HazeCLIP: Towards Language Guided Real-World Image Dehazing

Ruiyi Wang, Wenhao Li, Xiaohong Liu, *Member, IEEE*, Chunyi Li, Zicheng Zhang, Xiongkuo Min, *Member, IEEE*, and Guangtao Zhai, *Senior Member, IEEE*

***Abstract*—Existing methods have achieved remarkable performance in single image dehazing, particularly on synthetic datasets. However, they often struggle with real-world hazy images due to domain shift, limiting their practical applicability. This paper introduces HazeCLIP, a language-guided adaptation framework designed to enhance the real-world performance of pre-trained dehazing networks. Inspired by the Contrastive Language-Image Pre-training (CLIP) model's ability to distinguish between hazy and clean images, we utilize it to evaluate dehazing results. Combined with a region-specific dehazing technique and tailored prompt sets, CLIP model accurately identifies hazy areas, providing a high-quality, human-like prior that guides the fine-tuning process of pre-trained networks. Extensive experiments demonstrate that HazeCLIP achieves the state-of-the-art performance in real-word image dehazing, evaluated through both visual quality and no-reference quality assessments. Codes are available at https://github.com/Troivyn/HazeCLIP.**

***Index Terms*—Contrastive language-image pre-training, real-world image dehazing, language guidance**

I. Introduction

SINGLE image dehazing aims to restore clean images from hazy ones that suffer from reduced contrast and limited visibility. Similar to other restoration tasks [1], [2], this challenging task is crucial for high-level vision applications such as object detection [3], [4] and semantic scene understanding [5]–[7], making it a longstanding problem. Existing dehazing approaches include prior-based and learning-based methods. Early dehazing algorithms [8]–[13] typically estimate parameters of the atmospheric scattering model [14] using statistical priors. For example, He et al. [11] achieved impressive dehazing results using the Dark Channel Prior (DCP). However, while these prior-based models can perform well without training, their performance is often limited and fragile because hand-crafted priors cannot adapt to the diversity of real-world images.

With the availability of large-scale synthetic datasets and CNNs, numerous learning-based approaches have emerged [15]–[23]. Particularly, Liu et al. [24], [25] proposed an attention-based multi-scale grid network for image dehazing. However, due to the significant domain gap, their performance drops dramatically when applied to real-world hazy images. To overcome this challenge, several works have been proposed for real-world image dehazing [26]–[29]. Among them, many studies reintroduce prior knowledge. For example, Chen et al. [30] employed three statistical image priors for unsupervised fine-tuning, while Wu et al. [31] leveraged latent discrete priors in pre-trained VQGAN [32]. However, directly using handcrafted or learned priors cannot avoid the inherent flaws of prior-based methods.

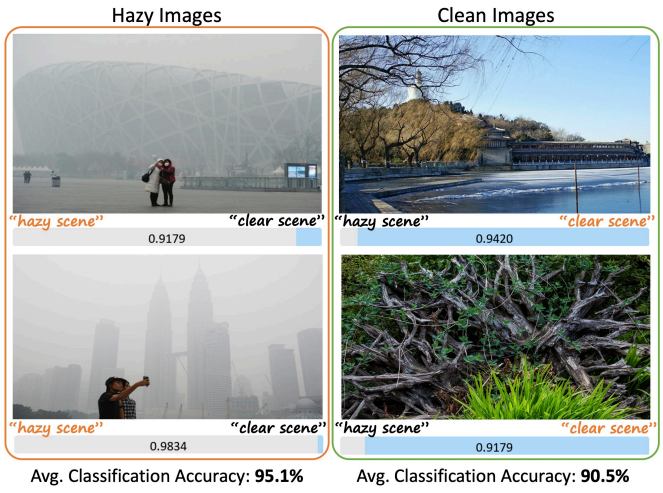


Fig. 1. CLIP model is capable of distinguishing between hazy and clean images. Classification probabilities for example images and average accuracy for hazy and clean image sets are reported.

Recently, a few methods have made breakthroughs in restoration tasks leveraging text information, including [33]–[36]. Specifically, Luo et al. [37] proposed a method to control the CLIP model to disentangle degradation factors from the image features. Zhang et al. [38] proposed an iterative prompt learning method for backlit image enhancement. Considering the success of these methods and the flexibility of natural language compared to statistical priors, we decided to leverage the power of CLIP model for image dehazing.

To overcome the drawbacks of existing dehazing methods, we present HazeCLIP, a language-guided adaptation framework for real image dehazing that generalizes dehazing networks to real-world domain. Inspired by CLIP model's ability to distinguish between hazy and clean images (Fig. 1), our approach leverages its rich visual-language prior instead of relying on the scattering model or statistical priors. Since CLIP model is trained on a diverse set of images, its prior is more robust than traditional statistical ones. Additionally,

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. Ruiyi Wang and Wenhao Li contributed equally to this work. (*Corresponding author: Xiaohong Liu*)

The authors are with the School of Electronics, Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. (e-mail: {thomas25, sevenhao, xiaohongliu, lcysyzxdxc, zzc1998, minxiongkuo, zhaiguangtao}@sjtu.edu.cn)

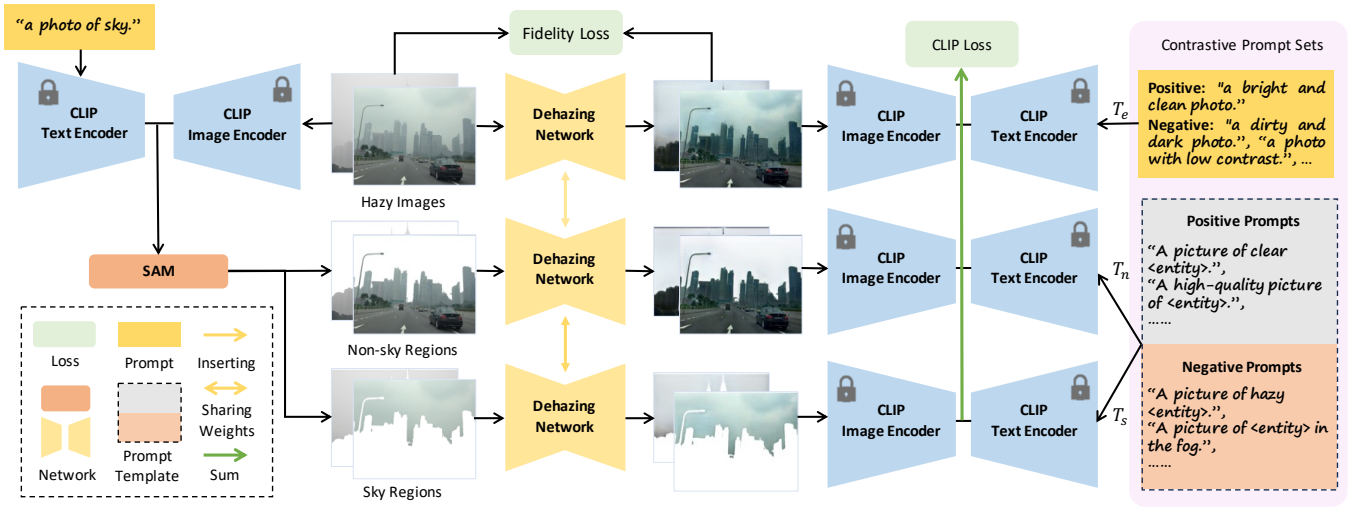


Fig. 2. Overview of the proposed HazeCLIP framework. Real-world hazy images are first separated into sky and non-sky regions using Segment Anything Model (SAM). Combined with the CLIP model, three contrastive prompt sets are applied to guide the adaptation process. The enhancing prompt set aims to improve overall image quality, and non-sky and sky dehazing prompt sets specifically guide dehazing in their respective regions.

CLIP model's perception aligns closely with human perception, allowing our method to generate images with superior visual quality. Nevertheless, potentially due to training data bias where images with haze captions often feature grey skies, CLIP model yields excessively high haze similarity scores for sky areas, impeding its ability to guide haze removal in other areas. To overcome this challenge, we propose a region-specific dehazing technique that separately processes the sky and non-sky regions during fine-tuning. This approach improves CLIP model's accuracy in identifying hazy areas, providing more effective guidance for haze removal in diverse scenes.

模型可能會錯誤地認為天空區域有霧霾，從而在去霧過程中對這些區域進行不必要的處理，而忽略了其他實際存在霧霾的區域。

We summarize our contributions as follows:

◇ We propose a general language-guided adapting framework for real-world image dehazing. This framework can be easily combined with many existing dehazing networks without modifying the network architecture.

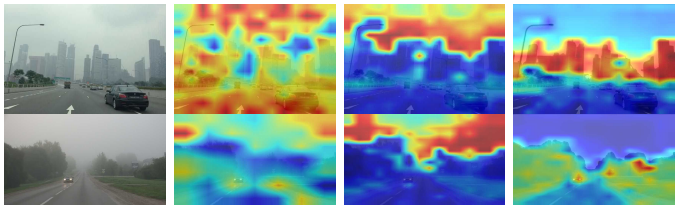
◇ We are pioneers to leverage the power of vision-language models in image dehazing. By designing contrastive prompt sets, the CLIP similarity score can guide the fine-tuning of the dehazing network. We also propose a region-specific dehazing technique to resolve the issue where CLIP model fails to accurately detect hazy regions.

◇ Extensive experiments demonstrate that our HazeCLIP achieves the SOTA real-world dehazing performance.

II. Proposed Method

A. Overview

As shown in Fig. 2, Our HazeCLIP is a fine-tuning framework that guides the pre-trained dehazing network towards real domain leveraging a frozen CLIP model. First we adopt a dehazing network $\mathcal{M}$ pre-trained with synthetic data as the backbone. Given that our approach constitutes a general framework, there are no specific requirements for the chosen dehazing network. Observing the random language-image similarity maps generated by the CLIP model, we employ CLIP surgery and a region-specific dehazing technique to achieve accurate haze detection. To enhance language guidance, we develop three contrastive prompt sets, denoted as T_e, T_n, T_s. We provide further details on the key components of our framework below.



(a) (b) (c) (d)

Fig. 3. Language-image similarity maps for hazy images and description of hazy images. By removing the sky, the CLIP model can focus more effectively on scene dehazing. (a) Hazy images, (b) Raw similarity maps, (c) Maps of CLIP surgery [39], (d) Maps with sky mask.

B. Region-Specific Dehazing

In Fig. 3, we present rough language-image similarity maps for hazy images, calculated by measuring the cosine similarity between the features of image patches and description of hazy images. While CLIP model effectively classifies hazy and clean images, it struggles to accurately locate hazy regions, resulting in random similarity maps. To address this limitation, we employ CLIP surgery [39], a technique designed to enhance CLIP model's explainability. However, a significant issue arises as sky region dominates the similarity with descriptions of hazy scenes. In consequence, when CLIP model evaluates the entire image, haze residuals in non-sky regions, such as hazy buildings, are overlooked, leading to sub-optimal dehazing performance.

(a)

(b)

To mitigate the adverse impact of the concentration on sky regions, we propose a region-specific dehazing technique in addition to CLIP surgery. During fine-tuning, the non-sky regions are handled separately from the sky regions. In

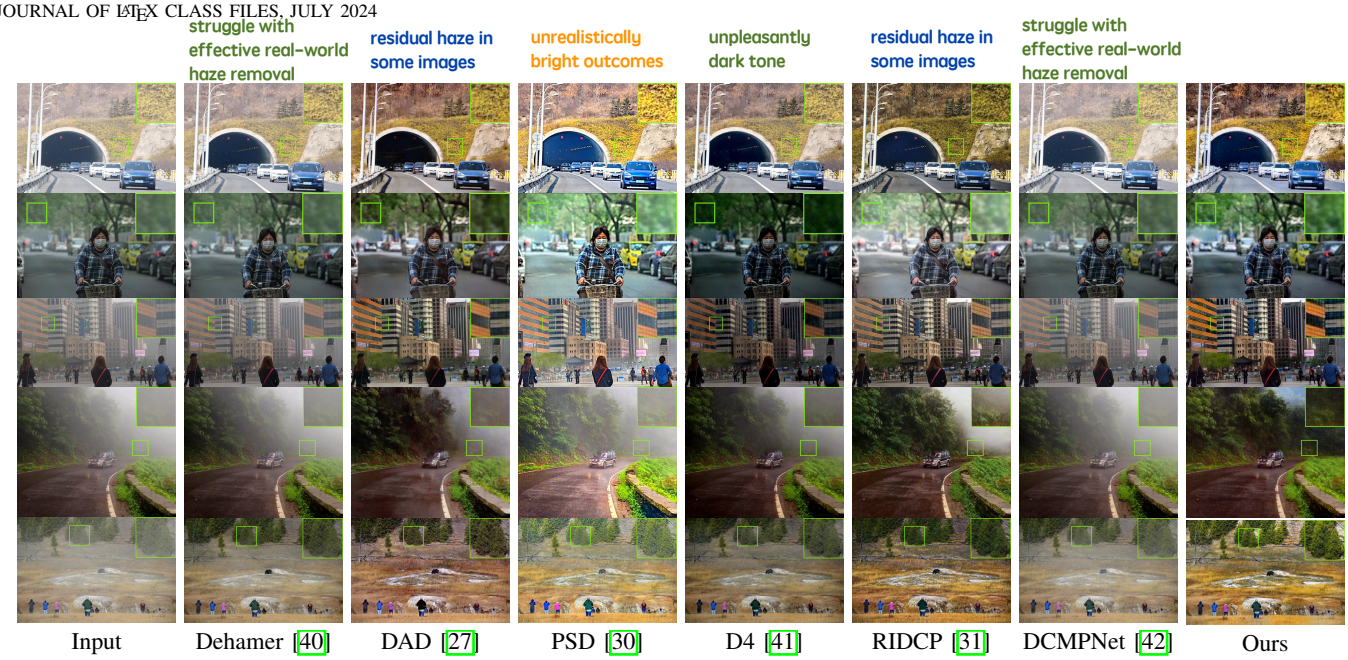


Fig. 4. Visual comparisons on RTTS [15] dataset. For better clarity, the region within the green rectangle is zoomed in and displayed in the top right corner.

Fig. 3(d), we present the similarity maps for hazy images with the sky masked out. The CLIP model successfully identifies hazy areas in non-sky regions, providing precise guidance. Sky masking is achieved by prompting SAM model with points that have high similarity to texts describing the sky.

C. Contrastive Prompt Set

To fully leverage the capabilities of CLIP model, we construct specific prompt sets to guide the dehazing network. Each contrastive prompt set consists of a positive prompt, denoted as T^p, which captures desired properties (e.g., clean), and one or more negative prompts, denoted as T^{n_1}, T^{n_2}, ..., T^{n_K}, which describe undesired properties (e.g., hazy). The text features of the positive and negative prompts are extracted by the text encoder Φ_t of the CLIP model. When evaluating a dehazed image I, the image encoder Φ_i of the CLIP model is employed to obtain the image features. Consequently, based on the text-image similarity in the CLIP latent space, we formulate the CLIP loss with respect to a prompt set as follows:

$$\mathcal{L}_T(I) = \frac{e^{cos(\Phi_i(I),\Phi_t(T^p))}}{\Sigma_{j\in\{p,n_1,...,n_K\}} e^{cos(\Phi_i(I),\Phi_t(T^j))}}, \quad (1)$$

where $T \in \{T_s, T_n, T_e\}$ represents the prompt set, which we introduce in the following section.

D. Prompt Ensemble

As presented in the right part of Fig. 2, to build suitable prompt sets as language guidance, we first construct a dehazing prompt template. Positive templates include expressions like follows:

```
a picture of <entity> in the fog.
a photo of a foggy <entity>.
```

Negative templates follow a similar structure. The sky-region dehazing prompt set T_s is created by setting `sky` as `<entity>`. The non-sky region dehazing set T_n is constructed by inserting objects like `building, people` as well as `scene` into the template. In T_s and T_n, all constructed prompts are averaged in the latent space to form one positive prompt and one negative prompt for calculating the CLIP loss.

In addition to the poor dehazing effect, the output of the pre-trained model also suffers from issues including dull colors and overall dirtiness. To mitigate this shortage, we use an additional image enhancing prompt set T_e to guide the model in producing high-quality results. The design of this prompt set is flexible, and with intentional design, the fine-tuning results can even exhibit a customized style. In this paper, we apply the prompts shown in Fig. 2 for the result presented.

E. Synthetic-to-Real Adaptation

With a pre-trained dehazing network and three contrastive prompt sets, we are able to generalize the model from synthetic to real-world domain. Given an unlabelled real-world hazy image I, sky region I_s and non-sky regions I_n are separated. The CLIP guidance loss is formulated as:

$$\mathcal{L}_c(I) = \mathcal{L}_{T_s}(\mathcal{M}(I_s)) + \mathcal{L}_{T_n}(\mathcal{M}(I_n)) + \lambda_1 \cdot \mathcal{L}_{T_e}(\mathcal{M}(I)), \quad (2)$$

where λ_1 represents the tradeoff weight, L_T is defined in Eq. 1, and $\mathcal{M}$ denotes the dehazing network.

To avoid catastrophic forgetting during fine-tuning, following [38], we implement a fidelity loss to constrain the dehazing result, with α_l representing the tradeoff weight of the l-th layer in the pre-trained CLIP image encoder Φ_i:

$$\mathcal{L}_f(I) = \Sigma_{l=0}^{4} \alpha_l \cdot \|\Phi_i^l(\mathcal{M}(I)) - \Phi_i^l(I)\|_2. \quad (3)$$

Eventually, the overall loss function $\mathcal{L}$ in fine-tuning is defined with respect to a tradeoff weight λ_2:

$$\mathcal{L}(I) = \mathcal{L}_c(I) + \lambda_2 \cdot \mathcal{L}_f(I). \quad (4)$$

III. Experiments

A. Experimental Settings

Dataset. For pre-training, we used synthetic data from RIDCP [31]. For fine-tuning, we selected the URHI split from

TABLE I
QUANTITATIVE COMPARISON ON RTTS DATASET [15]. THE BEST IS IN **RED** WHILE THE SECOND IS IN **BLUE**.

Method	FADE↓	BRISQUE↓	NIMA↑	MOS↑
Hazy image	2.484	37.011	4.3250	-
Dehamer [40]	1.895	33.866	3.8663	2.78
DAD [27]	1.130	32.727	4.0055	3.13
PSD [30]	**0.920**	25.239	4.3459	3.20
D4 [41]	1.358	33.206	3.7239	2.48
RIDCP [31]	0.944	**18.782**	4.4267	**3.57**
DCMPNet [42]	1.921	32.520	**4.4351**	2.85
HazeCLIP	**0.638**	**18.567**	**4.5510**	**3.60**

TABLE II
QUANTITATIVE COMPARISONS AMONG THE THREE SETTINGS OF ABLATION STUDY AND THE FULL VERSION OF HAZECLIP. [KEY: **BEST**]

without the region-specific dehazing technique

Metric	setting (a)	setting (b)	setting (c)	full version
FADE ↓	1.091	0.857	0.695	**0.638**
BRISQUE ↓	27.309	20.499	22.376	**18.567**
NIMA ↑	4.3961	4.4587	4.3965	**4.5510**

without HazeCLIP adaptation　　without the enhancing prompt set，只用一組

the RESIDE [15] dataset. Following previous works [31], we tested our method on the RTTS split from the RESIDE dataset, which contains over 4,000 real-world hazy images with diverse scenes and haze patterns.

Implementation Details. Unless otherwise specified, the experiments in this paper were conducted using the MSBDN [17] network. During pre-training, the network was trained for 200 epochs using the Lion optimizer. The initial learning rate was set to 3×10^{-5} with a cosine annealing scheduler. The network was later fine-tuned for 15 epochs. We empirically set $\lambda_1 = 0.5$ and $\lambda_2 = 0.1$.

B. Comparison with State-of-the-Art Methods

We compared the performance of HazeCLIP with several state-of-the-art dehazing approaches: Dehamer [40], DAD [27], PSD [30], D4 [41], RIDCP [31], and DCMPNet [42]. It's worth noting that since HazeCLIP doesn't introduce any new parameters, the computational cost during inference remains the same as that of the backbone network.

Visual Quality. We evaluated the visual quality of HazeCLIP on real-world hazy images from the RTTS dataset. As shown in Fig. 4, Dehamer and DCMPNet struggle with effective real-world haze removal. D4's results exhibit an unpleasantly dark tone, while PSD produces unrealistically bright outcomes. DAD and RIDCP leave residual haze in some images. In contrast, the proposed method demonstrates superior overall performance, effectively addressing the shortcomings of the other approaches.

No-Reference Image Quality Assessment. For quantitative comparison, we used the Fog Aware Density Evaluator (FADE) [43] to assess dehazing ability. We also included two widely-used image quality assessment metrics: BRISQUE [44] and NIMA [45]. Additionally, we conducted a user study that invited 22 subjects to obtain the subjective Mean Opinion Score (MOS). The results, reported in Table I, show that HazeCLIP achieved the best performance in all metrics. Notably, HazeCLIP demonstrated a 30.7% improvement in FADE, underscoring its superiority in real-world dehazing.

TABLE III
QUANTITATIVE COMPARISONS OF PRE-TRAINED AND FINE-TUNED VERSIONS FOR DIFFERENT NETWORKS. [KEY: **BEST**]

Method	FADE↓	BRISQUE↓	NIMA↑
GDN [24]	1.470	29.448	4.2502
GDN+HazeCLIP	**0.976**	**21.352**	**4.3252**
FFANet [16]	1.153	26.233	4.2817
FFANet+HazeCLIP	**0.913**	**19.522**	**4.4019**
本篇方式 MSBDN [17]	1.091	27.309	4.3961
MSBDN+HazeCLIP	**0.638**	**18.567**	**4.5510**

C. Ablation Study

To verify the effectiveness of each key component, we conducted a series of ablation experiments with the following variants to the whole framework: (a) without HazeCLIP adaptation; (b) without the region-specific dehazing technique (i.e., one dehazing prompt set applies to the whole image); and (c) without the enhancing prompt set. As shown in Table II, quantitative metrics demonstrate the necessity of the full framework. The pre-trained model performs poorly without HazeCLIP adaptation. Without the region-specific dehazing technique, the dehazing effect is impaired, as evidenced by FADE. Without the enhancing set, the overall image quality is reduced.

D. Framework Generalization

To further verify that HazeCLIP is a general framework capable of fine-tuning a wide range of image dehazing networks, we tested it on two additional popular models besides MSBDN: FFANet [16] and GDN [24]. We compared the performance of these pre-trained models with the same models fine-tuned using our HazeCLIP framework. As shown in Table III, HazeCLIP consistently improved the performance of the three networks across all evaluated metrics.

IV. DISCUSSION

Limitations. During development of HazeCLIP, we also identify two limitations. First, contrastive prompt sets can be constructed through a learned or more systematic approach. Additionally, robust evaluation metrics for real-world image dehazing are lacking, including both reduced-reference and no-reference metrics.

Conclusion. In this paper, we introduce HazeCLIP, an innovative adaptation framework designed to generalize dehazing networks pre-trained on synthetic data to real-world applications. By employing the region-specific dehazing technique and specially designed prompt sets, HazeCLIP accurately detect hazy regions and guide the fine-tuning process with precision. Both subjective evaluations and quantitative metrics demonstrate the superior performance of our proposed approach. HazeCLIP is a versatile fine-tuning framework, compatible with various dehazing networks, highlighting its practical value. Additionally, by modifying the contrastive prompt sets, this method can be extended to other image restoration tasks. We hope that HazeCLIP will inspire new directions in the integration of vision-language models with broader image restoration efforts.

REFERENCES

[1] X. Yin, X. Liu, and H. Liu, "Fmsnet: Underwater image restoration by learning from a synthesized dataset," in *Artificial Neural Networks and Machine Learning – ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part III*. Berlin, Heidelberg: Springer-Verlag, 2021, p. 421–432. [Online]. Available: https://doi.org/10.1007/978-3-030-86365-4_34

[2] K. Fu, Y. Peng, Z. Zhang, Q. Xu, X. Liu, J. Wang, and G. Zhai, "Attentionlut: Attention fusion-based canonical polyadic lut for real-time image enhancement," 2024. [Online]. Available: https://arxiv.org/abs/2401.01569

[3] Q. Qin, K. Chang, M. Huang, and G. Li, "Denet: Detection-driven enhancement network for object detection under adverse weather conditions," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Cham: Springer Nature Switzerland, 2023, pp. 491–507.

[4] S.-C. Huang, T.-H. Le, and D.-W. Jaw, "Dsnet: Joint semantic learning for object detection in inclement weather conditions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 8, pp. 2623–2633, 2021.

[5] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, p. 973–992, Mar. 2018. [Online]. Available: https://doi.org/10.1007/s11263-018-1072-8

[6] C. Sakaridis, D. Dai, S. Hecker, and L. Van Gool, "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[7] M. Hahner, D. Dai, C. Sakaridis, J.-N. Zaech, and L. V. Gool, "Semantic understanding of foggy scenes with purely synthetic data," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 3675–3681.

[8] D. Berman, T. Treibitz, and S. Avidan, "Non-local image dehazing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] R. Fattal, "Dehazing using color-lines," *ACM Trans. Graph.*, 2014.

[10] ——, "Single image dehazing," *ACM Trans. Graph.*, 2008.

[11] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[12] R. T. Tan, "Visibility in bad weather from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[13] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Transactions on Image Processing*, 2015.

[14] S. Narasimhan and S. Nayar, "Vision and the atmosphere," *International Journal of Computer Vision*, vol. 48, pp. 233–254, 07 2002.

[15] 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*, 2019.

[16] 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.

[17] 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 (CVPR)*, 2020.

[18] C. Li, C. Guo, J. Guo, P. Han, H. Fu, and R. Cong, "Pdr-net: Perception-inspired single image dehazing network with refinement," *IEEE Transactions on Multimedia*, 2020.

[19] X. Zhang, H. Dong, J. Pan, C. Zhu, Y. Tai, C. Wang, J. Li, F. Huang, and F. Wang, "Learning to restore hazy video: A new real-world dataset and a new method," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[20] 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*, 2016.

[21] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3194–3203.

[22] W. Dong, H. Zhou, R. Wang, X. Liu, G. Zhai, and J. Chen, "Dehazedct: Towards effective non-homogeneous dehazing via deformable convolutional transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024, pp. 6405–6414.

[23] C. Wang, R. Chen, Y. Lu, Y. Yan, and H. Wang, "Recurrent context aggregation network for single image dehazing," *IEEE Signal Processing Letters*, vol. 28, pp. 419–423, 2021.

[24] 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 (ICCV)*, 2019.

[25] X. Liu, Z. Shi, Z. Wu, J. Chen, and G. Zhai, "Griddehazenet+: An enhanced multi-scale network with intra-task knowledge transfer for single image dehazing," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 1, pp. 870–884, 2023.

[26] Y. Li, H. Chen, Q. Miao, D. Ge, S. Liang, Z. Ma, and B. Zhao, "Image hazing and dehazing: From the viewpoint of two-way image translation with a weakly supervised framework," *IEEE Transactions on Multimedia*, vol. 25, pp. 4704–4717, 2023.

[27] 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 (CVPR)*, 2020.

[28] T. Gao, Y. Liu, P. Cheng, T. Chen, and L. Liu, "Multi-scale density-aware network for single image dehazing," *IEEE Signal Processing Letters*, vol. 30, pp. 1117–1121, 2023.

[29] C. Yan, X. Zhang, X. Wang, G. Jiao, and H. He, "A novel simulation for polarization dehazing," *IEEE Signal Processing Letters*, vol. 31, pp. 341–345, 2024.

[30] 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 (CVPR)*, 2021.

[31] R.-Q. Wu, Z.-P. Duan, C.-L. Guo, Z. Chai, and C. Li, "Ridcp: Revitalizing real image dehazing via high-quality codebook priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[32] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[33] H. Sun, W. Li, J. Liu, H. Chen, R. Pei, X. Zou, Y. Yan, and Y. Yang, "Coser: Bridging image and language for cognitive super-resolution," *arXiv preprint arXiv:2311.16512*, 2023.

[34] Z. Chen, Y. Zhang, J. Gu, X. Yuan, L. Kong, G. Chen, and X. Yang, "Image super-resolution with text prompt diffusion," *arXiv preprint arXiv:2303.06373*, 2023.

[35] B. Zheng, J. Gu, S. Li, and C. Dong, "Lm4lv: A frozen large language model for low-level vision tasks," *arXiv preprint arXiv:2405.15734*, 2024.

[36] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Photo-realistic image restoration in the wild with controlled vision-language models," *arXiv preprint arXiv:2404.09732*, 2024.

[37] ——, "Controlling vision-language models for universal image restoration," *arXiv preprint arXiv:2310.01018*, 2023.

[38] Z. Liang, C. Li, S. Zhou, R. Feng, and C. C. Loy, "Iterative prompt learning for unsupervised backlit image enhancement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 8094–8103.

[39] Y. Li, H. Wang, Y. Duan, and X. Li, "Clip surgery for better explainability with enhancement in open-vocabulary tasks," 2023. [Online]. Available: https://arxiv.org/abs/2304.05653

[40] C.-L. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, and C. Li, "Image dehazing transformer with transmission-aware 3d position embedding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[41] Y. Yang, C. Wang, R. Liu, L. Zhang, X. Guo, and D. Tao, "Self-augmented unpaired image dehazing via density and depth decomposition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[42] Y. Zhang, S. Zhou, and H. Li, "Depth information assisted collaborative mutual promotion network for single image dehazing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 2846–2855.

[43] L. K. Choi, J. You, and A. C. Bovik, "Referenceless prediction of perceptual fog density and perceptual image defogging," *IEEE Transactions on Image Processing*, 2015.

[44] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, 2012.

[45] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, 2018.