FROM FOG TO FAILURE: THE UNINTENDED CONSEQUENCES OF DEHAZING ON OBJECT DETECTION IN CLEAR IMAGES

Ashutosh Kumar
School of Information
Rochester Institute of Technology
Rochester, NY 14623, USA
ak1825@cs.rit.edu

Aman Chadha*
Amazon Gen AI
Santa Clara, CA, USA
hi@aman.ai

ABSTRACT

This study explores the challenges of integrating human visual cue-based dehazing into object detection, given the selective nature of human perception. While human vision adapts dynamically to environmental conditions, computational dehazing does not always enhance detection uniformly. We propose a multi-stage framework where a lightweight detector identifies regions of interest (RoIs), which are then improved via spatial attention-based dehazing before final detection by a heavier model. Though effective in foggy conditions, this approach unexpectedly degrades the performance on clear images. We analyze this phenomenon, investigate possible causes, and offer insights for designing hybrid pipelines that balance enhancement and detection. Our findings highlight the need for selective preprocessing and challenge assumptions about universal benefits from cascading transformations. The implementation of the framework is available here[1].

1 INTRODUCTION

Low-visibility conditions, such as rain, snow, fog, smoke, and haze, pose significant challenges for deep learning applications in autonomous vehicles, security and surveillance, maritime navigation, and agricultural robotics. Object detection models struggle in these environments due to reduced contrast and obscured features, often leading to performance degradation. While image enhancement methods, including dehazing, improve visibility, they can also introduce artifacts or distortions that negatively impact downstream tasks. Overprocessing may lead to false positives and increased computational overhead, highlighting the need for more selective enhancement strategies. Motivated by real-world challenges, such as disruptions in airport operations where poor visibility delays taxiing and docking, this study proposes a vision-inspired deep learning framework tailored for adverse conditions, particularly fog.

To address these challenges, we introduce **Selective Region Enhancement**, a method that focuses on specific regions of interest rather than applying uniform dehazing. This approach reduces processing overhead and prevents unintended degradations that may introduce false positives. Additionally, we propose **Integration with Object Detection**, bridging image enhancement with object detection in a unified pipeline. This integration leverages the strengths of both techniques, overcoming limitations of traditional independent processing models. Our approach draws inspiration from human visual mechanisms, including selective attention, foveal and peripheral vision, adaptive eye responses, bottom-up sensory cues, and top-down goal-driven processing (see Appendix B).

The paper is structured as follows: Section 2 reviews prior work on low-visibility object detection and integration of human visual cues in deep learning applications. Section 3 presents our framework, including its vision-inspired design, data set selection, and experimental setup, followed by results and observed anomalies in Section 4.

*Work done outside position at Amazon.

[1]https://github.com/ashu1069/perceptual-piercing

2 RELATED WORK

Advancements in navigation and detection under low-visibility conditions have leveraged sensor fusion, visual cue integration, and computational techniques. Aircraft landing studies have explored sensor fusion of visible and virtual imagery (Liu et al., 2014) and visual-inertial navigation using runway features (Zhang et al., 2018). Multi-sensor fusion algorithms have improved odometry in GPS-denied environments (Khattak et al., 2019), while research on depth visualization has enhanced navigation and obstacle avoidance (Lieby et al., 2011). Synthetic Vision Systems and full-windshield Head-Up Displays aid drivers and pilots in low visibility (Kramer et al., 2014; Charissis & Papanastasiou, 2010). Image enhancement techniques for low-light conditions (Atom et al., 2020) and the fusion of visual cues with wireless communication improve road safety (Boban et al., 2012). Studies have also emphasized the role of geometrical shapes and colors in driving perception via Head-Up Displays (Zhan et al., 2023).

Despite these advances, challenges persist, including computational complexity (Zhang et al., 2018; Atom et al., 2020; Tang et al., 2022), performance issues under extreme conditions (Khattak et al., 2019; Boban et al., 2012), overfitting due to limited datasets (Zhang et al., 2018; Khattak et al., 2019), and insufficient real-world validation (Liu et al., 2014; Boban et al., 2012; Tang et al., 2022). Some works lack rigorous validation (Kramer et al., 2014; Zhan et al., 2023).

In visual recognition, research has explored human-like processing in computational models. Studies on brain mechanisms highlight hierarchical, feedforward object recognition (DiCarlo et al., 2012), while comparisons with deep neural networks (DNNs) reveal human superiority in handling distortions and attention mechanisms (Dodge & Karam, 2017; van Dyck et al., 2021). Eye-tracking data has been used to guide DNN attention with limited success (van Dyck et al., 2022). Approaches such as adversarial learning for feature discrimination (Yang et al., 2023a), biologically inspired top-down and bottom-up models (Malowany & Guterman, 2020), and retina-mimicking models for dehazing (Zhang et al., 2015) have been proposed. Foveal-peripheral dynamics have also been explored to balance computational efficiency and high-resolution perception (Lukanov et al., 2021).

Recent research has tackled low-visibility challenges like fog, low light, and sandstorms. The YOLOv5s FMG algorithm improves small-target detection with enhanced modules (Zheng et al., 2023), while novel MLP-based networks refine image clarity in hazy and sandstorm conditions (Gao et al., 2023). The PKAL approach integrates adversarial learning and feature priors for robust recognition (Yang et al., 2023b). Deformable convolutions and attention mechanisms enhance pedestrian and vehicle detection in poor visibility (Wu & Gao, 2023). Reviews highlight the limitations of non-learning and meta-heuristic dehazing methods in real-time applications (V et al., 2023), emphasizing the need for integrated low-level and high-level vision techniques (Yang et al., 2020). Innovations such as spatiotemporal attention for video sequences (Zhai & Shah, 2006), the PDE framework for simultaneous detection and enhancement (Li et al., 2022), spatial priors for saliency detection (Jian et al., 2021), and early visual cues for object boundary detection (Mély et al., 2016) further contribute to the field.

Despite advances, existing methods struggle with joint optimization of object detection and image enhancement, detection of low-contrast objects, and adaptation to dynamic visibility changes. This paper addresses these challenges by integrating human visual cues, such as attention mechanisms and contextual understanding, into object detection, enhancing both robustness and efficiency. Traditional approaches process entire images uniformly, increasing computational load, and sometimes degrading clear regions. Our method selectively enhances regions of interest, reducing unnecessary computations and improving responsiveness under varying conditions.

3 METHODOLOGY

The proposed methodology, illustrated in Figure 1, presents a deep learning framework that enhances object detection in low-visibility conditions by leveraging the atmospheric scattering model and human visual cortex principles. It integrates adaptive image enhancement with object detection, optimizing performance through different integration strategies. The pipeline starts with a lightweight detection model to identify regions of interest, guiding spatial attention in the dehazing process. This targeted enhancement preserves critical features while reducing computational overhead. A more robust detection model then refines and improves object recognition.

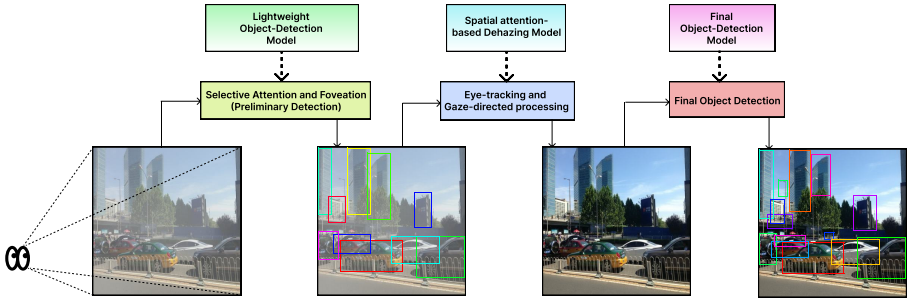


Figure 1: Overall architecture of Perceptual Piercing: (a) Preliminary detection using lightweight object detection model (b) Gaze-directed dehazing using spatial attention on region of interests (c) Final detection using a large and robust model.

For dehazing, we train and evaluate state-of-the-art models, including AOD-Net(Li et al., 2017b), UNet-Dehaze(Zhou et al., 2024), and DehazeNet(Cai et al., 2016), using the Foggy Cityscapes dataset (Sakaridis et al., 2018) (see AppendixA). Table 2 presents a comparative analysis of their dehazing performance, while Figure 3 illustrates their impact on object detection. Among these methods, AOD-Net demonstrated the best dehazing performance, prompting further architectural enhancements. This resulted in AOD-NetX, a spatial attention-enhanced version trained on the Foggy Cityscapes dataset. Detailed modifications to AOD-NetX are provided in Appendix D.4 and Figure 2. For object detection models, we have used pre-trained YOLOv5 and YOLOv8 (see Appendix E).

To ensure robust evaluation, we tested our pipeline on three datasets: Foggy Cityscapes, RESIDE-β, and RESIDE-OTS (see Appendix A) (Li et al., 2019). This comprehensive assessment highlights the adaptability and effectiveness of our approach across multiple low-visibility scenarios.

4 Results and Observed Anomalies

For in-distribution performance on the Foggy Cityscapes dataset, see Appendix F.1, while OOD evaluation on RESIDE-β (RTTS) and OTS datasets is detailed in Appendix F.2, demonstrating the pipeline's robustness across diverse hazy conditions. Evaluation metrics include SSIM and PSNR for dehazing and mAP for object detection (Appendix G). Figure 4 illustrates visibility improvement with AOD-Net, Figure 5 shows its impact on object detection after dehazing with AOD-NetX on Foggy Cityscapes, and Figure 6 presents performance on the RESIDE dataset.

Table 1: Comparison of mean Average Precision (mAP) on clear and foggy conditions for different architecture variants. The performance change column quantifies the relative drop (red) or gain (green) in detection accuracy when transitioning from clear to foggy conditions.

Architecture Variants	**mAP (Clear)**	**mAP (Foggy)**	**Performance Change**
YOLOv5x	0.5644	0.4850	-14.07%
AOD-Net+YOLOv5x	0.6813	0.5822	-14.53%
YOLOv5s+AOD-NetX+YOLOv5x	**0.4896**	**0.6152**	+25.68%
YOLOv8x	0.5243	0.4948	-5.63%
AOD-Net+YOLOv8x	0.6099	0.5900	-3.27%
YOLOv8n+AOD-NetX+YOLOv8x	**0.5150**	**0.6114**	+18.71%

An unexpected finding in our evaluation is the performance trend of models incorporating AOD-NetX. The proposed pipeline performed superior on foggy images and while conventional models like YOLOv5x, YOLOv8x and also their integration with AOD-Net show a natural drop in mAP when transitioning from clear to foggy conditions, architectures integrating AOD-NetX exhibit an inverse trend—performing better under foggy conditions than in clear ones (see Table 1).

Notably, **YOLOv5s+AOD-NetX+YOLOv5x** achieves a 25.68% relative mAP gain in foggy conditions, while **YOLOv8n+AOD-NetX+YOLOv8x** shows an 18.71% relative gain, highlighting the anomalies we found in our experiments.

5 DISCUSSIONS

The unexpected performance improvement of AOD-NetX-based models under foggy conditions (Table 1) raises critical questions about feature adaptation in hybrid object detection pipelines. Typically, object detectors show a decrease in accuracy when transitioning from clear to foggy conditions, as seen in conventional YOLOv5x and YOLOv8x models. However, our proposed AOD-NetX integration leads to an inverse trend, where object detection improves under foggy conditions relative to that under clear conditions. This suggests that AOD-NetX introduces an implicit domain adaptation effect, which makes the detection network more attuned to foggy environments at the cost of generalization to clear images.

One possible explanation lies in bias and overprocessing in the data set. Since AOD-NetX is trained primarily on foggy images, its learned feature space is optimized for haze removal, but lacks the necessary constraints to preserve features in clear conditions. Consequently, when applied to clear images, the model introduces distortions instead of enhancements, disrupting feature consistency for the object detector. This emphasizes the need for context-aware enhancement, where image-processing techniques are selectively applied based on scene conditions rather than indiscriminately.

Furthermore, the results challenge the assumption that cascading pipelines, where lightweight detection informs region-specific enhancement before a final robust detection, always improve performance. While effective in foggy settings, this multi-stage approach appears to introduce trade-offs, potentially harming accuracy in clear conditions. Future designs must strike a balance between specialization for adverse weather conditions and adaptability to diverse environments. Another consideration is the real-time feasibility of this approach. RoI-specific dehazing adds computational overhead, which could limit deployment in time-sensitive applications such as autonomous driving. Optimizing processing efficiency while retaining performance gains remains an open challenge.

6 LIMITATIONS & FUTURE WORK

A fundamental limitation of integrating dehazing into object detection is the feature space misalignment between foggy and clear images. Models trained primarily on foggy conditions lack the ability to preserve the natural characteristics of clear images, leading to unintended alterations that degrade detection performance. This highlights the importance of adaptive enhancement techniques that can determine when dehazing is necessary, rather than applying it universally. A potential solution is the integration of a haze-level estimation module, which could prevent unnecessary processing by triggering dehazing only when haze exceeds a certain threshold (Mao & Phommasak, 2014).

Another challenge is pretraining for scene differentiation. Since the AOD-NetX-enhanced models perform better in foggy conditions, their feature representations may be overfitting to haze-specific characteristics. Introducing joint training on both foggy and clear images could help mitigate this issue by aligning the feature space across different visibility conditions (Huang et al., 2024).

Additionally, unifying dehazing and object detection into a single model rather than having multi-stage framework may yield mutual benefits. For instance, detection-aware dehazing—where dehazing prioritizes regions of interest—could help the model preserve essential features for object detection, enhancing accuracy in both clear and foggy conditions (Fan et al., 2024).

Finally, computational efficiency remains a key concern for real-time applications. While the current pipeline enhances detection performance in low-visibility conditions, its multi-stage nature introduces latency. Future work should focus on optimizing inference speed, exploring lightweight architectures, and developing efficient knowledge distillation techniques to maintain accuracy while reducing processing overhead.

By implementing **adaptive processing strategies**, **improved pretraining**, and **joint optimization**, future object detection pipelines can become more resilient across diverse visibility conditions, ensuring robust performance without compromising clarity in optimal conditions.

REFERENCES

Y. Atom, M. Ye, L. Ren, Y. Tai, and X. Liu. Color-wise attention network for low-light image enhancement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2130–2139, Seattle, WA, USA, 2020. doi: 10.1109/CVPRW50498.2020.00261.

M. Boban, T. T. V. Vinhoza, O. K. Tonguz, and J. Barros. Seeing is believing—enhancing message dissemination in vehicular networks through visual cues. *IEEE Communications Letters*, 16(2): 238–241, February 2012. doi: 10.1109/LCOMM.2011.122211.112093.

Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE transactions on image processing*, 25(11):5187–5198, 2016.

Vassilis Charissis and Stylianos Papanastasiou. Human–machine collaboration through vehicle head up display interface. *Cognition, Technology Work*, 12:41–50, 2010. doi: 10.1007/s10111-008-0117-0.

J. J. DiCarlo, D. Zoccolan, and N. C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012. doi: 10.1016/j.neuron.2012.01.010.

S. Dodge and L. Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, 2017. doi: 10.1109/icccn.2017.8038465.

Yihua Fan, Yongzhen Wang, Mingqiang Wei, Fu Lee Wang, and Haoran Xie. Friendnet: Detection-friendly dehazing network. *arXiv preprint arXiv:2403.04443*, 2024.

Y. Gao, W. Xu, and Y. Lu. Let you see in haze and sandstorm: Two-in-one low-visibility enhancement network. *IEEE Transactions on Instrumentation and Measurement*, 72:1–12, 2023. doi: 10.1109/TIM.2023.3304668.

Xiaochen Huang, Xiaofeng Wang, Qizhi Teng, Xiaohai He, and Honggang Chen. Degradation type-aware image restoration for effective object detection in adverse weather. *Sensors*, 24(19):6330, 2024.

M. Jian, J. Wang, H. Yu, G. Wang, X. Meng, L. Yang, J. Dong, and Y. Yin. Visual saliency detection by integrating spatial position prior of object with background cues. *Expert Systems With Applications*, 168:114219, 2021. doi: 10.1016/j.eswa.2020.114219.

S. Khattak, C. Papachristos, and K. Alexis. Visual-thermal landmarks and inertial fusion for navigation in degraded visual environments. In *2019 IEEE Aerospace Conference*, pp. 1–9, Big Sky, MT, USA, 2019. doi: 10.1109/AERO.2019.8741787.

L. J. Kramer et al. Using vision system technologies to enable operational improvements for low visibility approach and landing operations. In *2014 IEEE/AIAA 33rd Digital Avionics Systems Conference (DASC)*, pp. 2B2–1–2B2–17, Colorado Springs, CO, USA, 2014. doi: 10.1109/DASC.2014.6979422.

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, 2017a.

B. Li et al. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2019. doi: 10.1109/TIP.2018.2867951.

Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *Proceedings of the IEEE international conference on computer vision*, pp. 4770–4778, 2017b.

Zhiying Li, Shuyuan Lin, Zhongming Liang, Yongjia Lei, Zefan Wang, and Hao Chen. Pde: A real-time object detection and enhancing model under low visibility conditions. *International Journal of Advanced Computer Science and Applications(IJACSA)*, 13(12), 2022. doi: 10.14569/IJACSA.2022.0131299.

P. Lieby et al. Substituting depth for intensity and real-time phosphene rendering: Visual navigation under low vision conditions. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 8017–8020, Boston, MA, USA, 2011. doi: 10.1109/IEMBS.2011.6091977.

C. Liu, Q. Zhao, Y. Zhang, and K. Tan. Runway extraction in low visibility conditions based on sensor fusion method. *IEEE Sensors Journal*, 14(6):1980–1987, June 2014. doi: 10.1109/JSEN.2014.2306911.

H. Lukanov, P. König, and G. Pipa. Biologically inspired deep learning model for efficient foveal-peripheral vision. *Frontiers in Computational Neuroscience*, 15, 2021. doi: 10.3389/fncom.2021.746204.

D. Malowany and H. Guterman. Biologically inspired visual system architecture for object recognition in autonomous systems. *Algorithms*, 13(7):167, 2020. doi: 10.3390/a13070167.

Jun Mao and Uthai Phommasak. Detecting foggy images and estimating the haze degree factor. *Journal of Computer Science & Systems Biology*, 7(06), 2014.

D. A. Mély, J. Kim, M. McGill, Y. Guo, and T. Serre. A systematic comparison between visual cues for boundary detection. *Vision Research*, 120:93–107, 2016. doi: 10.1016/j.visres.2015.11.007.

C. Sakaridis, D. Dai, and L. Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. doi: 10.1007/s11263-018-1072-8.

Rong Tang, Qian Li, and Shaoen Tang. Comparison of visual features for image-based visibility detection. *Journal of Atmospheric and Oceanic Technology*, 39, 2022. doi: 10.1175/JTECH-D-21-0170.1.

S. Krishna B V, B. Rajalakshmi, U. Dhammini, M. K. Monika, C. Nethra, and K. Ashok. Image dehazing techniques for vision based applications - a survey. In *2023 International Conference for Advancement in Technology (ICONAT)*, pp. 1–5, Goa, India, 2023. doi: 10.1109/ICONAT57137.2023.10080156.

L. E. van Dyck, R. Kwitt, S. J. Denzler, and W. R. Gruber. Comparing object recognition in humans and deep convolutional neural networks: An eye-tracking study. *Frontiers in Neuroscience*, 15, 2021. doi: 10.3389/fnins.2021.750639.

L. E. van Dyck, S. J. Denzler, and W. R. Gruber. Guiding visual attention in deep convolutional neural networks based on human eye movements. *Frontiers in Neuroscience*, 16, 2022. doi: 10.3389/fnins.2022.975639.

X. Wu and Z. Gao. Based on the improved yolov8 pedestrian and vehicle detection under low-visibility conditions. In *2023 2nd International Conference on Artificial Intelligence and Intelligent Information Processing (AIIIP)*, pp. 297–300, Hangzhou, China, 2023. doi: 10.1109/AIIIP61647.2023.00063.

J. Yang, J. Yang, L. Luo, Y. Wang, S. Wang, and J. Liu. Robust visual recognition in poor visibility conditions: A prior knowledge-guided adversarial learning approach. *Electronics*, 12(17):3711, 2023a. doi: 10.3390/electronics12173711.

J. Yang, J. Yang, L. Luo, Y. Wang, S. Wang, and J. Liu. Robust visual recognition in poor visibility conditions: A prior knowledge-guided adversarial learning approach. *Electronics*, 12:3711, 2023b. doi: 10.3390/electronics12173711.

W. Yang et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020. doi: 10.1109/TIP.2020.2981922.

Yun Zhai and Mubarak Shah. Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of the 14th ACM international conference on Multimedia (MM '06)*, pp. 815–824, New York, NY, USA, 2006. Association for Computing Machinery. doi: 10.1145/1180639.1180824.

Yu-Wei Zhan, Fan Liu, Xin Luo, Liqiang Nie, Xin-Shun Xu, and Mohan Kankanhalli. Generating human-centric visual cues for human-object interaction detection via large vision-language models. *arXiv*, 2023. arXiv:2311.16475.

H. Zhang and V.M. Patel. Densely connected pyramid dehazing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3194–3203, 2018.

L. Zhang, Z. Zhai, L. Bai, Y. Li, W. Niu, and L. Yuan. Visual-inertial state estimation for the civil aircraft landing in low visibility conditions. In *2018 International Conference on Networking and Network Applications (NaNA)*, pp. 292–297, Xi'an, China, 2018. doi: 10.1109/NANA.2018.8648726.

X.-S. Zhang, S.-B. Gao, C.-Y. Li, and Y.-J. Li. A retina inspired model for enhancing visibility of hazy images. *Frontiers in Computational Neuroscience*, 9, 2015. doi: 10.3389/fncom.2015.00151.

Y. Zheng, Y. Zhan, X. Huang, and G. Ji. Yolov5s fmg: An improved small target detection algorithm based on yolov5 in low visibility. *IEEE Access*, 11:75782–75793, 2023. doi: 10.1109/ACCESS.2023.3297218.

Hao Zhou, Zekai Chen, Qiao Li, and Tao Tao. Dehaze-unet: A lightweight network based on unet for single-image dehazing. *Electronics*, 13(11):2082, 2024.

A DATASETS

A.1 FOGGY CITYSCAPES

The Foggy Cityscapes dataset (Sakaridis et al., 2018) was developed to tackle the challenge of semantic foggy scene understanding (SFSU). While significant research has been conducted on image dehazing and semantic scene understanding for clear-weather images, SFSU remains relatively underexplored. Due to the challenges associated with collecting and annotating real-world foggy images, synthetic fog is introduced into clear-weather outdoor scenes. This synthetic fog generation process utilizes incomplete depth information to simulate realistic foggy conditions on images from the Cityscapes dataset, resulting in a dataset comprising 20,550 images. The dataset is divided into a training set of 2,975 images, a validation set of 500 images, and a test set of 1,525 images. Key characteristics of the dataset include:

- **Synthetic Fog Generation**: Synthetic fog is added to real clear-weather images using a dedicated pipeline that incorporates the transmission map.
- **Data Utilization**: The dataset supports both supervised and semi-supervised learning. A synthetic foggy dataset was generated using the synthetic transmission map, followed by supervised learning on the resulting foggy images.

A.2 RESIDE-β

The RESIDE-β Outdoor Training Set (OTS) is a comprehensive dataset curated to support research in outdoor image dehazing. It addresses the degradation caused by haze in outdoor scenes, which negatively impacts image quality and downstream tasks such as object detection and semantic segmentation. The dataset contains approximately 72,135 outdoor images with varying haze intensities, allowing for robust training of dehazing algorithms. For evaluation, we use the RESIDE-β (REalistic Single Image DEhazing) dataset (Li et al., 2019). A subset of RESIDE-β , the Real-Time Testing Set (RTTS), comprises 4,322 real-world hazy images with object detection annotations. The dataset is split into a training set of 3,000 images, a validation set of 500 images, and a test set of 1,500 images.

B HUMAN VISUAL CUES

Selective Attention and Foveation: The human eye does not perceive all areas of the visual field with equal clarity. Foveal vision, which corresponds to central vision, is highly detailed and is essential for tasks such as reading and object recognition. In contrast, peripheral vision is less detailed but more sensitive to motion. The visual system initially scans the entire scene using peripheral vision, akin to the preliminary detection phase in our approach. This broad scanning process helps identify regions requiring closer inspection, enabling a more detailed analysis through foveal vision. Similarly, the proposed method does not process every detail uniformly but prioritizes key areas of interest.

Adaptation to Environmental Conditions: The human visual system dynamically adjusts to varying lighting conditions and levels of visibility, such as adapting from bright sunlight to a dark room. Similarly, the adaptive dehazing method modulates its processing intensity and focus based on detection feedback and environmental context. This mechanism ensures optimal perception, mirroring the way human vision adapts to maintain clarity under diverse conditions.

Eye Tracking and Gaze-Directed Processing: Eye-tracking technology monitors gaze direction and identifies focal points of attention. This concept translates to strategically allocating resources toward regions of interest in computational visual processing. The proposed method follows a similar principle by directing dehazing and detailed object detection efforts to areas where objects are likely to be present. Just as human vision selectively fixates on specific regions when searching for an object, the system prioritizes certain parts of the image to enhance clarity and detection performance.

Integration of Bottom-Up and Top-Down Processes: Human vision combines bottom-up processing, driven by sensory input, with top-down processing, guided by prior knowledge, expectations,

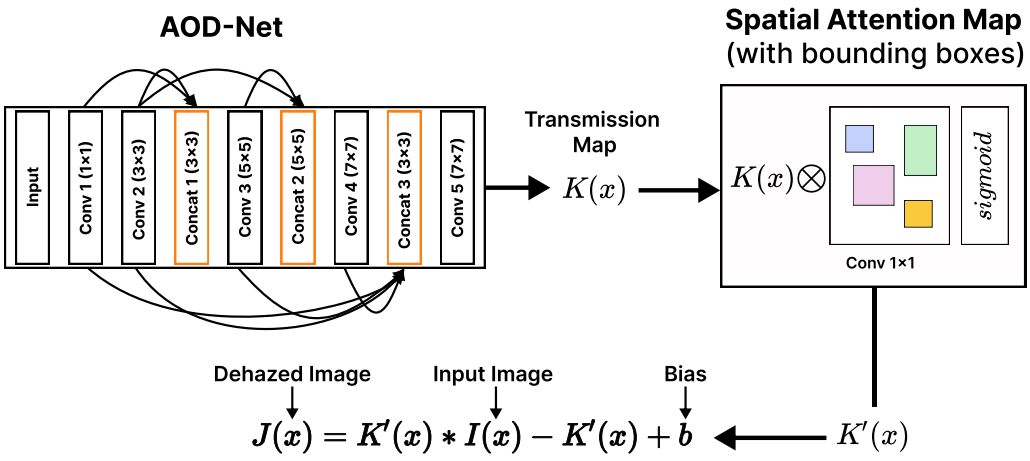


Figure 2: Architecture of AOD-NetX: The model takes the transmission map output, $K(x)$, from AOD-Net and applies a spatial attention layer to emphasize key regions of interest (bounding boxes) in the input image. The refined transmission map, $K'(x)$, is then utilized to dehaze the image.

and goals. The proposed model adopts a similar dual approach: it first employs a bottom-up strategy, where object detection algorithms identify potential areas of interest. This is followed by a top-down refinement process, where dehazing efforts are concentrated on flagged areas, leveraging previous learning. This interplay between data-driven signals and cognitive insights aligns with the way human perception integrates sensory input with contextual understanding.

C METHODOLOGY

Preliminary Detection: A lightweight and fast object detection algorithm, such as YOLOv5s or YOLOv8n, is employed to rapidly scan the image and identify potential regions of interest or active regions. These models flag image patches with a high probability of containing objects. While the smaller variants of YOLO models offer lower accuracy compared to their full-sized counterparts, they are significantly faster, making them well-suited for this initial detection phase.

Region-Based Dehazing: Dehazing algorithms are selectively applied to the active regions identified during the preliminary detection phase. The approach dynamically adjusts based on the depth or severity of haze within the detected regions, ensuring an adaptive and efficient dehazing process.

The proposed architecture, **AOD-NetX**, illustrated in Figure 2, builds upon the transmission map generated by the standard AOD-Net (Li et al., 2017a). This transmission map is integrated into a spatial attention map module, producing an attention-enhanced transmission map. The spatial attention map is derived from the bounding boxes or Regions of Interest (ROIs) detected by the lightweight object detection model (YOLOv5s/YOLOv8n) within the proposed framework. A sigmoid layer is applied to map the output probabilities to a range between 0 and 1. Unlike softmax, which normalizes outputs across multiple regions, sigmoid is preferred in this context since each bounding box holds independent significance.

D DEHAZING MODELS

D.1 AOD-NET

AOD-Net (All-in-One Dehazing Network) is a convolutional neural network (CNN) designed for haze removal by directly reconstructing the clean image in an end-to-end manner. Unlike traditional approaches that separately estimate transmission maps and atmospheric light, AOD-Net is based on a re-formulated atmospheric scattering model, allowing it to generate dehazed images without

intermediate computations. This lightweight architecture delivers superior performance in terms of Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) while also enhancing visual quality. Additionally, its modular design enables seamless integration into other deep learning models, such as Faster R-CNN, thereby improving object detection performance in hazy conditions.

D.2 DEHAZE-UNET

The Dehaze-UNet model employs an encoder-decoder architecture with skip connections to effectively restore dehazed images. The encoder reduces spatial dimensions through successive convolutional layers, extracting essential features while maintaining structural integrity. Each convolutional layer is followed by batch normalization and ReLU activation, enhancing feature representation. The decoder then upscales these features, reintegrating spatial details from the encoder via skip connections, which are critical for preserving fine-grained textures lost during downsampling. The final stage consists of double convolution layers in the decoder for feature refinement, followed by Max Pooling and Bilinear Interpolation to optimize feature representation and ensure smooth transitions in the reconstructed image. This makes U-Net particularly effective for dehazing tasks.

D.3 DEHAZENET

DehazeNet follows a structured four-stage approach, as illustrated in Figure 5 (Zhang & Patel, 2018). The first stage, feature extraction, employs 16 filters combined with four MaxOut layers, which use max pooling to reduce data dimensionality. The second stage, multi-scale mapping, captures both fine-grained and high-level features to ensure a comprehensive feature representation. The third stage, local extremum, utilizes a specialized max pooling operation to enhance spatial invariance while preserving image resolution. Finally, the non-linear regression stage incorporates Bilateral ReLU as the activation function, which constrains the output within a defined range to prevent oversaturation and maintain image clarity.

D.4 AOD-NETX

The proposed **AOD-NetX** architecture, depicted in Figure 2, extends AOD-Net by leveraging its transmission map within a spatial attention map module to generate an attention-focused transmission map. This spatial attention map is derived from the bounding boxes or Regions of Interest (ROIs) identified by the lightweight object detection model (YOLOv5s) within our framework. A sigmoid activation layer is applied to map the output probabilities to a range between 0 and 1. Unlike softmax, which normalizes outputs across multiple regions, sigmoid is preferred as each bounding box holds independent significance.

E OBJECT DETECTION MODELS

The detection pipeline incorporates various YOLO models, each optimized for specific applications. YOLOv5s is a lightweight variant designed for real-time detection with minimal computational overhead, while YOLOv8n (Nano) is optimized for high-speed processing on resource-constrained devices, such as mobile phones. In contrast, YOLOv5x, with its CSP backbone and advanced data augmentation techniques, delivers enhanced performance for more complex scenes, whereas YOLOv8x (Extra Large) achieves maximum accuracy when handling large-scale datasets.

The detection workflow begins by applying YOLOv5s or YOLOv8n to foggy images to generate initial object annotations. These annotations, along with the original image, undergo dehazing using AOD-NetX. The resulting dehazed image is then processed with YOLOv5x or YOLOv8x, ensuring precise and refined detection outcomes.

F ADDITIONAL RESULTS

The dehazing modules are trained independently on the provided datasets, while the object detection models (various YOLO versions) remain pre-trained on the MS-COCO dataset. This modular approach allows seamless integration of the dehazing module into existing detection pipelines without

Table 2: Average Loss and SSIM scores of Dehazing techniques on Foggy Cityscapes dataset

Dehazing Model	Average Loss	SSIM
AOD-Net	0.0468	0.994
UNet-Dehaze	0.0323	0.992
DehazeNet	0.0572	0.991

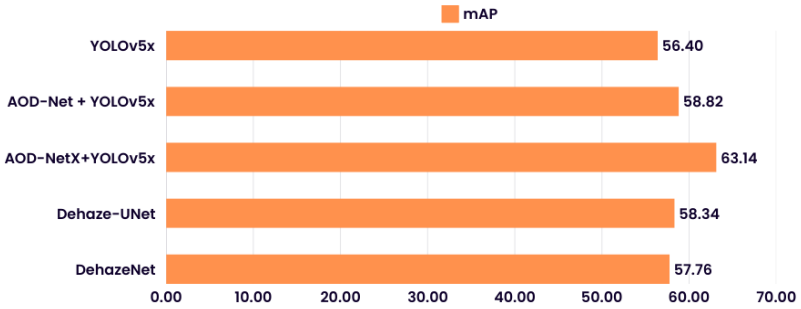


Figure 3: Comparison of mean Average Precision (mAP) for different dehazing and object detection module combinations.

requiring full retraining. However, fine-tuning the entire architecture on target datasets could yield further performance improvements, making it a promising direction for future ablation studies.

Table 3 presents the comparative results, demonstrating that AOD-NetX generally outperforms the standard AOD-Net in terms of SSIM and PSNR across most datasets. For Foggy Cityscapes and RESIDE-β OTS, AOD-NetX achieves higher SSIM and PSNR values, indicating superior structural similarity and signal fidelity. However, in the case of RESIDE-β RTTS, while AOD-NetX attains a slightly higher PSNR, AOD-Net exhibits a significantly higher SSIM score, suggesting better structural detail retention in this specific dataset. Overall, AOD-NetX proves more effective in most scenarios, particularly under complex foggy conditions.

F.1 In-Distribution Performance of Perceptual Piercing

The evaluation results of Perceptual Piercing variations, trained and tested on the Foggy Cityscapes dataset, are presented in Table 4. The integration of dehazing modules, such as AOD-Net and AOD-NetX (detailed in Appendix C), consistently enhances object detection in both clear and foggy conditions.

Among the tested variants, the AOD-Net + YOLOv5x configuration achieved the highest mAP under clear conditions (0.6813). In foggy conditions, YOLOv5s + AOD-NetX + YOLOv5x and YOLOv8n + AOD-NetX + YOLOv8x demonstrated the best performance, with mAP scores of 0.6152 and 0.6114, respectively. In contrast, baseline YOLO models (YOLOv5x and YOLOv8x) exhibited lower detection accuracy, highlighting the effectiveness of advanced dehazing techniques in low-visibility environments.

F.2 Out-of-Distribution Performance of Perceptual Piercing

The evaluation results in Table 5, where Perceptual Piercing variations were trained on Foggy Cityscapes and tested on the RESIDE-β OTS and RTTS datasets, highlight key performance trends. The YOLOv8x architecture achieved the highest mAP scores under foggy conditions, with 0.7125 on OTS and 0.6978 on RTTS. Among the YOLOv5 variants, the baseline YOLOv5x model performed best, achieving 0.6944 on OTS and 0.6655 on RTTS.

The addition of AOD-Net generally enhanced performance for YOLOv8 but had a diminishing effect on YOLOv5. Meanwhile, models incorporating AOD-NetX exhibited lower mAP values across

Table 3: Performance of dehazing methods: AOD-Net and AOD-NetX

Dataset	Dehazing Method	Evaluation Metrics	
		SSIM	PSNR
Foggy Cityscapes	AOD-Net	0.994	26.74
	AOD-NetX	**0.998**	**27.22**
RESIDE-β OTS	AOD-Net	0.920	24.14
	AOD-NetX	**0.945**	**25.80**
RESIDE-β RTTS	AOD-Net	**0.932**	27.59
	AOD-NetX	0.656	**27.62**

Table 4: **Train**- Foggy Cityscapes, **Test**- Foggy Cityscapes: Evaluation of various Perceptual Piercing variations based on mean Average Precision (mAP) under both clear and foggy conditions.

Architecture Variants	Conditions	Evaluation Metrics (mAP)
YOLOv5x	Clear	0.5644
	Foggy	0.485
AOD-Net+YOLOv5x	Clear	0.6813
	Foggy	0.5822
YOLOv5s+AOD-NetX+YOLOv5x	Clear	0.4896
	Foggy	0.6152
YOLOv8x	Clear	0.5243
	Foggy	0.4948
AOD-Net+YOLOv8x	Clear	0.6099
	Foggy	0.5900
YOLOv8n+AOD-NetX+YOLOv8x	Clear	0.5150
	Foggy	0.6114

both test datasets, suggesting that its integration may require further optimization. Overall, the results indicate that YOLOv8x is more robust in handling foggy conditions compared to other model variations.

(a) Foggy Cityscapes: Before Dehazing

(b) Foggy Cityscapes: After Dehazing (using AOD-NetX)

Figure 4: Dehazing performance on Foggy Cityscapes dataset.

G Evaluation Metrics

G.1 Structural Similarity Index Measure (SSIM)

The performance of dehazing methods is evaluated using the Structural Similarity Index Measure (SSIM), which quantifies the similarity between two images based on luminance, contrast, and structural components. It is defined as:

Table 5: **Train**- Foggy Cityscapes, **Test**- RESIDE-β OTS and RTTS: Evaluation of various Perceptual Piercing variations based on mean Average Precision (mAP) under foggy conditions.

Architecture Variants	Configuration	Evaluation Metrics (mAP)
YOLOv5x	Test: OTS	0.6944
	Test: RTTS	0.6655
AOD-Net+YOLOv5x	Test: OTS	0.6325
	Test: RTTS	0.6156
YOLOv5s+AOD-NetX+YOLOv5x	Test: OTS	0.5679
	Test: RTTS	0.5297
YOLOv8x	Test: OTS	0.7125
	Test: RTTS	0.6978
AOD-Net+YOLOv8x	Test: OTS	0.6458
	Test: RTTS	0.6125
YOLOv8n+AOD-NetX+YOLOv8x	Test: OTS	0.5779
	Test: RTTS	0.5312

(a) Foggy Cityscapes: Before Dehazing

(b) Foggy Cityscapes: After Dehazing (using AOD-NetX)

Figure 5: Dehazing performance on Foggy Cityscapes dataset.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{1}$$

where:

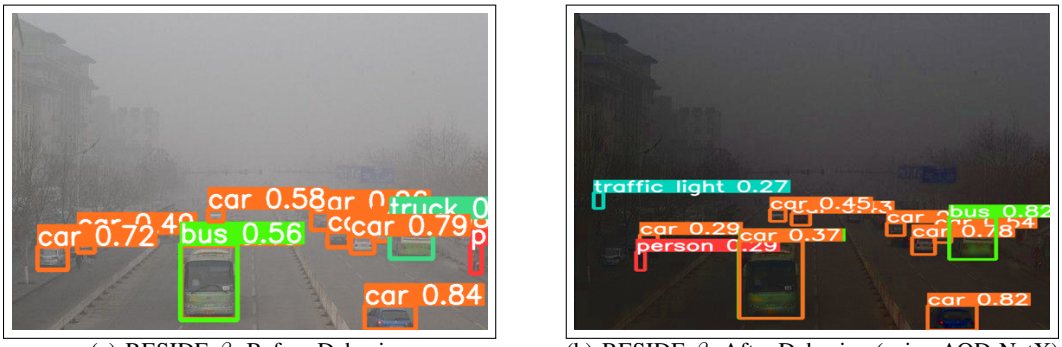
- μ_x and μ_y are the mean intensities of images x and y, respectively.
- σ_x^2 and σ_y^2 denote the variances of x and y.
- σ_{xy} represents the covariance between x and y.
- $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are stabilizing constants to prevent division by zero, where L is the dynamic range of pixel values (e.g., 255 for 8-bit images), and default values are $k_1 = 0.01$ and $k_2 = 0.03$.

G.2 Peak Signal-to-Noise Ratio (PSNR)

Peak Signal-to-Noise Ratio (PSNR) is a widely used metric to assess image reconstruction quality by comparing the original and processed images. It is expressed in decibels (dB) and is calculated as:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right), \tag{2}$$

where MAX is the maximum possible pixel value (e.g., 255 for 8-bit images), and MSE is the Mean Squared Error:



(a) RESIDE-β: Before Dehazing (b) RESIDE-β: After Dehazing (using AOD-NetX)

Figure 6: Dehazing performance on RESIDE-β dataset.

$$\text{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \left(I(i,j) - K(i,j) \right)^2 . \tag{3}$$

Here, $I(i,j)$ and $K(i,j)$ represent pixel values at position (i,j) in the original and reconstructed images, respectively. A higher PSNR value indicates better image quality, as it corresponds to lower distortion. PSNR is extensively used in evaluating dehazing, denoising, and image compression methods.

G.3 Mean Average Precision (mAP)

For object detection performance, we use mean Average Precision (mAP), which evaluates the precision-recall tradeoff. The Average Precision (AP) is computed as:

$$AP = \frac{\sum_{k=1}^{n} (P(k) \times \text{rel}(k))}{\text{number of relevant objects}} \tag{4}$$

where:

- $P(k)$ is the precision at rank k.
- $\text{rel}(k)$ is an indicator function, which is 1 if the object at rank k is relevant, and 0 otherwise.
- n is the total number of retrieved objects.

The mean Average Precision is computed as:

$$mAP = \frac{\sum_{q=1}^{Q} AP_q}{Q} \tag{5}$$

where AP_q is the Average Precision for the qth query, and Q is the total number of queries. Higher mAP values indicate better object detection performance across different classes.