2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Task-Driven Deep Image Enhancement Network for Autonomous Driving in Bad Weather

Younkwan Lee[1], Jihyo Jeon[2], Yeongmin Ko[1], Byunggwan Jeon[1], and Moongu Jeon[1,2]

***Abstract*— Visual perception in autonomous driving is a crucial part of a vehicle to navigate safely and sustainably in different traffic conditions. However, in bad weather such as heavy rain and haze, the performance of visual perception is greatly affected by several degrading effects. Recently, deep learning-based perception methods have addressed multiple degrading effects to reflect real-world bad weather cases but have shown limited success due to 1) high computational costs for deployment on mobile devices and 2) poor relevance between image enhancement and visual perception in terms of the model ability. To solve these issues, we propose a task-driven image enhancement network connected to the high-level vision task, which takes in an image corrupted by bad weather as input. Specifically, we introduce a novel low memory network to reduce most of the layer connections of dense blocks for less memory and computational cost while maintaining high performance. We also introduce a new task-driven training strategy to robustly guide the high-level task model suitable for both high-quality restoration of images and highly accurate perception. Experiment results demonstrate that the proposed method improves the performance among lane and 2D object detection, and depth estimation largely under adverse weather in terms of both low memory and accuracy.**

I. INTRODUCTION

Autonomous vehicles require comprehensive and accurate visual perception to safely navigate diverse driving conditions with little or no human effort. Currently, visual perception tasks are achieved by deep neural networks (DNN) which have demonstrated impressive performance on a wide range of high-level vision tasks, such as lane detection [1], [2], monocular depth estimation [3]–[5], and scene recognition [6]–[10]. The success of deep convolutional neural networks relies on a large number of high-quality images and a large computational cost on large-scale resource devices. However, existing models do not typically consider the degradations taken from bad weather conditions for training as well as low-resource devices to be deployed on mobile devices. Therefore, one needs to train complex visual degradations caused by bad weather with DNN having lower resource consumption.

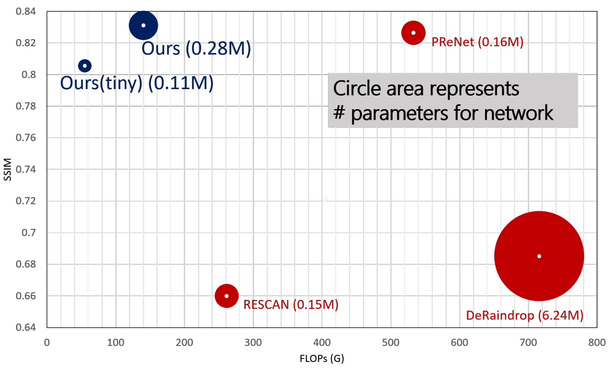


Fig. 1. TuSimple [11] SSIM vs FLOPs. The area of the circle plot represents the parameters of the model as described in parentheses. Compared to three methods (in blue): RESCAN [12], DeRaindrop [13] and PReNet [14], our proposed models (in red) are shown to achieve perception- and hardware-friendly performance, while reducing model size.

To solve this problem, image enhancement has been adopted as a key solution [14]–[19], enabling it to improve image quality by itself as a preprocessing method within independent components. Benefited from these methods, the latest deep learning-based enhancement models seem almost plausible at first glance. Unfortunately, however, we observe that existing enhancement methods still not suitable for high-level vision tasks in bad weather, and even worsen performance in some cases. Among them, we find its limitations in the following reasons. First, They usually rely on metrics based on the human visual system that are not correlated with visual perception models such as PSNR and SSIM. Thus, when the image is inferred by them, the recovered information is not sufficient or appropriate for high-level tasks. Second, they generally require a lot of computational power despite having to run autonomous driving with resource-limited platforms. Hence they are unaffordable for autonomous driving when integrated into the resource-constrained environments with other perception models.

In this paper, we introduce a novel task-driven image enhancement framework that benefits by exploring the mutual influence between visual perception and enhancement for safe and reliable autonomous driving in bad weather conditions. To this end, our model aims to have *perception- and hardware-friendly* characteristics against any bad weather situations as end-to-end learning, as shown in Fig. 1. To the best of our knowledge, this is the first attempt to connect

This work was partly supported by Ministry of Culture, Sports and Tourism and Korea Creative Content Agency(Project Number: R2020070004) and Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea Government (MSIT) (No. 2014-3-00077, Development of Global Multi-target Tracking and Event Prediction Techniques Based on Real-time Large-Scale Video Analysis), and the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (No. 2019R1A2C2087489).

[1] Machine Learning & Vision Laboratory, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, South Korea {brightyoun, koyeongmin, jbk34071, mgjeon}@gist.ac.kr

[2] Korea Culture Technology Institute (KCTI), Gwangju 61005, South Korea jihyo@gm.gist.ac.kr

978-1-7281-9077-8/21/$31.00 ©2021 IEEE

image enhancement and high-level vision for autonomous driving under multiple bad weather conditions. In summary, our work makes following key contributions:

- We propose a universal multiple bad weather removal framework that enables the high-level vision tasks to improve the robustness of existing models without degradation and retraining.
- We develop a task-driven enhancement network for less memory and computational cost, thus it is suitable for the embedded system when building ADAS (Advanced Driver Assistance System) for autonomous vehicles.
- We introduce a novel training strategy that minimizes the detrimental effects of image enhancement while improving the performance of high-level tasks in an end-to-end and task-driven manner.
- We experimentally validate the effectiveness of the proposed method when embeds high-level tasks such as lane detection, monocular depth estimation, and object detection. To our best knowledge, this work is one of few studies to apply the proposed image enhancement module for visual high-level tasks of autonomous driving under bad weather.

II. RELATED WORKS

A. Bad Weather Image Enhancement Algorithms

Many models and algorithms have been developed to deal with only one weather condition. For example, [12]–[14], [20], [21] have been proposed to recover rain effects, including rain streaks or raindrops. In [22], desnowing was designed with a multi-scale stacked densely connected CNN for detecting and removing snowflakes from a single snowy image. A few approaches for defogging/dehazing was proposed by non-local prior [23] or image-to-image translation network without relying on the physical scattering model [24]. However, they are not designed and trained for all the bad weather conditions, thus may not guarantee to build safe autonomous driving in bad weather. The above issue of universal bad weather enhancement has been addressed by hybrid all-in-one model [25]–[27]. In [26], a joint dehazing and deraining CNN was proposed with the classical atmospheric scattering model from the global context of a single image. In [27], generative adversarial networks were used by relying on task-specific encoders that only process a particular degradation type. Although these all-on-one methods have achieved impressive performance on bad weather image enhancement, most of them were only suited for one specific kinds of perception task, such as object detection [27] or semantic segmentation [26] without studying various high-level tasks. Moreover, their methods are computationally too inefficient for on-device embedding in autonomous vehicles and are not suitable for fast inference. To the best of our knowledge, our work is the first study to provide faster processing time and compressed parameters while being perception- and hardware-friendly to deal with a variety of bad weather conditions.

B. Limitation of High-Level Vision Models

When high-level vision tasks are conducted in bad weather conditions that they often encounter in autonomous driving, image enhancement is usually worked as an independent pre-processing stage, which might be poorly related to the task-specific goal [28]–[30]. Recently, limitation of deep learning-based high-level vision models has been investigated to reveal their inefficiency against bad weather conditions that they operate with image enhancement methods as the independent pre-processing stage. For example, [31] demonstrated that the existing image dehazing methods do not bring much benefit to help the image classification performance based on the analysis of the evaluation metric. Similarly, [18], [32] showed that existing image deraining models do not much improve the performance of recognition model, or worsened, based on images collected in the real world. To comprehend such problem, some researches [33]–[35] pointed out that visual enhancement works mainly focus on human perception quality [36], which becomes harmful by visual artifact patterns or noise perturbation.

Nevertheless, there are several studies to overcome the vulnerability of high-level vision models. In [25], re-formulated atmospheric scattering model that direct reconstructs haze-free images was studied by using an end-to-end learning scheme. In [37], various factors of image degradation were tackled by analyzing the semantic segmentation networks in autonomous driving scenes. In [26], [38], [39] image enhancement and high-level task were jointly designed as an end-to-end learning model, achieving improved performance over both tasks. However, most methods still consider some weather effects such as rain or fog, separately. In addition, their optimization is still not suitable for high-level visual tasks, and there is no consideration of the efficiency of the hardware. As far as we know, our method is the first attempt to propose a recognition- and hardware-friendly framework, taking into account various bad weather conditions.

III. PROPOSED METHOD

A. Problem Definition

Here we present the general setting of the problem prior to the illustration of the proposed method. We have a clean image I^{GT} and corresponding bad weather image I^{X}. We define that both images have the same high-level task label Y^{GT}. The bad weather input I^{X} is first fed into the image enhancement network E^{en} and outputs the recovered image I^{pred}, while the last layer before the final output of E^{en} represents f^{en}_{last}. Subsequently, the recovered image I^{pred} is fed forward through the high-level perception network E^{ht} and outputs the high-level perception result Y^{pred} with the last convolution layer f^{ht}_{last}. The parameters of each network are represented as θ_{en} and θ_{ht} with pre-trained for each task, where θ_{ht} is frozen while optimizing the proposed method. Note that we do not explicitly define a detailed network for high-level task, since our proposed is applicable to arbitrary high-level task baselines. Additionally, the last layers of the two networks mentioned above, f^{en}_{last} and f^{ht}_{last}, are respectively fed into feature identity extraction network with the learnable parameter ϕ. Fig. 2 shows the overall framework of the networks.

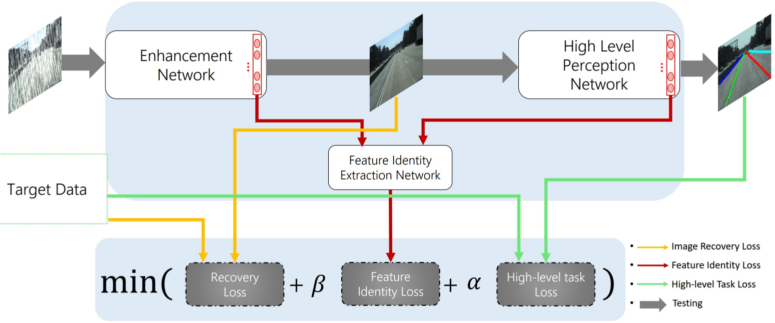

Enhancement Network
High Level Perception Network
Feature Identity Extraction Network
Target Data
min(Recovery Loss + β Feature Identity Loss + α High-level task Loss)
Image Recovery Loss
Feature Identity Loss
High-level Task Loss
Testing

Fig. 2. **Overview of our proposed enhancement framework.** Our framework comprises a low memory enhancement network, a task-specific high-level perception network, and a feature identity extraction network. We connect all networks into one pipeline and train in an end-to-end manner.

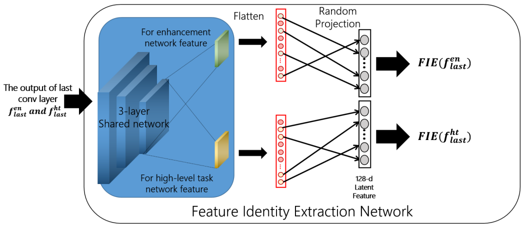

The output of last conv layer f^{en}_{last} and f^{ht}_{last}
3-layer Shared network
For enhancement network feature
For high-level task network feature
Flatten
Random Projection
128-d Latent Feature
$FIE(f^{en}_{last})$
$FIE(f^{ht}_{last})$
Feature Identity Extraction Network

Fig. 3. **The detailed structure in the feature identity extraction network (FIE).** The random projection shows the connections to 128-d latent feature space from the last flatten representation of each network.

B. Enhancement Network Architecture

Our network is inspired by DenseNet [40] as a feature encoding network. Feature encoding network has one efficient structure for high-resolution applications at the edge and outperforms existing image enhancement methods by leveraging HarDNet [41] based light-weight block for reducing the concatenation cost. Our enhancement network can be divided into two components: a Harmonic Dense Block (HBlock) for low memory computational cost and a Feature Identity Extraction Module (FIE) with feature fusion high-level perception task.

1) Harmonic Dense Block: To learn the recovery information, inspired by [41], we model HBlock with depth L layers. While the standard DenseNet passes the gradient from propagated all the layers, it leads to terrible large memory usage and heavy computational cost outweighing the gain. To solve these problems, the output of HBlock with depth-L is acquired through concatenation with L^{th} layer and all the previous odd-ordered layers. We also make the output of all even layers from 2 to L-2 to be removed once the HBlock is finished. Lastly, to adjust the dimension, we set the 32 channels in the last layer of each block. Each layer L has an output channel width k, and the number of its channels is calculated by $k \times 1.6^n$, where n is the maximum value at which the layer l is divided by the integer quotient by 2^m.

Additionally, we employ a bottleneck layer before every 4th convolution layer to further accelerate the parameter efficiency, and set its output channel to $\sqrt{\frac{c_{in}}{c_{out}}}$, where c_{in} and c_{out} are channels of input and output, respectively. To this end, we propose two versions of the network, each consisting of 71 layers (5 HBlocks) and 33 layers (3 HBlocks). The batch normalization is used after each convolution layer except the last layer. After that, ReLU is applied as an activation function. Finally, in order to achieve more high-quality recovery, a recursive enhancement structure is introduced with a total of 3 stages, which is gradually leading to perception-friendly quality at the final stage. The full description of our enhancement network is shown in Table

TABLE I
THE ARCHITECTURE OF ENHANCEMENT NETWORK (71 LAYERS).
INFO IS COMPOSED OF KERNEL AND STRIDE.

ID	Layer/Block	Info	Output
Input	Input$_{bad}$	-	-
Input$_{init}$	Input$_{x^0}$	-	-
Input$_{x^0}$	Input$_{bad}$		
for t=1 to T			
Concat1	Concat(Input$_{bad}$, $Input_{x^{t-1}}$)	-	1024x512x6
Conv1	Conv(Concat1)	3, 1	1024x512x32
HBlock1	HBlock(Conv1) : 8 layers	3, 1	1024x512x32
Concat2	Concat(Conv1, HBlock1)	-	1024x512x64
Conv2	Conv(Concat2)	1, 1	1024x512x32
Add1	Add(Conv1, Conv2)	-, -	1024x512x32
Conv3	Conv(Add1)	3, 1	1024x512x32
HBlock2	HBlock(Conv3) : 16 layers	3, 1	1024x512x32
Concat3	Concat(Conv3, HBlock2)	-	1024x512x64
Conv4	Conv(Concat3)	1, 1	1024x512x32
Add2	Add(Conv3, Conv4)	-, -	1024x512x32
Conv5	Conv(Add2)	3, 1	1024x512x32
HBlock3	HBlock(Conv5) : 16 layers	3, 1	1024x512x32
Concat4	Concat(Conv5, HBlock3)	-	1024x512x64
Conv6	Conv(Concat4)	1, 1	1024x512x32
Add3	Add(Conv5, Conv6)	-, -	1024x512x32
Conv7	Conv(Add3)	3, 1	1024x512x32
HBlock4	HBlock(Conv7) : 16 layers	3, 1	1024x512x32
Concat5	Concat(Conv7, HBlock4)	-	1024x512x64
Conv8	Conv(Concat5)	1, 1	1024x512x32
Add4	Add(Conv7, Conv8)	-, -	1024x512x32
Conv9	Conv(Add4)	3, 1	1024x512x32
HBlock5	HBlock(Conv9) : 4 layers	3, 1	1024x512x32
Concat6	Concat(Conv9, HBlock5)	-	1024x512x64
Conv10	Conv(Concat6)	1, 1	1024x512x32
Add5	Add(Conv9, Conv10)	-, -	1024x512x32
Conv11	Conv(Add5)	3, 1	1024x512x3
Recursive Output: Input$_{x^t}$ ($ForRestorationLearning$)			
end for			

1.

2) Feature Identity Extraction Module: The feature identity extraction module (FIE) is designed for correlating information from image enhancement and high-level visual perception features. The FIE is based on 3-layer CNN, which assigns exactly 128-dimensional latent features after the output of flattening the last layer of FIE with random projection instead of dense, as shown in Fig. 3. This allows unrestricted comparisons through random projection when the final layer output dimensions of FIE are different. Therefore, the FIE connects them by representing the mutual influence between image enhancement and visual perception in a unified framework.

C. Training Strategy

To learn the proposed network, we further integrate both image enhancement network and high-level network via three-stage. Our training strategy is divided into three parts: 1) image enhancement network learning, 2) high-level vision loss calculation, and 3) feature identity learning.

1) Image recovery loss: Existing state-of-the-art methods adopt the pixel-wise loss based on MSE (Mean-Squared error) to train enhancement network. However, the MSE optimization usually produces blurry visual information which results in perceptual unsatisfactory images with over-smooth content. To prevent this, we estimate the successive approximation to the bad weather distribution with the guidance of the Charbonnier penalty function [42], which is more robust to outliers. The recovery loss is expressed as:

$$L_R(I^{pred}, I^{GT}) = ||\sqrt{(I^{pred} - I^{GT})^2}||_2^2 + \varepsilon^2, \tag{1}$$

where ε is penalty coefficient and empirically set to 5×10^{-3}. We take one step further to give rich connectivity between the enhancement network and high-level perception.

2) High-level task loss: We use high-level task loss L_{HT} from a pre-trained high-level vision task network to provide the enhancement network with connectivity that promotes it to be perception-friendly. By default, perception networks for high-level tasks are pre-trained on benchmarks composed of clean images and are frozen while learning the proposed framework. In addition, even if our enhancement network is replaced with another model, it can be replaced without any additional tuning to the coefficients of objective function and retraining of the perception network. As far as we know, this is the first study to run a variety of high-level tasks while dealing with all bad weather, taking a step further in universality. To convey more strong perception-friendly property, we describe the feature identity loss in the next.

3) Feature identity loss: [43] propose to utilize a Euclidean distance which calculates identity information on image pairs, that proves to generate high-quality samples than the standard per-pixel losses. Their idea has been adopted mostly in image generation work, such as super-resolution, translation, and image recovery. Despite the fact, we observed that even when the recognition tasks other than image generation is involved, the identity information is still essential for stable optimization. To give the relevant information in the training process, we propose to use a feature identity loss that leads to the directly related to identity in hypersphere space, defined as:

$$L_{FI}(f_{last}^{en}, f_{last}^{ht}) = ||\widehat{FIE(f_{last}^{en})} - \widehat{FIE(f_{last}^{ht})}||_2^2, \tag{2}$$

where $FIE(f_{last}^{en})$ and $FIE(f_{last}^{ht})$ are the identity features extracted from (FIE) for input image I^X and recovered image I^{pred}, respectively. $\widehat{FIE(\cdot)}$ is the identity representation mapped to the hypersphere.

4) Objective Function: Based on the above introduction, we incorporate the above-mentioned losses as an objective function. We optimize the total objective function based on the stage-wise manner and can be trained by the following function:

$$\begin{aligned}\min_{\{\theta_{en},\phi\}} \frac{1}{N}\sum_{i=1}^{N}(L_R(I_i^{pred}, I_i^{GT}) + \alpha L_{HT}(Y_i^{pred}, Y_i^{GT})& \\ + \beta L_{FI}(f_{last}^{en}, f_{last}^{ht})),&\end{aligned} \tag{3}$$

where α and β are trade-off coefficients of the L_{HT} and L_{FI} respectively, θ_{en} and ϕ are learnable parameters from scratch with N samples.

IV. Experiments and Evaluation

A. Datasets

In computer vision, few image datasets contain comprehensive bad weather conditions specific to driving situations, and likewise none of the datasets mentioned above are available. In order to create bad weather effects, we adopt rain streak, raindrop, and haze simultaneously on each image in the dataset (2 rain streaks × 1 raindrop × 2 haze effects). For the realistic rain streak effect, we are motivated from [45] and thus create two versions of streak intensities (heavy and light rain) with randomly distributed orientation. For the raindrop effect, we adopt a simulation method from [46] to apply the water drop on the lens to all images. For the haze effect, we employ widely used atmospheric scattering model introduced in [47], [48] and generate two different hazy images under uniformly randomly chosen atmospheric lights and scattering coefficient as parameters. As a result, we obtain a paired dataset of clean target-bad weather images, where the split indexing of training and testing samples all follow the standard of the existing dataset.

For TuSimple dataset [11] as lane detection benchmark, we take 3,626 images for training and 2,782 images for testing. For monocular depth estimation, we also take 39,810 images for training and 4,424 images for testing from KITTI benchmark [44]. Lastly, object detection evaluation of our model is performed on the RID dataset [18] which contains real bad weather such as rain streaks and densely disrupted haze, it only provides 2,495 samples for testing without synthetic effects.

B. Training Configurations

The used datasets are resized to 1024 × 512 including both training and testing. All the networks are trained from scratch using Adam optimizer for 100 epochs with a total batch

TABLE II

QUANTITATIVE BAD WEATHER ENHANCEMENT EVALUATIONS WITH AVERAGE PSNR/SSIM ON SYNTHETIC IMAGES. THE BEST RESULTS IN ALL METHODS ARE MARKED IN BOLD. SECOND BEST ARE UNDERLINED.

Dataset	RESCAN [12]	DeRaindrop [13]	PReNet [14]	Ours(33-layer)	Ours(71-layer)
KITTI	22.95/0.7166	19.58/0.6845	25.19/0.7878	25.56/0.7932	**27.06/0.8227**
TuSimple	20.28/0.6581	20.97/0.7240	27.21/0.8266	27.79/0.8059	**28.37/0.8348**

TABLE III

QUANTITATIVE HIGH-LEVEL PERCEPTION EVALUATIONS ON TWO DATASETS (TUSIMPLE, KITTI) AND ONE REAL-WORLD DATASET RID. BEST RESULTS IN EACH CATEGORY ARE IN BOLD. SECOND BEST ARE UNDERLINED.

	metric	Bad Weather	RESCAN [12]	DeRaindrop [13]	PReNet [14]	Ours(33-layer)	Ours(71-layer)
Lane Detection	Acc ↑	95.86	95.16	95.38	95.54	96.19	**96.51**
	FP ↓	**2.85**	5.35	4.69	5.15	3.09	3.66
	FN ↓	3.70	5.59	5.04	4.79	3.34	**3.03**
Depth Estimation	RMSE ↓	7.360	7.549	10.246	6.245	5.501	**5.351**
	RMSE log ↓	0.343	0.322	0.487	0.262	0.219	**0.218**
Object Detection	mAP ↑	23.92	21.87	22.51	23.80	24.84	**29.55**

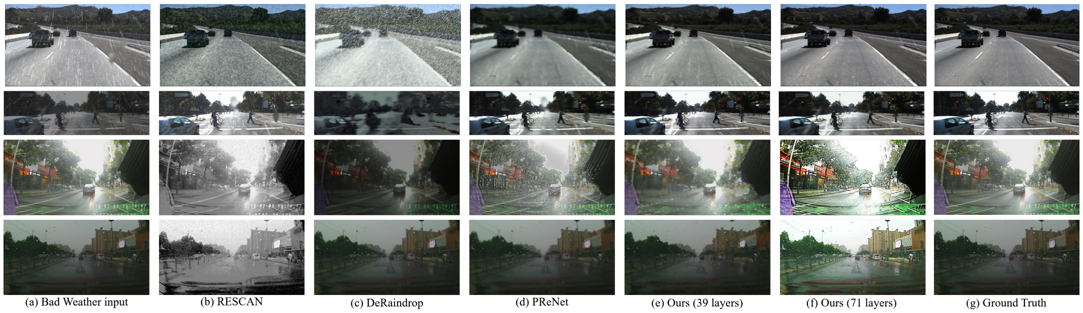

(a) Bad Weather input (b) RESCAN (c) DeRaindrop (d) PReNet (e) Ours (39 layers) (f) Ours (71 layers) (g) Ground Truth

Fig. 4. **Visual comparison of different enhancement results.** For the four bad weather images in the first column (a), columns (b-d) show that the enhancement results by state-of-the-art methods, respectively. The proposed method contributes to getting better restoration results in column (e-f). Ground truth is shown in the last column (g). Best viewed on the computer, in color, and zoomed in.

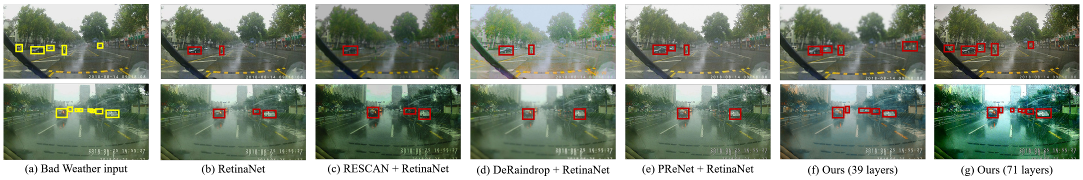

(a) Bad Weather input (b) RetinaNet (c) RESCAN + RetinaNet (d) DeRaindrop + RetinaNet (e) PReNet + RetinaNet (f) Ours (39 layers) (g) Ours (71 layers)

Fig. 5. **Visualization of object detection results on RID [18].** The yellow bounding box on the bad weather input (a) is ground truth. Other bounding boxes in (b-g) are predicted results. Best viewed on the computer, in color, and zoomed in.

size of 8. The learning rate is first initialized to 0.0001 and divided by 5 following milestones at the 30^{th}, 50^{th}, and 80^{th} epochs. The output channel width k is [14,16,20,20,40] for the 71-layer and [14,16,40] for the 33-layer. Additionally, by empirical finding, the coefficient α for high-level tasks is set to 0.01, 0.05, and 0.002 for lane detection, depth estimation, and object detection, respectively. All the experiments are performed by using one NVIDIA TITAN X GPU and one Intel Core i7-6700K CPU based on the PyTorch framework.

C. Experimental Configurations

For evaluation of our model, we test the effectiveness of our method on three representative high-level tasks: lane detection, monocular depth estimation, and object detection. For the perception baselines, we employ state-of-the-art baselines: PINet [2] for lane detection, Monodepth2 for monocular depth estimation [5], and RetinaNet [6] for object detection. We also utilize their pre-trained weights from the publicly available codes.

To quantitatively verify the effectiveness of the proposed method, we employ two types of metrics that measure task performance and image quality. For the image quality, we adopt PSNR and SSIM [49], which are standard [50], [51] in image recovery. However, they may become loosely related when it comes to other high-level task purposes [18], [26], [31]. Therefore, for the task-specific evaluation, we use standard metrics like the following: accuracy for lane detection, RMSE for monocular estimation, and mAP for object detection. Note that the proposed setting is evaluated without requiring manual data annotation in a comprehensive and fair setting.

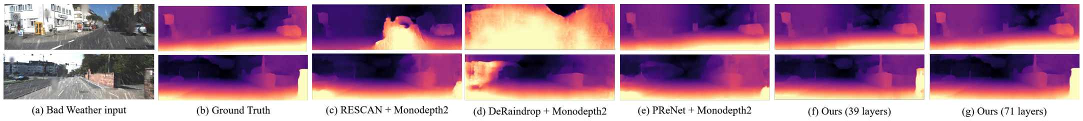
TABLE IV

COMPUTATIONAL COST ANALYSIS.

models	image size	FLOPs (G)	params (M)
RESCAN	1024x512	258.57	0.15
DeRaindrop	1024x512	716.39	6.24
PReNet	1024x512	531.51	0.16
Ours (33 layers)	1024x512	**57.02**	**0.11**
Ours (71 layers)	1024x512	146.16	0.28

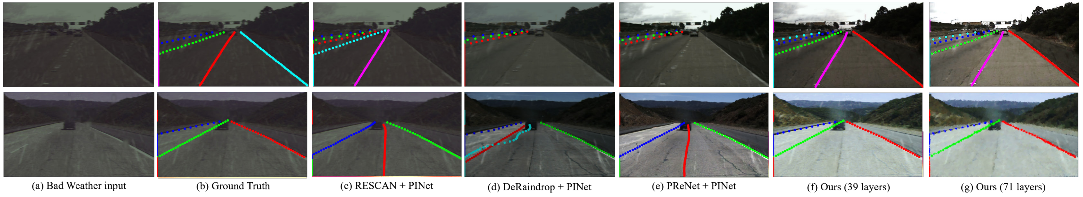
D. Image Enhancement Evaluations

Table 2 shows that our method achieves significant gains in terms of both PSNR and SSIM. As shown in Fig. 4, qualitative results reveal great effectiveness, while the result by DeRaindrop still contains visible bad weather elements.



(a) Bad Weather input (b) Ground Truth (c) RESCAN + Monodepth2 (d) DeRaindrop + Monodepth2 (e) PReNet + Monodepth2 (f) Ours (39 layers) (g) Ours (71 layers)

Fig. 6. **Visualization of monocular depth estimation results on KITTI [44].** The results in (b) is the ground truth extracted from the clean image. Other results in (c-f) are predicted results from bad weather input (a). Best viewed on the computer, in color, and zoomed in.



(a) Bad Weather input (b) Ground Truth (c) RESCAN + PINet (d) DeRaindrop + PINet (e) PReNet + PINet (f) Ours (39 layers) (g) Ours (71 layers)

Fig. 7. **Visualization of lane detection results on TuSimple [11]**. The results in (b) is the ground truth extracted from the clean image.Other results in (c-f) are predicted results from bad weather input (a). Best viewed on the computer, in color, and zoomed in.

Consequently, the visual quality enhancement by our methods is significant, while the results by existing methods still contain visible bad weather effects. Moreover, our lightweight model (33-layer) shows PSNR on par with the other three methods. To our best knowledge, our 71-layer model is the only bad weather enhancement method so far that can simultaneously achieve clean visual quality and hardware-friendly property. Such a large gain demonstrates our method generates promising image enhancement results, which are visually more clear.

E. High-Level Task Evaluations

To evaluate the effectiveness of our enhancement framework on high-level tasks in bad weather, we further show that the method yields meaningful results. From Table 3, our method outperforms the high-level perception performances in comparison for different methods using RESCAN, DeRaindrop, and PReNet as a pre-processing. As reported in [18], most models do not improve over the bad weather input in terms of high-level perception metrics. Instead, our method is observed to improve performance in high-level tasks without deteriorating. This proves that our method represents mutual influence between visual perception and enhancement, thus providing significant help for bad weather capabilities. Fig. 5, 6 and 7 also show that our proposed method outperforms the existing methods, confirming the perception-friendly ability of the model to bad weather.

Table 4 reports the computational cost of our enhancement network and some state-of-the-art methods. From the results, we can find that our method has less computational overload due to the harmonic dense block. Taking their hardware-friendly ability into account, it is appealing to still maintain perception performance when facing bad weather images.

F. Ablation Studies

To study the contribution of each network in our proposed framework, we alternatively remove it and identify the impact on the high-level perception performance. As can be seen in Table 5, our method with joint training (c-d) performs better than simple connection (b). Joint training (c) has a slightly lower baseline in Task1 and Task3, which is not surprising since it was not trained with the feature identity network. After applying the FIE (d), our method is the best-ranked approach that significantly outperforms the other three options (a-c), and we are encouraged to observe that the FIE brings the interconnection between the two networks. Finally, it is observed that all the networks in our proposed framework lead to important contribution in the final performance.

TABLE V

ABLATION STUDY. HLT REFERS TO THE BASELINES CORRESPONDING TO THE TASK. T1 TO 3 REFER TO LANE DETECTION, MONOCULAR DEPTH ESTIMATION, AND OBJECT DETECTION, RESPECTIVELY.

	metric	model	(a) HLT	(b) +EN (w/o training)	(c) +EN (training)	(d) +FIE (Ours)
T1	Acc	71L	95.86	94.23	96.37	**96.51**
		33L	95.86	95.43	95.95	**96.19**
T2	RMSE	71L	7.360	5.595	5.428	**5.351**
		33L	7.360	5.616	5.556	**5.501**
T3	mAP	71L	23.92	22.31	26.84	**29.55**
		33L	23.92	22.84	23.75	**24.84**

V. CONCLUSIONS

In this paper, we have proposed a novel task-driven image enhancement framework connected to visual perception for autonomous driving under the presence of bad weather. In particular, we have revealed that the existing methods are not practical for real-world autonomous driving in resource-constrained devices, and have aimed to improve them from two perspectives. First, our method is *perception-friendly* since it is optimized not only for the human-centric visibility but also for the high-level task models simultaneously. In addition, we developed a low-memory network architecture, focusing on a *hardware-friendly* ADAS system on the embedded system suitable for autonomous cars. Compared to previous methods, our method has verified improved performance in terms of both perception and hardware for autonomous driving despite bad weather. Future work will focus on modeling bad weather characteristics explicitly to remove artifacts and preserve details more effectively.

REFERENCES

[1] D. Neven, B. De Brabandere, S. Georgoulis, M. Proesmans, and L. Van Gool, "Towards end-to-end lane detection: an instance segmentation approach," in *2018 IEEE intelligent vehicles symposium (IV)*. IEEE, 2018, pp. 286–291.

[2] Y. Ko, Y. Lee, S. Azam, F. Munir, M. Jeon, and W. Pedrycz, "Key points estimation and point instance segmentation approach for lane detection," 2020.

[3] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.

[4] S. Pillai, R. Ambruş, and A. Gaidon, "Superdepth: Self-supervised, super-resolved monocular depth estimation," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9250–9256.

[5] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 3828–3838.

[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[7] Y. Lee, J. Jeon, J. Yu, and M. Jeon, "Context-aware multi-task learning for traffic scene recognition in autonomous vehicles," in *2020 IEEE Intelligent Vehicles Symposium (IV)*, 2020, pp. 723–730.

[8] D. Miller, L. Nicholson, F. Dayoub, and N. Sünderhauf, "Dropout sampling for robust object detection in open-set conditions," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–7.

[9] J. Yu, D. Y. Kim, Y. Lee, and M. Jeon, "Unsupervised pixel-level road defect detection via adversarial image-to-frequency transform," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1708–1713.

[10] D. Zhou, J. Fang, X. Song, L. Liu, J. Yin, Y. Dai, H. Li, and R. Yang, "Joint 3d instance segmentation and object detection for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1839–1849.

[11] "The tusimple lane challenge," in *http://benchmark.tusimple.ai/*.

[12] X. Li, J. Wu, Z. Lin, H. Liu, and H. Zha, "Recurrent squeeze-and-excitation context aggregation net for single image deraining," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 254–269.

[13] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2482–2491.

[14] D. Ren, W. Zuo, Q. Hu, P. Zhu, and D. Meng, "Progressive image deraining networks: A better and simpler baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 3937–3946.

[15] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[16] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *European conference on computer vision*. Springer, 2016, pp. 154–169.

[17] W. Ren, J. Tian, Z. Han, A. Chan, and Y. Tang, "Video desnowing and deraining based on matrix decomposition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4210–4219.

[18] S. Li, I. B. Araujo, W. Ren, Z. Wang, E. K. Tokuda, R. H. Junior, R. Cesar-Junior, J. Zhang, X. Guo, and X. Cao, "Single image deraining: A comprehensive benchmark analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3838–3847.

[19] P. Liu, J. Janai, M. Pollefeys, T. Sattler, and A. Geiger, "Self-supervised linear motion deblurring," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2475–2482, 2020.

[20] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley, "Removing rain from single images via a deep detail network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3855–3863.

[21] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *IEEE transactions on circuits and systems for video technology*, 2019.

[22] P. Li, M. Yun, J. Tian, Y. Tang, G. Wang, and C. Wu, "Stacked dense networks for single-image snow removal," *Neurocomputing*, vol. 367, pp. 152–163, 2019.

[23] D. Berman, S. Avidan, *et al.*, "Non-local image dehazing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1674–1682.

[24] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced pix2pix dehazing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8160–8168.

[25] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *Proceedings of the IEEE international conference on computer vision*, pp. 4770–4778.

[26] H. Sun, M. H. Ang, and D. Rus, "A convolutional network for joint deraining and dehazing from a single image for autonomous driving in rain," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 962–969.

[27] R. Li, R. T. Tan, and L.-F. Cheong, "All in one bad weather removal using architectural search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3175–3185.

[28] S. Hasirlioglu, A. Kamann, I. Doric, and T. Brandmeier, "Test methodology for rain influence on automotive surround sensors," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2016, pp. 2242–2247.

[29] M. Bijelic, T. Gruber, and W. Ritter, "Benchmarking image sensors under adverse weather conditions for autonomous driving," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1773–1779.

[30] D. Liu, B. Cheng, Z. Wang, H. Zhang, and T. S. Huang, "Enhance visual recognition under adverse conditions via deep networks," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4401–4412, 2019.

[31] Y. Pei, Y. Huang, Q. Zou, Y. Lu, and S. Wang, "Does haze removal help cnn-based image classification?" in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 682–697.

[32] C. H. Bahnsen and T. B. Moeslund, "Rain removal in traffic surveillance: Does it matter?" *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 8, pp. 2802–2819, 2018.

[33] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 427–436.

[34] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4792–4800.

[35] J. Wu, R. Timofte, Z. Huang, and L. Van Gool, "On the relation between color image denoising and classification," *arXiv preprint arXiv:1704.01372*, 2017.

[36] W.-S. Lai, J.-B. Huang, Z. Hu, N. Ahuja, and M.-H. Yang, "A comparative study for single image blind deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1701–1709.

[37] O. Zendel, K. Honauer, M. Murschitz, D. Steininger, and G. Fernandez Dominguez, "Wilddash-creating hazard-aware benchmarks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 402–416.

[38] Y. Lee, J. Lee, H. Ahn, and M. Jeon, "Snider: Single noisy image denoising and rectification for improving license plate recognition," in

Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Oct 2019.

[39] T. Son, J. Kang, N. Kim, S. Cho, and S. Kwak, “Urie: Universal image enhancement for visual recognition in the wild,” *arXiv preprint arXiv:2007.08979*, 2020.

[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[41] P. Chao, C.-Y. Kao, Y.-S. Ruan, C.-H. Huang, and Y.-L. Lin, “Hardnet: A low memory traffic network,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3552–3561.

[42] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Deep laplacian pyramid networks for fast and accurate super-resolution,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 624–632.

[43] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[44] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[45] S. S. Halder, J.-F. Lalonde, and R. d. Charette, “Physics-based rendering for improving robustness to rain,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10 203–10 212.

[46] “Raindrop on lens effect (role),” in *https://github.com/ricky40403/ROLE*.

[47] E. J. McCartney, “Optics of the atmosphere: scattering by molecules and particles,” *nyjw*, 1976.

[48] S. G. Narasimhan and S. K. Nayar, “Chromatic framework for vision in bad weather,” in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, vol. 1. IEEE, 2000, pp. 598–605.

[49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[50] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, “Dehazenet: An end-to-end system for single image haze removal,” *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5187–5198, 2016.

[51] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin, “Deep learning on image denoising: An overview,” *Neural Networks*, 2020.