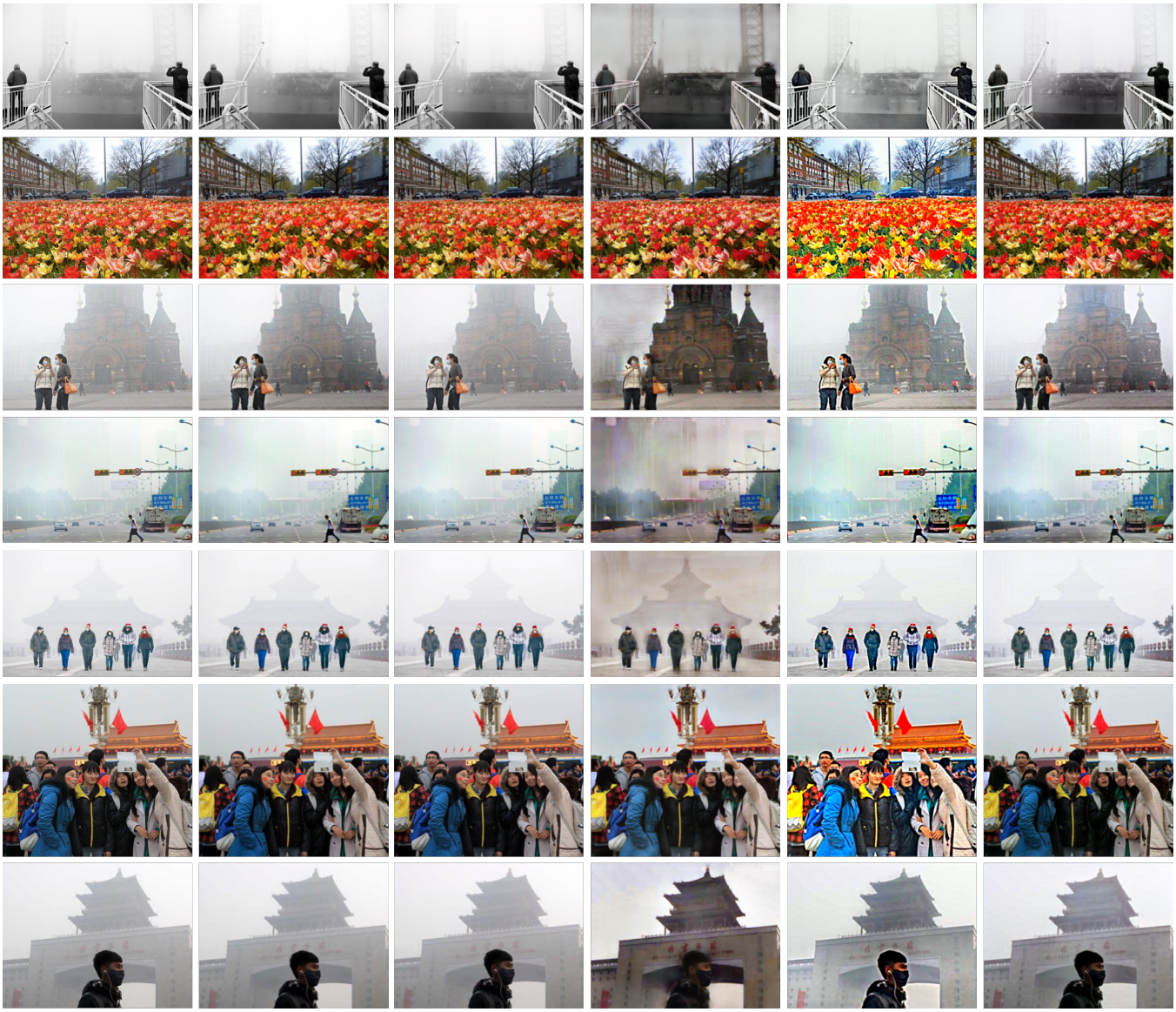


(a) Hazy Images (b) MSBDN [1] (c) DehazeFlow [2] (d) DA-Net [3] (e) PSD [4] (f) TCDAD(ours)

Fig. 1: Comparison of dehazing results on real hazy images from RTTS [5].

I. Experiments

To verify our method's effectiveness and generalization ability, we implemented the proposed framework on MSBDN [1] backbone and evaluated the framework on a real-world dataset. We compared our method with several state-of-the-art methods on visual quality and several commonly adopted metrics. Finally, we conducted ablation study and framework generalization to prove the effectiveness of TCDAD and extended InfoNCE loss.

TABLE I: Quantitative results using NR-IQA metrics on RTTS [5]. The best results are in bold.

Method	FADE [6]↓	BRISQUE [7]↓	PIQE [8]↓
Hazy	2.484	37.011	51.254
MSBDN [1]	1.363	28.743	50.657
DehazeFlow [2]	1.763	26.059	38.879
DA-Net [3]	1.130	32.456	50.787
PSD [4]	0.920	25.239	30.631
TCDAD(ours)	**0.915**	**23.177**	**18.480**

A. Implementation Details

1) Datasets: Our method is trained and evaluated on RESIDE [5] dataset. RESIDE [5] dataset is divided into five subsets, namely, ITS (Indoor Training Set), OTS (Outdoor Training Set), SOTS (Synthetic Object Testing Set), URHI (Unannotated real Hazy Images), and RTTS (real Task-driven Testing Set). We choose 3000 hazy image pairs from OTS and 1000 hazy images from URHI (same sample as [3]) for training. All the images are randomly cropped to patches of size 256×256, with normalized pixel values from -1 to 1 in the training phase.

(a) Hazy Images (b) DA-Net [3] (c) PSD [4] (d) TCDAD(ours)

Fig. 2: Some object detection results on RTTS [5]. The ground truths are circled with blue rectangles and the object detection results are circled with rectangles in other colors.

2) Training Details: We implement our framework in PyTorch and utilize ADAM optimizer with a batch size of 16 to train the networks on an Nvidia RTX3090. The number of encoder layers k = 4 and the temperature parameter τ = 1e-6 for both stages. In the synthetic-to-real adaptation stage, we train the translation networks for 90 epochs with the momentum β_1 = 0.5, β_2 = 0.999, and the learning rate is set as 5×10^{-5}. The trade-off weight is set as: α_1 = 0.1. In the hazy-to-clean adaptation stage, we train the dehazing networks for 40 epochs using the pre-trained translation network. The momentum and the learning rate are set as: β_1 = 0.95, β_2 = 0.999, lr = 10^{-4}. When computing the DC loss, we set the patch as 35 × 35. The trade-off weights are set as: λ_1 = 0.1, λ_2 = 0.05 and λ_3 = 0.01.

B. Comparisons with State-of-the-art Methods

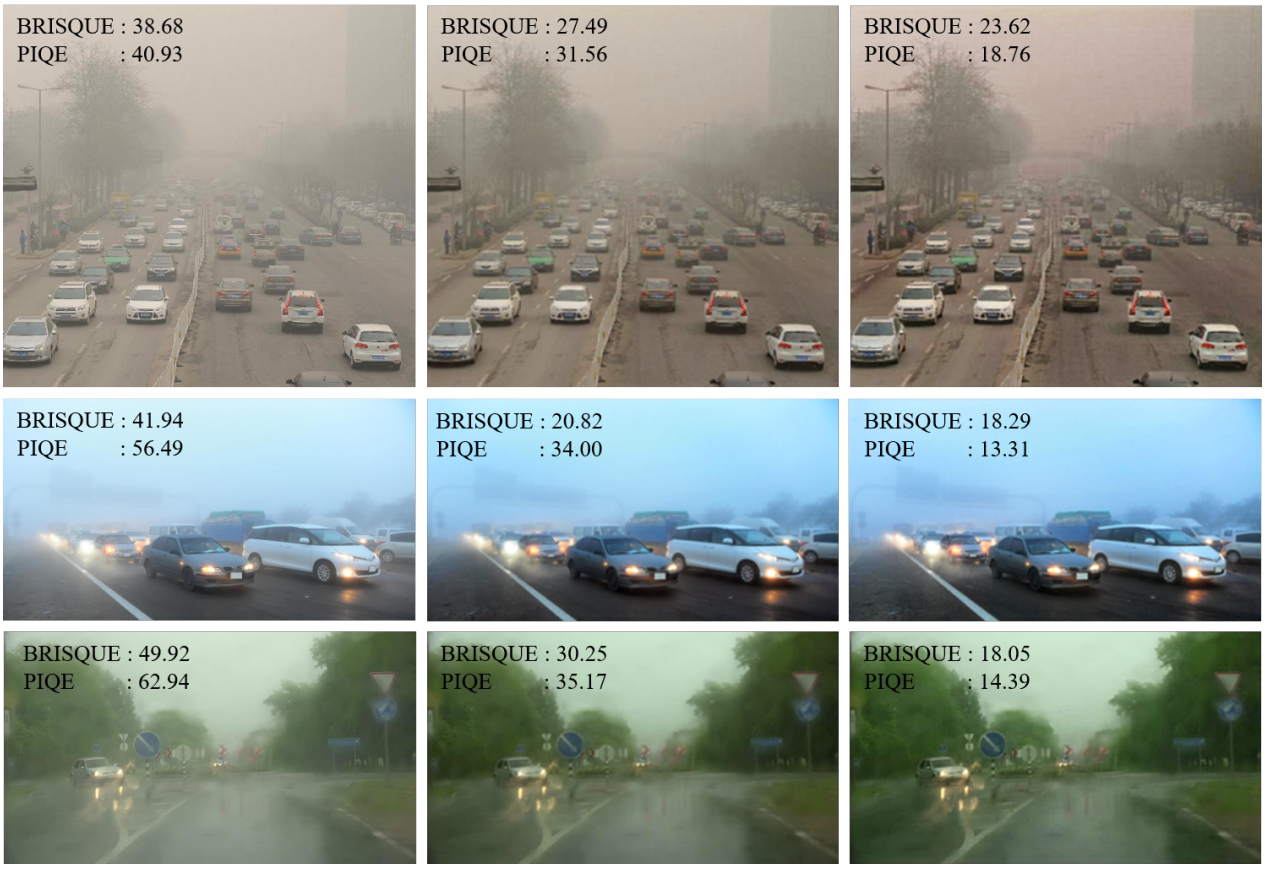
We compared the proposed framework with several state-of-the-art methods on a real-world dataset, qualitatively and quantitatively. For the sake of fairness, we prefer to use the metrics and the trained models provided by the authors.

1) Visual Quality Comparison: We first evaluate the visual quality of TCDAD on real hazy images from RTTS [5], which is a subset of the RESIDE dataset [5]. We compare results of TCDAD with four state-of-the-art methods including MSBDN [1] (the backbone), DehazeFlow [2], DA-NET [3] and PSD [4]. The results are shown in Fig. 1.

From Figs. 1, we can observe that images restored by MSBDN [1] and DehazeFlow [2] remain a bit hazy, especially in distant areas. MSBDN [1] and DehazeFlow [2] are trained on synthetic hazy-clean pairs, and thence lack the ability to remove haze in real-world scenes. DA-NET [3] suffers from details loss and spatial distortion. Moreover, images restored by DA-NET [3] have small shadows in some cases. Images restored by PSD [4] are brighter and have higher contrast due to the bright channel prior and contrast limited adaptive histogram equalization prior. However, We can also observe that the objects in the pictures appear edge blurring, especially in human images and in brighter scenes. What's more, the priors used by PSD [4] introduces many artifacts, making the results lack of naturalness. Compared with all these methods, TCDAD generates high-quality haze-free images with richer details and sharper edges. And the results of TCDAD have a similar style to real clean images. The comparisons on no-reference image quality assessment and high-level tasks provide more evidence.

2) No-Reference Image Quality Assessment: For quantitative comparison, we evaluate three no-reference image quality assessment indicators on RTTS [5] dataset. We employ the Fog Aware Density Evaluator (FADE) [6] to assess the haze density of dehazed images as in [4]. To assess the quality of dehazed images, we choose two well-known no-reference image quality assessment indicators: Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) [7] and Perception-based Image Quality Evaluator (PIQE) [8]. We compare TCDAD with MSBDN [1] (the backbone), DehazeFlow [2], DA-Net [3], and PSD [4].

Both BRISQUE [7] and PIQE [8] are blind image quality evaluators. BRISQUE [7] does not compute distortion-specific features, such as ringing, blur, or blocking, but instead uses scene statistics of locally normalized luminance coefficients to quantify possible losses of naturalness in the image due to the presence of distortions, thereby leading to a holistic measure of quality. PIQE [8] estimates quality only from perceptually significant spatial regions and the score relies on local features related to three distortions: blockiness, blur, and noise. As shown in Table I, TCDAD achieves the best on both two image quality assessment indicators and significantly exceeds other methods on PIQE [8]. The results on BRISQUE [7] and PIQE [8] meet the visual quality comparison, illustrating that TCDAD generates natural dehazed images by reserving more details and introducing fewer noises. Also, TCD surpasses PSD [4] and other methods in FADE [6] results, which means



(a) (b) (c)

Fig. 3: Results of the ablation study on the synthetic-to-real domain adaptation stage. (a) are hazy images, (b) are dehazing results from model without synthetic-to-real domain adaptation stage, (c) are dehazing results from TCDAD.

TABLE II: Object detection results on RTTS [5]. The best results are in bold. Note that mAP may not match the mean results of five categories due to rounding.

Method	bicycle	bus	car	motorbike	person	mAP	gain
Hazy	56.11	48.05	75.19	56.66	80.59	63.32	-
MSBDN [1]	57.82	48.75	76.70	60.77	81.82	65.16	+1.84
DehazeFlow [2]	58.12	49.39	75.79	57.82	81.98	64.62	+1.30
DA-Net [3]	57.91	**49.88**	77.40	59.69	81.13	65.20	+1.88
PSD [4]	59.86	49.79	76.90	60.10	82.57	65.84	+2.52
TCDAD(ours)	**59.93**	**49.88**	**78.07**	**61.53**	**83.18**	**66.52**	**+3.20**

that TCDAD has a stronger dehazing ability.

In total, we win all three metrics, further endorsing that TCDAD can generate haze-free images with high quality on real-world image dehazing.

3) Task-Driven Evaluation: Performance of high-level computer vision tasks, such as object detection and recognition, is severely constrained by image quality as pointed in [9], [10]. Haze, noise, and blur in images will lead to severe performance drop in high-level tasks. Thus, we take object detection as a downstream task and compare the results with MSBDN [1] (the backbone), DehazeFlow [2], DA-Net [3] and PSD [4] to evaluate the efficiency of our TCDAD. RTTS [5], which consists of 4,322 real-world hazy images annotated with bounding boxes and five object categories: bus, bicycle, car, motorbike and person, is taken as the test set. We take the fast and widely used YOLOv3 [11] as the object detection model. The model is pre-trained on coco [12] dataset which has eighty common object categories including bus, bicycle, car, motorbike, and person. Hazy images from RTTS [5] is first dehazed by the dehazing models. Then, object detection is performed on the dehazed images by YOLOv3 [11].

We compute the Average Precision (AP) of each of the five categories above as well as the mean Average Precision (mAP) on the detection results. As shown in Table II, TCDAD significantly improves the accuracy of object detection and achieves first place in all categories.

Also, some object detection results of hazy images from RTTS [5], dehazing results of DA-Net [3], PSD [4] and our TCDAD are shown in Fig. 2. As shown in the first row in Fig. 2, YOLOv3 [11] can not confirm whether the bus in the

Fig. 4: Results of the ablation study on the extended InfoNCE loss. (a) are hazy images, (b) are dehazing results from the model without extended InfoNCE loss, (c) are dehazing results from the model with extended InfoNCE loss only in the synthetic-to-real adaptation stage, (d) are dehazing results from the model with extended InfoNCE loss only in the hazy-to-clean adaptation stage, (e) are dehazing results from TCDAD.

TABLE III: Quantitative results using NR-IQA metrics of the ablation study on the synthetic-to-real stage and the extend InfoNCE loss. N-S2R denotes results from model without synthetic-to-real domain adaptation stage, NN-NCE denotes results from model without extended InfoNCE loss, ON-NCE denotes results from model with extended InfoNCE loss only in the synthetic-to-real adaptation stage, NO-NCE denotes results from model with extended InfoNCE loss only in the hazy-to-clean adaptation stage. The best results are in bold.

Method	FADE [6]↓	BRISQUE [7]↓	PIQE [8]↓
Hazy	2.484	37.011	51.254
N-S2R	1.314	25.416	31.036
NN-NCE	1.288	23.725	25.425
ON-NCE	1.041	23.350	**18.448**
NO-NCE	0.996	23.696	18.823
TCDAD(ours)	**0.915**	**23.177**	18.480

distance is a bus or a truck in the hazy image. As for the dehazed images of DA-Net [3] and PSD [4], YOLOv3 [11] falsely detect a non-existent truck. As a comparison, the object detection model successfully detects the two buses with high confidence on the dehazed image of TCDAD. In the second row in Fig. 2, YOLOv3 [11] successfully detect the three cars on the dehazed image of TCDAD while missed one on hazy and dehazed images of DA-Net [3] and PSD [4]. What's more, YOLOv3 [11] also detect other objects with higher confidence and more accurate border on the dehazed image of TCDAD than DA-Net [3] and PSD [4].

The comparison results on the object detection side prove that images processed by TCDAD reserve more details and have sharper edges.

C. Ablation Study

In order to demonstrate the effectiveness of the synthetic-to-real domain adaptation stage and our proposed extended InfoNCE loss, we conduct two ablation studies.

We first train a model without the synthetic-to-real domain adaptation stage (marked as N-S2R), which means that synthetic hazy-clean image pairs are fed into the dehazing network in the hazy-to-clean domain adaptation stage directly. Other settings of N-S2R are the same as full TCDAD. The visual results of N-S2R and full TCDAD are shown in Fig. 3, the dehazing results of N-S2R are much cleaner than the hazy images, but there are still some hazy areas. Compared to N-S2R, TCDAD further removes haze in input hazy images. Meanwhile, TCDAD also reserves more details and generates images with more harmonious color and sharper edges.

Then three different models based on TCDAD are trained to prove the effectiveness of the extended InfoNCE loss. We train a model without extended InfoNCE loss in both stages (marked as NN-NCE), a model with extended InfoNCE loss only in the synthetic-to-real domain adaptation stage (marked as ON-NCE), a model with extended InfoNCE loss only in the hazy-to-clean domain adaptation stage (marked as NO-NCE). The other settings of the three models are the same as full TCDAD. The visual results of the three models and full TCDAD are shown in Fig. 4. In Fig. 4 (b), without extended InfoNCE loss in both stages, the dehazing result suffers from significant low quality. In Fig. 4 (c), without extended InfoNCE loss in the hazy-to-clean adaptation stage, the dehazing result remains a bit hazy. In Fig. 4 (d), without extended InfoNCE loss in the synthetic-to-real adaptation stage, the colors are not coordinating. Visual results show that the extended InfoNCE loss improves the quality of dehazing images in both stages.

Fig. 5: Results on different backbones. (a) are hazy images, (b) and (c) are dehazing results from GridDehazeNet [13] and TCDAD upon GridDehazeNet [13] respectively, (d) and (e) are dehazing results from FFANet [14] and TCDAD upon FFANet [14] respectively.

We also evaluate the above models on three no-reference image quality assessment indicators (FADE [6], BRISQUE [7] and PIQE [8]) on RTTS [5]. The results are shown in Table III. We can observe that the full TCDAD exceeds N-S2R in all three metrics with a large margin, proving the necessity of the synthetic-to-real domain adaptation stage. Compared with NN-NCE, ON-NCE and NO-NCE both have an improvement in all three indicators, especially in FADE [6] and PIQE [8]. What's more, NO-NCE exceeds ON-NCE in FADE [6]. The comparison shows that the extended InfoNCE losses in both synthetic-to-real and hazy-to-clean domain adaptation stages can improve the quality of dehazing results. The extended InfoNCE loss in the hazy-to-clean domain adaptation stage helps more in haze removal. Compared with ON-NCE and NO-NCE, the full TCDAD has further improvement, especially in FADE [6]. The comparison shows that the extended InfoNCE losses in the synthetic-to-real and hazy-to-clean domain adaptation stage can work together to further improve the dehazing ability of the backbone dehazing network. Meanwhile, each of the extended InfoNCE loss in the synthetic-to-real and hazy-to-clean domain adaptation stages can significantly help the dehazing network to generate dehazing images with higher quality (more natural color, richer details, etc).

D. Generalization

Apart from the default MSBDN [1], we also test two other backbones, namely GridDehazeNet [13] and FFANet [14], to prove the generality of the proposed TCDAD. We get dehazing results of GridDehazeNet [13] and FFANet [14] from corresponding authors and denote them as GCA and FFA below, respectively. We conduct TCDAD upon GridDehazeNet [13] and FFANet [14] with the same settings as original TCDAD and denote them as GCA+ and FFA+ below, respectively. Since we just need to replace the dehazing network without any other changes on the framework, the transplant is easy and convenient.

TABLE IV: Quantitative results using NR-IQA metrics of the generalization on GridDehazeNet [13] and FFANet [14]. GCA denotes results from GridDehazeNet [13], GCA+ denotes results from TCDAD upon GridDehazeNet [13], FFA denotes results from FFANet [14], FFA+ denotes results from TCDAD upon FFANet [14].

Method	FADE [6]↓	BRISQUE [7]↓	PIQE [8]↓
Hazy	2.484	37.011	51.254
GCA	1.525	29.731	47.537
GCA+	**0.996**	**26.217**	**34.285**
FFA	2.074	34.439	48.199
FFA+	**1.275**	**30.119**	**47.367**

The visual results of GCA, GCA+, FFA and FFA+ are shown in Fig. 5. We can observe that compared with GCA, the dehazing results of GCA+ are significantly less hazy and have less vague areas. FFA+ also generates cleaner and more natural images than the corresponding backbone FFA.

A quantitative evaluation of GCA, GCA+, FFA and FFA+ is carried on three no-reference image quality assessment indicators (FADE [6], BRISQUE [7] and PIQE [8]) on RTTS [5]. As shown in Table IV, GCA+ outperforms GCA on all three metrics with a large margin and FFA+ also outperforms FFA on FADE [6] and BRISQUE [7] with a large margin. The excellent performance of GCA+ and FFA+ in the qualitative and quantitative comparisons shows that TCDAD is an effective, generic and portable framework. However, we can also find that FFA+ outperforms FFA on PIQE only with a small margin 0.832. We believe this is mainly because that FFA adopts Pixel Attention to specifically extract the edge and texture features of objects in the images. Thus the FFA dehazing network benefits less from the extended InfoNCE loss in terms of edge and detail preservation.

II. Conclusion

In this paper, we propose a novel Two-stage Contrastive Domain Adaptation Dehazing (TCDAD) framework for real-world scenes, which introduces extended InfoNCE into both the synthetic-to-real adaptation stage and the hazy-to-clean adaptation stage. TCDAD takes full advantage of synthetic hazy and clean images as well as real hazy and clean images to train a powerful dehazing network. And TCDAD is easy to transplant upon learning-based dehazing backbones. The extended InfoNCE accommodates the lack of strongly correlated positives in real-world dehazing. Moreover, We show that our extended InfoNCE can be used as an effective constraint in domain adaptation. Extensive experiments demonstrate the effectiveness and generality of the proposed method.

References

[1] H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, and M.-H. Yang, "Multi-scale boosted dehazing network with dense feature fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2157–2167.

[2] H. Li, J. Li, D. Zhao, and L. Xu, "DehazeFlow: Multi-scale conditional flow network for single image dehazing," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2577–2585.

[3] Y. Shao, L. Li, W. Ren, C. Gao, and N. Sang, "Domain adaptation for image dehazing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2808–2817.

[4] Z. Chen, Y. Wang, Y. Yang, and D. Liu, "Psd: Principled synthetic-to-real dehazing guided by physical priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7180–7189.

[5] 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.

[6] L. K. Choi, J. You, and A. C. Bovik, "Referenceless prediction of perceptual fog density and perceptual image defogging," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3888–3901, 2015.

[7] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[8] N. Venkatanath, D. Praneeth, M. C. Bh, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *Twenty First National Conference on Communications (NCC)*. IEEE, 2015, pp. 1–6.

[9] R. G. VidalMata, S. Banerjee, B. RichardWebster, M. Albright, P. Davalos, S. McCloskey, B. Miller, A. Tambo, S. Ghosh, S. Nagesh *et al.*, "Bridging the gap between computational photography and visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 12, pp. 4272–4290, 2020.

[10] W. Yang, Y. Yuan, W. Ren, J. Liu, W. J. Scheirer, Z. Wang, T. Zhang, Q. Zhong, D. Xie, S. Pu *et al.*, "Advancing image understanding in poor visibility environments: A collective benchmark study," *IEEE Transactions on Image Processing*, vol. 29, pp. 5737–5752, 2020.

[11] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[13] X. Liu, Y. Ma, Z. Shi, and J. Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7314–7323.

[14] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "FFA-Net: Feature fusion attention network for single image dehazing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 908–11 915.