

采用对比学习的多阶段 Transformer 图像去雾方法

高峰^{1,2}, 汲胜昌¹, 郭洁¹, 侯杰³, 欧阳超³, 杨彪⁴

(1. 西安交通大学电气工程学院, 710049, 西安; 2. 国网陕西省电力有限公司电力科学研究院, 710100, 西安;

3. 哈尔滨工业大学(深圳)计算机科学与技术学院, 518055, 广东深圳;

4. 哈尔滨工业大学(深圳)建筑学院, 518055, 广东深圳)

摘要: 为了解决现有图像去雾方法在图像局部去雾以及纹理细节恢复等方面始终不理想以及处理非均匀雾质始终不彻底的问题,提出了一种采用对比学习的多阶段自注意力模块(Transformer)的图像去雾 MSTCNet 方法。首先,利用信道级 Transformer 模块作为基本的特征提取模块,充分地捕获特征信道之间的长距离依赖关系;其次,通过提出的多监督对比学习方法最大限度地挖掘正负样本信息,使去雾图像在投影后的隐空间中更靠近清晰图像,同时远离有雾图像;最后,利用多阶段渐进式网络结构和可变形自注意力机制有效地整合图像局部细粒度特征和全局粗粒度信息。本文在 2 个合成数据集和 3 个真实数据集上对所提出的方法进行了大量的实验,结果表明:所提出的 MSTCNet 方法在 5 个数据集上的峰值信噪比(PSNR)分别提高了 1.49、1.45、0.11、1.45 和 0.22 dB,在通用数据集与非数据集的测试中均超越已有的方法,在浓雾质、非均匀雾质以及均匀雾质的测试中均表现出最佳的去雾视觉效果,并达到最高的客观评价指标值。

关键词: 图像去雾;对比学习;自注意力;渐进式网络

中图分类号: TP391 **文献标志码:** A

DOI: 10.7652/xjtuxb202301019 **文章编号:** 0253-987X(2023)01-0195-16

A Multi-Stage Transformer Network for Image Dehazing Based on Contrastive Learning

GAO Feng^{1,2}, JI Shengchang¹, GUO Jie¹, HOU Jie³, OUYANG Chao³, YANG Biao⁴

(1. School of Electrical Engineering, Xi'an Jiaotong University, Xi'an 710049, China;

2. Electric Power Research Institute, State Grid Shaanxi Electric Power Company Limited, Xi'an 710100, China;

3. School of Computer Science, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong 518055, China;

4. School of Architecture, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong 518055, China)

Abstract: A multi-stage Transformer network for image dehazing based on contrastive learning is proposed to solve the problem that existing image dehazing methods fail to achieve the desired results in local image dehazing and detail restoration, and the non-homogeneous haze cannot be removed completely all the way. First, the channel-wise Transformer block is utilized as the primary feature extraction block to adequately capture the mutual long-range dependencies among channels. Second, the multi-modality supervised contrastive learning is introduced to maximize the

收稿日期: 2022-06-15。 作者简介: 高峰(1984—),男,博士生;杨彪(通信作者),男,副教授,博士生导师。 基金项目: 国家自然科学基金资助项目(52078161);陕西电网重要输电通道抵御外部风险的监测和预警新技术研究资助项目(5226SX21002Q)。

网络出版时间: 2022-09-17

网络出版地址: <https://kns.cnki.net/kcms/detail/61.1069.T.20220916.1626.004.html>

capturing efficiency of information from the contrastive samples, so that the restored image is closer to the clear image in the embedding space while staying as far away from the hazy image as possible. Finally, a hierarchical multi-patch structure and deformable Transformer blocks are employed to effectively integrate the local and global structural information of the hazy image. Moreover, a large number of tests have been conducted on the proposed method by using two synthetic data sets and the three real data sets. The results show that the proposed MSTCNet achieves a higher peak signal-to-noise ratio (PSNR) gain of 1.49, 1.45, 0.11, 1.45 and 0.22 dB on five datasets, respectively. It outperforms existing methods in both general and non-data sets, shows the best visual effect of dehazing in removing the dense, non-homogeneous and uniform haze, and achieves the highest objective evaluation index value.

Keywords: image dehazing; contrastive learning; Transformer; multi-patch structure

雾霾作为一种常见的大气现象,显著降低了物体的可见度和对比度,导致观察到的物体出现模糊、颜色失真等视觉质量退化的现象。雾霾天气不仅直接对人类的视觉感知产生了严重的影响,也对众多高级计算机视觉系统构成挑战,例如物体检测和语义分割,这些系统会因为低质量的输入图像而导致系统性能的降低^[1]。因此,在将有雾的模糊图像送入到高级计算机视觉系统之前,对其进行去雾的预处理操作是非常必要的。

近年来,图像去雾作为一项典型的底层视觉任务,得到了研究人员越来越多的关注。早期提出的一些图像去雾方法大多是基于典型的大气散射模型^[2],该模型定义为

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

式中: x 表示像素点; $I(x)$ 表示观察到的雾霾场景; $J(x)$ 表示去雾场景; A 表示全局大气光强; $t(x)$ 表示介质传输函数,也称作透射率,它可以进一步表示为 $t(x) = e^{-\beta d(x)}$, β 表示大气散射系数, $d(x)$ 表示像素点 x 的场景深度。通常情况下,大气光强 A 和每个像素点的透射率 $t(x)$ 是不确定的,因此,图像去雾是一个严重的不适定问题(ill-posed problem)。针对大气散射模型,研究人员引入了一部分图像来先验估计未知的模型参数^[3-8],但是这不足以反映真实图像的特征。

研究人员提出了诸多端到端的基于卷积神经网络的去雾模型^[9-11],通过直接学习雾霾图像到清晰图像的转换来简化去雾问题。然而,基于卷积神经网络的方法存在两个主要问题:①卷积操作的感受野有限,不能随着网络的加深来建模远距离的依赖关系^[12];②在模型推理过程中,卷积滤波器的权值固定,不能灵活地适应不断变化的输入特征^[13]。针对上述问题,自注意力(self-attention)机制作为一

种更灵活的替代卷积运算的方法,主要通过计算所有位置的加权和得出像素点的推理值^[14]。基于此,本文提出了一种新的采用自注意力机制的端到端的单幅图像去雾网络;并通过所提出的多监督对比学习方法最大限度地利用雾霾图像、清晰图像和去雾图像之间的互信息,捕获图像域之间的内容和特征关系。具体来说,本文网络的构建主要考虑了以下两方面因素。

一方面,在雾霾分布均匀的场景中,不同场景深度下的物体会表现出不同的视觉退化程度;此外,在真实世界中,也存在非均匀分布的雾霾场景。因此,良好的去雾模型应该联合处理由以上两种不同的雾霾特征产生的视觉退化问题。目前常用的做法是在单阶段上提取图像的多尺度特征信息进行图像去雾^[10-11,15]。多阶段模型^[9,16-17]和尺度循环模型^[18]可以通过逐步整合不同空间尺度的图像特征实现从粗到细的学习,既可以学习到图像在粗尺度上的全局信息,也能捕获到图像的局部特征细节。因此,为了更好地恢复雾霾场景,本文采用多阶段网络设计,以有效地整合局部和全局信息。

另一方面,目前大多数现有的去雾工作只是利用清晰图像作为正向引导来监督模型训练^[9-11,16,19],而负样本的信息(例如雾霾图像)却没有得到很好地利用。最近,虽然已经有一些基于对比学习的图像恢复工作被发表,例如图像去雾^[15]和单幅图像超分辨率^[20],但是这些方法直接利用退化的输入图像或者数据集中其余的图像作为负样本,并将整张清晰图像作为正样本。显然,这些方法存在一定的缺陷^[21]:①用于训练的雾霾图像与模型重建的恢复图像之间的整体相似度低,整体上较为容易区分,难以充分发挥对比损失的作用;②现有的方法大多是利用预训练的深度卷积神经网络 VGG 网络,例如 VGGNet,作

为图像的嵌入特征提取网络,但是通过这种先验模型获得的特征往往是分类任务中的高级语义信息,无法分辨出图像细节的变化;③现有方法在对比样本的选择上比较简单且种类单一。因此,有必要研究一种适合去雾任务的对比学习策略,以最大限度地挖掘正负样本信息。

基于上述观点,本文利用由细到粗学习的思想^[17]提出了一种用于图像去雾任务的多阶段 Transformer 去雾网络(MSTCNet)。为了充分整合图像在不同空间尺度上的特征表示,本工作在各阶段之间进行了横向和纵向的加权信息融合;为了进一步优化去雾效果,在最后阶段添加了精细模块;为了充分发挥自注意力机制的优势,本工作选择基于信道级自注意力的模块^[14]作为主要的特征提取模块。此外,由于该模块没有对特征元之间的关系进行显式地建模,因此,本文在各阶段子网络的深度网络层额外嵌入了高效的可变形注意力 Transformer 模块^[22]作为对特征元之间关系的补充建模。

本文还提出了一种新的简化的对比学习框架^[21]与新的多监督对比学习方法。其中,将具有退化感知的马尔可夫判别器(PatchGAN)作为对比学习框架下的特征提取网络,并在训练过程中与主网络迭代优化。此外,本文采用多监督对比学习的方法在恢复图像与清晰图像上选取多个图像块作为正样本,并选取雾霾图像中的多个图像块作为负样本,通过图像块级的信息挖掘,最大限度地发挥对比学习的优势。

1 相关工作

1.1 单幅图像去雾

单幅图像去雾任务要求在不依赖附加图像数据的情况下,仅凭借一幅雾霾图像恢复出相应的清晰背景图像。这一方面,最早的方法大多是基于大气散射模型,而最近的方法主要是通过深度学习来预测图像的先验知识,或者通过深度学习直接学习雾霾图像和清晰图像之间的映射关系。

1.1.1 基于图像先验的方法

基于模型的图像去雾方法是在某些假设或特定的先验知识下估计式(1)中的透射率和全局大气光强,然后通过反推式(1)以求得去雾后的图像。He等^[3]发现了暗通道先验知识(DCP),即清晰图像中的像素至少在一个颜色信道中有接近于0的值,但是该工作只能在相对较暗的雾图上取得较好的去雾效果。高涛等^[6]对该算法做了一定的改进,针对单

幅图像去雾算法中透射率估计不准确和去雾不彻底的问题,提出了一种结合自适应亮度变换不等式估计透射率的图像去雾方法。Berman等^[4]发现了雾线先验,即图像的色彩值分布会因为雾霾的存在由簇状变为线状,并利用这一分布规律对模型中的透射率进行初步估计。虽然以上方法均取得了不同程度的去雾效果,但是这些先验知识都是通过对特定场景的观察和某些假设得出的,其普适性有待提高。因此,这些基于模型的传统方法的性能在本质上会受到先验知识特殊性的限制。

1.1.2 基于深度学习的方法

基于深度学习的方法通常有两条技术路线:一是通过深度神经网络直接估计大气散射模型中的透射率和大气光强^[23-24];二是直接学习从雾霾图像到清晰图像的映射关系^[9-11]。

Cai等^[23]提出的去雾网络利用三层卷积神经网络直接从雾霾图像中估计透射率;Ren等^[24]利用粗尺度和细尺度网络分别对透射率进行粗略和精细地估计。然而,上述的这类方法虽然是基于深度学习的,但是严重依赖于模型参数(投射图和大气光强)的真实值,而这些真实值在实际应用中是难以收集的。此外,参数的不准确估计也会在模型推理过程中引起累积误差。

由于依赖模型的去雾方法效果有限,最近的相关工作均集中于利用卷积神经网络直接学习雾霾图像到清晰图像的转换。例如,Qu等^[9]利用由粗到细学习的思想提出了增强条件生成去雾网络,即增强 pip2pix 网络。Qin等^[10]利用信道和像素注意力机制提出了特征融合注意力网络(FFA);Zhu等^[11]通过融合3个分辨率尺度的特征提出了多流融合网络(MSFNet)。可以发现,所有这些工作都只关注于提取输入分辨率下的全局多尺度特征,而忽略了对雾霾图像局部细节的直接重建。

1.2 对比学习

近年来,对比学习在无监督表征学习中表现出了巨大的潜力^[25]。其核心思想是利用数据之间的相关性和差异性来学习事物的本质。对于一个给定的锚样本,对比学习的目的是将它与正样本(语义相似)和负样本(语义不同)进行比较;通过在表征空间中将锚拉近正样本的同时推离负样本的操作以实现类似聚类的效果。此外,由于对比学习可以获得更高维、更本质的特征信息,因此,常被用于高级视觉任务中^[25]。然而,由于对比学习的正负样本构建的特殊性,直接将对比学习应用于图像去雾任务的研

究并不多见。例如, Wu 等^[15]针对图像去雾任务提出了新的构建对比样本的方法和新的像素级对比损失; Ji 等^[26]则提出了一种新颖的多视角对比学习策略,使网络能够敏感地捕捉到噪声分布。受以上工作的启发,本文针对图像去雾任务提出了一种新的构建对比样本的方法与新的图像块级的多监督对比损失。

1.3 视觉 Transformer

Transformer^[27]是一种完全基于注意力机制的深度神经网络,它的出现使自然语言处理(NLP)领域取得了许多突破。受 Transformer 强大表征能力的影响,研究人员尝试将 Transformer 应用于视觉任务中。然而,视觉 Transformer (ViT)^[28]的计算复杂度为像素数的平方,庞大的计算成本限制了其在视觉任务中的发展。为了解决上述问题,典型的工作是基于窗口的自注意力,即 Swin Transformer^[29]。该方法利用局域先验知识,通过不重叠的窗口来分解原始尺寸的特征,只在每个窗口内执行区域自注意力计算,

使得计算复杂度降低到了像素数的线性比例。然而,基于窗口的设计限制了局部区域内上下文特征信息的整合,仍然不能对全局依赖性进行有效地建模。基于此, Zamir 等^[14]提出了信道级自注意力机制,即 Restormer,它包括一种利用深度卷积的多头转置注意力机制(MDTA)。该机制没有直接对像素进行自注意力计算,而是在信道维度上建模全局上下文关系,大大降低了时间复杂度。因为 Restormer 在其他图像恢复任务中的优越性,所以在本工作中采用其作为网络的基本特征提取模块。

2 MSTCNet 去雾方法

本节将详细介绍所提出的 MSTCNet 图像去雾网络,其整体结构如图 1 所示。首先,详细阐述了 MSTCNet 网络的主要结构和网络信息流的传递过程;然后,介绍了所采用的两个关键模块;最后,介绍了所提出的基于图像块的多监督对比学习方法。

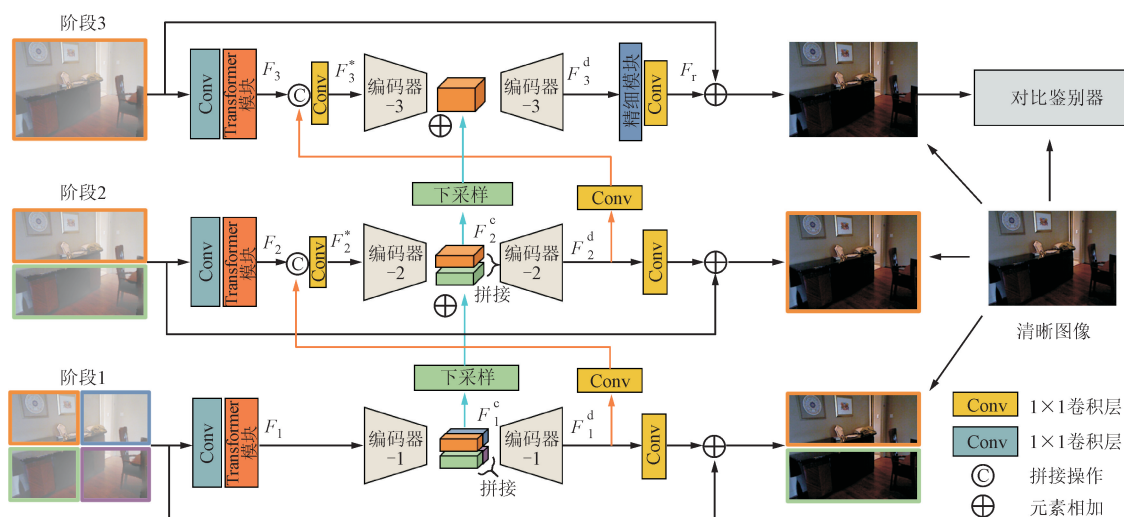


图 1 MSTCNet 整体结构图

Fig. 1 Overview of the proposed MSTCNet for single image dehazing

2.1 MSTCNet 的网络结构

为了克服现有方法在局部图像去雾以及去雾不彻底等方面的局限性,本文提出了一种三阶段的 Transformer 图像去雾网络,以逐步恢复雾霾图像,其网络结构主要由主去雾网络和具有退化感知的对比鉴别器组成。各个阶段都是一个基于编码器-解码器结构的子网络,用于分别处理不同空间尺度下的输入图像,并学习大量的上下文信息。这种设计利用了由细到粗学习的思想,使网络能够在较低阶段捕获局部图像的多尺度特征,并更多地关注于恢

复局部细节;在最后阶段捕获图像的全局多尺度特征,并更多地关注于整体的去雾效果。

由于该网络的设计目的是使底层阶段关注局部信息,并协助高层阶段进行粗尺度的去雾,因此有必要使各阶段之间进行信息的交互。在横向的信息交互中,本文方法没有像以前的工作^[17]那样简单地将各阶段之间的信息流相加,而是通过一个 1×1 的卷积层聚合前一阶段输出的像素级跨通道上下文,然后将它们与下一个阶段的浅层特征拼接并融合。此外,在纵向的信息交互中,将各阶段子网络的中间特

征进行加权的信息融合,以巩固和丰富前一阶段子网络的中间特征。

与之前工作^[17]类似的是,本文方法在各个阶段输入特征的选择上采用了多图像块的层次结构。具体地,在 3 个阶段分别输入尺寸不同、数量不同的不重叠的图像块集:首先将原始雾霾图像等分成 4 个图像块和 2 个图像块,并分别输入到第 1 阶段和第 2 阶段,然后将完整尺寸的原始雾图输入到第 3 阶段,如图 1 所示。这种方式允许网络在不同的空间尺度上分别学习具有相同语义的图像内容的纹理特征与噪声特征,从而保留更多的图像细节。

假设给定一张退化的雾霾图像 $I^h \in \mathbb{R}^{H \times W \times 3}$, 则第 i 个阶段的第 j 个输入图像块为 $I_{i,j}^h$, 其中 $H \times W$ 表示空间维度。具体而言,第 3 阶段的输入 I_3^h 即是原始尺寸的雾霾图像 I^h ; 在第 2 阶段,将 I^h 沿垂直方向等分为 $I_{2,1}^h$ 和 $I_{2,2}^h \in (H/2) \times W \times 3$, 并将两个图像块一起作为输入; 在第 1 阶段,将 $I_{2,1}^h$ 和 $I_{2,2}^h$ 沿水平方向进一步等分为 $I_{1,1}^h, I_{1,2}^h, I_{1,3}^h$ 和 $I_{1,4}^h$, 并将这 4 个尺寸大小都为 $(H/2) \times (W/2) \times 3$ 的图像块一起作为输入。第 i 个阶段的编码器和解码器分别表示为 Enc_i 和 Dec_i 。

整个去雾过程的信息流是自下而上的。对第 1 阶段的 4 个图像块分别进行初步的预处理得到浅层特征图

$$F_{1,j} = \text{Block}(\text{PE}(I_{1,j}^h)), j \in \{1, \dots, 4\} \quad (2)$$

式中: $\text{PE}(\cdot)$ 代表图像块嵌入特征提取操作, 实际上是一个核为 3 的卷积层; $\text{Block}(\cdot)$ 代表 Transformer 基本特征提取操作。将经过预处理得到的输入特征送入到编码器 Enc_1 中得到子网络的中间特征图

$$F_{1,j}^l = \text{Enc}_1(F_{1,j}), j \in \{1, \dots, 4\} \quad (3)$$

将空间分布上相邻图像块的中间特征图按水平维度拼接起来, 获得一个新的特征表示

$$F_{1,j}^c = [F_{1,2j-1}^l, F_{1,2j}^l], j \in \{1, 2\} \quad (4)$$

其中 $[\cdot, \cdot]$ 代表拼接操作。随后, 将拼接之后的特征表示送入到解码器 Dec_1 中以获得第 1 阶段的输出特征

$$F_{1,j}^d = \text{Dec}_1(F_{1,j}^c), j \in \{1, 2\} \quad (5)$$

信息流将分为两个分支, 其中一个分支通过 1×1 卷积层将输出特征的信道缩减到 3 个以获得残差信息, 进一步与输入图像融合得到两个去雾之后的图像块。将这两个去雾图像块按照空间分布进行拼接, 得到第一阶段的完整去雾图像

$$O_1 = [(\text{Conv}_{1 \times 1}(F_{1,1}^d) + I_{2,1}^h), (\text{Conv}_{1 \times 1}(F_{1,2}^d) + I_{2,2}^h)] \quad (6)$$

式中: $\text{Conv}_{1 \times 1}(\cdot)$ 代表核大小为 1 的卷积操作。

在另一个分支中, 为了使后一阶段能够获得精细阶段中推断出的所有信息, 需要将前一阶段经过像素级跨通道上下文聚合之后的特征 $S_{1,j}$ 送入到第 2 阶段。具体来说, 首先根据式(2)对第 2 阶段的输入图像块 $I_{2,1}^h$ 和 $I_{2,2}^h$ 进行预处理, 得到相应的浅层特征 $F_{2,j}, j \in \{1, 2\}$ 。然后, 将得到的浅层特征与第 1 阶段的卷积特征表示 $S_{1,j}$ 融合, 得到第 2 阶段编码器 Enc_2 的输入特征 $F_{2,j}^*$

$$S_{1,j} = \text{Conv}_{1 \times 1}(F_{1,j}^d), j \in \{1, 2\} \quad (7)$$

$$F_{2,j}^* = \text{Conv}_{1 \times 1}([F_{2,j}, S_{1,j}]) \quad (8)$$

一旦 Enc_2 的输出产生, 将其与第 1 阶段的拼接中间特征 $F_{1,j}^c$ 加权融合获得第 2 阶段的中间特征

$$F_{2,j}^t = \text{Enc}_2(F_{2,j}^*) + \alpha \text{Down}(F_{1,j}^c), j \in \{1, 2\} \quad (9)$$

其中 $\text{Down}(\cdot)$ 表示下采样操作。接下来的步骤基本与第 1 阶段相同。首先, 将从式(9)中得到的中间特征沿垂直方向拼接, 作为解码器 Dec_2 的输入, 经过解码后得到第 2 阶段的输出特征

$$F_2^d = \text{Dec}_2([F_{2,1}^t, F_{2,2}^t]) \quad (10)$$

特征 F_2^d 不仅用于第 2 阶段的雾霾图像恢复, 而且需要进一步辅助第 3 阶段的粗尺度处理。第 2 阶段的去雾图像通过如下公式获得

$$O_2 = \text{Conv}_{1 \times 1}(F_2^d) + I_3^h \quad (11)$$

在最后一个阶段, 信息流的传递过程几乎与第 2 个阶段完全相同。唯一不同的是, 在获得第 3 阶段子网络的输出特征 F_3^d 后, 本文设计了一个精细模块对特征表示进行后处理, 得到最后的残差特征 F_r 和最终的去雾图像 O_3

$$F_r = \text{Conv}_{1 \times 1}(\text{Refine}(F_3^d)) \quad (12)$$

$$O_3 = F_r + I_3^h \quad (13)$$

式中 $\text{Refine}(\cdot)$ 表示精细模块的后处理操作, 该模块由 3 个 Transformer 基本块组成。

2.2 子网络结构

由于自注意力机制在图像恢复任务上的成功^[14], 本文选择采用 Transformer 模块作为基本模块。图 2 给出了所提出的编码器-解码器子网络的结构。首先利用 Transformer 模块在每个阶段上进行多尺度特征提取, 然后在每个阶段的最后一个尺度上结合可变形注意力 Transformer^[22] 模块以进一步融合图像的空间结构信息, 增强模型的变换能力, 该部分将会在 2.2.2 节中详细介绍。最后, 解码器逐步将隐空间中的低分辨率特征表示恢复为高分辨率特征表示。本文将每个阶段上的编码器-解码器

之间的连接方式选择为简单的跳跃连接。此外,很多前期工作^[10,15]均利用转置卷积对图像特征进行上采样。但是,这种采样方式会导致最终恢复图像中出现明显的棋盘效应。为了缓解这种现象并提高去雾图像的质量,在所提出的网络中分别使用像素解重组(pixel-unshuffle)和像素重组(pixel-shuffle)采样操作对图像特征进行下采样和上采样。

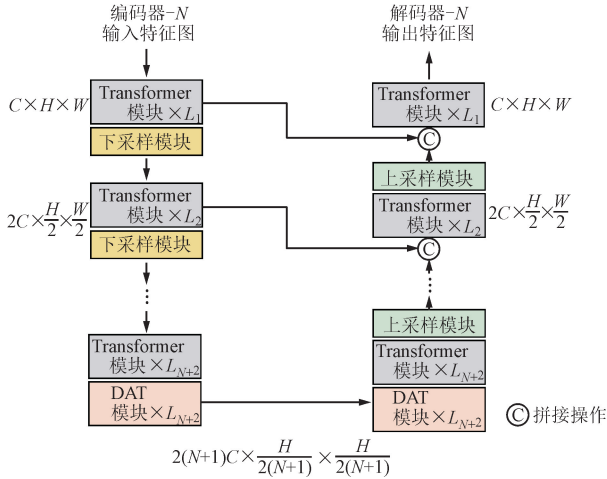


图 2 阶段 N 的编码器-解码器子网络结构

Fig. 2 Encoder-decoder sub-network at stage- N

多阶段的网络结构设计可以利用精细到粗糙的学习方式实现更好的局部去雾。特别是,较低阶段以小尺寸图像块作为输入,在精细尺度上更加关注图像的纹理和颜色细节;较高阶段以小图像块拼接成的大图像块作为输入,在粗尺度上强调图像的高级语义信息,以实现去雾的整体效果。因此,本文依据不同阶段所关注的重要信息将各阶段子网络设计为不同深度。具体而言,为了使模型的多个阶段分别关注于图像的细粒度特征信息和噪声相关的高级语义信息,本文选择在信息传递到更高阶段时进一步深化编码器-解码器结构。如图 2 所示, N 表示阶段数,在阶段 N 的子网络总层数表示为 $N+2$ 。具体地,第 1、2、3 阶段分别采用 3、4、5 层的编码器-解码器结构。在 3.3 节中通过实验证明了此种设计的优越性。

2.2.1 Transformer 模块结构

考虑到传统的基于卷积神经网络的方法不能有效地捕获远程依赖关系,而传统的视觉 Transformer 由于其庞大的计算量不能很好地应用于图像恢复任务。虽然近期的一些研究工作试图通过使用基于窗口的自注意力来降低计算复杂度,但是此方式仍然将上下文的聚合限制在局部窗口中。基于此,本文采用

包含 MDTA 和结合深度卷积的门控前馈网络(GD-FN)两个子模块的 Restormer^[14]作为主要的特征提取块,其结构分别如图 3(a)和图 3(c)所示。

在一般的自注意力机制中,会利用经过处理后的 3 种特征去计算加权之后的像素值,其中 Query 表示用来匹配其他单元的量,即自注意力机制中的查询特征;Key 表示用来被其他单元匹配,即键特征;Value 表示需要被提取的信息,即值特征,分别设为 Q, K, V 。MDTA 模块的核心思想不是直接计算空间维度上像素之间的自注意力,而是将 Q 和 K 转置,并计算跨信道的自注意力,具体计算过程如图 3(a)所示。具体而言,首先使用 1×1 的普通卷积(point-wise,表示为 $\text{Conv}_{1 \times 1}^B$)和 3×3 的深度卷积(depth-wise,表示为 $\text{Conv}_{3 \times 3}^{\text{depth}, B}$)分别对层归一化后的特征进行像素级跨信道上下文聚合和信道级局部上下文信息聚合,得到 Q, K 和 V

$$B = \text{Conv}_{3 \times 3}^{\text{depth}, I} (\text{Conv}_{1 \times 1}^I (\text{Norm}(F_{\text{in}}))) \quad (14)$$

$$B \in \{Q, K, V\}$$

式中: F_{in} 表示输入特征; $\text{Norm}(\cdot)$ 表示层归一化操作。将 Q, K 和 V 分别经过变形之后得到 $\hat{Q} \in \mathbb{R}^{H \times W \times C}$, $\hat{K} \in \mathbb{R}^{C \times H \times W}$ 和 $\hat{V} \in \mathbb{R}^{H \times W \times C}$,以计算信道级自注意力

$$\text{Attention}(\hat{Q}, \hat{K}, \hat{V}) = \hat{V} \text{Softmax}(A) \quad (15)$$

$$G = \hat{K} \hat{Q} / \alpha \quad (16)$$

式中: α 是一个可学习的缩放参数; G 代表获得具有 $\hat{C} \times \hat{C}$ 大小的转置注意力图,与标准的具有 $H \times W \times H \times W$ 大小的注意力图相比,参数量大大降低。将自注意力计算的结果与输入特征相加得到模块的输出特征

$$F_{\text{out}}^* = \text{Conv}_{1 \times 1} (\text{Attention}(\hat{Q}, \hat{K}, \hat{V})) + F_{\text{in}} \quad (17)$$

为了直观地说明信道级自注意力机制的优越性,本节将其与经典的基于卷积神经网络的信道注意力(CAB)^[10,30]进行比较,其结构如图 3(b)所示。可以发现,MDTA 和 CAB 都是计算与信道相关的注意力,并对特征信道之间的相关性进行建模。其中,MDTA 通过计算任意两个信道之间特征元的内积得到大小为 $\hat{C} \times \hat{C}$ 的注意力权重图。CAB 首先将整个信道的空间特征编码成一个全局特征,得到大小为 $\hat{C} \times 1 \times 1$ 的中间特征,然后通过两个全连接层学习信道之间的依赖关系,其中生成的中间特征大小为 $\hat{C}/r_1 \times 1 \times 1$ 。针对上述两种注意力权重的生成方式而言,CAB 在一开始就降低了空间维度,致使丢失了很多有价值的信息。MDTA 则通过内积的

计算方式充分考虑了每个特征元对所属信道相关权重的贡献。不仅如此,CAB通过一种类似门控的激励操作对信道间的全局相关性进行建模学习,而忽略了任意两个信道之间直接的显式建模,使得每个特征信道只有一个对应的全局相关性权值。与之不同的是,MDTA通过自注意力机制直接对任意两个信道(包括自身对自身)的相关性进行建模,并为每个特征通道生成一个 $\hat{C}$ 维的权值向量。因此,在MDTA生成的新特征图中,第 k 个信道的第 j 个位置的值 $F_{j,k}^*$ 是通过投影 $\mathbf{V}$ 中第 j 个位置的 $\hat{C}$ 维特征元向量 $\mathbf{F}_j$ 加权得到的

$$F_{j,k}^* = \mathbf{F}_j \mathbf{W}_k, j \in \{1, \dots, \hat{H}\hat{W}\}, k \in \{1, \dots, \hat{C}\} \quad (18)$$

其中 $\mathbf{W}_k \in \mathbb{R}^{\hat{C} \times 1}$ 表示第 k 个信道对应的权值向量。所以MDTA的相关性计算不仅体现在权值 $\mathbf{W}_k$ 的计算上,也体现在每一个特征元 $F_{j,k}^*$ 的计算上。然而,在CAB所产生的新特征中,特征元 $F_{j,k}^*$ 有如下计算过程

$$F_{j,k}^* = F_{j,k} W'_k \quad (19)$$

式中: $F_{j,k}$ 和 W'_k 都是标量,其中 $F_{j,k}$ 表示输入特征图中第 k 个信道的第 j 个位置的值, W'_k 表示第 k 个信道的权值。CAB的相关性计算仅体现在全局相关权重 W'_k 的计算上。由此可见,MDTA具有更强的信道之间的交互作用,这不仅体现在权值的生成过程中,也体现在新特征图的生成过程中。

另一方面,CAB只通过两个全连接层计算信道之间的相关性,但这两个全连接层的权值在推理过

程中是固定的。它不能根据不断变化的输入内容进行调整。相反,MDTA通过自注意力来计算任意两个信道之间的相关性。即使输入内容在推理过程中发生变化,它也能根据新的内容灵活计算出相应的注意力权值。最后,由于CAB采用了门控的思想,即根据每个特征信道对结果的重要性,强调有用的信道,抑制对结果用处不大的信道。该种机制可能会过度关注与雾霾相关的特征,而牺牲了背景的固有特征,在3.4节中通过实验证明了这一观点。对于MDTA的结果,更多地强调了信道之间的相关性,其在解耦雾霾相关的特征的同时,尽量保留了背景的固有特征,即使不重要的特征也不会被抑制。

与传统的前馈网络(FN)^[28]类似,GDFN也使用了两个 1×1 卷积分别对特征信道进行扩展和将信道数减少到原始数量。与FN不同的是,GDFN使用了 3×3 深度卷积来聚合局部上下文信息。此外,还引入了附加的门控机构。GDFN的输出特征为

$$F_{\text{out}} = \text{Conv}_{1 \times 1}(\text{Gating}(F_{\text{out}}^*)) + F_{\text{out}}^* \quad (20)$$

$$\text{Gating}(X) = \varphi(\text{Conv}_{3 \times 3}^{\text{depth}}(\text{Conv}_{1 \times 1}(\text{LN}(F_{\text{out}}^*)))) \odot \text{Conv}_{3 \times 3}^{\text{depth}}(\text{Conv}_{1 \times 1}(\text{LN}(F_{\text{out}}^*))) \quad (21)$$

其中: $\varphi(\cdot)$ 表示GeLU激活; $\text{LN}(\cdot)$ 是层归一化; $\odot$ 代表元素级相乘。此处使用的门控机制可以起到控制信息流的作用,使得模型能够专注于丰富特征的上下文信息,捕捉到对结果重要的像素特征。其整体思路与传统的像素注意力模块(PAB)类似,如图3(d)所示。PAB的做法是先获得大小为 $\hat{C}/r_2 \times \hat{H} \times \hat{W}$

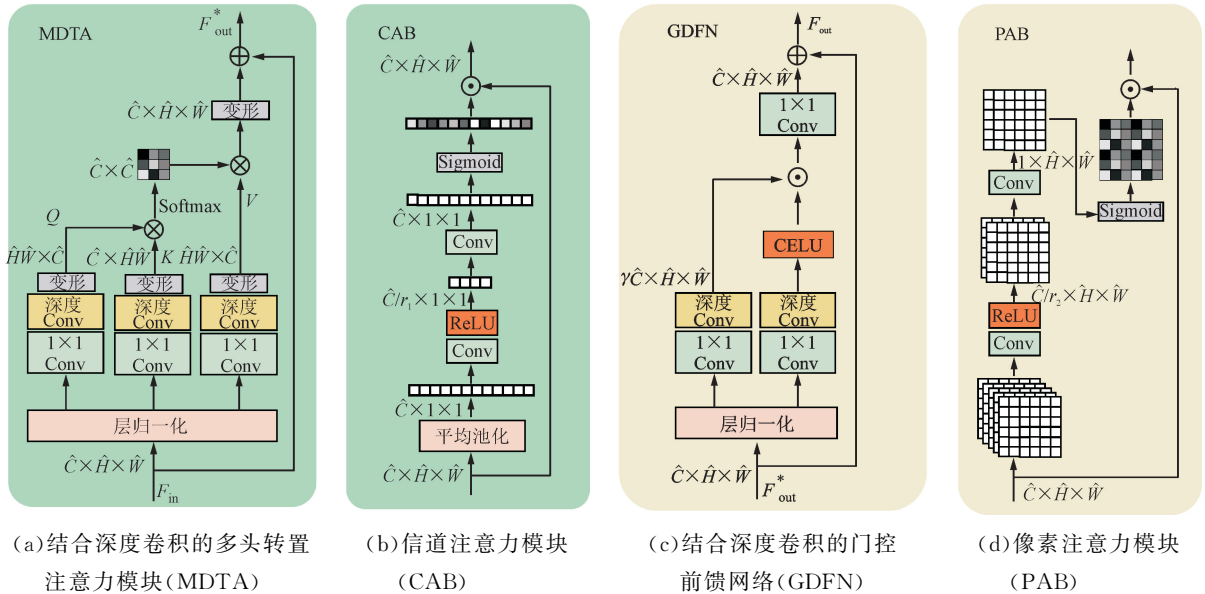


图 3 转置自注意力机制与传统注意力机制对比

Fig. 3 Transposed self-attention mechanism versus traditional attention mechanism

$\hat{H} \times \hat{W}$ 的中间特征,再获得大小为 $1 \times \hat{H} \times \hat{W}$ 的权值矩阵,通过两步压缩对不同通道中相同位置的特征点进行相同的权值处理。GDFN 则获得一个大小为 $\gamma \hat{C} \times \hat{H} \times \hat{W}$ 的权值图,对整个特征图中所有位置的特征点进行不同的权值处理。然而,GDFN 使用比 PAB 多 $\gamma \hat{C}$ 倍的权重参数,因此更具优越性。

2.2.2 可变形注意力 Transformer(DAT)

由于所采用的 MDTA 模块是在信道维度上计算自注意力,只是通过深度卷积来对像素之间的依赖关系进行隐式地建模,而忽略远距离的像素之间的显式建模。为了处理此问题,并避免使用密集注意力(如 ViT^[28])所增加的计算成本,需要使用一种高效的注意力机制来降低计算复杂度。可变形的自注意力机制则是最佳选择。

本文将 DAT 模块^[22]部署在各阶段的子网络深处,使网络能够自适应地专注于感兴趣区域的计算,提高网络的转换能力,融合更多的空间结构信息。如图 4 所示,ViT 和 DAT 中所有的查询特征都共享相同的感受野。然而,由于 ViT 使用的是全局注意力,因此感受野覆盖整个特征空间。与之不同的是,DAT 在保持所有的查询特征都与同一组键特征和值特征交互的同时,额外学习一组键特征和值特征的偏移量,将它们移动到重要的位置,使网络只关注于感兴趣的区域,大大降低了计算复杂度。

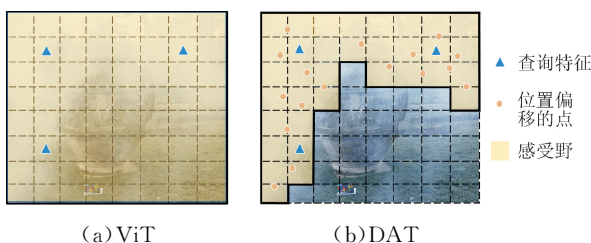


图 4 ViT 和 DAT 感受野大小对比

Fig. 4 Comparison of receptive field size between DAT and ViT

2.3 多监督对比学习

近年来,对比学习在高级视觉任务中取得了很多成果,但是在高级视觉任务中所获得的全局视觉表示并不适合纹理丰富的底层视觉任务。因此,将对学习的思想直接应用到去雾任务中是极具挑战性的。考虑到目前已有的去雾工作主要是受局部图像去雾不理想以及全局去雾不彻底的限制,本文拟通过一种图像块级的细粒度对比学习方法来改善这两个问题。受相关工作^[15,21,26]的启发,本文提出了一种新的多监督对比学习方法,即利用多个图像块

级的正样本信息来恢复雾霾图像,具体方法如图 5 所示。值得注意的是,与之前工作^[15,26]不同的是,本文提出的方法是针对单幅图像的,不需要数据集的其他图像数据,且对比样本数量较多。具体而言,从恢复后的图像中随机选取一个图像块作为锚样本,并从以下 3 个方面构建正样本:

(1)将清晰图像中与锚样本内容相同的图像块作为第 1 个正样本,以引导模型尽可能恢复与清晰图像中相同的细节纹理;

(2)将清晰图像中的其他随机图像块作为第 2 个正样本,鼓励模型不要过多地关注于图像内容,而是专注于学习与雾霾噪声相关的特征;

(3)将恢复图像中与锚样本内容不同的其他随机图像块作为第 3 个正样本,以保证图像不同区域去雾的一致性。

另外,负样本的选择策略是固定的,即是将雾霾图像中的所有图像块构造为负样本。总之,本文提出的对比学习方法强调让模型在更多地关注相似的图像内容的同时也专注于与雾霾相关的特征。对于图像块的嵌入特征提取,由于从预训练的 VGG 网络中获得的特征往往是用于分类任务的高级视觉特征,因此,本文没有像之前的工作^[15]那样使用其作为嵌入提取网络。

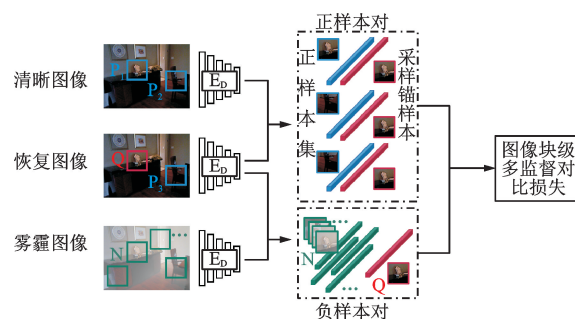


图 5 所提出的多监督对比学习方法

Fig. 5 Illustration of our proposed multi-supervised contrastive learning

由于生成对抗网络(GAN)中使用的鉴别器能够正确地分类输入图像是否为伪图像,因此,判别器是具有退化可知特性的^[21]。基于此,又考虑到本文使用的是图像块级的对比学习方法,因此,本工作采用 PatchGAN 鉴别器作为特征嵌入网络,即使恢复的图像非常接近于清晰的图像,判别器也可以正确地做出识别,所以对比损失也依然可以发挥作用。具体来说,分别将锚样本、正样本和负样本送入到特征嵌入网络 E_D 中以获得相应的嵌入特征 a_i 、 p_j 和 n_k 。然

后,利用所提出的多监督对比损失最大化图像块之间的互信息。对于第 i 个锚样本,有如下计算过程

$$L_i = -\frac{1}{J} \sum_{j=1}^J \ln \frac{\exp(a_i p_j / \tau)}{\exp(a_i p_j / \tau) + \sum_{k=1}^K \exp(a_i n_k / \tau)} \quad (22)$$

式中: J 是正样本数,本文 $J = 3$; K 是负样本数。假设锚样本的总数为 Y ,则总的对比损失 L_{SupCL} 表示为

$$L_{\text{SupCL}} = \frac{1}{Y} \sum_{i=1}^Y L_i \quad (23)$$

本工作采用类似 GAN 的框架来训练判别器,通过强制区分恢复图像和清晰图像以获得具有退化感知的特征嵌入网络。

2.4 损失函数

本工作对整个去雾模型进行了端到端的优化。模型的任意阶段 N 通过预测残差图像 R_N 以获得去雾后的图像 $O_N = I^h + R_N$,将清晰图像表示为 I^g ,则每个阶段 N 上的损失函数如下所示

$$L_N = \lambda_1 L_1(I^g, O_N) + \lambda_2 L_{\text{vgg}}(I^g, O_N) + \lambda_3 L_{\text{SupCL}}(E_D(I^g), E_D(I^h), E_D(O_N)) \quad (24)$$

其中感知损失 L_{vgg} 采用了在 ImageNet 上进行过预训练的 VGG16 进行特征提取

$$L_{\text{vgg}} = \sum_{j=1}^3 \frac{1}{C_j H_j W_j} \|\Phi_j(I^g) - \Phi_j(O_N)\|_2^2 \quad (25)$$

式中: C_j, H_j, W_j 分别为特征图的信道数、高和宽; Φ_j 是 j 层的激活函数,本文取 VGG16 中的“relu1_2”、“relu2_2”和“relu3_3”3 层。特别地,对于多阶段网络的训练,总的损失函数有

$$L = \sum_{N=1}^N (\lambda_1 \theta_N L_1 + \lambda_2 \eta_N L_{\text{vgg}} + \lambda_3 \mu_N L_{\text{SupCL}}) \quad (26)$$

式中: λ_1, λ_2 和 λ_3 分别代表每个损失项的权重,为损失系数; θ_N, η_N 和 μ_N 则分别代表每项损失在不同阶段的权重,为阶段系数。

3 实验验证

首先介绍训练网络所使用的数据集,其次列出网络的实现细节,然后通过消融实验来证明网络各组分的有效性,最后将提出的模型与当前最新的算法进行定性和定量的比较。

3.1 数据集和评价指标

3.1.1 RESIDE 合成数据集

RESIDE 数据集是一个应用广泛的合成数据集,它由 3 个主要的子数据集构成:室内训练集(ITS)、室

外训练集(OTS)和综合训练目标测试集(SOTS)。其中,室内训练集由 13 990 幅室内雾霾图像和 1 399 幅清晰图像组成,室外训练集由 313 915 幅室外雾霾图像和 8 971 幅清晰图像组成,测试集中分别包含 500 幅室内雾霾图像和 500 幅室外雾霾图像。

3.1.2 真实环境数据集

本工作进一步使用了 3 个真实环境下的小规模数据集来训练模型:Dense-Haze、NH-Haze 和 NH-Haze2。其中,Dense-Haze^[31]是 CVPR-NTIRE 2019 竞赛中去雾赛道使用的数据集,其雾的特点是密集且均匀的,它包含 45 对训练数据、5 对验证数据和 5 对测试数据。在本工作中,将训练集和验证集合并成训练集,将测试集用于模型的测试。NH-Haze^[32]是 CVPR-NTIRE 2020 竞赛中去雾赛道使用的数据集,它包含 45 对训练数据、5 对验证数据和 5 对测试数据,该数据集中雾的特点是不均匀分布。在本工作中,同样使用训练集和验证集一起作为训练集,将 5 对测试集数据用于测试。NH-Haze2^[33]为 NH-Haze 的扩展,是 CVPR-NTIRE 2021 竞赛中去雾赛道使用的数据集,但是它只包含了 25 对训练数据、5 对验证数据和 5 对测试数据,由于验证集和测试集还没有公开,所以本文按照之前工作的经验^[15]使用 1~20 号图像作为训练集,21~25 号图像作为测试集。

3.1.3 评价指标

为了评价所提出的模型及方法的有效性,本文使用常用的峰值信噪比(PSNR,设为 R_{PSN})和结构相似度指数(SSIM,设为 S_{SIM})来评价模型的性能。其中, R_{PSN} 或 S_{SIM} 值越高,代表恢复的效果越好。本文使用均值和标准差的形式处理测试集实验结果,若需要同时使用均值和标准差表示实验结果,则采用 $\pm$ 符号,即(均值 $\pm$ 样本标准差),在无特殊说明时,结果均以均值的形式给出。其中指标均值越高代表整体恢复效果越好,指标标准差不变或越小代表模型鲁棒性更好,发散程度越小,本实验以均值指标为主,标准差指标为辅。

3.2 实验设置

将每张图像随机裁剪为 192×192 大小的图像块用于训练,实现了随机旋转($90^\circ, 180^\circ, 270^\circ$)和随机水平翻转的数据增强。对于最多 5 个层级的编码器-解码器结构,各层使用 Transformer 的模块数为 $[1, 2, 2, 2, 4]$,隐空间层使用的 DAT 模块数设为 4,细化模块使用的 Transformer 模块数设为 3,训练使用的批大小设置为 4,优化器使用 AdamW 优化

器($\beta_1=0.9, \beta_2=0.999$, 权重衰减设置为 1×10^{-4}), 初始学习率设为 2×10^{-4} , 使用余弦退火策略逐渐将学习率减小至 1×10^{-6} , 总共训练 450 000 次。在合成数据集的实验设置中, 将式(26)中的各项系数设置为 $\lambda_1=1, \lambda_2=0.2, \lambda_3=0.01, \theta_N=\eta_N=0.33, \mu=\{\mu_N|0.8, 0.1, 0.1, N=1, 2, 3\}$ 。由于真实数据集的图像数非常少, 仅包含几十张数据。为了防止过拟合, 也为了更好地恢复图像, 针对这些小数据集, 本文减少模型的大小和参数数量, 将三阶段训练调整为二阶段训练, 训练迭代次数减少为 300 000 次。同时, 由于小规模真实数据集与大规模合成数据集在结构上差别较大, 所以损失函数中的各系数对应调整为 $\lambda_1=1, \lambda_2=1, \lambda_3=0.001, \theta_N=\eta_N=0.5, \mu=\{\mu_N|1.0, 0, N=1, 2\}$, 并将初始学习率调整为 5×10^{-4} 。

3.3 消融实验

为了证明所提出模型中每个组件和方法的必要性, 分别在合成数据集 ITS 和真实数据集 NH-Haze 上进行了消融研究。在消融实验中, 本文网络模型的训练设定为 150 000 次迭代。实验结果如表 1 所示。为了便于对比, 构建了一个基础网络作为去雾网络的基线模型。该网络是基于传统的卷积神经网络构建的, 如表 1 中构建的 M_1 模型所示, 提取特征的主要模块是 CAB 模块, 网络结构是 U-Net 结构。随后, 不断改变基础网络: M_2 模型将提取特征的模块用 Transformer 模块代替; M_3 模型在网络中添加 DAT 模块, 此时构成一阶段模型; M_4 模型训练时使用所提出对比学习损失; M_5 模型增加网络阶段, 即采用二阶段模型; M_6 模型在 M_5 的训练过程中使用所提出的对比学习损失; M_7 模型进一步增加网络阶段, 即采用三阶段模型; M_8 模型添加细化模块; M_9 模型在 M_8 的训练过程中使用所提出的对比学习损失。

3.3.1 转置自注意力机制的有效性

本工作中使用的转置自注意力可以敏感地捕捉到特征图中信道与信道之间的内部相互关系。因此, 如表 1 构建的 M_2 模型所示, 使用 Transformer 模块作为特征提取模块就可以在单阶段中将 R_{PSN} 均值指标提升 2.72 dB (RESIDE)、0.87 dB (NH-Haze)。而 R_{PSN} 标准差指标增大 0.29 dB (RESIDE)、0.65 dB (NH-Haze), 结果发散程度增大, 表明虽然测试集中的图像恢复质量都有提升, 但是每张图像的提升程度不同, 在只有 5 张测试数据的小数据集中尤为明显。

3.3.2 可变形自注意力机制的有效性

可变形注意力机制通过学习键特征和值特征的位置偏差以进一步关注局部区域内的重要特征。本文期望通过引入可变形注意力机制使网络在隐空间中更加准确地关注到与雾有关的噪声特征, 弥补信道级转置自注意力没有显式建模长距离像素依赖关系的不足。从表 1 构建的 M_3 模型的结果可知, 通过引入可变形自注意力机制, 可以在单阶段中将 R_{PSN} 均值指标提升 0.15 dB (RESIDE)、0.09 dB (NH-Haze)。与此同时, 在 RESIDE 数据集上的标准差指标增大 0.31 dB, 表明单阶段模型只针对个别有雾场景有了提升, 泛化性不强, 需要增加阶段数。对于 NH-Haze 数据集, 标准差指标下降了 0.28 dB, 发散程度进一步降低。

3.3.3 增加网络阶段以及精细模块的有效性

为了使网络更加关注图像细节特征, 本工作不断地增加网络阶段, 如表 1 中 M_3 、 M_5 和 M_7 模型的结果所示, 每增加一个阶段, R_{PSN} 均值指标都会有明显的提升。其中, 对于 RESIDE 数据集, 二阶段模型比一阶段提升 1.11 dB, 三阶段比二阶段提升 0.30 dB。标准差指标也都有明显的下降, 具体而言, 二阶段比一阶段下降 0.28 dB, 三阶段比二阶段下降 0.01 dB; 对于 NH-Haze 数据集, 为了防止小数据集的过拟合问题, 本工作只在二阶段上做了验证, 结果表明二阶段比一阶段提升了 0.33 dB 的均值指标, 同时降低了 0.11 dB 的标准差指标。以上结果都验证了网络多阶段设计的有效性。此外, 在三阶段的 M_8 模型中, R_{PSN} 均值指标提升 0.19 dB, 与此同时标准差指标有略微波动(增大 0.09 dB), 证明精细模块的有效性。

3.3.4 多监督对比学习方法的有效性

本节验证了对所提出的多监督对比学习方法的有效性。如表 1 构建的 M_4 、 M_6 和 M_9 模型所示, 为了保证实验的完整性, 在单阶段和多阶段上都对所提出的方法进行了验证。结果显示, 在单阶段上, R_{PSN} 均值指标分别在两个数据集上提升了 0.15 dB (RESIDE)、0.32 dB (NH-HAZE), 而标准差指标有略微波动甚至下降, 在 RESIDE 和 NH-HAZE 数据集上分别增大 0.01 dB 和减小 0.16 dB。此外, 可以观察到, 对于小规模数据集, 使用对比学习方法训练网络之后, 指标甚至与二阶段的结果相当, 表明针对非均匀雾质, 所提出的对比学习方法对结果的提升会更明显。在二阶段上, R_{PSN} 均值指标分别提升了 0.14 dB (RESIDE)、0.065 dB (NH-HAZE); 与此同时, 对于标准差指标而言, 在 RESIDE 数据集上有

上升波动,增大 0.20 dB,表明还有继续增大阶段数的必要,而在 NH-HAZE 数据集上下降 0.13 dB,进一步提升恢复性能。在三阶段上, R_{PSN} 均值指标在

RESIDE 数据集上提升了 0.14 dB,标准差指标下降 0.13 dB。以上结果都充分证明了所提出对比学习方法的有效性。

表 1 模型结构和损失函数在不同数据集上的消融实验结果

Table 1 Results of ablation experiments on model structures and loss functions on different datasets

模型编号	阶段	网络组件	精细模块	多监督对比损失	R_{PSN}	
					SOTS-室内	NH-HAZE
M_1	1	CAB	×	×	30.390 5±2.548 0	18.951 2±1.145 0
M_2	1	Transformer	×	×	33.115 3±2.844 5	19.818 8±1.793 1
M_3	1	Trans+DAT	×	×	33.268 5±3.162 9	19.907 9±1.514 5
M_4	1	Trans+DAT	×	✓	33.421 1±3.177 8	20.224 4±1.358 4
M_5	2	Trans+DAT	×	×	34.378 7±2.883 2	20.243 0±1.404 2
M_6	2	Trans+DAT	×	✓	34.519 2±3.082 4	20.308 0±1.277 3
M_7	3	Trans+DAT	×	×	34.681 0±2.871 0	
M_8	3	Trans+DAT	✓	×	34.869 0±2.906 3	
M_9	3	Trans+DAT	✓	✓	35.008 0±2.779 2	

注:✓表示相应的模型采用了该模块/损失;×表示相应的模型没有采用该模块/损失;加粗表示最优的结果。

3.3.5 各阶段子网络深度选择的优越性

为了证明本工作对各阶段子网络深度设计的优越性,本节在 RESIDE 数据集上对多种不同深度的设计进行了详细的实验,结果如表 2 所示。结果表明,三阶段模型在第 1、2 和 3 阶段的子网络层数(L_1, L_2, L_3)分别采用(3,4,5)的层数设计可以达到最优的 R_{PSN} 结果,次优且相近的 S_{SIM} 结果。因此,本工作使用(3,4,5)的子网络深度设计。

表 2 各阶段子网络深度设计的定量结果

Table 2 Quantitative results of sub-network depth design at each stage

(L_1, L_2, L_3)	R_{PSN}	S_{SIM}
(3,3,3)	34.14	0.984 3
(3,3,4)	34.35	0.983 9
(3,4,4)	34.55	0.984 4
(3,4,5)	35.01	<u>0.985 9</u>
(4,4,4)	<u>34.93</u>	0.986 2
(4,4,5)	34.31	0.983 9
(4,5,5)	34.85	0.985 9

注:加粗表示最优的结果;下划线表示次优的结果;所有结果以均值的形式给出。

3.4 对比实验

本小节中,在合成数据集和真实数据集上对所提出的方法与最新的去雾方法做了定性和定量比较。这些方法包括基于先验的方法,例如 DCP (TPAMI'10)^[3];基于物理模型的方法,例如 AOD-Net (ICCV'17)^[34];基于深度学习直接进行雾图到清晰图转换的方法,例如 FFA-Net (AAAI'20)^[10]、AECD (CVPR'21)^[15]、MSFNet (TIP'21)^[11];基于 Transformer 的最新方法,例如 U²-Former^[26],但是该工作的报道者未提供相关代码。

3.4.1 合成数据集的结果

表 3 中总结了在 ITS 和 OTS 数据集上,本文提出的模型与最新去雾模型的性能指标。与其他工作相比,本文方法实现了最佳的性能指标(ITS 上的 R_{PSN} 为(38.66±2.62) dB, S_{SIM} 为(0.990 9±0.003 1);OTS 上的 R_{PSN} 为(35.31±2.79) dB, S_{SIM} 为(0.977 3±0.007 0)。与性能次优的 AECD 相比,本文的模型在 ITS 上 R_{PSN} 均值指标提升了 1.49 dB,结果稳定,程度适中, S_{SIM} 相近且发散程度最低;在 OTS 上 R_{PSN} 提升了 1.45 dB, S_{SIM} 指标略有降低,且两种指标的发散程度均最低。此外,与同样是基于 Transformer 的工作 U²-Former^[26] 相比,本文模型性能也表现优异。

表3 在合成数据集上的定量对比实验结果

Table 3 Quantitative comparisons with state-of-the-art methods on synthetic datasets

方法	R_{PSN}		S_{SIM}	
	ITS	OTS	ITS	OST
DCP ^[3]	19.65±2.69	17.21±3.46	0.860 0±0.050 0	0.863 1±0.055 4
AOD-Net ^[31]	19.51±3.34	24.17±3.59	0.841 4±0.073 1	0.928 6±0.053 6
FFA-Net ^[10]	36.39±2.07	33.38±3.44	0.988 6±0.004 4	0.980 4±0.009 0
MSFNet ^[11]	34.74±2.79	32.10±3.35	0.989 5±0.004 3	0.984 9±0.008 2
AECR ^[15]	<u>37.17±2.19</u>	<u>33.86±3.40</u>	<u>0.990 1±0.003 6</u>	<u>0.984 6±0.009 1</u>
U ² -Former ^[26]	36.42	31.10	0.988 0	0.976 0
MSTCNet	38.66±2.62	35.31±2.79	0.990 9±0.003 1	0.977 3±0.007 0

注:加粗表示最优的结果;下划线表示次优的结果;所有结果以均值的形式给出。

3.4.2 真实数据集的结果

同样,本小节也将本文提出的模型与上述方法在真实数据集上做定量和定性的比较,性能指标如表4所示。可以发现,本工作在3个数据集上分别以 R_{PSN} (16.23±2.43)dB、(20.36±1.59)dB、(21.08±2.11)dB, S_{SIM} (0.586 1±0.161 3)、(0.688 0±0.036 5)、(0.854 8±0.052 2)的指标优于所有的对比方法。

本小节中还以Dense-Haze和NH-Haze数据集上的恢复结果为例,将提出的模型与上述方法在恢复图像的可视化质量上进行了定性的比较,如图6和图7所示。可以发现,与合成数据集相比,Dense-Haze浓雾数据集和NH-Haze非均匀雾质数据集具有更大的挑战性。具体地,如图6所示,在浓雾数据集上,早期的传统方法DCP和AOD-Net几

乎不能够达到去雾的效果;而 R_{PSN} 指标次优的FFA-Net虽然在一定程度上恢复了一些图像的纹理细节,但是在色彩恢复上严重失真,例如图6中红色方框标出的区域中色彩板的颜色完全没有恢复;MSFNet方法恢复的图像质量与FFA-Net相比稍有提高,但是仍然存在严重的色彩失真,对物体边缘周围的去雾效果也较差;AECR方法恢复的图像质量有明显的改善,但是存在过度去雾的问题,例如图6中第2张恢复图像中,树上的白霜特征也被当成雾去掉。此外,红色方框标出的区域中色彩板的颜色恢复效果依然不佳,也会产生斑纹伪影且恢复图像整体不够平滑,影响视觉效果;与上述工作相比,本文方法生成的恢复图像在视觉效果上进一步得到了改善,且恢复效果更佳,尤其是各个测试图像中用框线标注出的区域,细节纹理特征的恢复更加明显。

表4 在真实数据集上的定量对比实验结果

Table 4 Quantitative comparisons with state-of-the-art methods on real datasets

方法	R_{PSN}			S_{SIM}		
	Dense-Haze	NH-Haze	NH-Haze2	Dense-Haze	NH-Haze	NH-Haze2
DCP ^[3]	11.75±1.84	11.24±1.39	11.68±1.64	0.353 9±0.137 1	0.2354±0.116 7	0.647 6±0.128 6
AOD-Net ^[31]	13.99±2.56	15.66±0.98	17.02±2.58	0.499 9±0.154 2	0.5298±0.039 1	0.714 2±0.085 6
FFA-Net ^[10]	<u>16.12±3.20</u>	18.44±0.91	20.43±2.87	<u>0.568 8±0.148 5</u>	0.6384±0.033 6	0.823 7±0.057 5
MSFNet ^[11]	15.00±3.37	<u>18.91±1.14</u>	19.77±3.08	0.534 9±0.143 7	<u>0.6533±0.037 6</u>	0.839 0±0.061 1
AECR ^[15]	15.47±4.18	18.65±1.78	<u>20.86±1.23</u>	0.475 6±0.118 6	0.6495±0.045 4	<u>0.843 0±0.040 3</u>
MSTCNet	16.23±2.43	20.36±1.59	21.08±2.11	0.586 1±0.161 3	0.6880±0.036 5	0.854 8±0.052 2

注:加粗表示最优的结果;下划线表示次优的结果;所有结果以均值的形式给出。

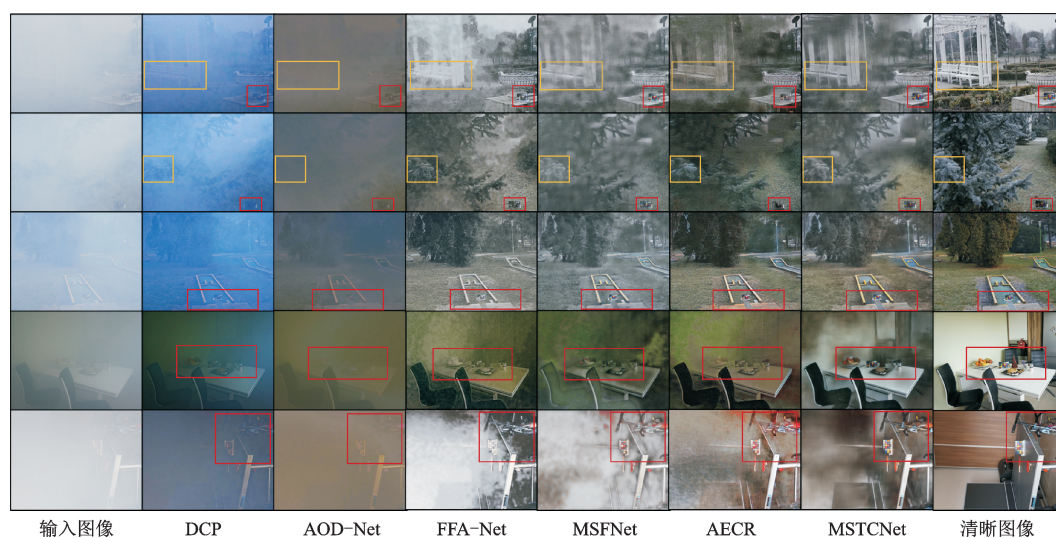


图 6 Dense-Haze 数据集上的可视化结果对比

Fig. 6 Comparison of visualization results on the Dense-Haze dataset



图 7 非数据集有雾图像的去雾效果

Fig. 7 Dehazing effect of non-dataset images

如图 8 所示,在非均匀雾质数据集上,传统方法 DCP 和 AOD-Net 具有和浓雾数据集上类似的问题,去雾效果不明显;而 FFA-Net、MSFNet 和

AECR 都具有局部去雾效果不彻底、物体边缘去雾效果差、整体图像色彩失真以及亮度变化等不足。与上述工作相比,本文模型生成的恢复图像进一步减轻

了上述出现的各种问题,尤其是红框标注出的区域,效果更佳。总的来说,本文提出的方法生成了最自然的恢复图像,在视觉效果上与清晰图像更为贴近。

此外,注意到 FFA-Net 与 AECR 两种方法都是基于门控信道注意力机制的方法,而这两种方法在 Dense-Haze 和 NH-Haze 数据集上都表现出了过度去雾,甚至将背景图像的特征当成雾霾去除的现象。

3.4.3 非数据集真实世界有雾图像的去雾结果

为了验证模型的泛化性与适应性,本节选用文献[11]中所使用的真实世界有雾图像作为输入(此文并未提供对应的清晰图像)。由于这些图像大多都是室外图像,因此本节选择在室外训练集(OTS)和非均匀雾质数据集(NH-Haze)两个数据集上经过预训练的模型作为测试的去雾模型,并挑选可视

化质量较高的图像作为展示。

非数据有雾图像的去雾效果如图 7 所示,由图 7 可以看出,传统的去雾算法 DCP 由于不需要网络模型的支撑,在非数据集图像的去雾泛化性较好。但是,对于天空区域的去雾仍然存在问题。AOD-Net 与 FFA-Net 两种模型的去雾效果较差,模型泛化性不强,基本无法达到可观的去雾效果,MSFNet 与 AECR 两种模型的去雾效果较好,但是仍存在人造伪影和去雾不彻底的现象。相比之下,本文提出的 MSTCNet 模型可以捕捉到人眼都无法分辨的区域,如图 7 中浓度较大的雾图以及红色方框标注出的区域。此外,MSTCNet 模型可以产生最佳的可视化效果以及颜色恢复效果,可以敏感地捕捉到细节纹理特征。因此,本文提出的方法在非数据集真实有雾场景中也可以达到最好的去雾效果。

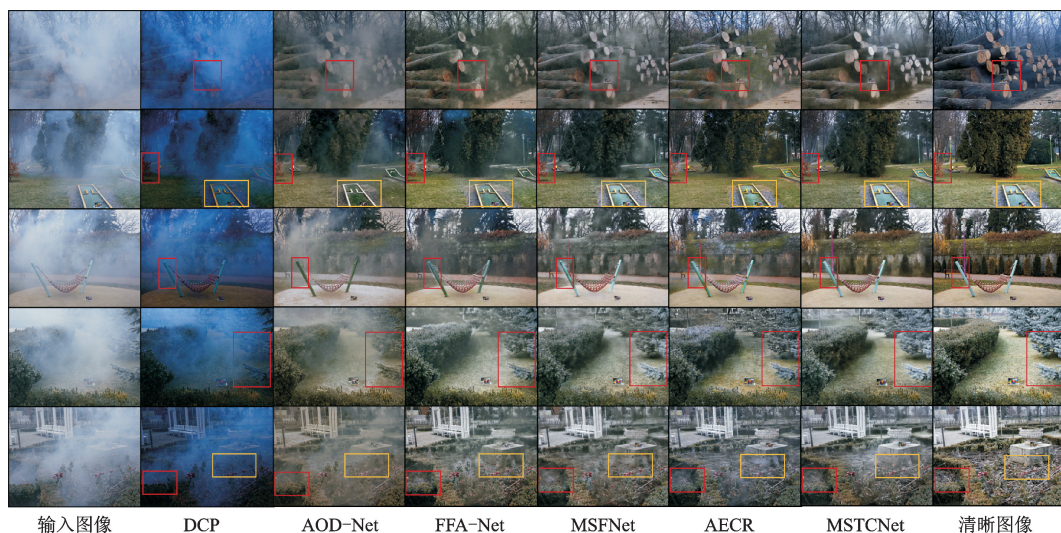


图 8 NH-Haze 数据集上的可视化结果对比

Fig. 8 Comparison of visualization results on the NH-HAZE dataset

4 结 论

本文提出了一种采用对比学习与自注意力机制的单幅图像去雾方法。所提出的去雾网络采用多阶段渐进式结构逐步改善去雾效果,其中基本的特征提取模块采用了转置 Transformer 模块。此外,为了使网络能够更好地感知雾霾图像中各区域的雾浓度,并弥补转置 Transformer 在显式建模上的不足,所提出的网络在各阶段的子网络最深层嵌入了可变形注意力 Transformer 模块。对于模型的训练,本文提出了一种多监督对比学习方法和相应的对比学习损失,该损失函数充分挖掘了单幅图像中图像块

之间的互信息,确保在嵌入空间中恢复后的图像块的特征表示离清晰图像块的特征表示更近,离雾霾图像块的特征表示更远。不仅如此,本文模型在训练中应用了多阶段监督学习策略,保证了各个阶段中特征学习的有效性。本文对所提出的去雾网络在 5 个数据集上进行了大量的实验与综合评估,并与其他方法进行比较。通过定性和定量分析的结果,证明了所提出的网络可以在去雾任务中达到最佳的效果。

参考文献:

- [1] HASSAN H, MISHRA P, AHMAD M, et al. Effects of

- haze and dehazing on deep learning-based vision models [J/OL]. *Applied Intelligence*, 2022 [2022-05-10]. DOI: 10.1007/s10489-022-03245-5.
- [2] NARASIMHAN S G, NAYAR S K. Vision and the atmosphere [J]. *International Journal of Computer Vision*, 2002, 48(3): 233-254.
- [3] HE Kaiming, SUN Jian, TANG Xiaoou. Single image haze removal using dark channel prior [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(12): 2341-2353.
- [4] BERMAN D, TREIBITZ T, AVIDAN S. Non-local image dehazing [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 1674-1682.
- [5] 黄文君, 李杰, 齐春. 低秩与字典表达分解的浓雾霾场景图像去雾算法 [J]. *西安交通大学学报*, 2020, 54(4): 118-125.
- HUANG Wenjun, LI Jie, QI Chun. A defogging algorithm for dense fog images via low-rank and dictionary expression decomposition [J]. *Journal of Xi'an Jiaotong University*, 2020, 54(4): 118-125.
- [6] 高涛, 刘梦尼, 陈婷, 等. 结合暗亮通道先验的远近景融合去雾算法 [J]. *西安交通大学学报*, 2021, 55(10): 78-86.
- GAO Tao, LIU Mengni, CHEN Ting, et al. A far and near scene fusion defogging algorithm based on the prior of Dark-light channel [J]. *Journal of Xi'an Jiaotong University*, 2021, 55(10): 78-86.
- [7] 杨燕, 陆鑫璇. 结合自适应亮度变换不等式估计透射率的图像去雾方法 [J]. *西安交通大学学报*, 2021, 55(6): 69-76.
- YANG Yan, LU Xinxuan. An image dehazing method combining adaptive brightness transformation inequality to estimate transmittance [J]. *Journal of Xi'an Jiaotong University*, 2021, 55(6): 69-76.
- [8] 黄鹤, 胡凯益, 宋京, 等. 雾霾线求解透射率的二次优化方法 [J]. *西安交通大学学报*, 2021, 55(8): 130-138.
- HUANG He, HU Kaiyi, SONG Jing, et al. A twice optimization method for solving transmittance with haze-lines [J]. *Journal of Xi'an Jiaotong University*, 2021, 55(8): 130-138.
- [9] QU Yanyun, CHEN Yizi, HUANG Jingying, et al. Enhanced Pix2pix dehazing network [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2019: 8152-8160.
- [10] QIN Xu, WANG Zhilin, BAI Yuanchao, et al. FFA-net: feature fusion attention network for single image dehazing [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI Press, 2020: 11908-11915.
- [11] ZHU Xinshan, LI Shuoshi, GAN Yongdong, et al. Multi-stream fusion network with generalized smooth L_1 loss for single image dehazing [J]. *IEEE Transactions on Image Processing*, 2021, 30: 7620-7635.
- [12] YAN Jingke, WANG Xin, CAI Jingye, et al. Medical image segmentation model based on triple gate Multi-Layer perceptron [J]. *Scientific Reports*, 2022, 12(1): 6103.
- [13] ZHAO Hengshuang, JIA Jiaya, KOLTUN V. Exploring self-attention for image recognition [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 10073-10082.
- [14] ZAMIR S W, ARORA A, KHAN S, et al. Restormer: efficient transformer for high-resolution image restoration [J/OL]. *arXiv*, 2022[2022-01-10]. DOI: 10.48550/arXiv.2111.09881.
- [15] WU Haiyan, QU Yanyun, LIN Shaohui, et al. Contrastive learning for compact single image dehazing [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2021: 10546-10555.
- [16] ZAMIR S W, ARORA A, KHAN S, et al. Multi-stage progressive image restoration [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2021: 14816-14826.
- [17] ZHANG Hongguang, DAI Yuchao, LI Hongdong, et al. Deep stacked hierarchical multi-patch network for image deblurring [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2019: 5971-5979.
- [18] TAO Xin, GAO Hongyun, SHEN Xiaoyong, et al. Scale-recurrent network for deep image deblurring [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 8174-8182.
- [19] LIU Xiaohong, MA Yongrui, SHI Zhihao, et al. GridDehazeNet: attention-based multi-scale network for image dehazing [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2019: 7313-7322.
- [20] WANG Yanbo, LIN Shaohui, QU Yanyun, et al. Towards compact single image super-resolution via con-

- rastive self-distillation [C]//Proceedings of the 30th International Joint Conference on Artificial Intelligence. Burlington, MA, USA; Morgan Kaufmann Publishers, 2021: 1122-1128.
- [21] WU Gang, JIANG Junjun, LIU Xianming, et al. A practical contrastive learning framework for single image super-resolution [J/OL]. arXiv, 2021[2022-02-01]. DOI: 10.48550/arXiv.2111.13924.
- [22] XIA Zhuofan, PAN Xuran, SONG Shiji, et al. Vision transformer with deformable attention [J/OL]. arXiv, 2022[2022-03-04]. DOI: 10.48550/arXiv.2201.00520.
- [23] CAI Bolun, XU Xiangmin, JIA Kui, et al. DehazeNet: an end-to-end system for single image haze removal [J]. IEEE Transactions on Image Processing, 2016, 25(11): 5187-5198.
- [24] REN Wenqi, LIU Si, ZHANG Hua, et al. Single image dehazing via multi-scale convolutional neural networks [C]//Computer Vision - ECCV 2016. Berlin, Germany: Springer International Publishing, 2016: 154-169.
- [25] HE Kaiming, FAN Haoqi, WU Yuxin, et al. Momentum contrast for unsupervised visual representation learning [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 9726-9735.
- [26] JI Haobo, FENG Xin, PEI Wenjie, et al. U2-former: a nested u-shaped transformer for image restoration [J/OL]. arXiv, 2021[2022-01-20]. DOI: 10.48550/arXiv.2112.02279.
- [27] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA; Curran Associates Inc., 2017: 6000-6010.
- [28] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale [J/OL]. arXiv, 2020[2021-09-18]. DOI: 10.48550/arXiv.2010.11929.
- [29] LIU Ze, LIN Yutong, CAO Yue, et al. Swin transformer: hierarchical vision transformer using shifted windows [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2021: 9992-10002.
- [30] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 7132-7141.
- [31] ANCUTI C O, ANCUTI C, SBERT M, et al. Dense-haze: a benchmark for image dehazing with dense-haze and haze-free images [C]//2019 IEEE International Conference on Image Processing (ICIP). Piscataway, NJ, USA: IEEE, 2019: 1014-1018.
- [32] ANCUTI C O, ANCUTI C, TIMOFTE R. NH-HAZE: an image dehazing benchmark with non-homogeneous hazy and haze-free images [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway, NJ, USA: IEEE, 2020: 1798-1805.
- [33] ANCUTI C O, ANCUTI C, VASLUIANU F A, et al. NTIRE 2021 NonHomogeneous dehazing challenge report [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway, NJ, USA: IEEE, 2021: 627-646.
- [34] LI Boyi, PENG Xiulian, WANG Zhangyang, et al. AOD-net: all-in-one dehazing network [C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2017: 4780-4788.

(编辑 武红江 亢列梅)