Unified Feature Fusion Network with Path Router for Multi-task Image Restoration

Jingyuan Zhou[1*], Chaktou Leong[2], Yiyang Luo[3], Minyi Lin[1], Wantong Liao[1], Congduan Li[1*]
[1] School of Electronics and Communication Engineering, Sun Yat-sen University, Guangzhou, China
[2] School of Computer Science, Wuhan University, Wuhan, China
[3] Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China
{zhoujy53, linmy23, liaowt6}@mail2.sysu.edu.cn, {ctleong18}@whu.edu.cn,
{licongd}@mail.sysu.edu.cn, {1155124470}@link.cuhk.edu.hk

***Abstract*—Image restoration is an important low-level vision task, which includes various sub-tasks such as deraining, dehazing, denoising, raindrop removal, etc. Although current researches have achieved significant results in various sub-tasks, only a few of them are designed for multiple degradation factors. However, in the actual natural environment, the weather is complex and changeable so the networks designed for a single task are usually inapplicable. In this paper, we propose an unified network that can effectively restore images in a variety of weather conditions (rain, haze, and raindrops). The network is mainly divided into three parts. The first part is the shared multi-expert feature extraction module. We introduce the multi-task learning with multi-gate mixture-of-experts(MMoE) architecture and propose the smooth dilated residual group to extract the low-level features. Furthermore, the gate fusion sub-network is proposed to weigh and sum the output of each expert, so the correlation and difference between tasks can be captured. The second part is the path router sub-network, which can select branches of different tasks, and only open one branch at a time. The third part is multi-gated feature fusion branch, which can extract high-level feature and fuse different levels of features. Finally, we add the output of the third part to the input image and get the clean image. Our model cam deal with a variety of weather conditions and the experiments show competitive results compared with state-of-the-art models for single tasks.**

***Index Terms*—Image restoration, Multi-gate mixture-of-experts(MMoE), Path router sub-network, Smooth dilated residual group**

I. Introduction

Image restoration is an important task in computer vision, which aims to enhance a degraded image to a certain extent. Image restoration plays an important role in automatic driving and security monitoring because high quality images provided by image restoration algorithms can boost the performance of other tasks like object detection and image segmentation. However, traditional image restoration methods are mostly based on image enhancement methods, statistics priors and physical models, for example, dark channel prior algorithm for image dehazing [1], Wiener filter for image denoising [2], etc. Currently, learning-based methods utilizing convolutional neural networks are widely used in low-level computer vision tasks, and have become state-of-the-art models in almost all the sub-fields, including rain streak removal [3]–[6], hazy removal [7]–[10], raindrop removal [11], [12], blur removal [13], etc. Image restoration network can be connected end-to-end in front of the network of high-level task to improve its performance, such as object detection and image segmentation.

Although existing models have achieved good results in almost all the single tasks, their generalization abilities are usually poor. In the real natural environment, the weather condition is complex so that models designed for a single task are usually inapplicable in the real world. At present, there are only a few researches on multi-task image restoration [14]–[16],which still have a lot of room to improve the performance. As it requires high generalization ability of the network and is more difficult in network design and data collection. Considering that the previous methods mainly use the attention mechanism to achieve multi-task image restoration implicitly, it demands specific operations of parameters initialization and training methods of the network. In order to simplify the training process, we introduce an explicit multi-task learning method MMoE [17] to recover each specific scene.

Most multi-task networks have multiple outputs for each forward propagation [18], but it is not suitable for multi-task image restoration. Since the network only needs to generate a clean image for the input of a degraded image, we propose a supervised path router sub-network to select the task branches, which ensures that only one task branch needs to be passed each time.

Feature extraction and fusion play significant roles in image restoration. According to the experimental results, it is difficult for a single feature extraction block to deal with multiple tasks. Therefore, we propose a structure of multiple parallel feature extraction blocks and the restoration performance has been greatly improved. Inspired by [9], we introduce smooth dilated convolution [19] combined with resblock as the backbone of the network. In order to solve this problem, we use the above block to form multiple groups to extract features of different levels. In each task, we need to keep the shared features and task specific features and filter out the unnecessary features. On this basis, we propose multi-gate sub-network for feature fusion, which can weight the feature map to retain the needed features.

In this paper, we propose an unified image restoration network for multiple scenes, which is mainly divided into

three parts. The first part is the multi expert feature extraction group, which can extract and fuse shared low-level features of different tasks. The second part is path router sub network, which can select the branches of different tasks. The third part is the branch network of different tasks called multi-gate feature fusion branches, which can extract and fuse the high-level specific features according to each weather condition, and finally get a clean image through skip-connection.
The contributions of this paper are summarized as follows:

- We propose an image restoration network which can deal with a variety of weather conditions(rain streak ,haze ,adherent raindrop).
- We regard different weather conditions as different tasks and introduce multi-expert feature extraction group to extract shared feature of different tasks. Then gate sub-network will fuse feature for each single task.
- We propose a path router sub network to select task branches for different weather scenarios. Different task branches can extract and fuse the specific features of each task.

II. Proposed Method

A. Overview

In this session, we will introduce the architecture of Multi-branch Path Router and Feature Fusion Network.
As shown in Fig 1., the image initially passes the shared encoder and the shared network of multi-expert, which are used to extract the common low-level features, and meanwhile passes the path router sub-network to classify the scene and select the branch. Specially, we add a layer of feature maps (size: H*W*1) after a first-order differential in the dimension of input, because edge features are beneficial to help training convergence according to [20].
The features extracted by the experts are weighted fusion in the gated fusion sub-network of the selected branch, and then pass two Smooth Dilated Resblock Group to extract the high-level features. The second gated fusion sub-network will weight the features from different layers and add them up. Finally the feature maps go through the decoder as the role of degraded residual and plus the input image to get the clean image.

B. Multi-expert Feature Extraction Groups

After sub-sampling the images, we obtain the initial feature maps, which are the low-level features of various weather conditions. Different tasks include both shared and specific features, so it is hard to handle this situation using a single shared bottom network. Based on the consideration ahead, we adopt multi-expert to extract different features for different tasks and fuse them. As shown in Fig 2.; several Smoothed Dilated Resblocks acting as fundamental blocks are piled together to form a group architecture, namely single expert.
[9] firstly introduced Smoothed Dilated Convolutions to low-level tasks and achieved great results. Dilated convolution is widely applied to pixel predicting tasks, such as image semantic segmentation and video modeling. As it can improve receptive field without increasing the calculation complexity. The formula is showed below, o indicates output, f indicates input and w indicates the weights of filter; r is dilation rate and this formula also means inserting $r-1$ zeros between every two adjacent weights in the standard convolution filter.

$$o[i] = \sum_{s=1}^{S} f[i - r * s] * w[i] \tag{1}$$

However, the two adjacent pixels in feature map after dilated convolution come from two different dilated convolutions of the feature map in the last layer, which will generate gridding artifacts. In order to solve this problem, [19] proposed a method that the separable shared convolutions are added before dilated convolutions, as shown in Fig 3. Each channel of separable shared convolutions shares the weight of the convolution kernel, which helps extract the local relation.

C. Path Router Sub-Network

We add a scene classifier after the encoder to select the scene. The structure of this classification is showed in Fig 4.(a); it only consists two convolution layers, a average pooling layer and a fully-connected layer. Since the encoder has extracted the low-level features already, only a few layers can achieve excellent result in scene classification. Path router sub-network produces a probability vector, whose argmax is the identification of the selected branch. We obtain this classification and feed the feature map into the branch we selected.

D. Multi-gated Feature Fusion Branches

For each branch, the feature maps firstly pass a gated fusion sub-network shown in Fig 4.(b) to weight fusion the shared multi-expert output. In [9], it has been proved that gated fusion network exerts excellent performance in feature fusion. Afterward we use two feature extracting groups to extract high-level features; finally use a gated fusion sub-network to weight fusion of the low-level and high-level features. At the last step, deconvolution acting as a decoder turns its size into H*W*3, which is added to the input RGB map to get the clean image.

E. Loss Function

Loss function is the key point in multi-task learning. As for the image restoration, we use smooth L1 and SSIM as the loss function. The gradient of Smooth L1 will decrease as the loss becomes small and differential coefficient remains 1 when the loss is rather large. So it will be more stable in the beginning of training than L2 and have better convergence than L1.
J indicates the hazy/rain/raindrop image. The outputs of our network include the restored image $\hat{I}$ and the predicted classification $\hat{Y}$.

$$\hat{I}, \hat{Y} = Network(J) \tag{2}$$

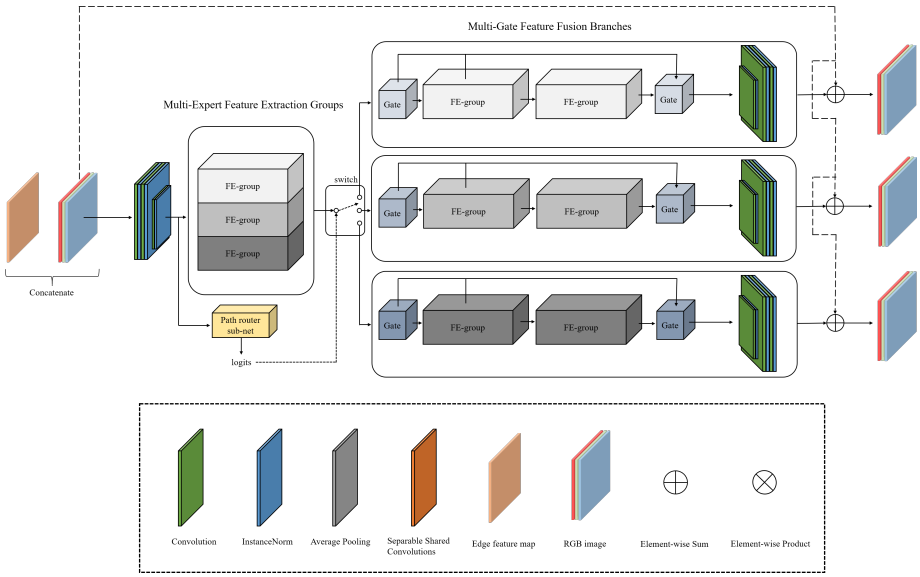


Fig. 1. The architecture of our network

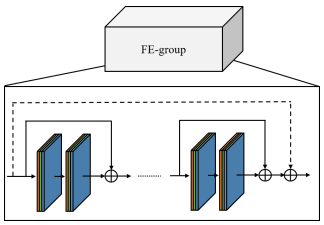


Fig. 2. the structure of FE group

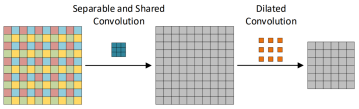


Fig. 3. An illustration of Smoothed Dilated Convolutions

$$\text{smooth}\_L_1(I,\hat{I}) = \begin{cases} 0.5(I-\hat{I})^2 & \text{if } |I-\hat{I}| < 1 \\ |I-\hat{I}| - 0.5 & \text{otherwise} \end{cases} \quad (3)$$

SSIM pays more attention to the whole brightness and contrast ratio of the image, instead of the pixels. Therefore, weighting these two loss functions as the loss of the image restoration can evaluate the image-wise and pixel-wise performance well.

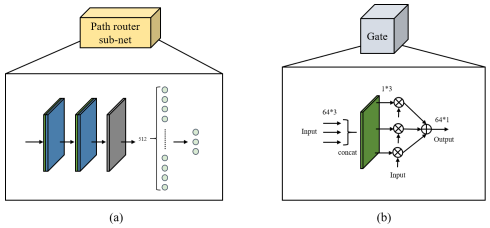


Fig. 4. (a)the structure of path router sub-net (b)the structure of gate

For the path selection, we use cross-entropy as the loss function. Considering that the learning rate is reduced in the later epoch of training and the classifier is roughly stable, we add a cosine decay coefficient to make it decrease smoothly.

$$Cross\_Entropy(Y,\hat{Y}) = \frac{1}{n}\sum_{X}\sum_{i=1}^{C} Y\log(\hat{Y}) \quad (4)$$

The loss function L of our network is as below. λ_1, λ_2 and λ_3 indicate the coefficients.

$$L = \lambda_1 Smooth\_L_1(I,\hat{I}) + \lambda_2(1 - SSIM(I,\hat{I})) + cosdecay(\lambda_3)CrossEntropy(Y,\hat{Y}) \quad (5)$$

III. Experiments

A. Dataset

We select a variety of degraded image datasets and merge them to make the model end-to-end training on it. The selected datasets include rain removal dataset "DDN" [6], hazy removal dataset "RESIDE V0" [21] and raindrop dataset [11]. The "DDN" dataset contains 12600 training samples and 1400 testing samples. "Reside V0" contains 13990 training samples. We use SOTS indoor as the testset, with 500 indoor test samples. The raindrop dataset contains 1119 raindrop and ground truth pairs, of which 58 real world images are selected as test images.

Considering the imbalance of size among datasets, we over-sample the raindrop dataset, and amplify all the images by rotation, random clipping and other methods. After processing, the number of samples in each epoch is 38644.

B. Training Details

The initial learning rate is set to 0.0001 and the number of epochs is 80. We use cosine decay to gradually reduce the

TABLE I
RESULTS OF TEST METHODS ON "DDN" DATASET

Metrics	DDN [6]	NLEDN [5]	MSPFN [4]	Ours
PSNR	28.24	29.79	**32.82**	30.68
SSIM	0.8654	0.8976	**0.9302**	0.9248

TABLE II
RESULTS OF TEST METHODS ON RESIDE V0 DATASET

Metrics	AOD [8]	PFFN [22]	MSBDN [23]	Ours
PSNR	19.67	24.78	**33.79**	30.48
SSIM	0.8065	0.8923	**0.9842**	0.955

learning rate. The formula is as follows. We use AdamW as optimizer and weight decay is set to $2 * 10^{-4}$.

$$lr = 0.5 * initial\_lr * (1 + cos(\pi \frac{global\_step}{decay\_steps})) \quad (6)$$

C. Results

1) Quantitative Results: PSNR(Peak Signal to Noise Ratio) and SSIM(Structural Similarity) are recognized image quality evaluation standards. We use them as the metrics to compare our network with some state of the art methods in various sub fields. It is important to note that the compared networks are trained on the dataset of a single task, and our network is trained on the multi task merging dataset.
We take some state of the art methods as the baseline for comparison, including **rain removal** : DDN [6], NLEDN [5], MSPFN [4]; **hazy removal**: AOD-Net [8], PFFN [22], MSBDN [23]; **raindrop removal**: AttentGAN [11], Quen et al. [12]. The results are shown in Table 1, Table 2 and Table 3. Even though our network can be used for multi-tasks, it still shows competitive results in PSNR and SSIM of the three tasks, which are close to the best results of the state of the art methods.

2) Qualitative Results: Fig 5., Fig 6. and Fig 7. show the output of different methods respectively. We can see that our restoration results have achieved competitive results in color restoration, and preserved the details well. In Fig 5., the girl's face: It can be seen that the texture of the skin and lip is well preserved rather than becoming smooth and the athlete: The number on the vest still remains clear and well-recognized. In Fig 6., We can see that the edges of the object and building haven't turned blunt, but still sharp enough to distinguish.

D. Ablation Study

In order to show the effectiveness of the multi-task module of our network, we conducted an ablation experiment. Single

TABLE III
RESULTS OF TEST METHODS ON RAINDROP DATASET

Metrics	AttentGAN [11]	Quen et al. [12]	Ours
PSNR	**31.57**	31.44	29.33
SSIM	0.9023	0.9263	**0.9272**

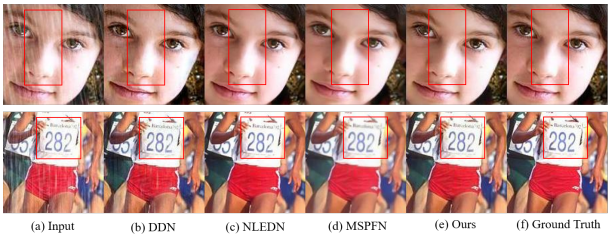


Fig. 5. Rain removal results of our method compared with state-of-the-art rain removal methods.

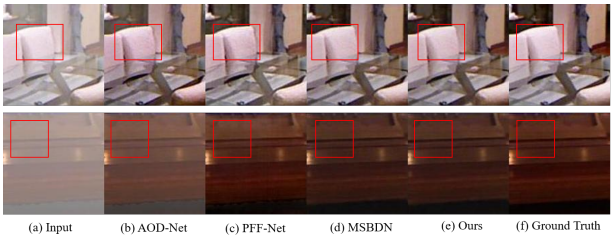


Fig. 6. Hazy removal results of our method compared with state-of-the-art hazy removal methods.

expert without gate, multi-expert with single gate and multi-expert with multi-gate (ours) are compared respectively. The structure of above modules are shown in Fig 8. The experimental results are shown in Table 4 and Fig 9. It can be seen that the performance of single expert without gate is the worst. Because of the differences among tasks, the feature map of single expert is difficult to contain the rich features of multi-task. For the results of multi-expert with single gate, we can see that the performance of dehazing is much better than that of no gate. That is because its task similarity is quite different from that of rain removal and raindrop removal. And gate sub-network can reduce the weight of irrelevant features. The results show that multi-expert with multi-gate can benefit

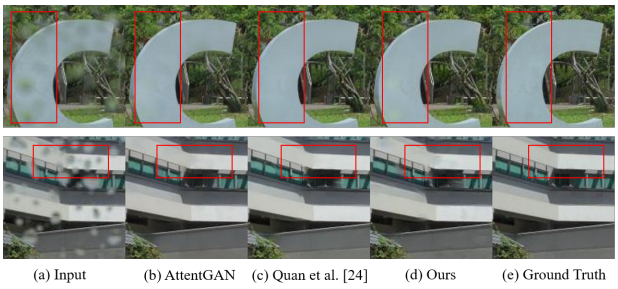


Fig. 7. Raindrop removal results of our method compared with state-of-the-art hazy Raindrop methods.

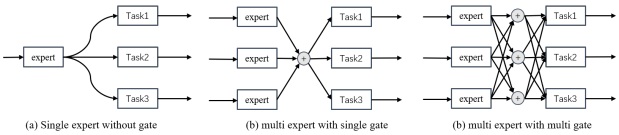
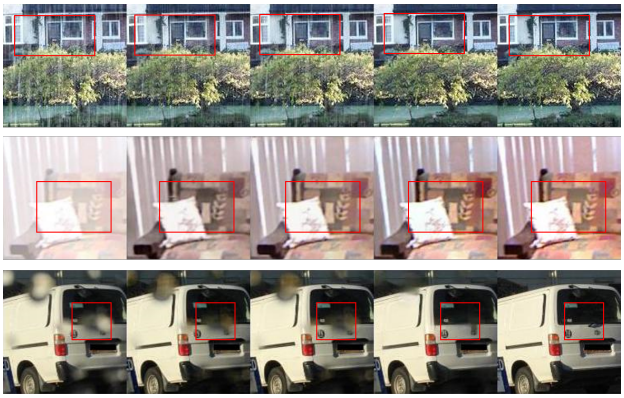


Fig. 8. The compared multi-task learning structures

TABLE IV
RESULTS OF TEST METHODS IN ABLATION STUDY

Task	Metrics	Single expert without gate	Multi-expert with single gate	Ours
Derain	PSNR	28.36	28.42	**30.48**
	SSIM	0.8850	0.8801	**0.9233**
	Accuracy	0.9907	0.9978	**1.0**
Dehaze	PSNR	22.90	24.18	**30.03**
	SSIM	0.8832	0.8897	**0.9552**
	Accuracy	0.998	**1.0**	**1.0**
Deraindrop	PSNR	27.05	26.34	**29.33**
	SSIM	0.8999	0.8917	**0.9272**
	Accuracy	**1.0**	**1.0**	**1.0**



(a) Input (b) Single expert without gate (c) Multi expert with single gate (d) Ours (e) Ground Truth

Fig. 9. Comparison between different networks in ablation study

the performance most.

IV. CONCLUSION

In this paper, we propose an unified network for multi-task image restoration of different degraded images. Competitive experimental results are obtained on the mixed dataset of rain, haze and raindrop removal. There are three main contributions of this paper. Firstly, the MMoE multi-task architecture is innovatively introduced into the image restoration task. The second point is proposing a path router sub-network to select branches through scene classification, so that the network can be trained and predicted end-to-end. Finally, we propose multi-gate sub-network to fuse the feature of different levels. With the development of multi task learning, more efficient network architecture can be migrated to multi-task image restoration with our framework. We believe that we can get a more robust and efficient model in the future.

REFERENCES

[1] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, 2011.

[2] P.-L. Shui, "Image denoising algorithm via doubly local wiener filtering with directional windows in wavelet domain," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 681–684, 2005.

[3] X. Li, J. Wu, Z. Lin, H. Liu, and H. Zha, "Recurrent squeeze-and-excitation context aggregation net for single image deraining," in *ECCV*, 2018.

[4] K. Jiang, Z. Wang, P. Yi, C. Chen, B. Huang, Y. Luo, J. Ma, and J. Jiang, "Multi-scale progressive fusion network for single image deraining," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8343–8352.

[5] G. Li, X. He, W. Zhang, H. Chang, L. Dong, and L. Lin, "Non-locally enhanced encoder-decoder network for single image de-raining," in *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, S. Boll, K. M. Lee, J. Luo, W. Zhu, H. Byun, C. W. Chen, R. Lienhart, and T. Mei, Eds. ACM, 2018, pp. 1056–1064.

[6] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley, "Removing rain from single images via a deep detail network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1715–1723.

[7] 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.

[8] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4780–4788.

[9] D. Chen, M. He, Q. Fan, J. Liao, L. Zhang, D. Hou, L. Yuan, and G. Hua, "Gated context aggregation network for image dehazing and deraining," *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1375–1383, 2019.

[10] D. Engin, A. Genç, and H. K. Ekenel, "Cycle-dehaze: Enhanced cyclegan for single image dehazing," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 938–9388, 2018.

[11] R. Qian, R. Tan, W. Yang, J. Su, and J. Liu, "Attentive generative adversarial network for raindrop removal from a single image," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2482–2491, 2018.

[12] Y. Quan, S. Deng, Y. Chen, and H. Ji, "Deep learning for seeing through window with raindrops," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2463–2471.

[13] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8183–8192.

[14] K. Yu, C. Dong, L. Lin, and C. C. Loy, "Crafting a toolchain for image restoration by deep reinforcement learning," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2443–2452, 2018.

[15] M. Suganuma, X. Liu, and T. Okatani, "Attention-based adaptive selection of operations for image restoration in the presence of unknown combined distortions," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9031–9040, 2019.

[16] R. Li, R. T. Tan, and L.-F. Cheong, "All in one bad weather removal using architectural search," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3172–3182.

[17] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, ser. KDD '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1930–1939.

[18] R. Caruana, *Multitask Learning*. Boston, MA: Springer US, 1998, pp. 95–133.

[19] Z. Wang and S. Ji, "Smoothed dilated convolutions for improved dense prediction," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, Jul 2018.

[20] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "A generic deep architecture for single image reflection removal and image smoothing," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3258–3267.

[21] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2018.

[22] K. Mei, A. Jiang, J. Li, and M. Wang, "Progressive feature fusion network for realistic image dehazing," in *Asian Conference on Computer Vision (ACCV)*, 2018.

[23] D. Hang, P. Jinshan, H. Zhe, L. Xiang, Z. Xinyi, W. Fei, and Y. Ming-Hsuan, "Multi-scale boosted dehazing network with dense feature fusion," in *CVPR*, 2020.