimation. As an overall loss for the depth estimation task, a joint depth loss is defined as follows:

$$\mathcal{L}_{joint-depth} = \mathcal{L}_{depth} + \mathcal{L}_{adv}. \tag{6}$$

E. Combined Loss

Our combined loss function for the overall architecture with three sub-modules: domain adaptation, semantic segmentation, and depth estimation, is formulated in three steps.

Method	Depth Error (lower, better)				Depth Accuracy(higher, better)		
	Abs. Rel.	Sq. Rel.	RMSE	RMSE log	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
Ours w/o domain adaptation	0.238	0.543	1.994	0.277	0.656	0.884	0.983
Ours w/ domain adaptation	0.238	0.733	2.130	0.280	0.654	0.892	0.980

TABLE II

QUANTITATIVE RESULTS OF DEPTH PREDICTION OVER THE *refined Foggy Cityscapes* [30] WITH AND WITHOUT DOMAIN ADAPTATION [37].

Firstly, adversarial loss for domain adaptation $\mathcal{L}_{adv}$ and cyclic-consistency loss ($\mathcal{L}_{cyc}$) functions are implemented. Secondly, we utilise the $\ell 1$ loss for depth estimation with the adversarial loss for depth ($\mathcal{L}_{adv}$) on *foggy* weather conditions. Finally, a cross-entropy loss is used as a semantic segmentation loss ($\mathcal{L}_{seg}$) as well as the adversarial loss for segmentation ($\mathcal{L}_{adv}$) on *foggy* scenes. The joint loss function on the overall architecture is thus as follows:

$$\mathcal{L} = \mathcal{L}_{domain-adapt} + \mathcal{L}_{joint-seg} + \mathcal{L}_{joint-depth}, \quad (7)$$

As manually adjusting weights is time-consuming, the weighted sum of losses in the combined loss is dynamically updated using the homoscedastic uncertainty technique to weight and balance the losses [18].

F. Implementation Details

Our implementation pipeline begins with the domain adaptation stage, followed by monocular depth estimation, then semantic segmentation stage. In domain adaptation, our goal is to generate corrected images from *foggy* scenes (defogging process) to be used later in semantic segmentation. Therefore, we train two generators proposed in [37] on two domains (*normal* and *foggy*), each generator on each domain, to generate corrected images from the *foggy* domain and close to the *normal*. Subsequently, we trained the monocular depth estimation component via two sub-models using RGB and luminance inputs, each model on each weather condition, to produce depth used as a complementary information in semantic segmentation. Ultimately, we train the semantic segmentation component via two sub-models. One model is trained on *normal* scenes using RGB, luminance and the generated depth map from the depth estimation stage. The other model is trained on the corrected images generated from *foggy* scenes using domain adaptation, luminance, and the complementary depth information provided form depth estimation stage.

Cityscapes [6] and the partially synthetic *Foggy Cityscapes* [30] have been used as target and source domains, with $2,975$ training and 500 testing image examples (at a resolution of 1024×2048). We applied data augmentation in training using random horizontal flip as well a down-sampled resolution of 128×256. In addition to *Foggy Cityscapes*, real-world datasets: *Foggy Driving* [30] and *Foggy Zurich* [7] were used to evaluate our approach. We implemented our approach in *PyTorch* [25]. For optimization, we employed ADAM [20] with an initial learning rate of 1×10^{-3} and momentum of $\beta_1 = 0.5, \beta_2 = 0.999$. Our model is optimized based on a joint loss discussed in Section III-E.

IV. EXPERIMENTAL RESULTS

We evaluated the performance of our proposed approach on publicly available datasets:- *Cityscapes dataset* [6], *Foggy Cityscapes dataset* [30], *Foggy Driving* [30], *Foggy Zurich*, and [7] for semantic segmentation under foggy weather conditions. With and without using domain adaptation [37], we assessed our approach using qualitative and quantitative comparisons against the state-of-the-art approaches. For semantic accuracy evaluation, we used the following evaluation measures: (1) class average accuracy, (2) global accuracy, and (3) mean intersection over union (mIoU). For the benchmark evaluation, we use the standard mIoU metric (Jaccard Index) which measures the percentage of mean intersections over union for predictions over all predicted classes. As an end-to-end pipeline, our model is trained to adapt *foggy* to *normal* weather conditions using [37] (Section III-A). Subsequently, depth estimation and semantic segmentation networks are trained. The detailed steps for the evaluation of our proposed architecture are as follows:

1) We first train the domain adaptation component (Section III-A) on the *Cityscapes* dataset (*normal* weather) [6] and Foggy Cityscapes (*adverse* weather) [30] to map from *adverse* scenes to *normal* weather conditions.
2) We train the depth estimation component (Section III-D) on both the *Cityscapes* dataset (*normal* weather) [6] and the Foggy Cityscapes datase (*adverse* weather) [30] (two models for each with shared weights as set out in Section II-B).
3) Mirroring step 2, we train the semantic segmentation component (Section III-C), but using the corrected images from *foggy* scenes generated from step 1 and incorporating the generated depth maps from step 2.
4) Models obtained from steps 1, 2, and 3 were fine-tuned using *refined Cityscapes* [30] (a sub set that includes 498 training and 52 testing images examples with better quality).
5) The fine-tuned models in step 4 were evaluated on both synthetic and real-world datasets including *Foggy Zurich* [7] and *Foggy Driving* [30].

In the rest of this section, we discuss the results of semantic segmentation (Section IV-A) and monocular depth estimation (Section IV-B).

A. Semantic Segmentation

We evaluated the performance of semantic segmentation on the following benchmark foggy weather conditions datasets: *Foggy Driving* [30] and *Foggy Zurich* [7]. This was a challenging task as our model has not seen a single image from the aforementioned datasets. As an initial stage, we evaluated our model directly using *foggy* scenes (with no *domain adaptation*) from the aforementioned datasets. As seen in Table I, our model failed to obtain any favourable quantitative and qualitative results compared with no *domain adaptation*. However, using *domain adaptation*, our model clearly provides an improved performance of the mean intersection over union (mIoU) scores across all classes: from 13.9% to **26.1**$\%$ on *Foggy Zurich* [7] and from 17.8% to **31.6**$\%$ of *Foggy Driving* [30] (Table I). By contrast, we evaluate our model on a test set from (*Foggy Cityscapes* [30]) having also trained on this dataset, which leads to improved segmentation: from 59.4% to **60.3**$\%$ (Table I). Figure 4 shows qualitative results on *Foggy Driving* [30], *Foggy Zurich* [7] and *Foggy Cityscapes* [30] through different scenarios using our proposed approach. Overall, we consider that *domain adaption*, as a method, influences the semantic segmentation performance under *foggy* weather conditions.

As a comparison with the state-of-the-art semantic segmentation under foggy weather conditions, our approach with domain adaptation outperforms the work of [7], [12] on *Foggy Zurich*. When tested on *Foggy Driving* [30] our approach was able to surpass the work of [7]. In addition, our model outperforms the work of [12] with the three fine-tuned methods on: *refined Foggy Cityscapes* [6] (498) images, *Foggy Synscapes* [12] $(24,000)$ images, and the combination of *Foggy Cityscapes* [6] and *Foggy Synscapes* [12]. However, our proposed approach remains competitive with the proposed approaches in [7], [12], [29], [30]. Table I presents a comparison of our proposed approach against the-state-of-the-art foggy semantic segmentation.

Overall, we observe that our proposed approach provides a competitive performance against state-of-the-art techniques despite the complexity involve of using multi-task modality. In contrast, each component of the overall model has less computational complexity. As clearly seen in Table I, the semantic segmentation component uses fewer parameters (2.4M) when compare with existing approaches, enabling the possibility of real-time performance. Another important aspect that underlines the superiority of our model is that all comparators use off-the-shelf complex segmentation networks such as RefineNet [22], DeepLab [5], and Dilated Convolution Network [35], which constrained the practical application of their approaches in real-time performance.

B. Monocular Depth Estimation

Even though monocular depth estimation is not the primary focus, we assess the efficacy of our model in monocular depth estimation using *Cityscapes* [6] and *Foggy Cityscapes* which provide a disparity dataset labelled using Semi-Global Matching [15]. Unlike the semantic segmentation network, the monocular depth component was not dependent on the

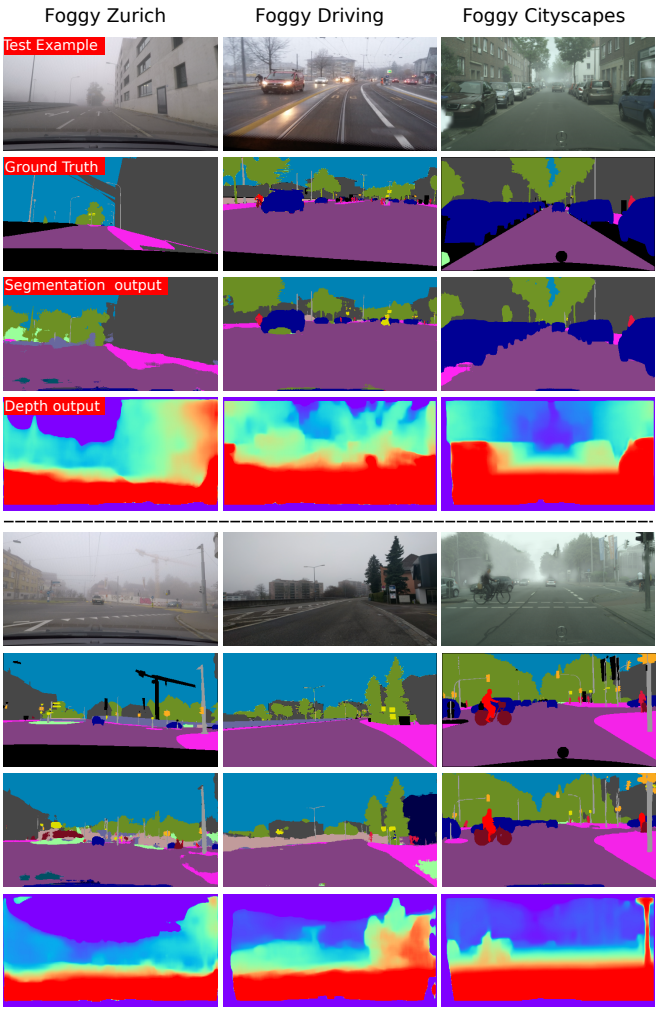

Fig. 4. Segmentation and depth predictions on Foggy Zurich [30], Foggy Driving [7] and Foggy Cityscapes [30] using our approach.

domain adaptation sub-model. In other words, we evaluate our model directly using *Foggy* images, with and without *domain adaptation* sub-model working alongside the depth estimation network (*i.e.* no defogging or dehazing processing was used). Here, we assume that there is a similarity between fog and depth in terms of objects localization in which objects close to the camera are clearly visible. In the same vein, depth is used within the literature [12], [30] as a key input for fog simulation, whilst fog and noise lead to better depth estimation [2]. We evaluate our approach on monocular depth estimation quantitatively and qualitatively using two methods. Firstly, we use a single model to perform the following three tasks: (1) domain adaptation, (2) semantic segmentation, and (3) monocular depth estimation. Secondly, we use the same model but without the domain adaptation component. Measurement metrics are based on [8]. As seen in Table II, our approach provides monocular depth estimation results that are close to each other using the two aforementioned methods.

V. Conclusion

We propose a novel multi-task approach for automotive semantic segmentation and depth estimation under *foggy*

weather conditions. Our approach is designed via multi-modality to produce optimal performance through using domain adaptation (GAN-based) [37] to correct images with poor visibility taken in *foggy* weather conditions. By using synthetic and real-world datasets, depth estimation and semantic segmentation components are trained with a unified framework providing promising results. With dense-connectivity, skip-connections, and fusion-based techniques, we propose a competitive encoder-decoder for semantic segmentation and depth estimation were proposed. Our overall approach is characterized by a complexity that allows multi-task learning. In addition, each component was designed with a lightweight architecture allowing real-time performance. Using extensive experimentation, we show the performance of our approach achieves significant results over the state-of-the-art semantic segmentation under adverse weather condition [7], [12], [30] as well as providing extra tasks (*i.e.,* monocular depth estimation).

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. 1, 2

[2] A. Atapour-Abarghouei and T. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2018. 2, 5, 7

[3] A. Atapour-Abarghouei and T. P. Breckon, "Veritatem dies aperit-temporally consistent depth prediction enabled by a multi-task geometric and semantic scene understanding approach," in *Proc. Conf. on Computer Vision and Pattern Recognition*, 2019, pp. 3373–3384. 2, 3, 5

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for scene segmentation," *IEEE Trans. on Pattern Analysis and Machine Intel.*, 2017. 1, 2

[5] 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, 5, 7

[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. Conf on Comp. Vision and Pattern Recog.*, 2016. 5, 6, 7

[7] 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, 5, 6, 7, 8

[8] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Adv. in neural info. proc. sys.*, 2014, pp. 2366–2374. 2, 7

[9] R. Garg, B. V. Kumar, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *Euro. Conf. on Comp. Vis.* Springer, 2016, pp. 740–756. 2

[10] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Conf. on Comp. Vision and Pattern Recog.*, July 2017. 2, 4

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *neural inf. proces. sys.*, 2014, pp. 2672–2680. 2, 3, 5

[12] M. Hahner, D. Dai, C. Sakaridis, J.-N. Zaech, and L. Van Gool, "Semantic understanding of foggy scenes with purely synthetic data," in *Int. Transp. Sys. Conf.*, 2019, pp. 3675–3681. 2, 5, 7, 8

[13] O. K. Hamilton and T. P. Breckon, "Generalized dynamic object removal for dense stereo vision based scene mapping using synthesised optical flow," in *Int. Conf. on Image Proc.*, 2016, pp. 3439–3443. 2

[14] 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*. Springer, 2016, pp. 213–228. 3, 4, 5

[15] 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. 7

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

[17] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. Conf. on Comp. Vis. and Patt. Recog.*, 2017, pp. 1125–1134. 2, 3, 5

[18] A. Kendall, Y. Gal, and R. Cipolla, "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. 5, 6

[19] T. Kim, Y. Tai, and S. Yoon, "Pca based computation of illumination-invariant space for road detection," in *Winter Conf. on Applications of Computer Vision*, 2017, pp. 632–640. 1, 2

[20] P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *In Proc. Int. Conf. Learning Represent.*, 2014. 6

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

[22] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Conf. Comp. Vis. Patt. Recog.*, 2017. 1, 2, 5, 7

[23] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2024–2039, 2015. 2

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Conference on Computer Vision and Pattern Recognition*, June 2015. 2

[25] 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. 6

[26] E. Romera, J. M. lvarez, 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, Jan 2018. 2

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241. 2

[28] C. Sakaridis, D. Dai, and L. V. Gool, "Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," in *International Conference on Computer Vision*, 2019, pp. 7374–7383. 1, 2

[29] C. Sakaridis, D. Dai, S. Hecker, and L. V. 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, 7

[30] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *Int. Journal of Computer Vision (IJCV)*, vol. 126, no. 9, pp. 973–992, 2018. 1, 2, 5, 6, 7, 8

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 2

[32] Y. Tsai, W. Hung, S. Schulter, K. Sohn, M. 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. 2, 3, 5

[33] P. Wenzel, Q. Khan, D. Cremers, and L. Leal-Taixe, "Modular vehicle control for transferring semantic information between weather conditions using gans," in *Conference on Robot Learning*, 2018, pp. 253–269. 1, 2, 3

[34] 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 conf. on comp. vision*, 2018, pp. 325–341. 5

[35] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015. 5, 7

[36] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Conf. on Comp. Vision and Pattern Recog.*, 2017. 1, 2

[37] J. 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, 3, 5, 6, 8