Semi-UFormer: Semi-supervised Uncertainty-aware Transformer for Image Dehazing

Ming Tong¹, Xuefeng Yan¹,²,* , Yongzhen Wang¹, Mingqiang Wei¹

¹ School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China
² Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, China
{tongming, yxf, wangyz}@nuaa.edu.cn, {mingqiang.wei}@gmail.com

Abstract—Image dehazing is fundamental yet not well-solved in computer vision. Most cutting-edge models are trained in synthetic data, leading to the poor performance on real-world hazy scenarios. Besides, they commonly give deterministic dehazed images while neglecting to mine their uncertainty. To bridge the domain gap and enhance the dehazing performance, we propose a novel semi-supervised uncertainty-aware transformer network, called Semi-UFormer. Semi-UFormer can well leverage both the real-world hazy images and their uncertainty guidance information. Specifically, Semi-UFormer builds itself on the knowledge distillation framework. Such teacher-student networks effectively absorb real-world haze information for quality dehazing. Furthermore, an uncertainty estimation block is introduced into the model to estimate the pixel uncertainty representations, which is then used as a guidance signal to help the student network produce haze-free images more accurately. Extensive experiments demonstrate that Semi-UFormer generalizes well from synthetic to real-world images.

Index Terms—Uncertainty-aware, Semi-supervised, CNN-Transformer, Image dehazing

I. INTRODUCTION

Images captured in hazy weather often suffer from substantial degradation in visibility, color distortion, and contrast reduction. These factors further undermine the effectiveness of downstream vision-based systems, including object detection, autonomous driving, and traffic surveillance [1]. Hence, the restoration of clean images from their hazy counterparts holds significant importance.

Existing dehazing efforts can be categorized into two main types: prior-based [2], [3] and learning-based [4], [5] methods. Traditional prior-based research relies on hand-crafted image priors, such as the dark channel prior (DCP) [2] and non-local color prior [3], to address the challenge of image dehazing based on the atmospheric scattering model [6]. While these algorithms have the potential to enhance image visibility, their reliability is restricted by the heavy reliance on assumptions. Furthermore, to attain optimal dehazing performance, these methods demand laborious parameter adjustment, imposing significant costs in terms of time and effort.

The recent advancements in deep learning have opened up significant opportunities for tackling image dehazing tasks, resulting in a proliferation of learning-based approaches [4], [5], [7], [8]. These algorithms exhibit high efficiency and demonstrate promising results on well-established benchmarks. However, the performance of dehazing networks heavily relies on the distribution of the training set, often leading to a decline when faced with out-of-distribution samples. Additionally, a substantial portion of these methods necessitates paired data for training, which is impossible to collect in practice. The discrepancy between real and synthetic data gives rise to the issue of domain shift, thereby limiting their applicability in real-world scenarios. In recent years, several unsupervised methods [9], [10] have attempted to address the domain gap by training models on real images. Nonetheless, these approaches exhibit limited dehazing capabilities since the ground-truth of real-world hazy images cannot be utilized as reconstruction loss to effectively guide the network training. To overcome these challenges, this paper adopts a semi-supervised learning approach, aiming to mitigate the domain shift problem within the dehazing task while simultaneously ensuring proficient dehazing performance.

In addition, most learning-based dehazing efforts only produce the final dehazed images without addressing the uncertainty associated with the results, which plays a vital role in recovering the edge and texture regions in hazy images. Current dehazing models typically treat all pixels equally, neglecting the fact that pixels in edge and texture regions contain more visual information than those in smooth regions, and these pixels often exhibit significant levels of uncertainty. Furthermore, due to the opaque nature inherent in deep learning models, the dehazing results they produce are challenging to fully rely on, given our limited knowledge about the extent of uncertainty in the outcomes. On the other hand, uncertainty estimation can offer confidence that facilitates informed decision-making by the model. For instance, in [11] and [12], the models achieved superior outputs by leveraging uncertainty estimation. In these studies, uncertainty estimation is considered as a conditional input, and further exploration is necessary to effectively employ the acquired uncertainty information for enhancing the representation of pixels in uncertain regions.

To address these issues, we propose a novel semi-supervised uncertainty-aware transformer network (Semi-UFormer) for image dehazing. Specifically, the proposed Semi-UFormer is built upon the knowledge distillation framework, employing...
KL divergence loss to align feature distributions between synthetic and real data. This alignment facilitates the transfer of dehazing knowledge acquired from synthetic images to real-world dehazing scenarios. The network comprises a supervised branch trained on synthetic data, learning the mapping from hazy images to clean images, and an unsupervised branch trained on real data to capture the feature distribution of real-world hazy images and enhance model robustness. Furthermore, considering the non-uniform haze density and variations in visual information distribution across images, different regions should be given varying levels of attention. In this regard, we introduce an uncertainty estimation block based on the attention mechanism to estimate pixel-level epistemic uncertainty. The uncertainty serves as guiding information and, together with prior knowledge, acts as regularization constraints for the network, encouraging focused feature learning in specific regions. Consequently, the network produces dehazed images with improved details and enhanced clarity. In summary, our contributions can be summarized in three aspects:

- A novel semi-supervised uncertainty-aware transformer network called Semi-UFormer is proposed for image dehazing, which leverages both real-world data and uncertainty guidance information to boost the model’s dehazing ability.
- An uncertainty estimation block is exploited to predict the epistemic uncertainty of the dehazed images, which is then used to guide the network to better reconstruct the image texture and edge regions.
- We leverage knowledge distillation technology to align the feature distributions between synthetic and real data, which can help the network generalize well in real-world scenarios.

II. RELATED WORK

A. Single image dehazing

Image dehazing is a well-researched topic in the image restoration community. Numerous methods have been proposed to tackle this issue. These dehazing methods can be mainly categorized into the following two categories.

The first category comprises prior-based methods, which utilize hand-crafted priors to estimate the parameters of the atmospheric scattering model [13] for restoring hazy images. He et al. [2] proposed the dark channel prior (DCP) to achieve remarkable dehazing results, which assumes that pixels in non-sky local areas always contain at least one channel with a low value. Fattal et al. [14] introduced a color-line prior that removes haze by leveraging the fact that pixels in clean image patches exhibit a one-dimensional distribution in RGB-space. Berman et al. [3] employs the non-local color prior (NCP) for dehazing, assuming that a clear image can be approximated by hundreds of different colors. Although traditional prior-based methods can enhance the visibility of hazy images without requiring training with paired images, they are limited in their applicability to different scenes, often resulting in inaccurate transmission map estimates and unsatisfactory dehazing outcomes.

The second category is learning-based methods. Early methods often utilize CNNs to recover hazy-free images by estimating transmission maps and atmospheric light in the scattering model [15]–[17]. However, these estimates can be inaccurate and may result in artifacts. Subsequently, several end-to-end methods have been proposed, which directly learn to generate clean images [5], [7], [18], [19]. These CNN algorithms typically employ various techniques, including smoothed dilation convolution [5], attention mechanisms [7], progressive feature fusion [20], and more, to enhance model performance. In recent years, the introduction of Transformer [8], [21] into the image restoration community has garnered attention due to its superior performance. While these methods achieve impressive results with synthetic datasets, they struggle to perform well in real-world scenarios due to the domain gap.

To narrow the domain gap, there is increasing adoption of semi-supervised [22]–[24] and unsupervised [9], [10], [25] strategies. Several unsupervised methods, relying on the physical scattering model [9] or cycle-GAN [10], [25], have been proposed. However, these methods often exhibit limited performance due to the lack of hazy-free counterparts for learning. Therefore, we have developed a semi-supervised network that is trained with both synthetic and real images to adapt to the real image domain.

B. Uncertainty Estimation

In Bayesian modeling, uncertainty estimation can be divided into two main types: aleatoric uncertainty and epistemic uncertainty [26]. Uncertainty modeling has been explored since the 1990s [27], but it hadn’t been effectively applied until 2015. Gal [28] et al. estimate uncertainty by using dropout training in CNNs as an approximate Bayesian model. Recently, Bayesian modeling has been introduced into the computer vision tasks, such as underwater image enhancement [29], image dehazing [30], image derain [31], and image denoising [12], to improve the restored results by exploiting the uncertainty of the deep learning methods. Our methods introduce an uncertainty strategy to predict pixel-wise uncertain maps to improve robustness on real hazy images.

III. PROPOSED METHOD

Beyond existing image dehazing wisdom, Semi-UFormer extensively investigates the utilization of knowledge distillation technology and uncertainty guidance information to improve the network’s capacity for generating clearer images with more confidence. Fig. 1 illustrates the overview of our Semi-UFormer, where the teacher and student network share the same architecture. These two proposed networks comprise two branches with identical structures and shared parameters. The supervised branch is trained using paired images and supervised loss functions, while the unsupervised branch is trained using real-world hazy images and unsupervised loss functions.
Real-world images typically adhere to specific rules and exhibit global characteristics, including contrast ratio and dark channel sparsity. In order to obtain comprehensive global information and achieve precise local dehazing, we integrate parallel mix Transformer-CNN blocks into our network to effectively leverage Transformer’s global modeling capability and CNN’s local modeling capability [33], [34].

As exhibited in Fig. 2, the Transformer-based dehazing network is an enhanced 7-stage U-Net, which consists of three modules: a shallow feature extraction, the Mix Dehaze Blocks, and a reconstruction module.

Shallow feature extraction. A $3 \times 3$ convolutional layer $C_3(\cdot)$ is first applied to extract shallow feature information from the hazy image $I \in R^{H \times W \times 3}$:

$$F_{\text{shallow}} = C_3(I)$$  \hfill (1)

Mix Dehaze Block (MDB). Next, feature $F_{\text{shallow}}$ will be sent to Mix Dehaze Blocks for extracting image global and non-local features. In MDB, we process the features $F \in R^{H \times W \times C}$ using a $1 \times 1$ convolutional layer $C_1(\cdot)$ and then evenly divide them into two parts, $F_{\text{trans}} \in R^{H \times W \times \frac{C}{2}}$ and $F_{\text{cnn}} \in R^{H \times W \times \frac{C}{2}}$. These divided parts are then separately input into the transformer block $Trans(\cdot)$ and the residual block $Res(\cdot)$ to obtain features $F_{\text{trans}}$ and $F_{\text{cnn}}$, respectively. This operation decreases the number of input channels into both the Transformer and residual blocks, leading to a reduction in the model’s parameter count. It also enhances the network’s ability to process local and non-local features in a more efficient parallel manner. Subsequently, we concatenate $F_{\text{trans}}$ and $F_{\text{cnn}}$ and then apply $1 \times 1$ convolution $C_1(\cdot)$ to fuse the local and non-local features. Finally, we implement residual learning and compensate for distant information using skip connections. The detailed procedure is presented in the formula below:

$$F_{\text{trans}}, F_{\text{cnn}} = \text{Split}(C_1(F))$$  \hfill (2)

$$F_{\text{trans}}', F_{\text{cnn}}' = Trans(F_{\text{trans}}), Res(F_{\text{cnn}})$$  \hfill (3)

$$F_{\text{out}} = F + C_1(Concat(F_{\text{trans}}', F_{\text{cnn}}'))$$  \hfill (4)

Reconstruction module. Finally, a $3 \times 3$ convolutional layer $C_3(\cdot)$ and a pixel shuffle layer $P(\cdot)$ are used to produce the haze-free image $J \in R^{H \times W \times 3}$ from the extracted $F_{\text{out}}$:

$$J = P(C_3(F_{\text{out}}))$$  \hfill (5)

C. Uncertainty Estimated Block and Uncertain Loss

Theoretically, there are two types of uncertainty in Bayesian modeling: aleatoric uncertainty from the data and epistemic uncertainty from the model. The former is prevalent in dehazing models, yet existing methods overlook its exploration. Consequently, we leverage an uncertainty estimation block (UEB) to predict the uncertainty of dehazing results, allowing the model to prioritize regions with abundant visual information, such as edge areas, for enhanced final restoration outcomes.
Uncertainty Estimated Block (UEB). To enhance the ability to model uncertainty in network outputs, we propose an uncertainty estimation block for predicting uncertainty in network outputs. The uncertainty estimation block utilizes a residual block with pixel-wise attention and channel-wise attention [7], which improves the feature extraction capabilities of the network. The added attention mechanisms enable the network to adaptively focus on important parts of the image, such as object contours and texture features. The residual learning approach is employed using skip connections and summation operations to compensate for the original features. Finally, a three-layer convolutional operation with the activation function ELU is used to integrate and refine the features. The specific structure of the uncertainty estimation block is shown in the following equation:

$$ F_{\text{out}}^\prime = C_3(\sigma(C_3(F_{\text{out}}))) + F_{\text{out}} $$

$$ F_{\text{out}}^\prime = CA(PA(F_{\text{out}}')) $$

$$ \text{uncertainty} = C_3(\sigma(F_{\text{out}}')) $$

where $\sigma(\cdot)$ denotes the activation function.

By using the uncertainty estimation block and uncertain loss, we can obtain pixel-level uncertainty for each image. In the following, we will explain how to improve the first-stage network output dehazing results using the predicted uncertainty to enhance the model’s dehazing performance.

The prediction process for uncertainty $\theta$ [35] can be expressed as:

$$ \hat{J}_i = G_1(I_i) + \epsilon \theta_i $$

where $\hat{J}_i$, $G_1(I_i)$, $\epsilon$ and $\theta_i$ denote the ground-truth, coarse dehazed image generated by teacher network, Laplace distribution, and aleatoric uncertainty from the synthetic dehazed image, respectively. For a more accurate prediction of $\theta_i$, we introduce Jeffrey’s prior [36] into the uncertainty estimation process. For $\hat{J}_i$, $G_1(I_i)$, the Laplace distribution-characterized log-likelihood function and uncertainty estimation loss $L_{ue}$ can be expressed as:

$$ \ln p(\hat{J}_i|I_i) = -\frac{\|\hat{J}_i - G_1(I_i)\|_1}{\theta_i} - 2\ln \theta_i - \ln 2 $$

$$ L_{ue} = \frac{1}{N} \sum_{i=1}^{N} exp(-\ln \theta_i) \left\| \hat{J}_i - I_i \right\|_1 + 2\ln \theta_i $$

We employ $L_{ue}$ to estimate the $\theta$ more accurately. Then, through the guidance of $\theta$, we apply the uncertainty-guided loss $L_{ugs}$ to push the network to concentrate more on the reconstruction error area with large uncertainty in the dehazed image, to obtain accurate and confident dehazed results. The formula is shown in (12). In addition, inspired by the identity loss [37], we incorporate the uncertainty-guided loss $L_{ugu}$ into the unsupervised branch, as shown in (13).

$$ L_{ugs} = \frac{1}{N} \sum_{i=1}^{N} (\ln \theta_i - \min(\ln \theta_i)) \left\| \hat{J}_i - G_2(I_i) \right\|_1 $$

$$ L_{ugu} = \frac{1}{N} \sum_{j=1}^{N} (\ln \theta_j - \min(\ln \theta_j)) \left\| J_j - G_2(J_j) \right\|_1 $$

where $G_2(\cdot)$, $J_j$, $\theta_j$ represents the student network, real-world images, and uncertainty from real-world dehazed images, respectively.

D. Loss Functions

Network training comprises two stages: during the first stage, the teacher network is trained; and during the second stage, the student network is initialized with the trained teacher model and subsequently retrained. In both training stages, the supervised and unsupervised branches are constrained by supervised and unsupervised losses, respectively.
Teacher network. The overall loss functions for the teacher network are formulated as:

\[ L_{tea} = L_{ts} + L_{tu} \]  \hspace{1cm} (14)

where \( L_{ts}, L_{tu} \) refers to the loss functions of the supervised and unsupervised branches, respectively.

\[ L_{ts} = \lambda_1 \ast L_{ue} \]  \hspace{1cm} (15)

\[ L_{tu} = \lambda_2 \ast L_{ide} + \lambda_3 \ast L_{dc} + \lambda_4 \ast L_{tv} \]  \hspace{1cm} (16)

where \( L_{ide}, L_{dc}, L_{tv} \) represent identity loss [37], total variation loss and dark channel loss [22].

Identity loss: In order to minimize the structural and color differences between hazy images and dehazed images, Identity loss [37] is adopted to constrain the generated images:

\[ L_{ide} = E_{J_r \sim P_{data}(I_r)}[\|J_r - I_r\|_1] \]  \hspace{1cm} (17)

where \( J_r, I_r \) refers to the real clean images and real hazy images, respectively.

TV loss: In order to preserve the edge details of the image object, total variation \( L_{tv} \) is used to constrain the spatial smoothness of the dehazed images:

\[ L_{tv} = \frac{1}{N} \sum_{i=1}^{N} (\|\nabla_h J_i\| + \|\nabla_v J_i\|) \]  \hspace{1cm} (18)

where \( \nabla_h, \nabla_v \) represent the matrix of differential operation in horizontal and vertical directions, respectively.

DC loss: The DCP [2] reveals the presence of at least one color channel in the non-sky region of a clean image with remarkably low grayscale values for its pixels, which can be expressed as:

\[ D(I) = \min_{y \in N(x)} \min_{c \in \{r,g,b\}} \text{Softmax}(y) \]  \hspace{1cm} (19)

where \( x, y \) refers to the coordinates, \( N(x) \) and \( I^c \) refers to the patch centered at \( x \) and \( c \)-th color channel. In order to generate images with similar statistical characteristics as clean images, we use dark channel loss [22] \( L_{dc} \) to enhance the dark channel sparsity of dehazed images:

\[ L_{dc} = \frac{1}{N} \sum_{i=1}^{N} \|D_j\|_1 \]  \hspace{1cm} (20)

where \( D_j \) is the dehazed image from the real hazy image.

Student network. The overall loss functions for the student network are formulated as:

\[ L_{stu} = L_{ss} + L_{su} \]  \hspace{1cm} (21)

where \( L_{ss}, L_{su} \) refers to the loss functions of the supervised and unsupervised branches, respectively.

\[ L_{ss} = \lambda_1 \ast L_{ugs} \]  \hspace{1cm} (22)

\[ L_{su} = \lambda_2 \ast L_{ugu} + \lambda_3 \ast L_{dc} + \lambda_4 \ast L_{tv} + \lambda_5 \ast L_{kl} \]  \hspace{1cm} (23)

where \( L_{kl} \) denotes the KL divergence loss.

KL loss: Since the intermediate structure of the network tends to extract high-dimensional haze-related features, we choose the 4-th MDB to supply the synthetic hazy image embedding \( V_{syn} \) and the real-world image embedding \( V_{real} \). And by taking \( V_{syn} \) as the pseudo-label of \( V_{real} \), these two hazy distribution features are transformed into high-dimensional vectors to calculate \( L_{kl} \) [23], [29], which can be expressed as:

\[ L_{kl} = KL(\text{Softmax}(V_{real}), \text{Softmax}(V_{syn})) \]  \hspace{1cm} (24)

by which to enhance the similarity of haze distribution features between synthetic and real-world images.

IV. EXPERIMENT

A. Implementation Details

Experiments are implemented on Pytorch 1.7 with NVIDIA RTX 3090 GPU and Adam optimizer with parameters \( \beta_1 = 0.9, \beta_2 = 0.99, \epsilon = 10^{-8} \) to train the network. The teacher network is trained for 100 epochs, in which we update the unsupervised branch once after updating the supervised five times. The student network is trained for 60 epochs, in which we update the supervised branch once after updating the unsupervised five times. In each stage, the learning rate is set to 10^{-4} for the first half and then decays linearly to 0 at the end. The batch size is set to 2. The images are randomly cropped to the size of 256 × 256, and the pixel values of the images are normalized to [-1,1]. The patch size is set to 35×35 when training with the Dark Channel loss. The loss weights are set to: \( \lambda_1 = 1, \lambda_2 = 10^{-2}, \lambda_3 = 2, \lambda_4 = 10^{-2}, \lambda_5 = 10^{-5}, \lambda_6 = 10^{-6} \).

We randomly selected 10,000 paired synthetic images and 2,000 real-world hazy images from the RESIDE dataset [38] to create the training dataset. The synthetic images were sourced from the outdoor training sets (OTS), while the real-world images were sourced from the unlabeled real hazy images (URHI). Considering that haze typically occurs outdoors, only the outdoor synthetic data from the synthetic object testing set (SOTS) and the Hybrid Subjective Testing Set (HSTS) were selected for the synthetic testing dataset.

B. Comparison Results

Results on Synthetic Datasets. Our Semi-UFormer was compared with other methods on SOTS and HSTS datasets. The visual outcomes are depicted in Fig. 3. And Tab. I showcases the objective results of all tested dehazing algorithms on synthetic datasets. Alongside commonly used metrics like PSNR and SSIM, we include CIEDE2000 [41] to assess dehazing algorithms from a color difference perspective. The proposed Semi-UFormer demonstrates superior performance in PSNR and CIEDE2000. This suggests that our method produces more natural and realistic results, visually resembling a real clean image more closely.

Results on Real-world Images. To demonstrate the effectiveness of the proposed method in real scenarios, we compare it with other methods on 50 randomly selected real-world hazy images. The visual results are shown in Fig. 4. To quantitatively compare the results, we use three no-reference image quality evaluation metrics: SSEQ [42], \( \sigma \) [43]and HCC
Fig. 3. Visual comparison of dehazing results from SOTS (Synthetic Object Testing Set) [38] outdoor dataset.

| Method          | Type                     | SOTS outdoor | HSTS         |
|-----------------|--------------------------|--------------|--------------|
|                 | PSNR↑ | SSIM↑ | CIED2000↓ | PSNR↑ | SSIM↑ | CIED2000↓ |
| DCP [2]         | Prior               | 18.83  | 0.819   | 10.199 | 17.01  | 0.803   | 9.186   |
| NCP [3]         | Prior               | 18.07  | 0.802   | 10.420 | 17.62  | 0.798   | 10.722  |
| Dual-ScaleNet [39] | Prior + Supervised | 21.76  | 0.909   | 6.394  | 24.94  | 0.912   | 5.187   |
| AOD [4]         | Supervised           | 20.08  | 0.861   | 7.287  | 19.68  | 0.835   | 7.646   |
| GCA [5]         | Supervised           | 21.66  | 0.867   | 7.314  | 21.37  | 0.874   | 8.107   |
| EPDN [18]       | Supervised           | 22.57  | 0.863   | 6.847  | 23.17  | 0.873   | 7.337   |
| GFN-IJCV [19]   | Supervised           | 24.21  | 0.849   | 4.336  | 23.17  | 0.835   | 5.987   |
| YOLY [9]        | Unsupervised         | 20.39  | 0.889   | 8.557  | 21.02  | 0.905   | 7.910   |
| SLAdhezasing [40]| Self-supervised      | 24.33  | 0.932   | 5.753  | 24.183 | 0.893   | 5.423   |
| Semi-dehazing   | Semi-supervised      | 24.79  | 0.892   | 4.856  | 24.36  | 0.889   | 5.312   |
| PSD [24]        | Semi-supervised      | 20.49  | 0.844   | 14.292 | 19.37  | 0.824   | 14.820  |
| Semi-UFormer    | Semi-supervised      | 26.63  | 0.923   | 4.173  | 28.33  | 0.934   | 2.967   |

TABLE II

Quantitative comparisons (SSEQ/σ/HCC) with SOTA approaches on 50 real-world images. The 1st and 2nd rankings are visually denoted by RED and BLUE.

| Method         | SSEQ↓ | σ↓  | HCC↑  |
|----------------|-------|-----|-------|
| DCP            | 38.402| 0.0033 | -0.1609 |
| NCP            | 41.506| 0.0154 | -0.0220 |
| GCA            | 38.400| 0.0067 | -0.0284 |
| EPDN           | 38.632| 0.0009 | -0.0757 |
| Semi-dehazing  | 38.104| 0.0089 | 0.2124 |
| PSD            | 40.912| 0.0017 | 0.6085 |
| Ours           | 37.762| 0.0003 | 0.2248 |

TABLE I

Averaged PSNR, SSIM, and CIEDE2000 with state-of-the-art dehazing algorithms on synthetic datasets. The 1st rank are visually denoted in BOLD.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
improving the dehazing performance of the model. In particular, the introduction of uncertainty aids in the network’s adaptive focus on areas with uncertain haze predictions, thereby optimizing the dehazing results.

**TABLE III**
Ablation Analysis on Semi-UFormer.

| Variants   | Base | V1   | V2   | V3   |
|------------|------|------|------|------|
| MDB w/o    | ✓    | ✓    | ✓    | ✓    |
| Uncertainty w/o | w/o | ✓    | ✓    | ✓    |
| KL loss w/o | w/o  | w/o  | ✓    | ✓    |
| PSNR       | 24.51| 25.11| 25.93| 26.63|
| SSIM       | 0.900| 0.905| 0.914| 0.923|

The impact of different quantities of real data on Semi-UFormer’s dehazing performance. To investigate the relationship between the dehazing performance of Semi-UFormer and the quantity of real data, we conducted comprehensive experiments. Throughout the experimental process, we varied the number of real data samples in the unsupervised branch, setting them to 0, 500, 1000, and 2000, while keeping other training conditions unchanged. As shown in Tab. IV, Semi-UFormer trained with real data performs worse than the supervised learning model on the synthetic dataset. Moreover, the performance of Semi-UFormer on the synthetic dataset is minimally affected by the varying quantities of real data. However, Semi-UFormer trained with real data outperforms the supervised model on the real dataset, and the model’s performance improves with an increase in the quantity of real data.

**D. Runtime Analysis**

We compared the average running time of our Semi-UFormer with other methods on the HSTS dataset. All tests were conducted in the same environment consisting of an Intel(R) Core(TM) i9-10920X CPU, 32 GB RAM, and an NVIDIA GeForce RTX 3090 GPU. Tab. V shows the comparison results of the runtime for different methods.

**TABLE IV**
Quantitative evaluation of the impact of training with different quantities of real data on dehazing performance.

| The number of real images | SOTS outdoor | Real-world images |
|---------------------------|--------------|-------------------|
|                           | PSNR↑ | SSIM↑ | SSEQ↓ | σ↓ | HCC↑ |
| 0                         | 27.47 | 0.931 | 38.183 | 0.0016 | 0.0951 |
| 500                       | 26.27 | 0.928 | 38.192 | 0.0009 | 0.1016 |
| 1000                      | 26.67 | 0.930 | 38.133 | 0.0008 | 0.1819 |
| 1500                      | 26.03 | 0.925 | 38.071 | 0.0006 | 0.2089 |
| 2000                      | 26.63 | 0.923 | 37.762 | 0.0003 | 0.2248 |

**TABLE V**
Average runtime (in seconds) of different approaches tested on the HSTS dataset.

| Method       | Platform | Average time |
|--------------|----------|--------------|
| DCP          | Python(CPU) | 1.41          |
| NCP          | Python(CPU) | 1.45          |
| AOD          | PyTorch(GPU) | 0.11          |
| GCA          | PyTorch(GPU) | 0.21          |
| EPDN         | PyTorch(GPU) | 0.23          |
| GFN-IJCV     | PyTorch(GPU) | 0.43          |
| Semi-dehazing| PyTorch(GPU) | 0.32          |
| PSD          | PyTorch(GPU) | 0.39          |
| YOLY         | PyTorch(GPU) | 40.56         |
| Ours         | PyTorch(GPU) | 0.20          |

V. Conclusion

In this work, a novel semi-supervised uncertainty-aware transformer network called Semi-UFormer is proposed for image dehazing, which leverages both real-world data and uncertainty guidance information to facilitate the dehazing tasks. To bridge the gap between synthetic and real data, we build our Semi-UFormer on a knowledge distillation framework and apply a two-branch network to train our model on
both synthetic and real-world images. Moreover, we exploit an uncertainty estimation block (UEB) to predict the pixel uncertainty of the coarse dehazed results and then guide the network to better restore the image edges and structures. Experiments on both synthetic and real-world images fully validate the effectiveness of our Semi-UFormer.

ACKNOWLEDGMENT

This work is supported by the Basic Research for National Defense under Grant Nos. JCKY2020605C003.

REFERENCES

[1] Boyi Xiao, Jin Xie, and Jing Nie. Ddnet: Density and depth-aware network for object detection in foggy scenes. In IJCNN, pages 1–7. IEEE, 2023.

[2] Kaiming He, Jian Sun, and Xiaou Tang. Single image haze removal using dark channel prior. TPAMI, 33(12):2341–2353, 2010.

[3] Dana Berman, Shai Avidan, et al. Non-local image dehazing. In CVPR, pages 1674–1682, 2016.

[4] Boyi Li, Xiuliang Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In ICCV, pages 4770–4778, 2017.

[5] Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Dongdong Hou, Lu Yuan, and Gang Hua. Gated context aggregation network for image dehazing and deraining. In WACV, pages 1375–1383, 2019.

[6] Srinivasa G Narasimhan and Shree K Nayar. Vision and the atmosphere. IJCV, 48(3):233–254, 2002.

[7] Xu Qin, Zhihui Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In AAAI, volume 34, pages 11908–11915, 2020.

[8] Yuda Song, Zhaqing He, Hui Qian, and Xin Du. Vision transformers with holistic edges. In CVPR, pages 128:240–259, 2020.

[9] Yanyun Qu, Yizi Chen, Jingying Huang, and Yuan Xie. Enhanced pix2pix dehazing network. In IJCV, pages 8160–8168, 2019.

[10] Qian Ning, Weisheng Dong, Xin Li, Jinjian Wu, and Guangming Shi. Uncertainty-driven loss for single image super-resolution. NeurIPS, 34:16398–16409, 2021.

[11] Lucas Teixeira, Martin R Oswald, Marc Pollefeys, and Margarita Chli. Aerial single-view depth completion with image-guided uncertainty estimation. IEEE Robotics and Automation Letters, 5(2):1055–1062, 2020.

[12] Boyu Lu, Jingyi Liu, Guanyue Liu,Wenhao Song, Zhenqi Fu, Wu Wang, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Uncertainty inspired underwater image enhancement. In ECCV, pages 465–482. Springer, 2022.

[13] Yanyun Qu, Yizi Chen, Jingying Huang, and Yuan Xie. Enhanced pix2pix dehazing network. In IJCV, pages 1375–1383, 2019.

[14] Srinivasa G Narasimhan and Shree K Nayar. Vision and the atmosphere. IJCV, 48(3):233–254, 2002.

[15] Lucas Teixeira, Martin R Oswald, Marc Pollefeys, and Margarita Chli. Aerial single-view depth completion with image-guided uncertainty estimation. IEEE Robotics and Automation Letters, 5(2):1055–1062, 2020.

[16] Boyu Lu, Jingyi Liu, Guanyue Liu, Wenhao Song, Zhenqi Fu, Wu Wang, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Uncertainty inspired underwater image enhancement. In ECCV, pages 465–482. Springer, 2022.

[17] Zhenqi Fu, Wu Wang, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Uncertainty inspired underwater image enhancement. In ECCV, pages 465–482. Springer, 2022.

[18] Yanyun Qu, Yizi Chen, Jingying Huang, and Yuan Xie. Enhanced pix2pix dehazing network. In IJCV, pages 1375–1383, 2019.

[19] Srinivasa G Narasimhan and Shree K Nayar. Vision and the atmosphere. IJCV, 48(3):233–254, 2002.

[20] Boyi Li, Xiuliang Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In ICCV, pages 4770–4778, 2017.

[21] Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Dongdong Hou, Lu Yuan, and Gang Hua. Gated context aggregation network for image dehazing and deraining. In WACV, pages 1375–1383, 2019.

[22] Srinivasa G Narasimhan and Shree K Nayar. Vision and the atmosphere. IJCV, 48(3):233–254, 2002.