Trinity-Net: Gradient-Guided Swin Transformer-based Remote Sensing Image Dehazing and Beyond

Kaichen Chi, Yuan Yuan, *Senior Member, IEEE*, and Qi Wang, *Senior Member, IEEE*

***Abstract*—Haze superimposes a veil over remote sensing images, which severely limits the extraction of valuable military information. To this end, we present a novel trinity model to restore realistic surface information by integrating the merits of both prior-based and deep learning-based strategies. Concretely, the critical insight of our Trinity-Net is to investigate how to incorporate prior information into CNNs and Swin Transformer for reasonable estimation of haze parameters. Then, haze-free images are obtained by reconstructing the remote sensing image formation model. Although Swin Transformer has shown tremendous potential in the dehazing task, which typically results in ambiguous details. We devise a gradient guidance module that naturally inherits structure priors of gradient maps, guiding the deep model to generate visually pleasing details. In light of the generality of image formation parameters, we successfully promote Trinity-Net to natural image dehazing and underwater image enhancement tasks. Notably, the acquisition of large-scale remote sensing hazy images and natural hazy images in military scenes is not feasible in practice. To bridge this gap, we construct a *Remote Sensing Image Dehazing Benchmark* (RSID) and a *Natural Image Dehazing Benchmark* (NID), including 1000 real-world hazy images with corresponding ground truth images, respectively. To our knowledge, this is the first exploration to develop dehazing benchmarks in the military field, alleviating the dilemma of data scarcity. Extensive experiments on three vision tasks illustrate the superiority of our Trinity-Net against multiple state-of-the-art methods. The datasets and code are available at https://github.com/chi-kaichen/Trinity-Net.**

***Index Terms*—Image dehazing, swin transformer, haze thickness prior, image formation model.**

I. Introduction

IMAGES captured in hazy scenarios inevitably suffer from low contrast and poor visibility because of interference from turbid media (*e.g.*, haze, underwater, and sand) [1]–[3]. Such quality-degraded images are typically not available for computer vision applications, such as object detection [4], object tracking [5], image recognition [6], and semantic segmentation [7]. Thus, an effective dehazing method is of practical value for enhancing the perceptual quality of remote sensing images.

Over the past few years, a range of dehazing strategies has been developed, roughly divided into two categories: prior-based and deep learning-based methods. Early explorations

This work was supported by the National Natural Science Foundation of China under Grant U21B2041, 61825603, National Key R&D Program of China 2020YFB2103902. *(Corresponding author: Qi Wang.)*

K. Chi, Y. Yuan, and Q. Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: chikaichen@mail.nwpu.edu.cn, y.yuan.ieee@gmail.com, crabwq@gmail.com).

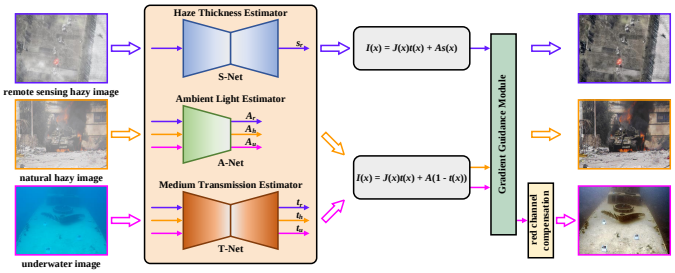


Fig. 1. The schematic illustration of Trinity-Net. Trinity-Net can achieve satisfactory performance on three computer vision tasks by incorporating the imaging mechanism into deep models.

mainly center around the atmospheric scattering model [1], [8], [9], mathematically formulated as:

$$I(x) = J(x)t(x) \oplus A(1 - t(x)), \tag{1}$$

where x represents the pixel location, $I(x)$ represents the hazy image, $J(x)$ represents the haze-free image, A and $t(x)$ denotes the haze parameters, *i.e.*, ambient light and medium transmission. Unfortunately, the estimation of haze parameters is an ill-posed problem because both ambient light and medium transmission are unknown. Diverse prior assumptions have been designed to tackle this problem, such as dark channel prior (DCP) [1], haze-lines [10], elliptical boundary prior [11], and dark-object subtraction [12]. However, prior-based methods tend to produce unsatisfactory dehazing results when facing dense and non-homogeneous hazy scenes. This is because the accurate estimation of multiple haze parameters is knotty for prior-based methods.

Currently, deep learning technology has achieved significant progress in dehazing [13], [14]. Some end-to-end models have been presented to alleviate the deep dependence on predefined prior information, but they usually have limitations in interpretability [15]. In addition, some deep networks follow the atmospheric scattering model, and utilize the powerful feature representation to obtain ambient light and medium transmission [16]–[18]. However, these methods typically apply deep models originally devised for natural image dehazing to remote sensing images, which do not work well in some cases [11]. This is because the design of current dehazing models ignores the haze parameter specific to the remote sensing image dehazing task, *i.e.*, haze thickness.

In this work, we present Trinity-Net, which consists of three parameter estimator networks, *i.e.*, haze thickness estimator (S-

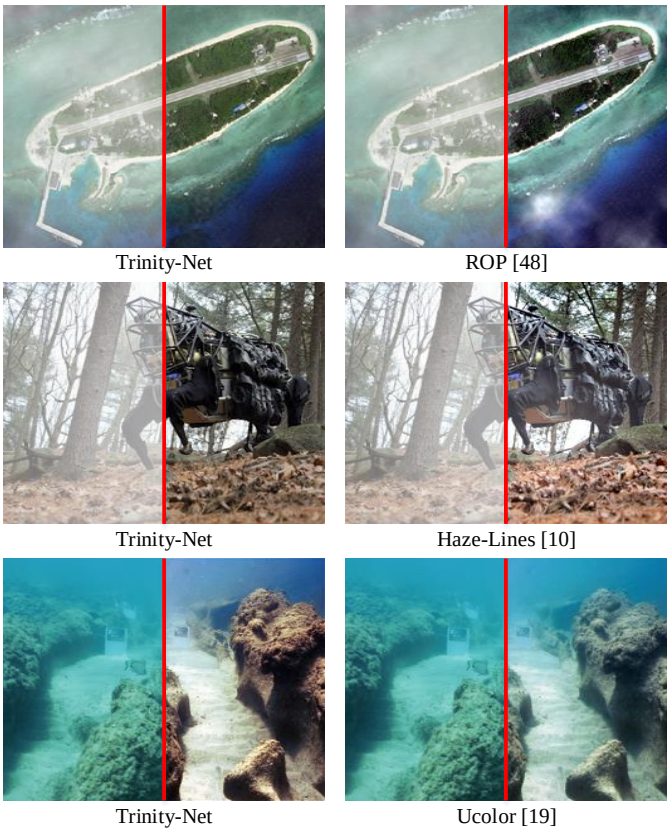


Fig. 2. Visual comparisons on remote sensing hazy, natural hazy, and underwater images. The quality-degraded images and enhanced results are on either side of the red line.

Net), ambient light estimator (A-Net), and medium transmission estimator (T-Net). The schematic illustration is depicted in Fig. 1. Concretely, we attempt to integrate the remote sensing image formation model and deep learning strategy, taking the best of both worlds to restore the visibility of remote sensing images. To this end, we incorporate the dark channel prior into Swin Transformer to calculate the haze thickness through nonoverlapping local windows. Meanwhile, multi-scale information from each level of Swin Transformer is adaptively extracted and aggregated to deal with inaccurate parameter estimation due to fixed patch sizes. Such a manner learns the spatial distribution of haze. To downplay the interference of non-hazy bright pixels (*e.g.*, flames), a local minimum filtering is embedded in A-Net. Coupled with the feature attention mechanism, ambient light-related feature representations are highlighted adaptively. More importantly, ambient light is intuitively associated with high-intensity pixels, which accurately reflects the physical properties of ambient light. Simultaneously, we leverage the complementarity of fine-scale and coarse-scale information flow to estimate the transmission map. Since Trinity-Net is purely data-driven, our method is tolerant of errors due to inaccurate haze parameter estimation [19]. Besides, inspired by the conclusion that gradient maps help deep models concentrate more on geometric structures [20], we devise a gradient guidance module to unearth richer local details. Without bells and whistles, the introduction of gradient maps allows us to overcome the inadequacy of the local modeling capability of Swin Transformer [21], [22].

Due to the flexibility of the well-designed architecture in parameters estimation and application, we extend Trinity-Net to handle other related computer vision tasks. In Fig. 2, we show examples of remote sensing image dehazing, natural image dehazing, and underwater image enhancement tasks. The compared methods either cannot remove the haze or maintain the greenish color deviation. In contrast, Trinity-Net achieves visually promising results in several scenarios. In a nutshell, the main contributions of this article are as follows.

- We contribute two large-scale real-world dehazing benchmarks (*i.e.*, RSID and NID). With the constructed benchmarks, deep models can be easily trained. Besides, RSID and NID provide a platform that can be employed to evaluate the performance of various dehazing strategies.
- We establish a general and reasonable learning framework called Trinity-Net, which effectively restores realistic surface information from dense and non-homogeneous hazy remote sensing images. Trinity-Net embeds the remote sensing image formation model into the deep model to jointly learn haze parameters. More importantly, the proposed method is the first to incorporate the power of Swin Transformer into haze parameters estimation compared with pure CNN-based dehazing models.
- We propose a gradient guidance module to provide additional structure priors and enforce the deep model to generate richer perceptual details. It explores the complementary advantages between structure prior knowledge and Swin Transformer.
- As a crucial byproduct, we promote Trinity-Net to tackle related vision tasks, such as natural image dehazing and underwater image enhancement, which verifies the flexibility and generality of our work.
- Our Trinity-Net achieves nontrivial performance on three computer vision tasks for both visual perception and evaluation metrics.

II. Related Work

Prior-based Methods. To reconstruct under-constrained image formation models, prior-based methods employ certain handcrafted priors to estimate haze parameters. Berman *et al.* [23] assumed that hundreds of tight color clusters approximate the colors of clear scenarios, then proposed a haze-lines prior for dehazing. Peng *et al.* [8] employed regression analysis to calculate the depth-dependent color casts, then utilized light differential to estimate the medium transmission. In [24], a sphere model was designed to accurately estimate the medium transmission, which helps to remove the non-homogeneous haze and thin cloud. Shen *et al.* [25] proposed a globally non-uniform ambient light model to predict spatially varied ambient light and designed a bright pixel index to correct the transmission. With the predicted haze parameters, they reversed the atmospheric scattering model to restore visibility.

Although prior-based methods can improve the perceptual quality, handcrafted prior assumptions employed do not always hold [19]. In addition, the accurate estimation of haze parameters in complex environments challenges current prior-based

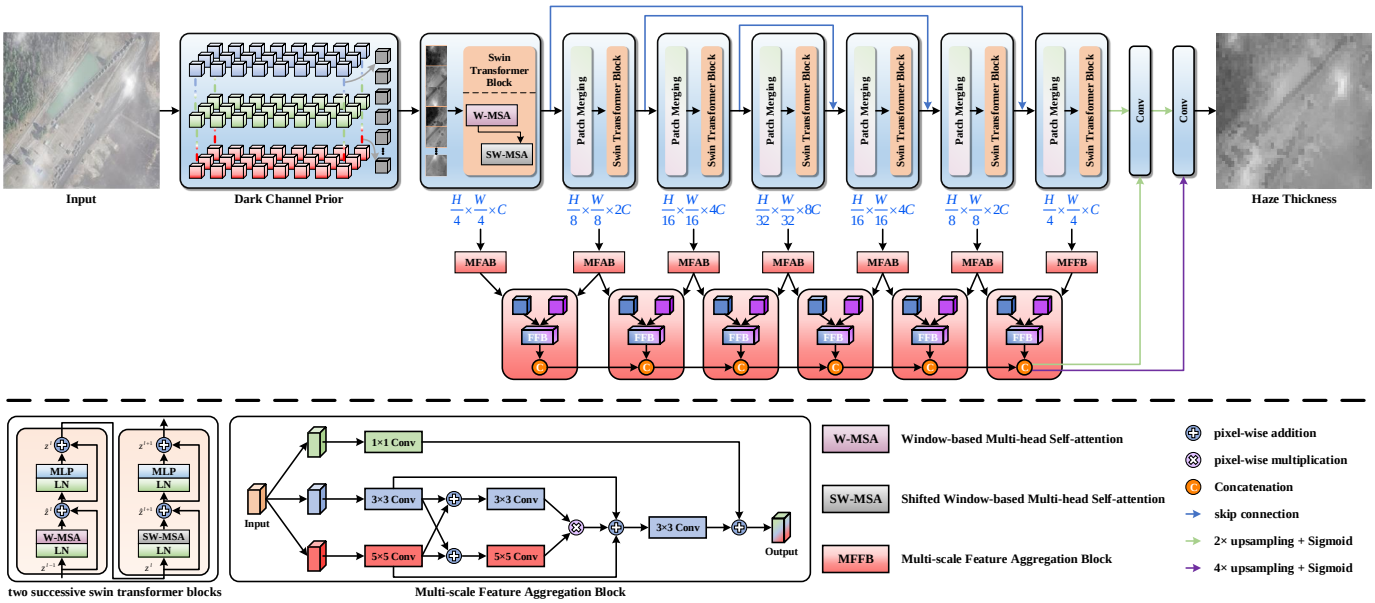


Fig. 3. Detail structure of the haze thickness estimator network. The haze thickness estimator network searches dark pixels using the dark channel prior. After patch partition, the nonoverlapping dark pixel patches are fed to two successive swin transformer blocks to estimate the haze thickness map through shifted windows. Meanwhile, considering that the patch size affects the performance of the dark channel prior, information at different scales is extracted and aggregated to guide Swin Transformer to select the patch size adaptively.

methods [8], [23]–[25]. In contrast, Trinity-Net can accurately estimate haze parameters by leveraging the complementary advantages of prior-based and deep learning-based methods.

Deep Learning-based Methods. With the success of CNNs, there are several attempts to estimate haze parameters through deep learning technology [26], [27]. In [16], the densely connected pyramid was used to jointly estimate ambient light and medium transmission. Built on the layer disentanglement, Li *et al.* [17] designed a zero-shot model, which performs regularization on the estimation of haze parameters to improve the stability of deep models. Nie *et al.* [28] explored the correlation between binocular images to alleviate the parameter estimation errors caused by depth variations. Ullah *et al.* [29] proposed a mathematical expression of haze imaging to reconstruct haze-free scenarios.

These deep learning models typically apply the atmospheric scattering model for natural hazy images to remote sensing hazy images and ignore the spatial distribution of haze. Specifically, since the haze thickness in remote sensing scenes does not vary with depth, the commonly used atmospheric scattering model cannot accurately characterize the formation process of haze-affected remote sensing images [3], [11]. To reasonably describe haze imaging, a remote sensing image formation model was proposed [11]:

$$I(x) = J(x)t(x) \oplus As(x), \tag{2}$$

where $s(x)$ represents the haze thickness and describes the structured haziness. Some traditional methods propose haze thickness estimation priors to obtain this key parameter [11], [12]. However, the adjustable variables in traditional methods need to be well-designed for different scenarios, and their robustness is limited and unsatisfactory. Therefore, the inspired design in this study aims to employ prior information to provide guidance for global context modeling capability toward the reasonable estimation of haze thickness. This is rarely explored in the context of remote sensing image dehazing.

Compared with existing learning-based methods, Trinity-Net provides the following unique characteristics: 1) with both the global modeling ability of Swin Transformer and the local representation ability of CNNs, Trinity-Net can accurately learn haze parameters; 2) the plug-and-play gradient guidance module alleviates the problem that Swin Transformer does not possess the locality by incorporating structure priors into deep models; 3) Trinity-Net does not require pre-processing strategies and employs a supervised learning manner, thereby restoring more stable results; 4) Trinity-Net achieves remarkable performance on several vision applications. These innovations contribute a promising direction for investigating the complementary advantages of deep learning strategy and domain knowledge of remote sensing imaging.

III. Proposed Method

In this section, we detail Trinity-Net, which learns haze parameters in Eq. (2) by building three parameter estimator networks. Additionally, we design a gradient guidance module to enforce the detail restoration for vision tasks in hazy scenarios. In what follows, we elaborate on the key components of Trinity-Net, including the haze thickness estimator, the ambient light estimator, the medium transmission estimator, the gradient guidance module, and the loss function.

A. Haze Thickness Estimator Network

Haze thickness is typically variable in the scenario, creating a challenge for its calculation. In view of the fact that Swin Transformer performs self-attention within nonoverlapping

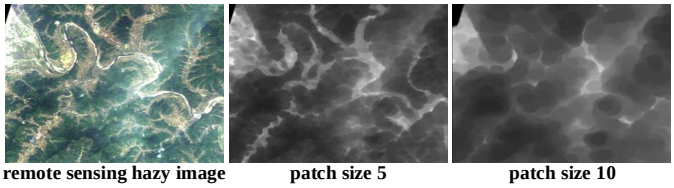


Fig. 4. Visual comparisons on dark channels produced at patch sizes 5 and 10, respectively.

local windows for efficient modeling, consistent with the dark pixel-based haze thickness estimation strategy [12]. Inspired by this, we embed the haze thickness-related prior reflected by dark pixels into the Swin Transformer backbone to learn the haze thickness map, as shown in Fig. 3. Concretely, the dark channel of remote sensing images can be expressed as:

$$DCP(I) = \min_{y\in\Omega(x)} \left(\min_{c\in\{r,g,b\}} I^c(y) \right), \tag{3}$$

where $\Omega(x)$ represents a patch centered at x, and $\min_{c\in\{r,g,b\}}$ is performed to search the minimum color channel (RGB) values. However, DCP tends to produce unstable prior information in bright scenarios. The main reason is that DCP employs a fixed patch size to search dark pixels [30]. As shown in Fig. 4, a small patch size generates a detailed dark channel, but may lead to non-zero dark channel values in regions obscured by haze. In comparison, a large patch size generates a coarse dark channel that ensures the dark channel is zero, but fails to present rich prior information [30]. To guide S-Net adaptively select the patch size, we establish a CNN backbone parallel to the Swin Transformer backbone, which accurately predicts the patch map through a hybrid learning strategy integrating feature extraction and feature fusion.

The framework of the Swin Transformer backbone consists of a patch partition and seven swin transformer blocks. In addition, it establishes skip paths between the corresponding encoder and decoder feature maps to enhance the communication from low-level visual information to high-level semantic information. Patch partition is used to divide $DCP(I)$ into nonoverlapping patches. The swin transformer block is connected sequentially by a window-based multi-head self-attention block (W-MSA) and a shift-window-based multi-head self-attention block (SW-MSA). In W-MSA, the feature $\mathbf{z}^{l-1}$ goes through successive Layer Normalization (LN) and W-MSA layers, coupled with a skip connection to acquire the output feature $\hat{\mathbf{z}}^l$. Then, $\hat{\mathbf{z}}^l$ goes through successive LN and multi-layer perceptron (MLP) layers, finally acquiring the output feature $\mathbf{z}^l$ via another skip connection. The framework of the SW-MSA block is similar to that of the W-MSA block, except that a shifted window partitioning strategy is embedded in the SW-MSA layer to enable the modeling of global spatial dependencies. The process can be described as:

$$\begin{aligned}
\hat{\mathbf{z}}^l &= \mathrm{W_{MSA}}(\mathrm{LN}(\mathbf{z}^{l-1})) \oplus \mathbf{z}^{l-1}, \\
\mathbf{z}^l &= \mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{z}}^l)) \oplus \hat{\mathbf{z}}^l, \\
\hat{\mathbf{z}}^{l+1} &= \mathrm{SW_{MSA}}(\mathrm{LN}(\mathbf{z}^l)) \oplus \mathbf{z}^l, \\
\mathbf{z}^{l+1} &= \mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{z}}^{l+1})) \oplus \hat{\mathbf{z}}^{l+1},
\end{aligned} \tag{4}$$

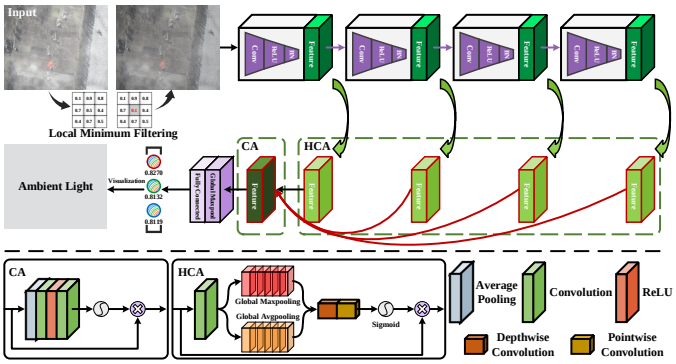


Fig. 5. Detail structure of the ambient light estimator network. A local minimum filtering is embedded in the ambient light estimator network to remove non-hazy bright pixels. Moreover, a feature attention mechanism is used to emphasize features in haze-affected regions and suppress less useful features, thus estimating ambient light robustly from haze-affected regions.

The framework of the CNN backbone consists of multi-scale feature aggregation blocks and feature fusion blocks. Inspired by [30], the multi-scale feature aggregation block generates a special scale patch map through multi-scale technology, since the patch size is related to the scale information [30]. Concretely, we feed the output item of SW-MSA to a three-stream network with diverse kernels, and employ feature aggregation to make the multi-scale information flow and the original feature information flow complement each other. The process can be described as:

$$\begin{aligned}
f_{3\times3} &= \delta(Conv_3(Conv_3(\mathbf{z}^{l+1}) \oplus Conv_5(\mathbf{z}^{l+1}))), \\
f_{5\times5} &= \delta(Conv_5(Conv_5(\mathbf{z}^{l+1}) \oplus Conv_3(\mathbf{z}^{l+1}))), \\
f_{fuse} &= f_{3\times3} \otimes f_{5\times5} \oplus Conv_5(\mathbf{z}^{l+1}) \oplus Conv_3(\mathbf{z}^{l+1}), \\
f_{agg} &= \delta(Conv_3(f_{fuse})) \oplus \delta(Conv_3(\mathbf{z}^{l+1})),
\end{aligned} \tag{5}$$

where $\delta(\cdot)$ denotes the ReLU layer. Then, aggregated features corresponding to adjacent swin transformer blocks are fed to the feature fusion block (T-Net presents the detail structure of FFB), which copes with scale variation by leveraging the cross-level correlations. Finally, the fused features are concatenated and propagated to the sigmoid function as a feature-level attention map to guide Swin Transformer adaptively select the patch size. Thus, S-Net reasonably estimates haze thickness by exploring the complementary advantages of prior knowledge, patch optimization, and local-window self-attention.

B. Ambient Light Estimator Network

Non-hazy bright objects such as flames lead to inaccurate estimation of ambient light, resulting in overexposed or underexposed results [9]. To this end, we devise an ambient light estimator network that inserts a local minimum filtering to downplay non-hazy bright pixels, as depicted in Fig. 5:

$$g(x,y) = \min_{(s,t)\in I_{x,y}} \{I(s,t)\}, \tag{6}$$

where $I_{x,y}$ represents the neighborhood window centered on pixel $I(x,y)$ and $g(x,y)$ denotes the filtered pixel. The intuition behind our method is to replace the grayscale value of pixel (x,y) with the minimum value within the neighborhood

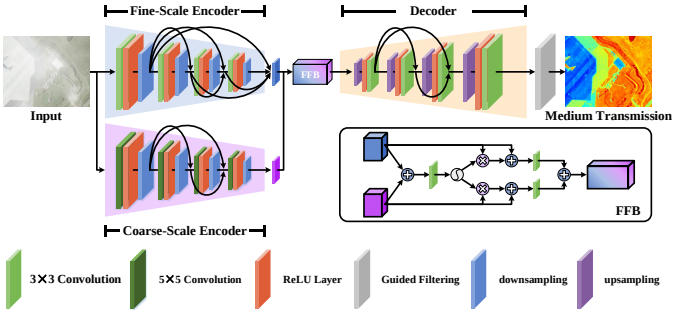


Fig. 6. Detail structure of the medium transmission estimator network. The medium transmission estimator network integrates both local visual cues and global contextual information to learn the medium transmission map.

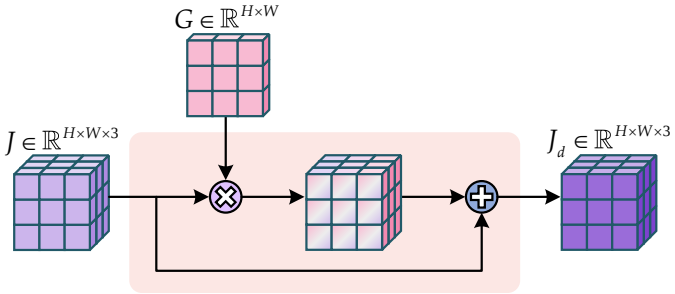


Fig. 7. Detail structure of the gradient guidance module. The gradient map as a feature selector is utilized to unearth more generic structure priors, thus obtaining dehazing images with rich details and fewer geometric distortions.

window. Considering that ambient light typically lies in the farthest region and that these are typically the haze-affected region [31]. We devise a feature attention mechanism to treat different pixels unequally and focus more attention on haze-affected regions, which consists of hybrid convolutional attention and channel attention. Concretely, we feed the hierarchical features captured by hierarchical learning to the hybrid convolutional attention block. These high-level features are rich in semantic information that contributes to preserving luminance and color details, but also retain similar or redundant features [32]. Inspired by [15], the hybrid convolutional attention block captures common and distinctive semantic information through global average pooling and global maximal pooling, then employs depth-wise separable convolution to avoid information loss caused by pooling layers while spatially transforming redundant features into valuable ones. Such a manner is able to adaptively highlight haze-related features:

$$\begin{aligned} W_{hca} &= \sigma(W_p\delta(W_d f_{avg}) \oplus W_p\delta(W_d f_{max})), \\ f_{hca} &= W_{hca} \otimes g, \end{aligned} \tag{7}$$

where $\sigma(\cdot)$ denotes the sigmoid function, f_{avg} and f_{max} denote the average pooling features and maximal pooling features, $W_p \in R^{C\times 1}$ and $W_d \in R^{1\times C}$ denote the depth-wise separable convolution weights, and $W_{hca} \in R^{H\times W}$ denotes the hybrid convolutional attention map. Then, we extend the output items of the hybrid convolutional attention block at different scales from spatial space to channel space, bringing rich information compensation:

$$\begin{aligned} W_{ca} &= \sigma(\mathrm{Conv}(\delta(\mathrm{Conv}(\mathrm{AP}(f^i_{hca}))))), \quad 1 \le i \le 4, \\ f_{ca} &= W_{ca} \otimes f^i_{hca}, \end{aligned} \tag{8}$$

where $\mathrm{AP}(\cdot)$ denotes the average pooling function. Considering that ambient light is an aggregation of single values of color channels, we employ a global maximal pooling at the end of the A-Net to intuitively pool single maximum intensity. Unlike previous encoder-decoder frameworks [16]–[18], such a manner associates ambient light with high-intensity pixels, consistent with the high luminance properties of ambient light.

C. Medium Transmission Estimator Network

Medium transmission indicates the percentage of scene radiation that reaches the sensor after reflecting in turbid media, which represents the degree of quality degradation for diverse areas. However, the ground truth medium transmission maps of remote sensing images are not available, training deep models for estimating medium transmission is a challenge. Inspired by the zero-shot learning that avoids employing information beyond the remote sensing image itself covered [17], we devise a fine-coarse scale complementary framework to estimate the medium transmission map, as shown in Fig. 6. In the encoder network, on the one hand, the fine-scale path employs densely connected modules with small kernels to capture fine-scale structures to refine structural details of the medium transmission. On the other hand, the coarse-scale path employs larger kernels to capture long-range context that complements the fine-scale branch through a feature fusion strategy. At the end of decoder network, we incorporate guided filtering into the T-Net to suppress unwanted textures while preserving edges. With the haze thickness s, ambient light A, and medium transmission t, haze-free images are acquired by reversing the remote sensing image formation model:

$$J(x) = \frac{I(x) - As(x)}{\max(t(x), t_0)}, \tag{9}$$

where t_0 is empirically set to 0.1 to increase exposure.

D. Gradient Guidance Module

Although Swin Transformer has shown impressive results in image restoration tasks by exploiting self-attention to model long-range dependencies, it is difficult to perform local modeling [33], [34]. To solve this issue, we propose a simple but effective gradient guidance module, which aims at enjoying mutual benefits between structure priors and Swin Transformer. In Fig. 7, we employ the Sobel operator to calculate the gradient map of remote sensing images and integrate it directly into the gradient guidance module. Then, we use the gradient map as the pixel-wise attention map, assigning larger attention weights to high-quality degradation pixels, and highlighting regions that require more attention to sharpness, thus enforcing the deep network to generate haze-free images with clear details:

$$J_d = J \oplus J \otimes G, \tag{10}$$

where G denotes the gradient map and J_d denotes the haze-free image with perceptual-pleasant details. We treat attention

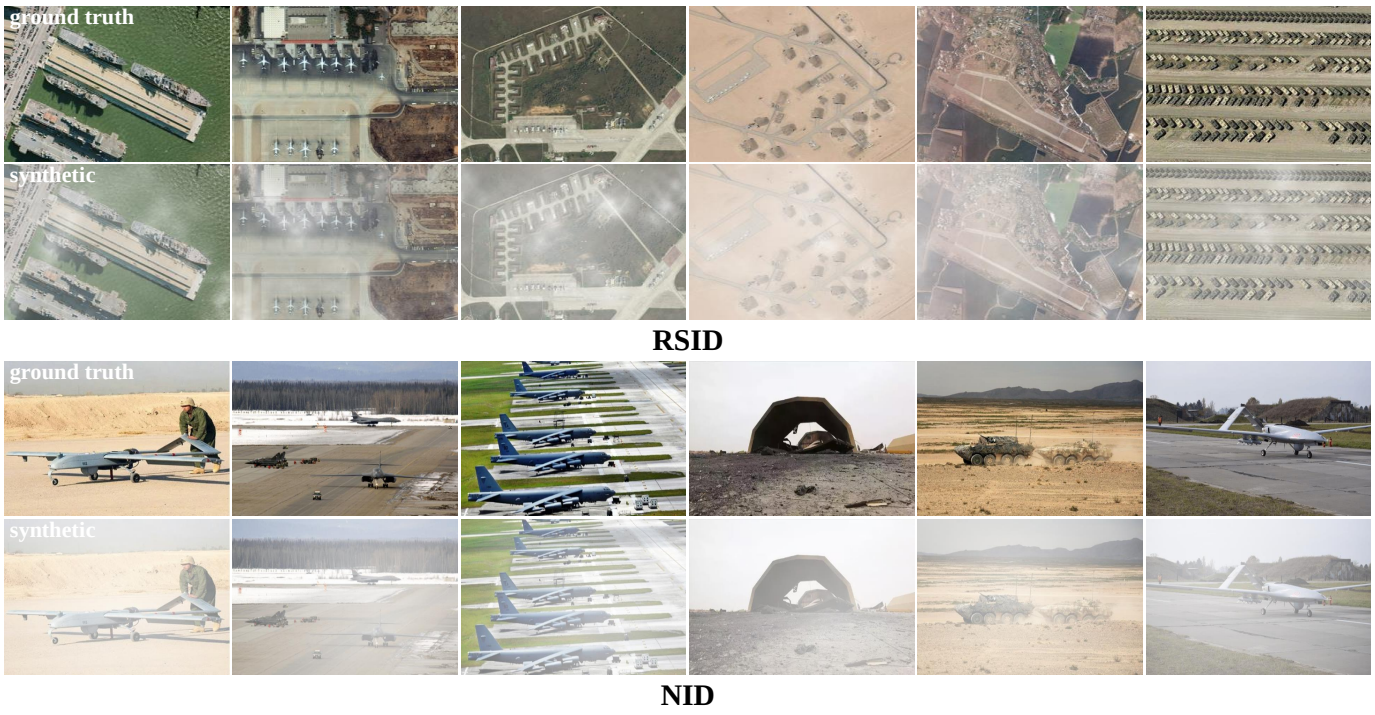


Fig. 8. Samples from the RSID and NID datasets. These images suffer from the obvious dense and non-homogeneous haze and are captured in a diversity of military scenes.

weights as identity connections to tolerate detail loss caused by the lack of local perception capability. Therefore, our gradient guidance module and Swin Transformer are complementary.

E. Loss Function

Following the previous work [35], we employ the linear combination of ℓ_1 loss $\mathcal{L}_{\ell_1}$, MS-SSIM loss $\mathcal{L}_M$, and gradient loss $\mathcal{L}_G$ to balance both visual perception and quantitative assessments, and the overall loss $\mathcal{L}_{total}$ is described as:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{\ell_1} + \lambda_2 \mathcal{L}_M + \lambda_3 \mathcal{L}_G, \tag{11}$$

where λ_1, λ_2, and λ_3 are empirically set to 1, 2, and 1 to balance the scales of multiple losses. Concretely, the ℓ_1 loss is applied to pixel-level blocks for calculating the ℓ_1 distance between the dehazing image J_d and the corresponding ground truth image $\hat{J}$:

$$\mathcal{L}_{\ell_1} = \| J_d - \hat{J} \|_1 . \tag{12}$$

The MS-SSIM loss integrates the variations of resolution and visualization conditions to consider structural differences:

$$\mathcal{L}_M(J_d, \hat{J}) = 1 - \text{MS-SSIM}(J_d, \hat{J}). \tag{13}$$

Since the gradient map can intuitively reflect structural information of remote sensing images, we employ the gradient loss as a second-order constraint, which helps to refine the edge of dehazing results:

$$\mathcal{L}_G = \mathbb{E}_{G\sim Q(d), \hat{G}\sim Q(g)} \| G - \hat{G} \|_1 . \tag{14}$$

where $\hat{G}$ and G represent the gradient map of $\hat{J}$ and J_d, respectively. $Q(g)$ and $Q(d)$ represent the distribution of $\hat{G}$ and G, respectively.

F. Extensions on Other Vision Tasks

Different image enhancement problems typically share some similar image formation parameters [36]. Following this idea, we extend the proposed Trinity-Net to two different applications, including natural image dehazing and underwater image enhancement. As is well known, both vision tasks employ the atmospheric scattering model to model the quality degradation process [1]. Similar to Eq. (2), the solution of the atmospheric scattering model requires two factors, *i.e.*, ambient light A and medium transmission t. With the solution strategy already devised, solving two factors is similar to remote sensing image dehazing. After that, the final enhancement images are restored by:

$$J(x) = \frac{I(x) - A}{\max(t(x), t_0)} + A, \tag{15}$$

Note that the attenuation process of light in underwater scenes selectively affects the wavelength spectrum, resulting in the greenish-bluish appearance of underwater images [37]. Inspired by the conclusion that the green channel contains color information from opponent channels [37], we propose to add a portion of green channel to red channel for compensating the loss of red channel:

$$J_{water}(x) = J_R(x) + (\bar{J}_G - \bar{J}_R) \otimes (1 - J_R(x)) \otimes J_G(x). \tag{16}$$

where J_R and J_G denote red and green color channels of image J after normalization, respectively. $\bar{J}_G$ and $\bar{J}_R$ represent the mean value of J_R and J_G. The operation of red channel compensation will increase the scene adaptability of our Trinity-Net to some extent.

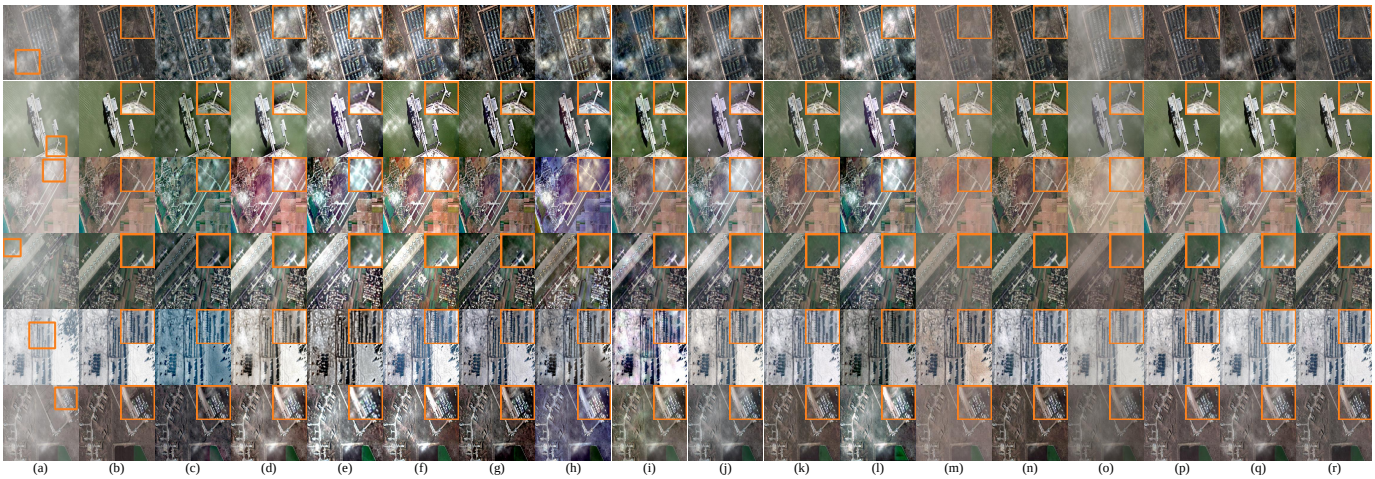


Fig. 9. Visual comparisons on synthetic remote sensing hazy images sampled from **R100** dataset. (a) input. (b) ground truth. (c) SDCP [24]. (d) ROP [48]. (e) STD [49]. (f) IDeRs [50]. (g) EVPM [51]. (h) GRS-HTM [52]. (i) ZID [17]. (j) zero-restore [18]. (k) FCTF-Net [53]. (l) TCN [54]. (m) dehaze-cGAN [55]. (n) FFA-Net [31]. (o) Cycle-SNSPGAN [56]. (p) UHD [13]. (q) DeHamer [21]. (r) Trinity-Net.

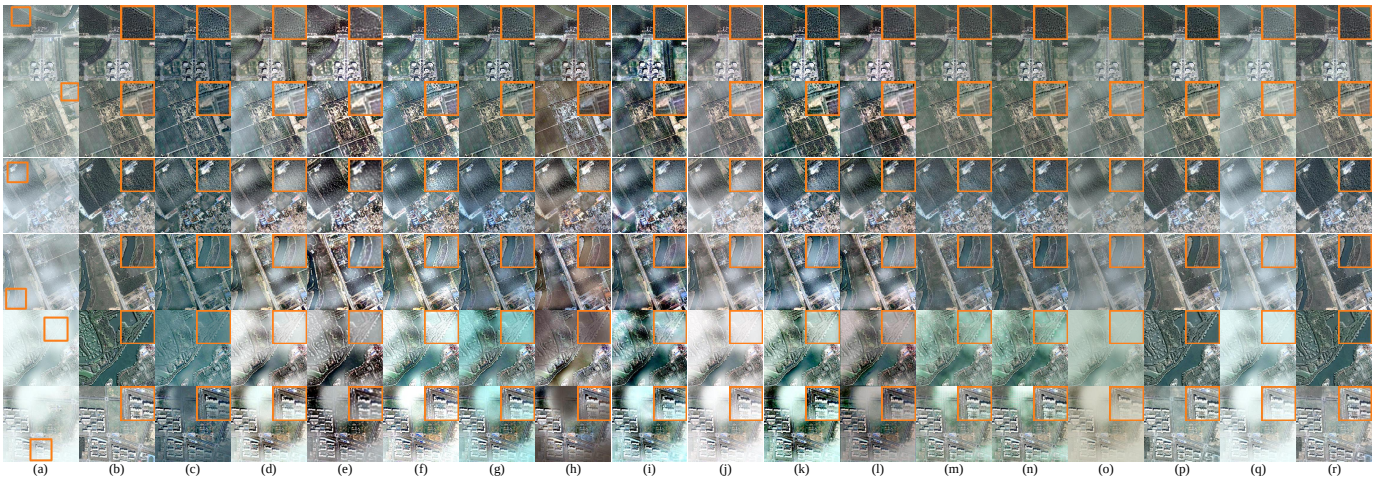


Fig. 10. Visual comparisons on synthetic remote sensing hazy images with different haze densities sampled from **S100** dataset. (a) input. (b) ground truth. (c) SDCP [24]. (d) ROP [48]. (e) STD [49]. (f) IDeRs [50]. (g) EVPM [51]. (h) GRS-HTM [52]. (i) ZID [17]. (j) zero-restore [18]. (k) FCTF-Net [53]. (l) TCN [54]. (m) dehaze-cGAN [55]. (n) FFA-Net [31]. (o) Cycle-SNSPGAN [56]. (p) UHD [13]. (q) DeHamer [21]. (r) Trinity-Net.

IV. Experiments

A. Data Generation

For the training of deep networks, a critical issue is how to capture pairs of hazy data and corresponding ground truth data with a consistent scenario, viewpoint, and time. To alleviate the data shortage, we collected a large number of real-world remote sensing and natural images from Google and related papers [38]–[40] in military scenes, covering ships, airfields, *etc*. Then, we employed hazy image synthesis algorithms to construct RSID and NID datasets. Some example images of the constructed datasets are presented in Fig. 8.

RSID dataset. Following [3], we synthesized remote sensing hazy images based on Eq. (2). Based on extensive statistics of haze parameters, the ambient light $A \in [0.7, 1]$, the medium transmission $t \in [0.25, 0.65]$, and the haze thickness $s \in [0.35, 0.75]$ were randomly sampled to generate poor visibility samples. Each hazy sample has a resolution of 256×256.

NID dataset. Natural hazy image synthesis algorithms rely on the pre-estimation of depth information. Thus, we first performed a depth prediction algorithm [41] to estimate depth maps of natural images, then synthesized natural hazy images based on a haze simulation strategy [42]. Each natural hazy image also has a resolution of 256×256.

B. Implementation Details

For training, we randomly selected 900 image pairs from RSID. Besides, we incorporated 860 image pairs from SateHaze1k [43] to enable the training data to cover diverse scenes and comprehensive image content. SateHaze1k provides 960 synthetic images with different haze densities and corresponding ground truth images, including thin haze, moderate haze, and dense haze. Please note that we cropped images to size of 128×128 to augment the training data.

The Trinity-Net was implemented using PyTorch and trained on an NVIDIA RTX 3090 GPU. We used ADAM for

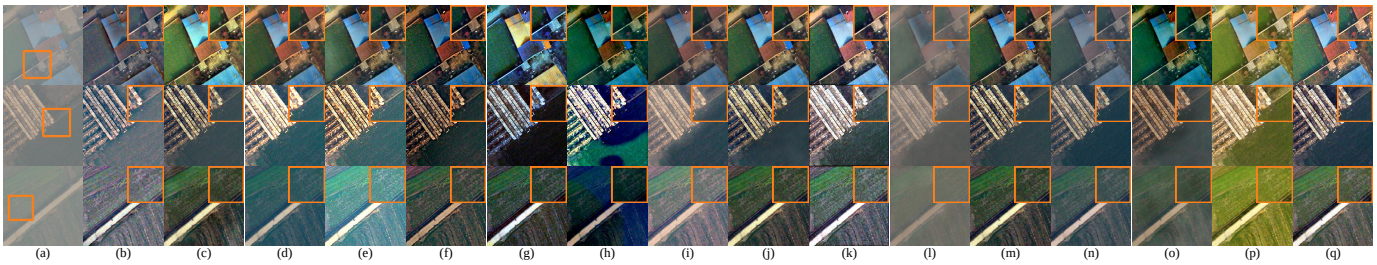


Fig. 11. Visual comparisons on real-world remote sensing hazy images sampled from **UAV** dataset. (a) input. (b) SDCP [24]. (c) ROP [48]. (d) STD [49]. (e) IDeRs [50]. (f) EVPM [51]. (g) GRS-HTM [52]. (h) ZID [17]. (i) zero-restore [18]. (j) FCTF-Net [53]. (k) TCN [54]. (l) dehaze-cGAN [55]. (m) FFA-Net [31]. (n) Cycle-SNSPGAN [56]. (o) UHD [13]. (p) DeHamer [21]. (q) Trinity-Net.

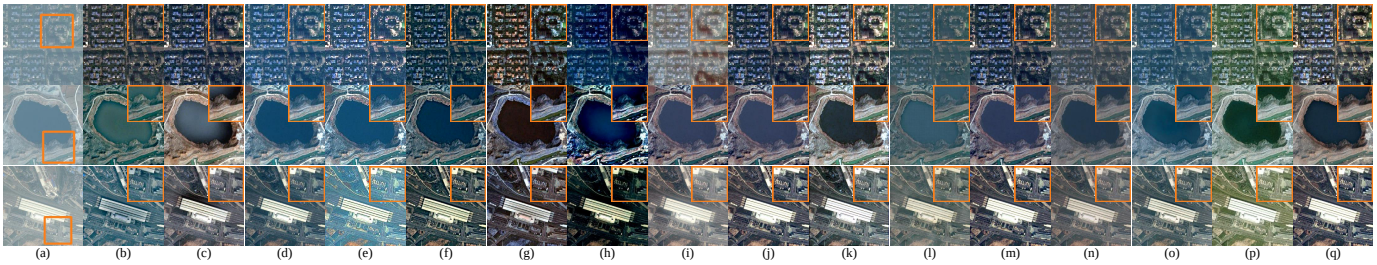


Fig. 12. Visual comparisons on real-world remote sensing hazy images sampled from **RS-Three** dataset. (a) input. (b) SDCP [24]. (c) ROP [48]. (d) STD [49]. (e) IDeRs [50]. (f) EVPM [51]. (g) GRS-HTM [52]. (h) ZID [17]. (i) zero-restore [18]. (j) FCTF-Net [53]. (k) TCN [54]. (l) dehaze-cGAN [55]. (m) FFA-Net [31]. (n) Cycle-SNSPGAN [56]. (o) UHD [13]. (p) DeHamer [21]. (q) Trinity-Net.

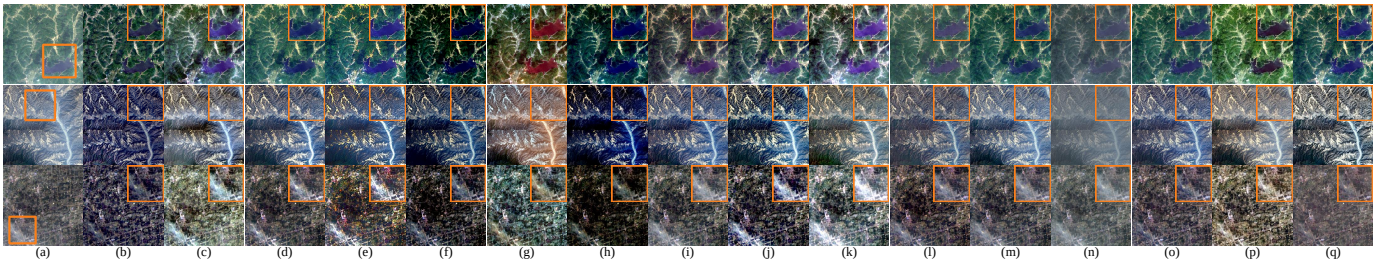


Fig. 13. Visual comparisons on real-world remote sensing hazy images sampled from **Landsat-8** dataset. (a) input. (b) SDCP [24]. (c) ROP [48]. (d) STD [49]. (e) IDeRs [50]. (f) EVPM [51]. (g) GRS-HTM [52]. (h) ZID [17]. (i) zero-restore [18]. (j) FCTF-Net [53]. (k) TCN [54]. (l) dehaze-cGAN [55]. (m) FFA-Net [31]. (n) Cycle-SNSPGAN [56]. (o) UHD [13]. (p) DeHamer [21]. (q) Trinity-Net.

model optimization and set the initial learning rate to $5 \times e^{-4}$. The batch size and epoch were set to 4 and 300, respectively. In terms of hyperparameters, the window size of the swin transformer block was 8 and the number of multi-head was 8.

C. Experiment Settings

Benchmarks. To test Trinity-Net, we utilized the rest 100 image pairs from RSID, denoted as **R100** dataset, while the rest 100 image pairs from SateHaze1k, denoted as **S100** dataset. Besides, we performed comprehensive experiments on the **UAV** [44], **RS-Three**, and **Landsat-8** datasets. UAV is a real-world benchmark containing 150 remote sensing hazy images captured by unmanned aerial vehicles. RS-Three integrates 200 remote sensing hazy images from three real-world benchmarks (*i.e.*, NWPU-RESISC45 [45], RSICD [46], and RSOD [47]). Landsat-8 contains 50 hazy images cropped from the Landsat-8 Operational Land Imager.

Compared Methods. We compared Trinity-Net with the following state-of-the-art methods.

- **Prior-based methods:** SDCP [24], ROP [48], STD [49], IDeRs [50], EVPM [51], GRS-HTM [52].
- **Deep learning-based methods:** ZID [17], zero-restore [18], FCTF-Net [53], TCN [54], dehaze-cGAN [55], FFA-Net [31], Cycle-SNSPGAN [56], UHD [13], DeHamer [21].

Evaluation Metrics. For R100 and S100 datasets, we performed full-reference evaluations utilizing PSNR, SSIM, MSE, FSIM, and FSIMc. A higher PSNR and SSIM value or a lower MSE value indicates that dehazing images are closer to corresponding ground truth images regarding image content. FSIM and FSIMc incorporate phase congruency and gradient magnitude to measure feature similarity, and a higher score denotes better visual quality. For UAV, RS-Three, and Landsat-8 datasets, we performed no-reference evaluations utilizing FADE [57] and Entropy. A lower FADE score indicates better visibility, while a higher Entropy score indicates that the image presents more detail.

TABLE I
THE AVERAGE PSNR (DB), SSIM, MSE ($\times 10^3$), FSIM, FSIMC, FADE, AND ENTROPY SCORES ON **R100**, **S100**, **UAV**, **RS-THREE**, AND **LANDSAT-8** DATASETS. THE BEST SCORE IS IN **RED**, THE SECOND-BEST IS IN **GREEN**, AND THE THIRD-BEST IS IN **BLUE**.

Methods	Publication	R100					S100					UAV		RS-Three		Landsat-8	
		PSNR ↑	SSIM ↑	MSE ↓	FSIM ↑	FSIMc ↑	PSNR ↑	SSIM ↑	MSE ↓	FSIM ↑	FSIMc ↑	FADE ↓	Entropy ↑	FADE ↓	Entropy ↑	FADE ↓	Entropy ↑
SDCP [24]	GRSL'19	16.0546	0.6913	2.0014	0.8604	0.8531	15.7528	0.7818	1.9563	0.9106	0.8997	0.5012	6.2653	0.2870	7.0512	0.1279	6.9234
ROP [48]	TPAMI'22	15.5747	0.7500	2.5020	0.8794	0.8754	15.9837	0.8299	2.3627	0.9082	0.9026	0.3658	7.2588	0.3083	7.4788	0.2091	7.3784
STD [49]	GRSL'21	16.2578	0.5592	3.3235	0.7740	0.7681	15.9349	0.6159	2.0537	0.7943	0.7876	0.4717	6.7680	0.2775	7.2853	0.1441	7.3515
IDeRs [50]	INS'19	13.6044	0.6439	3.6657	0.7662	0.7619	14.3124	0.7544	2.9211	0.8346	0.8281	0.2887	7.2904	0.2977	7.2594	0.1277	7.2350
EVPM [51]	INS'22	15.5794	0.6889	2.1345	0.7982	0.7950	17.8998	0.8246	1.2839	0.9241	0.9138	0.2659	6.9241	0.2783	7.0723	**0.1232**	5.1894
GRS-HTM [52]	SP'17	14.7996	0.5190	2.7928	0.8508	0.8335	16.1294	0.7177	1.7770	0.9026	0.8836	**0.2550**	5.8586	0.2931	6.8678	0.1318	7.3096
ZID [17]	TIP'20	18.9916	0.7267	0.9462	0.8662	0.8601	16.6307	0.7611	1.7300	0.8930	0.8811	0.2842	6.4705	0.3359	6.3373	0.1278	6.2448
zero-restore [18]	CVPR'21	16.6476	0.7173	2.0185	0.9068	0.9019	15.0654	0.7815	3.1844	0.8769	0.8717	0.8401	6.6588	0.3857	7.1970	0.1586	7.3991
FCTF-Net [53]	GRSL'21	19.3058	0.8559	0.7242	0.9088	0.9076	18.6402	0.8267	1.0084	0.9322	0.9283	0.3501	7.0295	**0.2618**	7.3476	**0.1269**	6.9161
TCN [54]	TMM'21	14.2077	0.6058	3.2109	0.8335	0.8238	16.7335	0.7995	1.5668	0.8905	0.8833	**0.2640**	**7.3385**	0.2685	**7.5590**	0.1885	**7.4290**
dehaze-cGAN [55]	CVPR'18	18.7025	0.7430	1.2163	0.9038	0.8969	21.4942	0.8877	0.6513	0.9365	0.9328	0.8692	5.7297	0.4938	6.6075	0.2811	6.9509
FFA-Net [31]	AAAI'20	**24.0516**	**0.8993**	**0.3337**	0.9387	0.9373	**25.0442**	**0.9013**	**0.3004**	0.9415	0.9373	0.2926	7.1178	**0.2644**	7.4290	0.1540	7.3747
Cycle-SNSPGAN [56]	TITS'22	18.3439	0.7288	1.0449	0.8381	0.8330	20.3632	0.8519	0.7061	0.9012	0.8997	0.5451	6.5692	0.4763	6.8223	0.3577	6.7658
UHD [13]	CVPR'21	**26.6588**	**0.9234**	**0.2535**	**0.9598**	**0.9581**	**27.2713**	**0.9460**	**0.1461**	**0.9660**	**0.9637**	0.5361	6.2487	0.3416	7.0498	0.1704	7.2237
DeHamer [21]	CVPR'22	23.7518	0.8985	0.4411	**0.9525**	**0.9520**	21.5605	0.8848	0.6280	**0.9501**	**0.9482**	0.2662	**7.3025**	0.2686	**7.5000**	0.1305	**7.4259**
Trinity-Net	–	**27.2431**	**0.9344**	**0.1596**	**0.9709**	**0.9700**	**29.5058**	**0.9553**	**0.1054**	**0.9779**	**0.9757**	**0.2490**	**7.3118**	**0.2660**	**7.5476**	**0.1273**	**7.4270**

D. Visual Comparisons

For visual comparison, we first show the results on the synthetic dataset in Fig. 9. FCTF-Net [53] achieves relatively satisfactory dehazing results, but the non-homogeneous haze is not completely removed. Other competing methods either fail to remove dense and non-homogeneous haze or produce undesirable color deviation. In contrast, UHD [13] and Trinity-Net are closest to the corresponding ground truth image regarding detail and color. Although ZID [17] and zero-restore [18] design deep model architectures based on the atmospheric scattering model, their dehazing performance is poorer than Trinity-Net, which indicates that considering the spatial distribution of haze is significant.

We then show the results of synthetic images with different haze densities in Fig. 10. SDCP [24] restores the structural details of hazy images but introduces extra greenish color deviation. UHD [13] effectively removes haze but differs slightly from the ground truth image in terms of color. For dense hazy images, other competing methods fail to achieve visually pleasing results. Some of them even produce reddish artifacts, such as GRS-HTM [52] and TCN [54]. In comparison, Trinity-Net not only alleviates the dense haze but also recovers the relatively realistic color, thus improving the visual perception of hazy scenes. The main reason is that we incorporate the domain knowledge of remote sensing imaging into the deep model. In addition, our purely data-driven architecture tolerates the inaccuracy of haze parameters.

We also present comparisons on remote sensing hazy images sampled from real-world datasets in Figs. 11, 12, and 13. In terms of dehazing, ROP [48], zero-restore [18], dehaze-cGAN [55], and Cycle-SNSPGAN [56] remove most haze, but residuals remain. In terms of color, some compared methods change the color of remote sensing scenes from an overall perspective, such as GRS-HTM [52] and ZID [17]. Although UHD [13] shows excellent dehazing capability on synthetic datasets, it fails to improve the visibility of real data. In contrast, Trinity-Net shows impressive scene adaptability, improving the visual quality of both synthetic and real-world images. Besides, our method restores more clear texture detail compared with competing methods, which is credited to the reasonable design of the gradient guidance module.

E. Quantitative Comparisons

To perform fair quantitative comparisons, we train each competing model utilizing our training data and achieve the best evaluation scores. The quantitative results on synthetic and real-world datasets are reported in Table I. In terms of full-reference metrics, our Trinity-Net achieves an average percentage gain of 49%/27%/91%/11%/12%, 58%/17%/92%/8%/8% in terms of PSNR/SSIM/MSE/FSIM/FSIMc on R100 and S100 datasets compared with competing models. In terms of no-reference metrics, our Trinity-Net achieves an average percentage gain of 41%/8%, 17%/5%, and 25%/5% in terms of FADE and Entropy on UAV, RS-Three, and Landsat-8 datasets compared with competing models. These results show that Trinity-Net can better recover the visibility of remote sensing images. There is an interesting finding from quantitative comparisons. Although traditional methods use prior information to estimate haze parameters, similar to our Trinity-Net, the performance is quite dissimilar. Such a finding again demonstrates that domain knowledge and deep models complement each other and improve the performance of remote sensing image dehazing.

F. Ablation Study

We conduct comprehensive ablation studies to verify the effectiveness of core components, including the haze thickness estimator, the ambient light estimator, the medium transmission estimator, and the gradient guidance module (GGM). Besides, we analyze the combination of ℓ_1 loss, MS-SSIM loss, and gradient loss.

- w/o CNN represents S-Net without the CNN backbone.
- w/o LMF represents A-Net without the local minimum filtering.
- w/o FA represents A-Net without the feature attention mechanism.
- w/o GF represents T-Net without the guided filtering.
- w/o FSE represents T-Net without the fine-scale encoder.
- w/o CSE represents T-Net without the coarse-scale encoder.
- w/o GGM represents Trinity-Net without the gradient guidance module.

TABLE II
A ABLATION STUDY FOR CORE COMPONENTS OF TRINITY-NET. THE BEST SCORE IS IN **BOLD**.

Module	Baseline	R100		S100	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
S-Net	w/o CNN	24.2662	0.9228	26.1249	0.9286
A-Net	w/o LMF	24.5027	0.9243	26.2094	0.9298
	w/o FA	24.3498	0.9263	25.8583	0.8962
T-Net	w/o GF	22.4332	0.9038	21.4189	0.8299
	w/o FSE	23.1253	0.9140	24.3779	0.9124
	w/o CSE	23.8734	0.9208	26.0697	0.9265
GGM	w/o GGM	24.1189	0.9281	25.2862	0.9211
	full model	**27.2431**	**0.9344**	**29.5058**	**0.9553**

TABLE III
A ABLATION STUDY FOR LOSS FUNCTION OF TRINITY-NET. THE BEST SCORE IS IN **BOLD**.

Loss Function			R100		S100	
$\mathcal{L}_{\ell_1}$	$\mathcal{L}_M$	$\mathcal{L}_G$	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
✓	–	–	23.2532	0.8876	24.2238	0.8700
–	✓	–	22.3160	0.9193	22.7623	0.9171
–	–	✓	20.4186	0.8085	19.7586	0.7451
✓	✓	–	24.8704	0.9287	26.3798	0.9252
✓	–	✓	23.6248	0.9145	24.9721	0.9097
–	✓	✓	23.5559	0.9296	24.1798	0.9229
✓	✓	✓	**27.2431**	**0.9344**	**29.5058**	**0.9553**

The quantitative scores of the ablation study are shown in Tables II and III. The visual comparisons of the contributions of core components and the effectiveness of the loss function are presented in Figs. 14, 15, 16, 17, 18, 19, and 20. The conclusions derived from the ablation study are listed below.

1) As reported in Tables II and III, the full Trinity-Net achieves superior quantitative scores on R100 and S100 datasets compared with ablated models, implying the effectiveness of the combination of well-designed components.

2) In Fig. 14, the full model accurately reflects the spatial distribution of haze by selecting the patch size adaptively, resulting in better defogging performance. Moreover, the quantitative scores of Trinity-Net are superior to the ablated model w/o CNN, indicating that it is crucial to consider the effect of patch size on the dark channel prior.

3) The ablation model w/o LMF generates an overexposure result, especially an impulse caused by flames appearing in the bright pixel region, as shown in Fig. 15. In contrast, Trinity-Net removes the impulse and achieves visually pleasing illumination, which is credited to the effective application of local minimum filtering. Besides, we employ the feature attention to pay more attention to hazy regions and highlight task-relevant features. The visualization results are depicted in Fig. 16.

4) In Fig. 17, the ablation model w/o GF generates blurry textures. This is probably caused by removing guided filtering, which has excellent edge-preservation properties. In comparison, T-Net contributes to learning more discriminative texture representations. This is demonstrated by the canny operator detecting more edge textures. Besides, the mutual complementarity between the fine-scale and coarse-scale paths likewise contributes to the recovery of texture representations, as shown in Fig. 18.

5) The ablation model w/o GGM fails to obtain satisfactory results in terms of details, as shown in Fig. 19. In contrast, Trinity-Net presents perceptual-pleasant details. The main reason is that the gradient guidance module provides additional structure priors and gradient information for dehazing process.

6) In Fig. 20, it can be observed that utilizing a single loss function for training is meaningless to obtain visually pleasing results. In terms of visual perception, our Trinity-Net trained with the overall loss $\mathcal{L}_{total}$ outperforms ablation models. Thus, training with the full loss function is of great significance to improve the visual quality of the dehazing results.

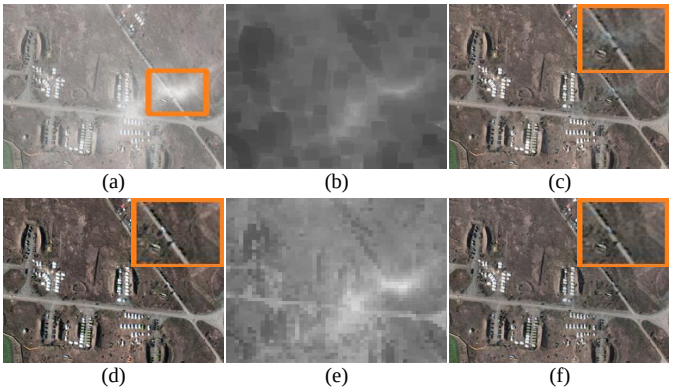


Fig. 14. Ablation study on the CNN backbone. (a) input. (b) haze thickness map estimated by w/o CNN. (c) w/o CNN. (d) ground truth. (e) haze thickness map estimated by S-Net. (f) Trinity-Net.

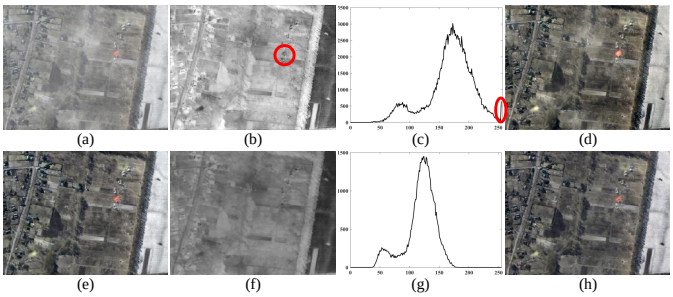


Fig. 15. Ablation study on the local minimum filtering. (a) input. (b) dark channel of w/o LMF. (c) pixel distributions of w/o LMF. (d) w/o LMF. (e) ground truth. (f) dark channel of Trinity-Net. (g) pixel distributions of Trinity-Net. (h) Trinity-Net.

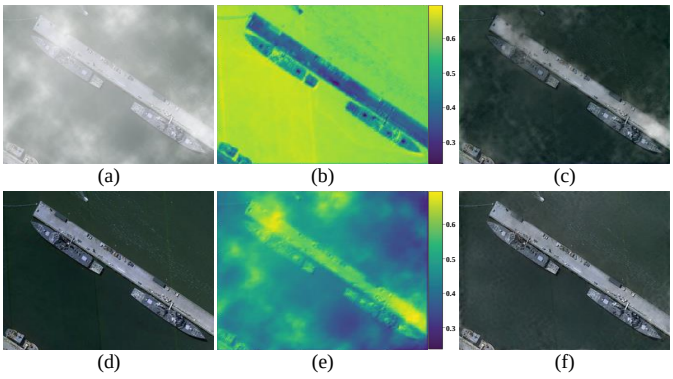


Fig. 16. Ablation study on the feature attention mechanism. (a) input. (b) attention map of w/o FA with corresponding attention weight. (c) w/o FA. (d) ground truth. (e) attention map of Trinity-Net with corresponding attention weight. (f) Trinity-Net.

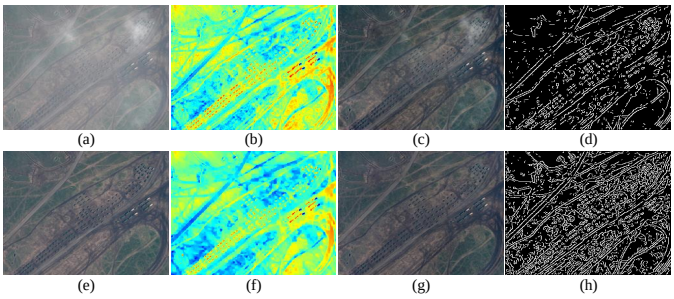


Fig. 17. Ablation study on the guided filtering. (a) input. (b) medium transmission map of w/o GF. (c) w/o GF. (d) canny edge of w/o GF. (e) ground truth. (f) medium transmission map of Trinity-Net. (g) Trinity-Net. (h) canny edge of Trinity-Net.

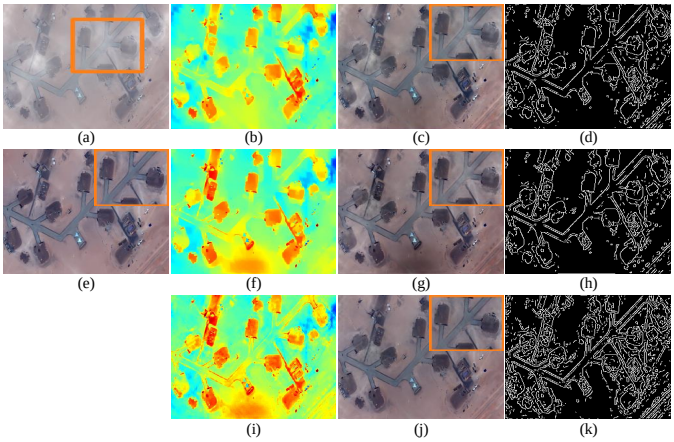


Fig. 18. Ablation study on the fine-scale encoder and coarse-scale encoder. (a) input. (b) medium transmission map of w/o FSE. (c) w/o FSE. (d) canny edge of w/o FSE. (e) ground truth. (f) medium transmission map of w/o CSE. (g) w/o CSE. (h) canny edge of w/o CSE. (i) medium transmission map of Trinity-Net. (j) Trinity-Net. (k) canny edge of Trinity-Net.

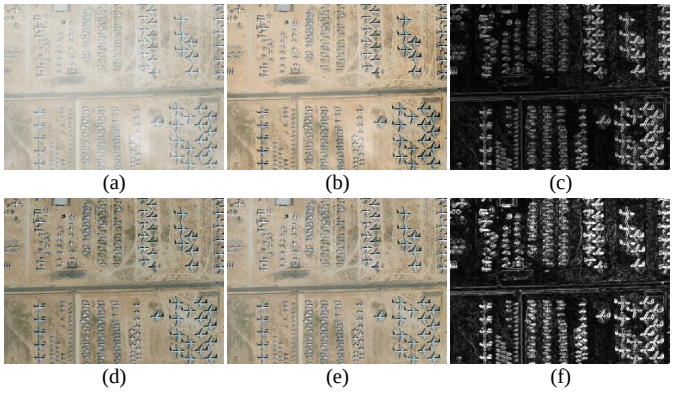


Fig. 19. Ablation study on the gradient guidance module. (a) input. (b) w/o GGM. (c) gradient map of w/o GGM. (d) ground truth. (e) Trinity-Net. (f) gradient map of Trinity-Net.

G. Extension to Natural Image Dehazing

In this section, we apply Trinity-Net to natural image dehazing to demonstrate its strong universality. For training, we randomly selected 900 image pairs from NID and utilized the rest 100 image pairs as the test set, denoted as N100. In Fig. 21, all competing methods fail to achieve satisfactory dehazing results. Some of them cannot completely remove the dense haze, especially areas framed in red and yellow, such as ROP [48], DEFADE [57], Haze-Lines [10], ZID [17], zero-restore [18], and DeHamer [21]. Additionally, GDCP [8], ALC [9], and TCN [54] present poor brightness, resulting in the loss of details and distorted colors. Although UHD [13] achieves good performance, it suffers from color deviations. In comparison, Trinity-Net not only improves visibility but also restores realistic color. Next, we perform a quantitative comparison on N100. In Table IV, our Trinity-Net achieves the best or second best average results across seven metrics, demonstrating the superiority of our dehazing strategy.

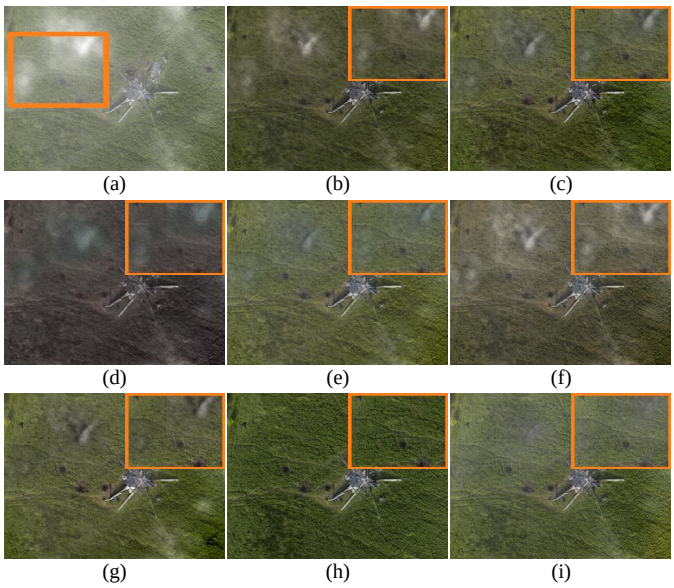


Fig. 20. Ablation study on the loss function. (a) input. (b) ℓ_1 loss. (c) MS-SSIM loss. (d) gradient loss. (e) ℓ_1 loss + MS-SSIM loss. (f) ℓ_1 loss + gradient loss. (g) MS-SSIM loss + gradient loss. (h) ground truth. (i) Trinity-Net.

H. Extension to Underwater Image Enhancement

To demonstrate the scene adaptability, we perform visual comparisons and numerical analysis on underwater images. For training, we randomly selected 790 image pairs from UIEB [58] and utilized the rest 100 image pairs as the test set, denoted as U100. In Fig. 22, the compared approaches either cannot alleviate tricky turbidity or improve color deviations. Our method removes the greenish color deviation while restoring visibility. Simultaneously, the quantitative results in Table V demonstrate that Trinity-Net is quite promising for challenging underwater scenarios. Interestingly, some competing methods are also extended to related vision tasks. However, the performance of Trinity-Net on each vision task is substantially better than these methods, as shown in Fig. 23. This is because they ignore the unique property of vision tasks, such as the spatial distribution of haze in remote sensing scenarios and the attenuation of red channel in underwater scenarios.

I. Failure Case

Trinity-Net as well as other competing methods may not work well when faced with remote sensing hazy images with

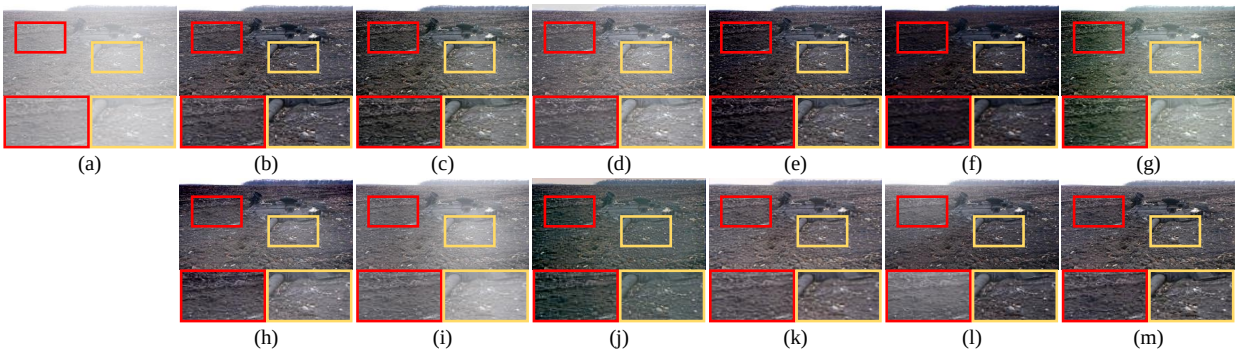


Fig. 21. Visual comparisons on a challenging natural hazy image sampled from **N100** dataset. (a) input. (b) ground truth. (c) ROP [48]. (d) DEFADE [57]. (e) GDCP [8]. (f) ALC [9]. (g) Haze-Lines [10]. (h) ZID [17]. (i) zero-restore [18]. (j) TCN [54]. (k) UHD [13]. (l) DeHamer [21]. (m) Trinity-Net.

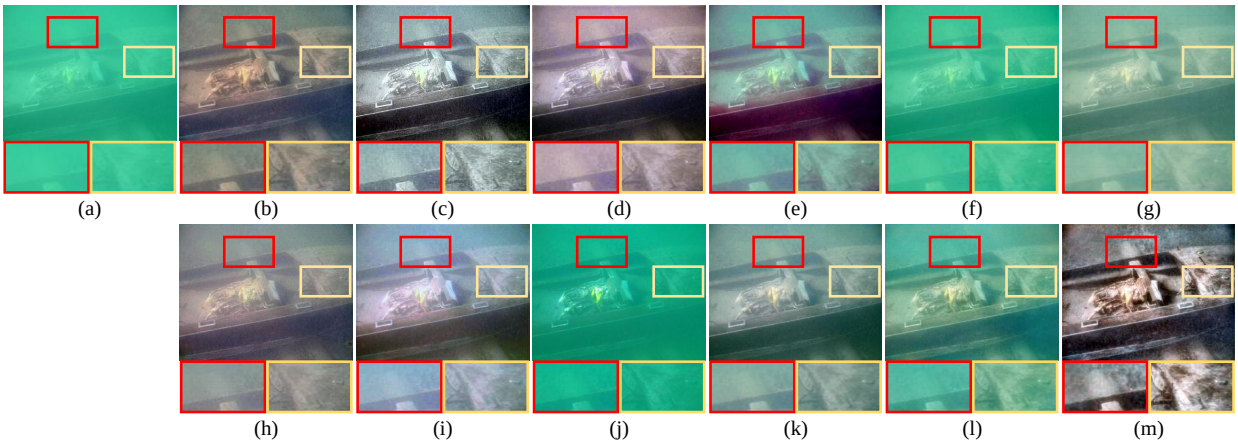


Fig. 22. Visual comparisons on a challenging underwater image sampled from **U100** dataset. (a) input. (b) reference. (c) MMLE [59]. (d) Fusion [37]. (e) GDCP [8]. (f) IBLA [60]. (g) TOPAL [61]. (h) TACL [62]. (i) Water-Net [58]. (j) LCNet [63]. (k) Ucolor [19]. (l) UICoE-Net [64]. (m) Trinity-Net.

TABLE IV
THE AVERAGE PSNR (DB), SSIM, MSE ($\times 10^3$), FSIM, FSIMC, FADE, AND ENTROPY SCORES ON **N100** DATASET. THE BEST SCORE IS IN **RED**, THE SECOND-BEST IS IN **GREEN**, AND THE THIRD-BEST IS IN **BLUE**.

Dataset	Metric	Methods										
		ROP [48] (TPAMI'11)	DEFADE [57] (TIP'15)	GDCP [8] (TIP'18)	ALC [9] (TCSVT'20)	Haze-Lines [10] (TPAMI'20)	ZID [17] (TIP'20)	zero-restore [18] (CVPR'21)	TCN [54] (TMM'21)	UHD [13] (CVPR'21)	DeHamer [21] (CVPR'22)	Trinity-Net
N100	PSNR ↑	16.2987	15.8032	15.9969	19.0275	17.9559	19.6239	16.0975	16.3645	**23.7157**	**23.1134**	**26.2210**
	SSIM ↑	0.8087	0.8383	0.7433	0.8499	0.8378	0.8270	0.7849	0.7112	**0.9182**	**0.9369**	**0.9685**
	MSE ↓	1.8937	2.3352	2.9394	1.0600	1.1967	0.7956	1.9159	1.7248	**0.2921**	**0.3959**	**0.2581**
	FSIM ↑	0.9599	0.9342	0.9262	0.9586	0.9502	0.9292	0.9550	0.9178	**0.9738**	**0.9700**	**0.9906**
	FSIMc ↑	0.9535	0.9328	0.9169	0.9560	0.9451	0.9238	0.9482	0.9047	**0.9715**	**0.9690**	**0.9902**
	FADE ↓	0.5336	0.6529	0.5895	0.3719	0.4364	0.3884	0.7560	**0.3452**	**0.3622**	0.3749	**0.3565**
	Entropy ↑	7.5408	7.2378	7.0513	7.5245	7.5411	**7.5569**	7.3812	**7.5544**	7.3918	7.4930	**7.5853**

TABLE V
THE AVERAGE PSNR (DB), SSIM, MSE ($\times 10^3$), FSIM, FSIMC, UCIQE, AND UIQM SCORES ON **U100** DATASET. THE BEST SCORE IS IN **RED**, THE SECOND-BEST IS IN **GREEN**, AND THE THIRD-BEST IS IN **BLUE**

Dataset	Metric	Methods										
		MMLE [59] (TIP'22)	Fusion [37] (TIP'18)	GDCP [8] (TIP'18)	IBLA [60] (TIP'17)	TOPAL [61] (TCSVT'22)	TACL [62] (TIP'22)	Water-Net [58] (TIP'20)	LCNet [63] (TMM'21)	Ucolor [19] (TIP'21)	UICoE-Net [64] (TCSVT'22)	Trinity-Net
U100	PSNR ↑	17.2665	20.1737	13.6194	15.8193	21.1205	**23.6329**	21.5209	18.8619	**22.0281**	21.8582	**24.5167**
	SSIM ↑	0.6869	0.7937	0.5764	0.5879	0.8034	**0.8633**	0.8022	0.7183	0.8393	**0.8413**	**0.8701**
	MSE ↓	1.5997	0.8739	3.5027	2.7863	0.7318	**0.4350**	0.6640	1.1539	**0.5302**	0.6473	**0.2627**
	FSIM ↑	0.8324	0.9486	0.9144	0.8922	**0.9555**	0.9299	0.9317	0.9406	0.9520	**0.9633**	**0.9701**
	FSIMc ↑	0.8186	0.9336	0.8826	0.8664	0.9402	0.9212	0.9167	0.9190	**0.9414**	**0.9503**	**0.9687**
	UCIQE ↑	**0.6128**	0.5926	0.6109	0.5963	0.5794	**0.6123**	0.5929	0.5638	0.5626	0.5817	**0.6195**
	UIQM ↑	**1.4349**	1.3550	**1.4286**	1.4076	1.3036	1.4013	1.3164	1.3319	1.3558	1.3763	**1.4314**

limited luminance. In Fig. 24, Trinity-Net and supervised-learning based methods (*e.g.*, UHD [13] and FFA-Net [31]) cannot cope with such images. The main reason lies in the few low-light images in the training benchmark. The stronger performance and richer training data that tackle such images will be our future goal.

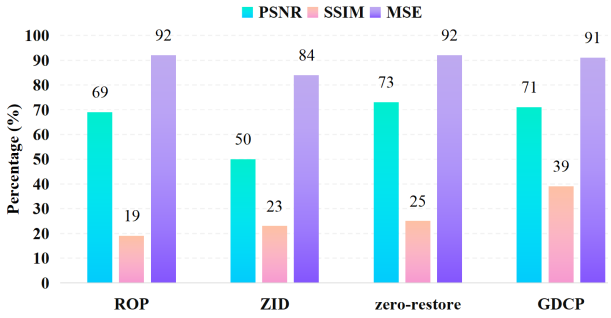


Fig. 23. The percentage gain of quantitative metrics relative to competing methods.

Fig. 24. Failure Case. Although our method fails to improve the visibility of the image, it does not introduce limited illumination like competing methods. (a) input. (b) UHD [13]. (c) FFA-Net [31]. (d) Trinity-Net.

V. Conclusion

This paper presented Trinity-Net for remote sensing image dehazing. Unlike previous works that either devised frameworks based on prior knowledge or constructed deep models heuristically, Trinity-Net performed haze parameter estimation by aggregating priors to CNNs and Swin Transformer. Then, the structure prior is incorporated into the network by exploring the complementary advantages of gradient information and Swin-Transformer to generate rich details. In addition, we demonstrate that Trinity-Net can also be promoted to natural image dehazing and underwater image enhancement tasks. Experimental results on diverse datasets indicated the effectiveness of our solution.

References

[1] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[2] Q. Yi, J. Li, F. Fang, A. Jiang, and G. Zhang, "Efficient and accurate multi-scale topological network for single image dehazing," *IEEE Trans. Multimedia*, vol. 24, pp. 3114–3128, Jun. 2022.

[3] L. Zhang and S. Wang, "Dense haze removal based on dynamic collaborative inference learning for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, Sep. 2022.

[4] C. Zhang, K.-M. Lam, and Q. Wang, "CoF-Net: A progressive coarse-to-fine framework for object detection in remote-sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–17, Jan. 2023.

[5] X. Yang, Y. Wang, N. Wang, and X. Gao, "An enhanced siammask network for coastal ship tracking," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, Oct. 2022.

[6] Z. Peng, Z. Li, J. Zhang, Y. Li, G.-J. Qi, and J. Tang, "Few-shot image recognition with knowledge transfer," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2019, pp. 441–449.

[7] Z. Li, Y. Sun, L. Zhang, and J. Tang, "CTNet: Context-based tandem network for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9904–9917, Dec. 2022.

[8] Y.-T. Peng, K. Cao, and P. C. Cosman, "Generalization of the dark channel prior for single image restoration," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2856–2868, Jun. 2018.

[9] Y.-T. Peng, Z. Lu, F.-C. Cheng, Y. Zheng, and S.-C. Huang, "Image haze removal using airlight white correction, local light filter, and aerial perspective prior," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 5, pp. 1385–1395, May. 2020.

[10] D. Berman, T. Treibitz, and S. Avidan, "Single image dehazing using haze-lines," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 720–734, Mar. 2020.

[11] Q. Guo, H.-M. Hu, and B. Li, "Haze and thin cloud removal using elliptical boundary prior for remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9124–9137, Nov. 2019.

[12] A. Makarau, R. Richter, R. Müller, and P. Reinartz, "Haze detection and removal in remotely sensed multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5895–5905, Sep. 2014.

[13] Z. Zheng *et al.*, "Ultra-high-definition image dehazing via multi-guided bilateral learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16180–16189.

[14] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced pix2pix dehazing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8152–8160.

[15] P. Li, J. Tian, Y. Tang, G. Wang, and C. Wu, "Deep retinex network for single image dehazing," *IEEE Trans. Image Process.*, vol. 30, pp. 1100–1115, Dec. 2020.

[16] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3194–3203.

[17] B. Li, Y. Gou, J. Z. Liu, H. Zhu, J. T. Zhou, and X. Peng, "Zero-shot image dehazing," *IEEE Trans. Image Process.*, vol. 29, pp. 8457–8466, Aug. 2020.

[18] A. Kar, S. K. Dhara, D. Sen, and P. K. Biswas, "Zero-shot single image restoration through controlled perturbation of Koschmieder's model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16200–16210.

[19] C. Li, S. Anwar, J. Hou, R. Cong, C. Guo, and W. Ren, "Underwater image enhancement via medium transmission-guided multi-color space embedding," *IEEE Trans. Image Process.*, vol. 30, pp. 4985–5000, May. 2021.

[20] C. Ma, Y. Rao, J. Lu, and J. Zhou, "Structure-preserving image super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7898–7911, Nov. 2022.

[21] C. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, and C. Li, "Image dehazing transformer with transmission-aware 3D position embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5802–5810.

[22] T. Ren *et al.*, "Reinforced swin-convs transformer for simultaneous underwater sensing scene image enhancement and super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, Sep. 2022.

[23] D. Berman, T. Treibitz, and S. Avidan, "Non-local image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1674–1682.

[24] J. Li, Q. Hu, and M. Ai, "Haze and thin cloud removal via sphere model improved dark channel prior," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 472–476, Mar. 2019.

[25] H. Shen, C. Zhang, H. Li, Q. Yuan, and L. Zhang, "A spatial-spectral adaptive haze removal method for visible remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6168–6180, Sep. 2020.

[26] Y. Song, J. Li, X. Wang, and X. Chen, "Single image dehazing using ranking convolutional neural network," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1548–1560, Jun. 2018.

[27] J. Guo, J. Yang, H. Yue, C. Hou, and K. Li, "Landsat-8 OLI multispectral image dehazing based on optimized atmospheric scattering model," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10255–10265, Dec. 2021.

[28] J. Nie, Y. Pang, J. Xie, J. Han, and X. Li, "Binocular image dehazing via a plain network without disparity estimation," *IEEE Trans. Multimedia*, early access, Aug. 17, 2022, doi: 10.1109/TMM.2022.3199553.

[29] H. Ullah *et al.*, "Light-DehazeNet: A novel lightweight CNN architecture for single image dehazing," *IEEE Trans. Image Process.*, vol. 30, pp. 8968–8982, Oct. 2021.

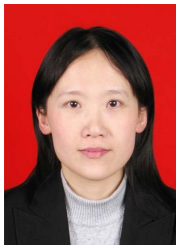
[30] W.-T. Chen, H.-Y. Fang, J.-J. Ding, and S.-Y. Kuo, "PMHLD: Patch map-based hybrid learning dehazenet for single image haze removal," *IEEE Trans. Image Process.*, vol. 29, pp. 6773–6788, May. 2020.

[31] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "FFA-Net: Feature fusion attention network for single image dehazing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11908–11915.

[32] X. Zhang, J. Wang, T. Wang, and R. Jiang, "Hierarchical feature fusion with mixed convolution attention for single image dehazing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 510–522, Feb. 2022.

[33] J. Xiao, X. Fu, A. Liu, F. Wu, and Z.-J. Zha, "Image de-raining transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jun. 16, 2022, doi: 10.1109/TPAMI.2022.3183612.
[34] Z. Huang, J. Li, Z. Hua, and L. Fan, "Underwater image enhancement via adaptive group attention-based multiscale cascade transformer," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–18, Jul. 2022.
[35] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imaging*, vol. 3, no. 1, pp. 47–57, Mar. 2017.
[36] R. Liu, S. Li, J. Liu, L. Ma, X. Fan, and Z. Luo, "Learning hadamard-product-propagation for image dehazing and beyond," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1366–1379, Apr. 2021.
[37] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert, "Color balance and fusion for underwater image enhancement," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 379–393, Jan. 2018.
[38] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, Aug. 2019.
[39] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3974–3983.
[40] J. Chen, K. Chen, H. Chen, Z. Zou, and Z. Shi, "A degraded reconstruction enhancement-based method for tiny ship detection in remote sensing images with a new large-scale dataset," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, Jun. 2022.
[41] Z. Li and N. Snavely, "MegaDepth: Learning single-view depth prediction from internet photos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2041–2050.
[42] Y. Zhang, L. Ding, and G. Sharma, "HazeRD: An outdoor scene dataset and benchmark for single image dehazing," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3205–3209.
[43] B. Huang, Z. Li, C. Yang, F. Sun, and Y. Song, "Single satellite optical imagery dehazing using SAR image prior based on conditional generative adversarial networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1795–1802.
[44] K. Zhang, S. Ma, R. Zheng, and L. Zhang, "UAV remote sensing image dehazing based on double-scale transmission optimization strategy," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Sep. 2022.
[45] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no, 10, pp. 1865–1883, Oct. 2017.
[46] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.
[47] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May. 2017.
[48] J. Liu, R. W. Liu, J. Sun, and T. Zeng, "Rank-One Prior: Real-time scene recovery," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 02, 2022, doi: 10.1109/TPAMI.2022.3226276.
[49] Z. Mi, Y. Li, J. Jin, Z. Liang, and X. Fu, "A generalized enhancement framework for hazy images with complex illumination," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, May. 2021.
[50] L. Xu, D. Zhao, Y. Yan, S. Kwong, J. Chen, and L.-Y. Duan, "IDeRs: Iterative dehazing method for single remote sensing image," *Inf. Sci.*, vol. 489, pp. 50–62, Jul. 2019.
[51] J. Han, S. Zhang, N. Fan, and Z. Ye, "Local patchwise minimal and maximal values prior for single optical remote sensing image dehazing," *Inf. Sci.*, vol. 606, pp. 173–193, Aug. 2022.
[52] Q. Liu, X. Gao, L. He, and W. Lu, "Haze removal for a single visible remote sensing image," *Signal Process.*, vol. 137, pp. 33–43, Aug. 2017.
[53] Y. Li and X. Chen, "A coarse-to-fine two-stage attentive network for haze removal of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 10, pp. 1751–1755, Oct. 2021.
[54] J. Shin, H. Park, and J. Paik, "Region-based dehazing via dual-supervised triple-convolutional network," *IEEE Trans. Multimedia*, vol. 24, pp. 245–260, Jan. 2021.
[55] R. Li, J. Pan, Z. Li, and J. Tang, "Single image dehazing via conditional generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8202–8211.
[56] Y. Wang *et al.*, "Cycle-SNSPGAN: Towards real-world image dehazing via cycle spectral normalized soft likelihood estimation patch GAN," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 20368–20382, Nov. 2022.
[57] L. K. Choi, J. You, and A. C. Bovik, "Referenceless prediction of perceptual fog density and perceptual image defogging," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3888–3901, Nov. 2015.
[58] C. Li *et al.*, "An underwater image enhancement benchmark dataset and beyond," *IEEE Trans. Image Process.*, vol. 29, pp. 4376–4389, Feb. 2020.
[59] W. Zhang, P. Zhuang, H.-H. Sun, G. Li, S. Kwong, and C. Li, "Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement," *IEEE Trans. Image Process.*, vol. 31, pp. 3997–4010, Jun. 2022.
[60] Y.-T. Peng and P. C. Cosman, "Underwater image restoration based on image blurriness and light absorption," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1579–1594, Apr. 2017.
[61] Z. Jiang, Z. Li, S. Yang, X. Fan, and R. Liu, "Target oriented perceptual adversarial fusion network for underwater image enhancement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6584–6598, Oct. 2022.
[62] R. Liu, Z. Jiang, S. Yang, and X. Fan, "Twin adversarial contrastive learning for underwater image enhancement and beyond," *IEEE Trans. Image Process.*, vol. 31, pp. 4922–4936, Jul. 2022.
[63] N. Jiang, W. Chen, Y. Lin, T. Zhao, and C.-W. Lin, "Underwater image enhancement with lightweight cascaded network," *IEEE Trans. Multimedia*, vol. 24, pp. 4301–4313, Sep. 2021.
[64] Q. Qi *et al.*, "Underwater image co-enhancement with correlation feature matching and joint learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1133–1147, Mar. 2022.

Kaichen Chi received the B.E. degree in electronic and information engineering and the M.E. degree in communication and information system from Liaoning Technical University, Huludao, China, in 2019 and 2022 respectively. He is currently working toward the Ph.D. degree in the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include image processing and deep learning.



Yuan Yuan (M'05-SM'09) is currently a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS AND PATTERN RECOGNITION, as well as the conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and remote sensing.