Pamukkale Univ Muh Bilim Derg, 31(3), 373-383, 2025

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Not all fog removers are equal: Unmasking the impact of dehazing on object detection

Tüm sis gidericiler aynı değildir: Sis gideriminin nesne tespiti üzerindeki etkisinin ortaya çıkarılması

Ahmet Selman Bozkır[1*], *Nurçiçek Özenç*[2]

[1]Department of Computer Engineering, Faculty of Engineering, Hacettepe University, Ankara, Türkiye.
selman@cs.hacettepe.edu.tr

[2]Data and Information Engineering, Institute of Informatics, Hacettepe University, Ankara, Türkiye.
nurcicekozenc@gmail.com

Received/Geliş Tarihi: 24.02.2024 Revision/Düzeltme Tarihi: 06.07.2024 doi: 10.5505/pajes.2024.05784
Accepted/Kabul Tarihi: 20.08.2024 Research Article/Araştırma Makalesi

Abstract

Dehazing is an important branch of computational photography aiming to enhancing image clarity by removing atmospheric haze and scattering effects, crucial for improving visibility in applications such as unmanned aerial vehicles, traffic control, and autonomous driving. However, most of the studies in this particular field lack an assessment of the developed algorithm in context of object detection (OD). In this study, we aim to quantify and evaluate the contribution of several state-of-the-art dehazing methods (C2PNet, D4, Dehamer, gUNet) on OD using YOLOv8, known for its superior performance. For this purpose, we utilized the test portion of the VisDrone-DET dataset including 548 haze-free aerial images as the data source. For a more comprehensive assessment, we evaluated these approaches to object detection under different haze levels and resolutions. Since it is inherently impossible to obtain hazy and clean images simultaneously, we (1) generated synthetically hazed images involving varying haze densities and (2) resized to 640p and 1280p resolutions. Next, we used YOLO8 and YOLO10 models to evaluate the OD performance in (i) haze-free ground truth, (ii) three different hazed versions, and (iii) their dehazed counterparts through several metrics. Our experiments showed that the gUNET approach, incorporating a variant of the U-Net model inspired by GCANet and GridDehazeNet outperformed the others in terms of OD performance. Surprisingly, the Dehamer negatively affected the OD performance due to the artifacts it produced. This assessment not only provides valuable findings into the effectiveness of these methods but also sheds light on how to benefit them when it comes to object detection under hazy atmospheric conditions.

Keywords: Object detection, YOLO, Image dehazing, Synthetic haze.

Öz

Sis giderimi insansız hava araçları, trafik kontrolü ve otonom sürüş gibi uygulamalarda hayati önemdeki görünürlüğü iyileştirmek amacıyla atmosferik pus ve saçılım etkilerini ortadan kaldırmayı hedefleyen hesaplamalı fotografinin önemli bir dalıdır. Ancak bu alandaki çalışmaların birçoğu geliştirilen algoritmanın nesne tespiti (NT) bağlamında değerlendirilmesinden yoksundur. Bu çalışmada üstün performansıyla bilinen YOLOv8 üzerinden son teknoloji ürünü çeşitli sis giderici yöntemlerin (C2PNet, D4, Dehamer, gUNet) katkısının NT bağlamında ölçülmesi ve değerlendirilmesini amaçlanmıştır. Bu amaçla veri kaynağı olarak VisDrone-DET veri kümesindeki 548 sissiz gökyüzü görüntüsü içeren test kısmından faydalandık. Daha kapsamlı bir değerlendirme için farklı sis seviyeleri ve çözünürlükler altında NT bağlamında bu yaklaşımları değerlendirdik. Sisli ve temiz imgeleri doğal olarak aynı anda elde etmek mümkün olmadığındavn, (1) değişen sis yoğunlukları içeren sentetik sisli imgeler oluşturduk ve (2) 640p ve 1280p çözünürlüklerinde yeniden boyutlandırdık. Ardından (i) sissiz kesin referans, (ii) üç farklı sislendirilmiş sürüm ve (iii) bunların sisi giderilmiş muadillerinde YOLO8 ve YOLO10 modelini kullanarak NT performansını çeşitli ölçütler üzerinden değerlendirdik. Deneylerimiz GCANet ile GridDehazeNet'ten esinlenen ve U-Net modelinin bir varyantını içeren gUNET yaklaşımının NT performansı açısından diğerlerinden daha iyi başarım gösterdiğini ortaya koymuştur. Dehamer yöntemi saşırtıcı şekilde üretilen "artifakt" nedeniyle NT başarımını olumsuz etkilemiştir. Bu değerlendirme ilgili yöntemlerin etkinliği hakkında değerli bulgular sunmakla kalmayarak sisli hava koşullarında NT söz konusu olduğunda bu yöntemlerden nasıl faydalanılacağına da ışık tutmaktadır.

Anahtar kelimeler: Nesne tespiti, YOLO, İmge sis giderimi, Sentetik sis.

1 Introduction

Haze is a phenomenon resulting from the scattering of aerosol particles into the atmosphere, posing a significant challenge to image quality. This circumstance encourages the development of clarity-focused haze removal techniques necessary for tasks like segmentation and object detection in computer vision under hazy weather conditions [1]. In addition to all these complexities, successfully navigating through such challenging tasks is crucial. The intricacies of outdoor scenes, characterized by low visibility and color shift, compound the challenges of haze removal, transforming it into a multifaceted restoration issue [2]. This complexity not only influences real-time object detection, encompassing localization and classification [3], but also introduces instability in modern detectors where factors such as noise, blurriness, and vibration significantly hinder object detection performance [4],[5]. Despite advancements in the field, a persistent obstacle in haze removal studies is the impractical creation of datasets reflecting both hazy and clean images. This challenge persists even under constant environmental conditions, emphasizing the necessity for a cost-effective restoration of an image in the absence of a ground-truth reference.

*Corresponding author/Yazışılan Yazar

Moreover, successfully navigating through all these intricacies is critical in tasks like object detection, which are challenging. Object detection plays a crucial role as a vital preprocessing module in various applications, including surveillance for pedestrian detection, person re-identification, and tracking. Additionally, it enhances the robustness of autonomous driving systems and the reliability of high-level challenging computer vision tasks [6]-[8]. In defense systems, military object detection is not only deemed essential but also presents unique challenges [9]. A notable lack in the literature is the inadequate consideration of the success of dehazing approaches in different downstream tasks like object detection. Image Quality Assessment (IQA) metrics, relying on objective criteria, commonly employ full-reference metrics like PSNR and SSIM, and occasionally MSE, which necessitate a ground truth image to calculate the difference or error from the target image [10]. However, it has been observed that these metrics inadequately characterize both human perceptual quality and effectiveness in computer vision problems [2].

In this study, the impact and possible contributions of several state of the art haze removal methods including C2PNet [11], D4 [12], Dehamer [13], and gUNet [7] on object detection is systematically investigated by employing (a) the VisDrone dataset since it involves aerial drone images having a varying range of object sizes, (b) fined YOLO8 and YOLO10 small models which were trained on VisDrone dataset, (c) a depth aware synthetic haze generator. Our comprehensive experiments report the results by also shedding light on the impact of these methods. In addition, we discuss whether any dehazing mechanism should be coupled with an object detection method without performing a prior test on OD.

The rest of the article is organized as follows; Section 2 introduces the literature review whereas Section 3 describes the methodological aspects involving the used dataset and approaches. Next, Section 4 explains our evaluation approach while Section 5 presents our results and discusses the findings. Finally, Section 6 concludes the paper.

2 Related work

Studies related to haze removal in the literature can be grouped in two primary categories: (a) early/conventional and (b) contemporary convolutional neural network (CNN) based deep learning approaches. Understanding the evolution of these approaches is crucial for the development of applications such as haze removal. Popular studies in the literature on this subject are briefly summarized in this section.

The early works focused on exploring information obtained from statistical analysis and observations [2]. These studies primarily concentrate on the direct prediction of the transmission map and atmospheric light. However, due to uncertainties in the prediction of the transmission map and overall atmospheric light or certain prediction biases, these methods can lead to restoration errors and significant reconstruction errors between hazy and clear images. It has been observed that transferring information from a clean image to a dehazing network may lead to a kind of cumulative error [14]. For instance, the Dark Channel Prior (DCP) yields invalid results, especially when the images being worked on are similar to the atmospheric light and are free of shadows because it measures the intensity of pixels in color channels in outdoor images and takes this measure as a kind of statistic [15]. Another classic method, non-local image dehazing approach based on the assumption that haze-free images exhibit distinct color clusters in RGB space, allowing the algorithm to recover both distance maps and haze-free images efficiently. Method fail in scenes where the airlight is significantly brighter than the scene. In such cases, most pixels will point in the same direction and it will be difficult to detect the haze lines [16].

On contrary of classic methods, learning-based models involve directly learning the transformation from a hazy image to a haze-free image. These models are data-driven and typically use deep neural networks within the physical scattering model to predict the transmission map and atmospheric light, representing latent images that capture the data [2], [17]. CNNs that have achieved universal success have played a significant role in computer vision tasks and have recently found applications in haze removal methods. Instances of learning-based models in this field include DehazeNet, AOD-Net, Multi-Scale CNN, Feature Pyramid Network (FPN), FFANet, and Transformer architectures, which emerge as a trainable end-to-end model specifically for medium transmission predictions.

Image dehazing methods, particularly DehazeNet and Multi-Scale CNNs, are acknowledged pioneers in the realm of learning-based approaches [7]. DehazeNet introduces a trainable model aiming to predict the transmission matrix from hazy images, utilizing a multi-scale CNN (MSCNN) that initially generates a coarse-scale transmission matrix and progressively refines it. Despite the success of Multi-Scale CNNs in object detection by leveraging multi-scale features, their drawbacks include high computational demands, extended processing times, and susceptibility to performance degradation in scenarios with unresolved scale inconsistencies [18].

AOD-Net and DehazeNet may exhibit a preference for 'under-dehazed' images, potentially sacrificing details and suffering from method artifacts. Furthermore, these methods might struggle with generalization to real-world hazy images [2]. Notably, AOD-Net exhibits very fast processing exhibiting a suitable solution for video dehazing.

FFANet introduces feature attention (FA) blocks that enhance haze removal using both channel and pixel attention mechanisms [14]. Despite the small computational cost of the channel attention module, the introduced parameters and delay are non-negligible [7].

In the realm of Transformer methods, it has been observed that while they combine long-term and local attention in CNN features, they tend to overlook the physical properties of the haze generation process [14]. As is known, vision transformers have recently outperformed most CNN architectures in high-level vision tasks, employing a flat Transformer architecture. Additionally, pioneering features in haze removal have demonstrated success in eliminating non-uniform haze in images [7].

3 The methodology

3.1 VisDrone dataset

VisDrone dataset was created to serve the drone-based computer vision research community at the Machine Learning and Data Mining Laboratory of Tianjin University in China. In total, dataset includes carefully annotated ground truth data consist of 288 video clips formed by 261908 frames and 10209 static images, captured by various drone-mounted cameras [19]. In our study we have employed a portion of validation split of this dataset containing 548 outdoor images and 38759 bounding boxes [20].

The images were processes at resolutions of 640p and 1280p in YOLO8 and YOLO10 small models. The images in our study cover 10 different object classes. To provide a more detailed analysis, synthetic hazy images were generated by adding different haze densities (for ß = 1.0, ß = 1.5, and ß = 2.0) to the same images through MonoDepth [21]. As shown in Table 1 below, both clean and hazy images were produced at resolutions of 640p and 1280p, with hazy images having distinct haze densities from clean images. Subsequently, using four different methods, namely C2PNet, D4, Dehazer, and gUNET, dehazing was performed on these images having varying haze densities and resolutions. The dataset was prepared in this manner to facilitate the comparison of clean, hazy, and dehazed images during the evaluation stage, allowing for the measurement of object detection performance.

Table 1. Data set obtained after the applied resolution and haze addition processes

		640p	1280p
Clean Images		548	548
Hazy Images	ß = 1.0	548	548
	ß = 1.5	548	548
	ß = 2.0	548	548

3.2 YOLOv8 and YOLOv10

YOLOv8 incorporates various architectural and developer experience changes compared to YOLOv5, aiming to make the application process more accessible, attracting interest from a broader audience [23],[24].

YOLOv8 stands out due to its exceptional speed compared to other real-time detectors [26], the ability of its model architecture to directly and rapidly provide the position and class of the bounding box during object detection [27], and its function to minimize errors by globally processing the entire image during prediction [26]. In addition to these advantages, a key reason for its preference is that YOLOv8 is a fast and single-shot model for real-time object detection, in contrast to models that perform separate processes for each class, allowing for effective use in multi-class detection [28].

YOLOv8 involves 53 convolutional layers exhibiting cross-state partial connections to enable communication among different layers. In addition, it involves the self-attention feature which helps focus and combine relevant features. The incorporated feature pyramid network [29] enables YOLO8 to detect varying-sized objects more accurately in a multi-scale fashion. It should be noted that along with being an anchor-free approach, YOLO8 leverages mosaic augmentation for better generalization capability in cluttered and complex scenes.

On the other hand, the most recent version of YOLO series, namely YOLO10 brings new features such as elimination of the need for Non-Maximum-Supression (NMS) during training and an improved architecture consisting of (a) Cross-Stage-Partial Network for feature extraction, (b) Path Aggregation Network for multi-scale feature fusion and (c) *one-to-one* and *one-to-many* head design [22]. By avoiding NMS, YOLOv10 reduces inference latency, which can be especially beneficial for real-time applications. In this study, we have employed two different YOLO algorithms in order to understand whether new improvements are beneficial for dehazed images.

3.3 Domain shift problem and creating synthetic hazy images with monodepth

The purpose of dehazing is to recover a haze-free image from a hazy one [7]. In computer vision, haze in images typically degrades the quality of detected images. This phenomenon affects the reliability of models in high-level vision tasks and transforms image dehazing into a meaningful low-level vision task [6]. Researchers often use atmospheric scattering models for this purpose [7]. However, the decrease in object recognition performance is associated with a phenomenon known as domain shift. This issue arises when there is a separation between training and test datasets, leading to differences in metric results. Specifically, it occurs when the model is trained on a dataset that does not adequately represent the conditions encountered during testing. For example, variations in environmental factors such as good weather conditions during training and poor weather conditions during testing can result in different performance outcomes. The domain shift problem underscores the difficulty of ensuring the robustness and generalization ability of the model across different real-world scenarios. Addressing this issue is important for improving the reliability and applicability of object detection systems under different environmental conditions, including those affected by haze [4],[30].

The Atmospheric Scattering Model, shown in Figure 1 above, is also known as Koschmieder's law or haze model. [1],[4],[7],[8],[14].

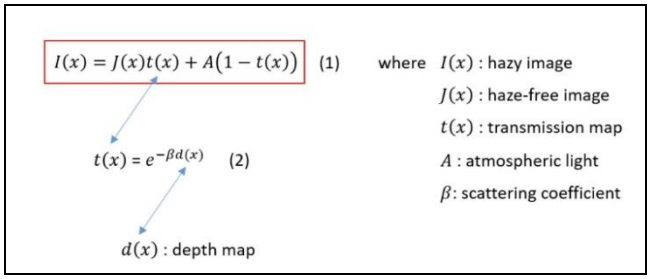

Figure1. Atmospheric scattering model based Koschmieder's law. Adopted from [4].

$$I(x) = J(x)t(x) + A(1 - t(x)) \quad (1)$$

Eq. (1) above illustrates that x denotes the pixel position, $I(x)$ is the intended hazy image, $J(x)$ is the existing clean image. Here, A represents the atmospheric light, typically set to 1 whereas $t(x)$ denotes the transmission matrix. To derive $I(x)$, knowledge of $t(x)$ is necessary. When the atmospheric light is homogeneous, $t(x)$ can be expressed as stated [4],[8]:

$$t(x) = e^{-ßd(x)} \quad (2)$$

The parameter β represents the scattering coefficient of the atmosphere. Based on this, the required parameter is the depth map $d(x)$.

Monodepth, a contemporary method for predicting scene depth from a single image, has undergone significant improvements in its successor Monodepth2, the version utilized in this study [21]. The thickness of artificially generated hazy images is adjusted by manipulating the atmospheric scattering coefficient with the same variety of β values as suggested in the article [4]. The β parameter in (2) is a random real number between 1.0 and 3.0, corresponding to the scattering coefficient that is essential for the transmission matrix. Synthetic hazy

data, representing different atmospheric conditions is generated using β values of 1.0, 1.5, and 2.0, as given in Figure 2. The study utilizes the Monodepth2 model to create synthetic hazy images with three different β values, as shown in Figure 2, depicting the data synthesis module. The β values are adjusted to create scenarios modeling different atmospheric conditions, providing a versatile approach for the proposed data synthesis [4].

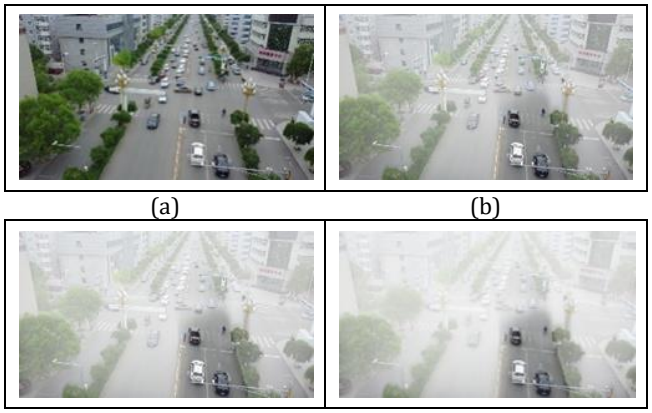

Figure 2. Synthetic hazy image generation via Monodepth2. (a): Original image. (b): β =1.0, (c): β =1.5, (d): β =2.0.

3.4 Dehamer

Dehamer stands out as a method addressing the challenges of equivariance and locality in dehazing tasks [32]. As illustrated in Figure 3, the model comprises five key modules: a transmission-aware 3D position embedding module, a Transformer module, a CNN encoder module, a feature modulation module, and a CNN decoder module [13].

It is widely known that CNNs lead to uncertainties and coarse details in dehazing, while the Transformer based methods on the other hand suffer from performance degradation by neglecting variational haze densities. Dehamer addresses these two issues by organizing features through learning modulation matrices. With this method, the long-range modeling capability of the Transformer is integrated into the image enhancement process, and this capability, combined with the local representation ability of CNNs, providing an effective solution in the problem domain [4].

As an alternative to traditional image processing methods, vision transformers (VT) focus on learning spatially distant relationships among the parts more effectively [33]. This approach offers advantages in more effective feature extraction. The transformer architecture draws attention with its parallel processing capability, where each element of the input data can be processed independently of the others [34]. Unlike traditional models, processing each element while considering global context allows for more effective modeling of distant relationships [35]. However, VT's disadvantage lies in their challenge to seamlessly integrate into existing schemes for image dehazing due to their inherent lack of local representation capability and unsuitable position embedding for this specific task [13].

The training dataset utilized consists of hazy images from the Indoor Training Set (ITS) and Outdoor Training Set (OTS) subsets within the Realistic Single Image Dehazing (RESIDE) dataset. Additionally, DenseHaze and NH-HAZE datasets were included in the experiments.

Dehamer method outperforms the classical DCP method and other deep learning based methods including DehazeNet, AODNet, GridDehazeNet, FFANet, MSBDN, and UHD in terms of PSNR and SSIM metrics.

3.5 C2PNet

In contrastive models, the issue of insufficiently constraining the solution space arises due to the distant representation of hazy images from cleaned images in the embedding manifold. C2PNet is a deep learning method aiming for a consensual contrastive solution space. In this method, negative images (hazy ones) are typically compared with positive images (cleaned ones) during the haze removal process.

C2PNet enhances feature space interpretability by adopting a dual-branch network structure based on atmospheric scattering models and physics awareness. While Dong et al.'s Feature Dehazing Unit (FDU) model focuses on minimizing cumulative errors in the raw space by incorporating physics models in the feature space [3], it lacks a mechanism for evaluating diverse physical features. To overcome this, the physics-aware dual-branch unit (PDU) is introduced, inspired by re-evaluating the physics model for haze removal. The PDU aims to provide interpretability in haze without relying on the actual values of transmission matrix (t) and atmospheric light (A) by applying physical priorities in the feature space, addressing the challenges posed by the unknown factors in the atmospheric scattering model.

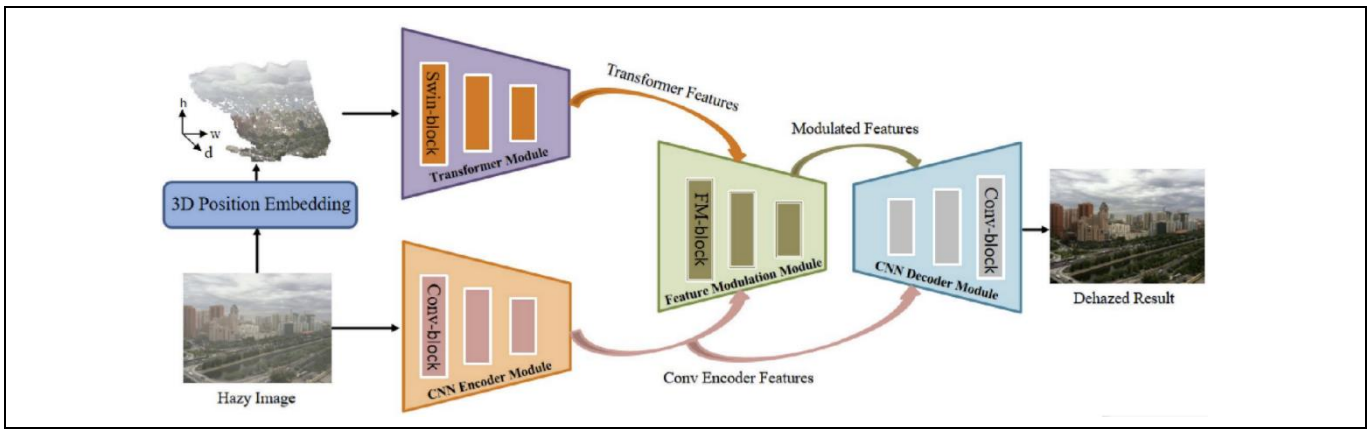

Figure 3. Diagram of Dehamer method's neural architecture. Adopted from [13].

The study primarily has two objectives. First, it aims to increase the interpretability of the feature space used in the haze removal process. Second, the goal is to create a more concise solution space using contrastive examples. To achieve these objectives, the aim is to minimize the distance to the anchor (prediction) of L1 positive images and maximize the distance to negative images. In this context, the difficulty of different negative images is defined, and as part of the learning strategy, three levels of difficulty - easy, hard, and ultra-hard-are determined. Continuous hazy input is used for easy negative images, while the difficulty levels of other negative images are dynamically determined during training. The diagram representing network architecture is shown in Figure 4 below.

During the training of C2PNet, the utilized dataset consists of hazy images from the Indoor Training Set (ITS) and Outdoor Training Set (OTS) subsets within the Realistic Single Image Dehazing (RESIDE) dataset for synthetic image dehazing. Additionally, DenseHaze and NH-HAZE2 datasets including real-world hazy images were utilized.

3.6 D4

The D4 [12] method focuses on the density and depth properties in the dehazing process by targeting to explore scattering coefficients and depth information in hazy and clear images rather than just estimating transmission maps. Emphasis is placed on the tendency of dehazing models trained on synthetic images to generalize to real-world hazy images.

The difficulty arises from the challenge of obtaining pairs of clean and their corresponding hazy images. The D4 method predicts the depth information of the clean input image and focuses on synthesizing hazy images at different densities. Afterwards, the model is tested on both synthetic and original hazy images. Unlike the RefineDNet, the D4 approach exhibits superior performance in haze removal. In contrast to CycleGAN-based methods, which can only produce haze with a fixed density for a specific clean image, the generated haze lacks consistency with the depth information.

3.7 gUNet

The gUNET approach [36], in comparison to the recently developed Vision Transformer-based Dehamer approach, adopts a simpler architecture with minimal modifications to create a U-Net model variant, aiming to simplify the implementation, integration, and usage. Inspired by GCANet and GridDehazeNet, the proposed network architecture focuses on predicting residual between a clean image and a hazy image, instead of estimating global atmospheric light A and transmission matrix $t(x)$ in the atmospheric scattering module. Pixel and channel attention modules from FFA-Net are incorporated into this structure. The gUNET architecture uses depthwise separable convolution layers to merge spatial information and transform features effectively. The global information extraction process is carried out by another module based on the SK module, which dynamically merges feature maps channel-wise. The visual representation of these neural architectural can be observed in the accompanying Figure 5.

When examining the performance advantages and features of gUNET, especially the smaller and lighter models (particularly GUNet-S), it is observed that gUNET outperforms other methods on the RESIDE dataset. It should be noted that gUNET has fewer parameters and a lighter model architecture suggests its suitability for faster and broader applications. Ablation studies shows that performance gain of the model mainly stem from attention mechanisms, nonlinear activation functions, global information extraction, normalization layers, and the number of training epochs. The channel attention mechanism is emphasized for its effectiveness in extracting global information and predicting atmospheric light [7].

3.8 Evaluation metrics

Object detection aims to successfully estimate both the localization and classification of objects within a given image [37]. Location of the object is generally given in the form of a bounding rectangle called bounding box. In object detection, precision and recall are crucial metrics used to evaluate how well a model performs. While seemingly similar, they measure different aspects of the model's accuracy. Precision shows the proportion of correctly detected objects among all those the model identified. High precision means that the model seldomly misidentifies an object when flagging it, while low precision indicates a higher frequency of false positives, such as mistakenly identifying a cat as a dog. The formula of precision is given in (3) below [38]:

$$P = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \tag{3}$$

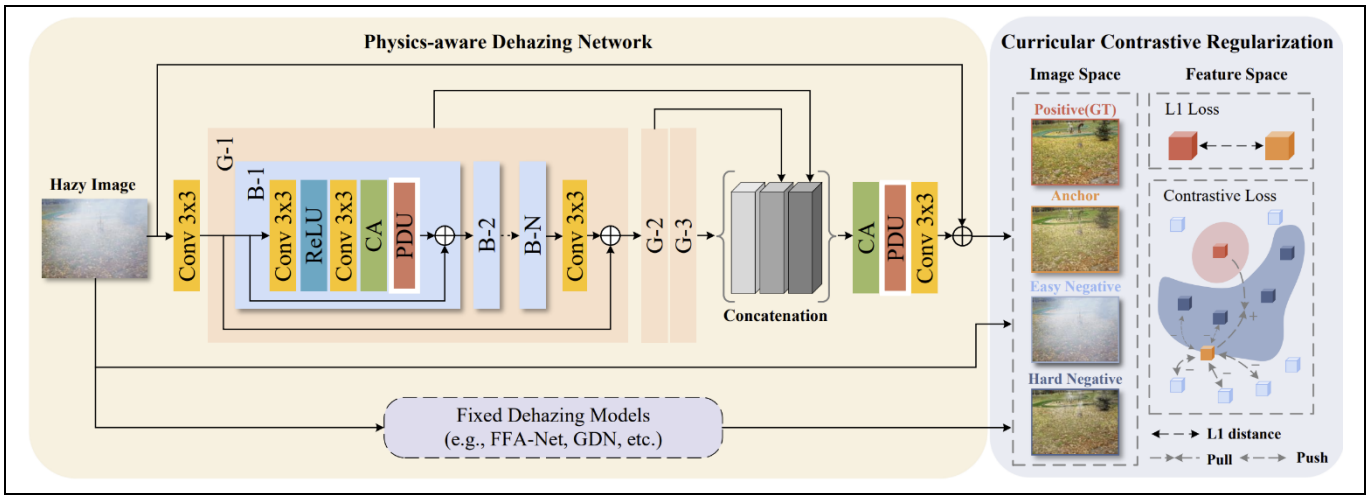

Figure 4. Architecture diagram of C2PNet. Adopted from [14].

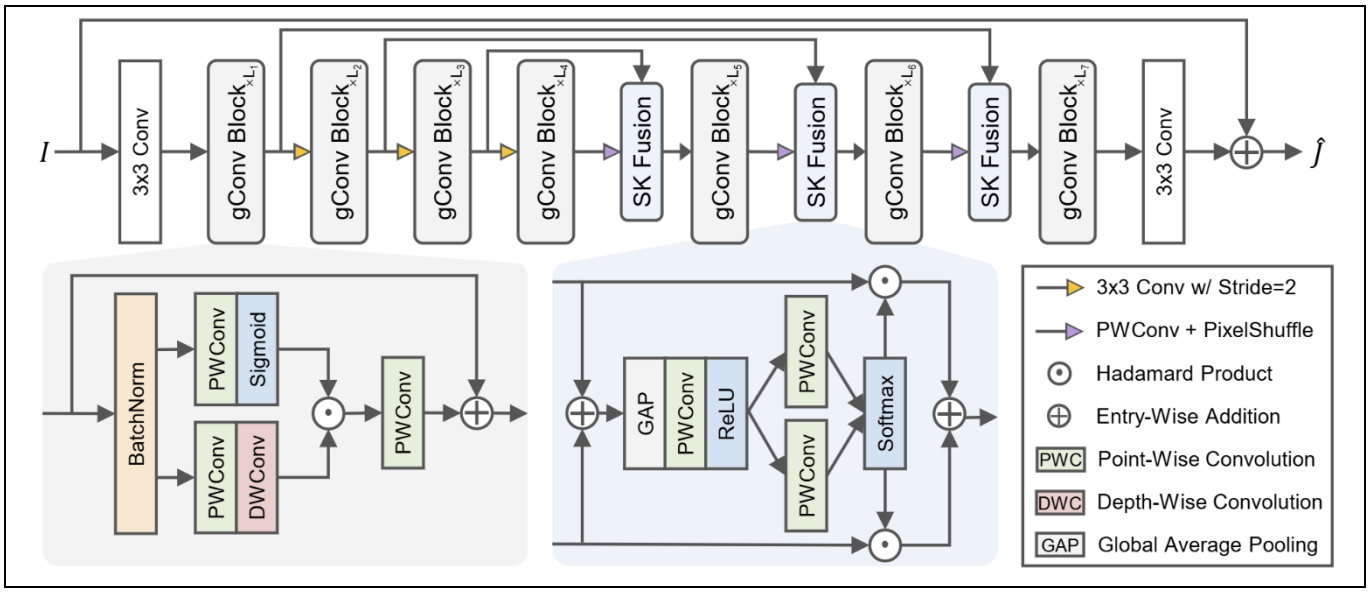

Figure 5. The neural network flow of gUNet. Adopted from [7].

On the other hand, in the context of object detection, recall refers to the ability of a model to correctly identify or detect all relevant objects in a given dataset. In essence, it assesses the ratio of correctly identified objects (true positive detections) to all instances of the object class present in the dataset. Put simply, recall gauges how effectively a model captures all occurrences of a specific object class within the dataset. The computation of recall is given in (4) [38]:

$$R = \text{TP} / (\text{TP} + \text{FN}) \tag{4}$$

Both Recall and Precision metrics range between 0 and 1, with higher values indicating greater success of the metric.

Precision and recall have an inherent trade-off. A model with very high recall might sacrifice precision by detecting many false positives. AP (Average Precision) and mAP (Mean Average Precision) account for this by calculating an average performance across different levels of precision and recall. This gives a more holistic picture of the model's ability to balance correctly identifying true objects (high recall) with minimizing false positives (high precision), providing a comprehensive assessment of the detection quality. In addition, AP and mAP are computed across different confidence thresholds for detections, helping to identify the optimal operating point for the model.

Calculation of AP involves two steps (a) generating the Precision-Recall curve and (b) computing Precision at different Recall levels. Next, for each recall interval we multiply the precision value by the recall range width followed by summing the values obtained for all recall intervals. This sum represents the area under the Precision-Recall Curve, which is the Average Precision (AP).

The mAP (mean Average Precision) score, on the other hand, is often reported at different Intersection over Union (IoU) thresholds such as 0.5 and 0.95. These thresholds represent the level of overlap required between the predicted bounding boxes and the ground truth bounding boxes to consider a detection as correct. The computation of mAP is given in (5) below.

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n} \tag{5}$$

4 The approach

In our study, dehazing models, namely Dehamer, C2PNet, D4, and gUNet, have been adopted as the main techniques for experimental purposes. The choice of algorithms was shaped by their yet-to-be-evaluated aspects concerning object detection performance, their recent publication dates, and their presentation at significant conferences such as CVPR.

Our study mainly aims to compare the results obtained with these four different dehazing algorithms to determine whether and to what extent those algorithms improve object detection performance. Subsequently, the performance of the dehazed images is evaluated through the three common metrics in object detection where YOLO8 and YOLO10 small models are used. Followingly, we have compared the obtained results with haze-free images.

One highlighted aspect of the study is the reliance on Image Quality Assessment (IQA) metrics in most of the dehazing studies, particularly based on metrics like PSNR and SSIM, and the fact that the success of object detection is often overlooked. It is a known fact that some dehazing models such as AOD-NET and DCP produce artifacts while recovering the original image signal. Thus, we, in essence, discuss the effectiveness of selected state-of-the-art dehazing methods from the point of object detection view. This study, hence, particularly emphasizes the critical role of measuring the success of object detection algorithms in challenging computer vision tasks such as video surveillance, identification, detection, and autonomous driving systems in challenging weather conditions.

In line with these objectives, one of the key reasons for the preference of YOLO series is its ability to quickly provide the position and class of the bounding box during object detection, minimizing errors by globally processing the entire image during prediction. Standing out as a rapid and single-shot model for real-time object detection, allowing for efficient use in multi-class detection, unlike models that follow a two-stage process, we have chosen to work with YOLOv8 for this reason.

On the other hand, the new non-maximum suppression free detection and Cross Stage Partial Network equipped architecture of YOLOv10 have played a key role to utilize it. The rationale behind the selection of the small variant is related to its decent detection quality together with requiring less process time which is a key consideration in industrial applications and edge-based production environments.

Along with their object annotations, a total of 548 images were collected from the validation set of the "VisDrone-DET dataset" [20]. Further, we employed the YOLO8s model which was previously trained on the VisDrone-DET dataset's training portion. Synthetic hazy images were generated through the Monodepth2 module, across three different haze densities (i.e. β: 1, 1.5, and 2) and two resolutions (640p and 1280p). The results were evaluated for each four methods via the precision, recall, mAP @50 and mAP @50-95 metrics. We hypothesized that the scores related to object detection success for dehazed images would be lower than the ground-truth haze-free images and higher than synthetically hazed ones, with values falling between these two extremes.

5 Results and discussion

We evaluated the OD performance through four metrics namely precision, recall, mAP50, and mAP50-95. As seen from Tables 2 and 3 below, the scores for C2PNet, D4, and gUNET methods align with our hypothesis. Haze-free ground-truth image based detection yielded the highest performance, while hazy image based detection demonstrated the lowest performance. The outcomes of C2PNet, D4, and gUNET methods fall between ground truth and hazy images. According to the results listed in Table 2 below, it is observed that for 640p images, the gUNET model outperforms other methods in terms of both dehazing and object detection success on hazy images, considering objective evaluations of quality. Subsequently, the second successful method emerged as C2PNet, followed by method D4 whereas the Dehamer performed the worst results.

Table 2. Computed object detection results for 640p input regime in YOLO8 small model. The bold results are the best recovered results for each β coefficient.

Images	P	R	mAP50	mAP 50-95
Ground truth	0.523	0.388	0.404	0.242
Hazy image (β =1.0)	0.457	0.333	0.339	0.201
Hazy image (β =1.5)	0.432	0.292	0.295	0.174
Hazy image (β =2.0)	0.392	0.247	0.247	0.145
C2PNet (β =1.0)	0.475	0.357	0.366	0.219
C2PNet (β =1.5)	0.447	0.321	0.326	0.193
C2PNet (β =2.0)	0.42	0.27	0.274	0.162
D4 (β =1.0)	0.497	0.37	0.38	0.226
D4 (β =1.5)	0.456	0.338	0.339	0.2
D4 (β =2.0)	0.401	0.269	0.269	0.157
Dehamer (β =1.0)	0.434	0.313	0.301	0.131
Dehamer (β =1.5)	0.411	0.289	0.277	0.121
Dehamer (β =2.0)	0.367	0.241	0.229	0.099
gUNET (β =1.0)	**0.512**	**0.381**	**0.393**	**0.235**
gUNET (β =1.5)	**0.485**	**0.372**	**0.377**	**0.225**
gUNET (β =2.0)	**0.45**	**0.324**	**0.329**	**0.195**

As seen in Table 3 below, the gUNET method was found to be superior in terms of both dehazing and object detection performance, considering objective evaluations of quality, for 1280p images as well. It is observed that the performance in all methods is measured higher compared to that in 640p. This finding indicates the existence of a positive correlation between OD performance and the input image resolution.

Table 3. Computed object detection results for 1280p input regime in YOLO8. The bold results are the best recovered results for each β coefficient.

Images	P	R	mAP50	mAP 50-95
Ground truth	0.568	0.47	0.486	0.297
Hazy image (β =1.0)	0.503	0.391	0.399	0.243
Hazy image (β =1.5)	0.47	0.336	0.341	0.207
Hazy image (β =2.0)	0.41	0.274	0.276	0.168
C2PNet (β =1.0)	0.522	0.416	0.427	0.259
C2PNet (β =1.5)	0.468	0.359	0.365	0.221
C2PNet (β =2.0)	0.435	0.286	0.292	0.178
D4 (β =1.0)	0.541	0.453	0.459	0.278
D4 (β =1.5)	0.484	0.383	0.389	0.235
D4 (β =2.0)	0.437	0.285	0.295	0.177
Dehamer (β =1.0)	0.447	0.347	0.334	0.138
Dehamer (β =1.5)	0.422	0.316	0.299	0.124
Dehamer (β =2.0)	0.437	0.285	0.295	0.177
gUNET (β =1.0)	**0.569**	**0.456**	**0.474**	**0.288**
gUNET (β =1.5)	**0.537**	**0.445**	**0.455**	**0.277**
gUNET (β =2.0)	**0.499**	**0.382**	**0.391**	**0.238**

As can be seen from Tables 4 and 5, YOLOv10 *small* model based detections have shown that YOLOv10 performs better than YOLOv8. The difference between YOLO v8 and v10 has significantly escalated especially when the input image is given in 1280p compared to 640p. YOLOv10 outperformed YOLOv8 in all image sets involving ground truth, hazed versions and their dehazed counterparts in all three beta coefficients. We believe that this improvement sources from the fact that YOLOv10 optimizes various components from both efficiency and accuracy perspectives, including lightweight classification heads, spatial-channel decoupled downsampling, and rank-guided block design [25].

Table 4. Computed object detection results for 640p input regime in YOLO10 small model. The bold results are the best recovered results for each β coefficient.

Images	P	R	mAP 50	mAP50-95
Ground Truth	0.527	0.389	0.404	0.243
Hazy Images (β =1.0)	0.475	0.332	0.346	0.205
Hazy Images (β =1.5)	0.457	0.296	0.310	0.184
Hazy Images (β =2.0)	0.428	0.254	0.266	0.158
C2PNet (β =1.0)	0.478	0.361	0.366	0.219
C2PNet (β =1.5)	0.456	0.323	0.348	0.198
C2PNet (β =2.0)	0.441	0.279	0.289	0.172
D4 (β =1.0)	0.485	0.370	0.378	0.226
D4 (β =1.5)	0.469	0.338	0.343	0.203
D4 (β =2.0)	0.423	0.284	0.286	0.168
Dehamer (β =1.0)	0.447	0.312	0.306	0.132
Dehamer (β =1.5)	0.412	0.298	0.283	0.122
Dehamer (β =2.0)	0.374	0.253	0.241	0.103
gUnet (β =1.0)	**0.510**	**0.379**	**0.391**	**0.234**
gUnet (β =1.5)	**0.493**	**0.365**	**0.373**	**0.224**
gUnet (β =2.0)	**0.470**	**0.325**	**0.335**	**0.199**

Table 5. Computed object detection results for 1280p input regime in YOLO10. The bold results are the best recovered results for each β coefficient.

Images	P	R	mAP 50	mAP 50-95
Ground Truth	0.626	0.522	0.554	0.354
Hazy Images (β =1.0)	0.601	0.462	0.496	0.314
Hazy Images (β =1.5)	0.559	0.427	0.450	0.284
Hazy Images (β =2.0)	0.540	0.372	0.398	0.250
C2PNet (β =1.0)	0.602	0.484	0.513	0.325
C2PNet (β =1.5)	0.574	0.441	0.468	0.295
C2PNet (β =2.0)	0.527	0.384	0.405	0.255
D4 (β =1.0)	0.609	0.505	0.532	0.338
D4 (β =1.5)	0.595	0.455	0.485	0.307
D4 (β =2.0)	0.545	0.379	0.406	0.255
Dehamer (β =1.0)	0.519	0.409	0.402	0.172
Dehamer (β =1.5)	0.503	0.375	0.375	0.162
Dehamer (β =2.0)	0.465	0.321	0.318	0.138
gUnet (β =1.0)	**0.628**	**0.510**	**0.544**	**0.346**
gUnet (β =1.5)	**0.604**	**0.499**	**0.527**	**0.335**
gUnet (β =2.0)	**0.582**	**0.451**	**0.480**	**0.304**

In Figure 6. the YOLOv10 F1-Confidence curves are depicted for hazed images with β =1, 1.5 and 2.0 (on top row) and their dehazed counterparts through gUNET. The detections are done for 1280p images. The F1 scores are obtained with precision and recall values. As can be seen from the Figure 6, the detection confidence scores of all ten object classes have increased. Thus, gUNET (the most successful method in this study) has brought the greatest reconstruction rate when in all classes. This can be also seen from the overall blue curve depicting the average F1-confidence score (between 0.5 to 0.95 mAP).

Regarding the Dehamer model, the implemented transformer architecture is strengthened by combining long-range features with local attention sourcing from CNN features. However, due to the design of the approach, it has been inferred that this methodology still does not take into account the physical characteristics of the natural hazing process [14] well and consequently, it lags in the manner of performance. Nevertheless, as can be seen from the visuals in Figure 7, it is considered successful in mitigating some amount of haze, despite a serious color distortion problem. At this point, we argue that a dehazing mechanism should not be considered "successful" by just evaluating its PSNR and SSIM scores due to the produced artifacts during the process. Qualatitavly, a dehazer algorithm may produce visually appealing results. Nevertheless, the unexpected finding here once again points out the subtle perceptional difference between the human visual system and vision algorithms. The metrics of PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) might report a performance score oriented toward the human visual system whereas the pixel distortions may affect other computer vision tasks adversely. Meanwhile, higher the PSNR score we have, better the image quality is obtained. Likewise, SSIM is an index that evaluates image similarity, indicating the level of similarity between two images. The higher the value, the more similar the two images are. When both metrics have high values, it is concluded that the image quality is also high. These two convetional metrics often involve values obtained by comparing an output image with a reference (ground truth) counterpart.

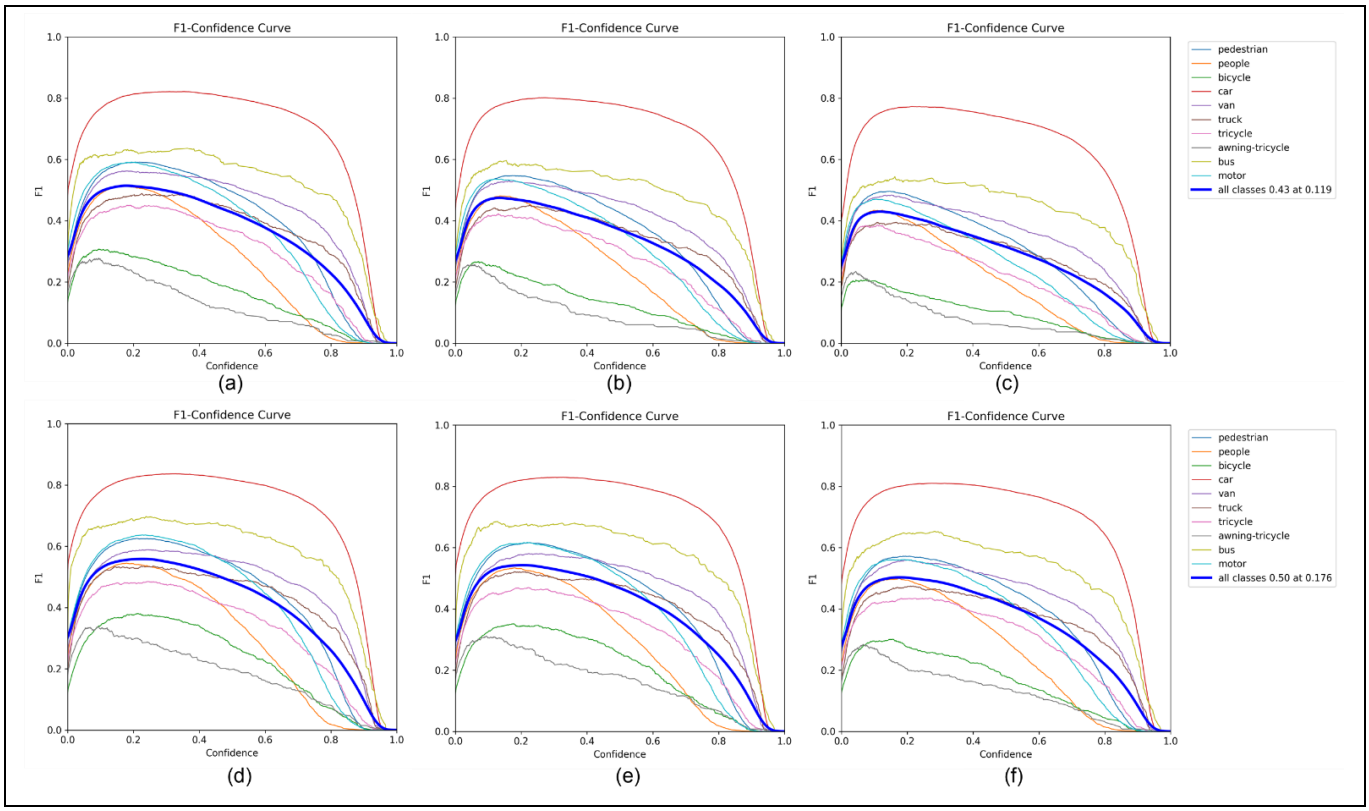

Figure 6. F1-Confidence curves obtained after 1280p/YOLOv10 based detections with images of (a): Hazed with β =1.0, (b): Hazed with β =1.5, (c): Hazed with β =2.0, (d): Detections after gUNET for β =1.0 hazed images, (e): Detections after gUNET for β =1.5 hazed images and (f): Detections after gUNET for β =2.0 hazed images.

From this point of view, it can be said that the gUNET method is relatively successful. Nonetheless, from our perspective, for a more comprehensive analysis, the OD performance should also be considered whether the dehazer algorithm is artifact free and suitable for other vision tasks. In addition to the listed evaluations, all image instances related to one frame are shared with the readers for subjective assessment in Figure 7 below.

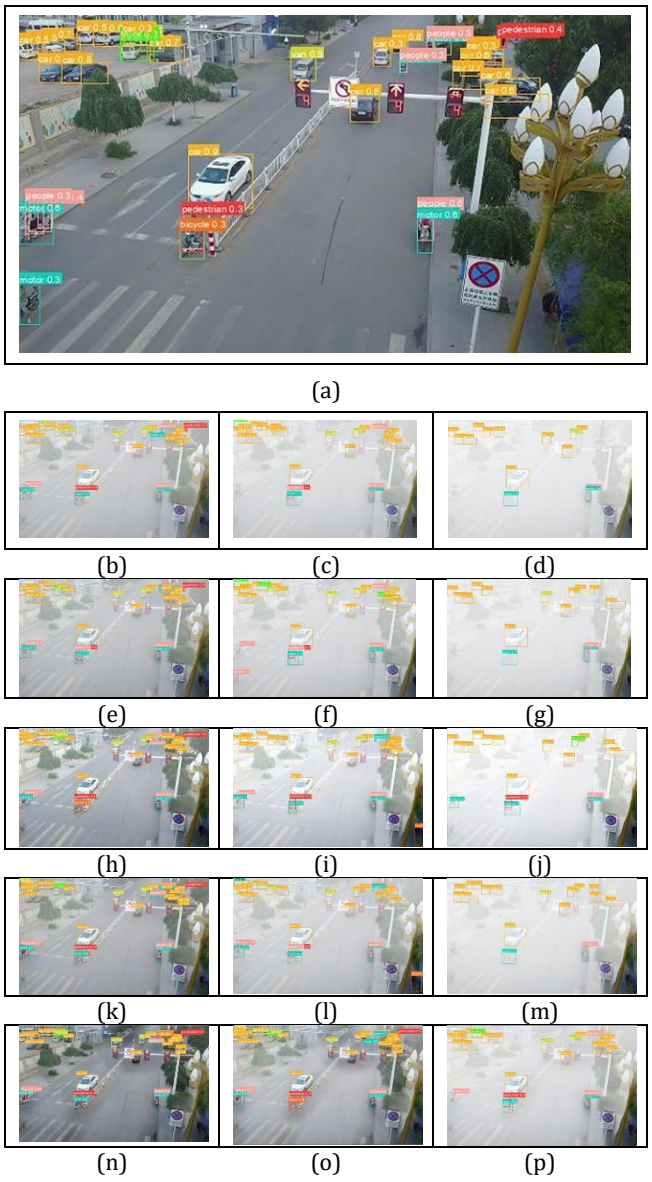

Figure 7(a): Original. (b): Hazy image for β =1.0, (c): Hazy image for β =1.5, (d): Hazy image for β =2.0, (e): C2PNet for β =1.0, (f): C2PNet for β =1.5, (g): C2PNet for β =2.0, (h): D4 for β =1.0 (i): D4 for β =1.5, (j): D4 for β =2.0, (k): Dehamer for β =1.0, (l): Dehamer for β =1.5, (m): Dehamer for β =2.0, (n): gUNET for β =1.0, (o): gUNET for β =1.5, (p): gUNET for β =2.0

Apart from the discussions presented above, we list our general findings below:

- YOLOv10 performed better than YOLOv8 in all data regimes. The improvement with YOLOv10 has become more evident, especially as the input image's resolution increases,
- gUNET outperformed other methods for both dehazing and object detection on hazy images at both 640p and 1280p, considering objective quality evaluations,
- The second most successful method was C2PNet, followed by D4, and the least successful was the Dehamer method,
- As the input image size for OD module increases from 640p to 1280p, the number of detected bounding boxes also increases proportionally regardless of the OD method. This expected behavior is found to be valid for dehazed images as well,
- The OD performance, as expected, decreases as the applied hazing factor (β coefficient) increases. Thus, using larger images is of utmost importance when the expected environmental haze level is high,
- The Dehamer model, despite a serious color distortion and artifact generation problem, shows some success in OD. This success, however, remains behind the results obtained with pure hazed images. This is likely be related to the disparity between previously learned filters of CNN and the computed activation maps during inference,
- Due to the dataset's exclusive focus on outdoor scenes, augmenting it with indoor images and training the model accordingly could improve the depth perception of the model and introduce diversity.

6 Conclusion

In this study, the feasibility of dehazing methods in object detection tasks is investigated. Numerous studies in the dehazing field, however, lack OD assessment. This is crucial for many above-mentioned applications since they need to be run in harsh environmental conditions. For instance, studies such as [39] rely on high quality aerial images for task specific models. In this direction, we experimented with four different transformer based state-of-the-art dehazing algorithms. In addition, we have utilized 548 annotated images taken by UAVs due to high relevance to the industrial use case. During the experiments, those images were synthetically hazed with varying densities constituting a challenge for both dehazing and object detection algorithms.

According to our experiments, we once again conclude that success in haze removal may not necessarily translate to equal proficiency in object recognition since one out of four approaches fails to improve OD performance. This is because of (1) the possible image artifact generation and (2) the widely used success metrics in dehazing do not often consider the natural pixel distributions which are very important for CNN based vision schemes. Further, the input resolution matters for both OD and dehazing. As a consequence, it is highly suggested to (i) increase the input frame resolution and (ii) inspect the OD performance when dehazing comes into prominence.

Due to the nature of dehazing, it is far apart from OD task in which it takes an input and renders a dehazed output image whereas OD takes an image and infers region proposals. For this reason, it is currently the most viable idea to cascade two different neural network architectures (e.g, gUNET – YOLO) although this creates the computation burden without any optimization. Nonetheless, CNN or Transformer blocks can be trained to have multi-task learning. Therefore, in the future, it should be studied to have implicit dehazing capabilities in

potential object regions by avoiding the use of dehazing algorithms as a pre-processing step and incorporating dehazing operations in less resource-requiring CNN activation maps. As a possible future work, diversifying haze densities, incorporating indoor scenes, and experimenting with more dehazing methods could contribute to the focus of this paper.

7 Authors contribution statements

Within the scope of this study, Nurçiçek Özenç contributed to the literature review, data collection, obtaining results, analysis of results and writing the paper. Ahmet Selman Bozkır contributed to the formation of idea, interpretation of the results, writing the paper and revision of the article.

8 Ethics committee approval and conflict of interest statement

"There is no need to obtain permission from the ethics committee for the article prepared".

"There is no conflict of interest with any person / institution in the article prepared".

9 References

[1] Yang Y, Wang C, Liu R, Zhang L, Guo X, Tao D. "Self-augmented unpaired image dehazing via density and depth decomposition". *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 19–24 June 2022.

[2] Li B, Ren W, Fu D, Tao D, Feng D, Zeng W, Wang Z. "Benchmarking single-image dehazing and beyond". *IEEE Transactions on Image Processing*, 28(1), 492-505, 2019.

[3] Chahal KS, Dey K. "A survey of modern object detection literature using deep learning". *arXiv*, 2018. https://arxiv.org/pdf/1808.07256

[4] Medium, "Synthesize Hazy/Foggy Images using Monodepth and Atmospheric Scattering Model". https://towardsdatascience.com/synthesize-hazy-foggy-image-using-monodepth-and-atmospheric-scattering-model-9850c721b74e (08.08.2024).

[5] Tran LA, Do TD, Park DC, Le MH. "Robustness enhancement of object detection in advanced driver assistance systems (ADAS)". *arXiv,* 2021. https://arxiv.org/pdf/2105.01580.

[6] Song Y, He Z, Qian H, Du X. "Vision transformers for single image dehazing". *IEEE Transactions on Image Processing*, 32, 1927-1941, 2023.

[7] Song Y, Zhou Y, Qian H, Du X. "Rethinking performance gains in image dehazing networks". *arXiv* 2022. https://arxiv.org/pdf/2209.11448

[8] Thakur N, Nagrath P, Jain R, Saini D, Sharma N, Hemanth J. "Object detection in deep surveillance". *Research Square*, 2021. https://doi.org/10.21203/rs.3.rs-901583/v1

[9] Ali S, Abdullah Athar A, Ali M, Hussain A, Kim HC. "Computer vision-based military tank recognition using object detection technique: an application of the YOLO framework". *1st International Conference on Advanced Innovations in Smart Cities*, Jeddah, Saudi Arabia, 23-25 January 2023.

[10] Rahadianti L, Azizah A Y, Deborah H. "Evaluation of the quality indicators in dehazed images: color, contrast, naturalness, and visual pleasingness". *Heliyon*, 7(9), 1-12, 2021.

[11] Wu H, Qu Y, Lin S, Zhou JJ, Qiao R, Zhang Z, Xie Y, Ma L. "Contrastive learning for compact single image dehazing". *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, 19–25 June 2021.

[12] Yang Y, Wang C, Liu R, Zhang L, Guo X, Tao D. "Self-augmented unpaired image dehazing via density and depth decomposition", *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 19–24 June 2022

[13] Guo C, Yan Q, Anwar S, Cong R, Ren W, Li C. "Image dehazing transformer with transmission-aware 3D position embedding". *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 19–24 June 2022.

[14] Zheng Y, Zhan J, He S, Dong J, Du Y. "Curricular contrastive regularization for physics-aware single image dehazing". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 18-22 June 2023.

[15] He K, Sun J, Tang X. "Single image haze removal using dark channel prior". *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 20-25 June 2009.

[16] Berman D, Treibitz T, Avidan S. "Non-local image dehazing". *2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 26 June-1 July 2016.

[17] Li B, Peng X, Wang Z, Xu J, Feng D. "AOD-Net: all-in-one dehazing network". *IEEE International Conference on Computer Vision*, Venice, Italy, 22-29 October 2017.

[18] Ancuti CO, Ancuti C. "Single image dehazing by multi-scale fusion". *IEEE Transactions on Image Processing*, 22(8), 3271-3282, 2013.

[19] Ultralytics. "VisDrone". https://docs.ultralytics.com/tr/datasets/detect/visdrone/#citations-and-acknowledgments (08.02.2024).

[20] GitHub. "VisDrone/VisDrone-Dataset". https://github.com/VisDrone/VisDrone-Dataset (08.07.2024).

[21] GitHub. "tranleanh/haze-synthesis". https://github.com/tranleanh/haze-synthesis (09.05.2024).

[22] Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, Ding, G. "Yolov10: Real-Time End-To-End Object Detection". *arXiv* 2024. https://arxiv.org/pdf/2405.14458

[23] Hussain M. "YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection". *Machines*, 11(7), 677, 2023.

[24] Roboflow Blog. "Your Comprehensive Guide to the YOLO Family of Models". https://blog.roboflow.com/guide-to-yolo-models/ (08.02.2024).

[25] Ghosh A. "YOLOv10: The Dual-Head OG of YOLO Series". https://learnopencv.com/yolov10/ (01.07.2024).

[26] Marium A, Srinivasan D G, Shetty S A. "Literature survey on object detection using YOLO". *International Research Journal of Engineering and Technology*, 7(6), 3082-3088, 2020.

[27] Jiang P, Ergu D, Liu F, Cai Y, Ma B. "A review of YOLO algorithm developments". *Procedia Computer Science*, 199, 1066-1073, 2022.

[28] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg, AC. "SSD: single shot multibox detector". *Computer Vision–ECCV 2016: 14th European Conference*, Amsterdam, The Netherlands, 11–14 October 2016.

[29] Deng C, Wang M, Liu L, Liu Y, Jiang Y. "Extended feature pyramid network for small object detection". *IEEE Transactions on Multimedia*, 24, 1968-1979, 2022.
[30] Hnewa M, Radha H. "Multiscale domain adaptive YOLO for cross-domain object detection". *2021 IEEE International Conference on Image Processing*, Anchorage, Alaska, USA, 19-22 September 2021.
[31] Sirisha U, Praveen SP, Srinivasu PN, Barsocchi P, Bhoi AK. "Statistical analysis of design aspects of various YOLO-based deep learning models for object detection". *International Journal of Computational Intelligence Systems*, 16(126), 1-29, 2023.
[32] GitHub "Li-Chongyi/Dehamer". https://github.com/Li-Chongyi/Dehamer (08.02.2024).
[33] Wu B, Xu C, Dai X, Wan A, Zhang P, Yan Z, Tomizuka M, Gonzalez J, Keutzer K, Vajda P. "Visual transformers: token-based image representation and processing for computer vision". *arXiv* 2020. https://arxiv.org/pdf/2006.03677
[34] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. "Attention is all you need". *arXiv*, 2017. https://arxiv.org/pdf/1706.03762
[35] Dosovitskiy A. "An image is worth 16x16 words: transformers for image recognition at scale". *arXiv*, 2020. https://arxiv.org/pdf/2010.11929
[36] GitHub. "IDKiro/gUNet". https://github.com/IDKiro/gUNet (08.02.2024).
[37] Shah T. "Measuring object detection models - mAP - what is mean average precision?". https://tarangshah.com/blog/2018-01-27/what-is-map-understanding-the-statistic-of-choice-for-comparing-object-detection-models/ (08.02.2024).
[38] LearnOpenCV. "Mean average precision (mAP) in object detection". https://learnopencv.com/mean-average-precision-map-object-detection-model-evaluation-metric/ (08.02.2024).
[39] Altun M, Türker M. "Vehicle detection in urban areas from very high resolution UAV color images". *Pamukkale University Journal of Engineering Sciences*, 26(2), 371-384, 2020.