J. Vis. Commun. Image R. 90 (2023) 103722

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Full length article

Densely connected convolutional transformer for single image dehazing ☆

Anil Singh Parihar *,[1], Abhinav Java[1]

Machine Learning Research Laboratory, Department of Computer Science and Engineering, Delhi Technological University, Delhi, India

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Image Dehazing
Transformers
Attention

ABSTRACT

Image Dehazing is an important low-level vision task that aims to remove the haze from an image. In this paper, we proposed Densely Connected Convolutional Transformer (DCCT) for single image dehazing. DCCT is an efficient architecture that combines the multi-head Performer with the local dependencies. To prevent loss of information between features at different levels, we propose a learnable connection layer that is used to fuse features at different levels across the entire architecture. We guide the training of DCCT through a joint loss considering a supervised metric learning approach that allows us to consider both negative and positive features for a multi-image perceptual loss. We validate the design choices and the effectiveness of the proposed DCCT through ablation studies. Through comparison with the representative techniques, we establish that the proposed DCCT is highly competitive with the state of the art.

1. Introduction

Suspended atmospheric particles that cause light scattering are responsible for haze formation. As a result of haze formation, severe information loss and degradation occurs in the captured images. This negatively influences the performance of high-level vision tasks like classification [1–3], object detection [4–6], and segmentation [7–9]. The critical nature of the applications of these high level vision tasks in areas like [10], lane-detection [11] and surveillance [12] make it imperative to restore haze-free images.

Seminal methods [13,14] developed hand-crafted priors to estimate the unknown variables, J (true scene radiance) and A (ambient light intensity) in the atmospheric scattering model [15,16]. However, the complex distribution of haze across diverse set of inputs cannot always be captured by such priors.

The recent surge of Deep Learning has encouraged the development of more sophisticated pipelines that capture the relationship between haze-free and the hazy image from the data. With the availability of large-scale datasets [17], the deep learning based methods easily outperform their heuristic counterparts. Broadly, these data-driven methods either leverage the atmospheric scattering model [18–21] or directly estimate the clear image [22–24].

Despite the recent success of the learning-based methods, we identify the following *key challenges and avenues* in the existing methods to motivate our approach:

- Convolutional Neural Networks (CNNs) have been predominantly used to extract haze relevant features for clear image estimation. Some techniques [22,23] also use localized attention in convolutional backbones to improve performance in the dehazing task. More recently, self-attention driven transformers [25] have shown to effectively capture global information in images compared to CNNs; and have achieved state of the art performance in high level vision tasks (like classification) due to their representation capability. However, it is noteworthy that *the importance of transformers in image dehazing is still largely unexplored.*
- Recent single image dehazing methods [21–23,26,27] leverage *skip connections* to learn deep haze relevant features. Skip connections allow the combination of lower and higher level features in a neural network through non-discriminative addition. Consider a set of features f_l and f_h, simple addition based fusion $f_h + f_l$ does not account for their individual saliency for the dehazing task making the optimization harder. This can inadvertently lead to the abatement of more important feature sets. *We posit that a more robust approach to fusing high and low level features is required for the same.*
- Typically, a dehazing network is guided by two losses; spatial information loss (L1/L2) and a regularization term like perceptual (feature-level) loss. Perceptual losses for image dehazing attempt to make the predicted features similar to the ground truth example. *However establishing both similarity and dissimilarity*

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail address: parihar.anil@gmail.com (A.S. Parihar).

[1] Equal Contribution.

https://doi.org/10.1016/j.jvcir.2022.103722
Received 1 March 2022; Received in revised form 15 August 2022; Accepted 3 December 2022
Available online 7 December 2022

between multiple images enables learning generic representations [28], highlighting the need for a better contrastive constraint for single image dehazing.

In this paper, we propose *Densely Connected Convolutional Transformer (DCCT)*, a novel architecture for single image dehazing. The proposed architecture addresses the key limitations and challenges identified in the existing work as pointed out above. *First,* the proposed *DCCT* combines the local features with the global attention maps allowing better representation learning for image dehazing. *Second, DCCT* encourages better feature fusion through Dense Learnable Connection Layers. *Third,* we propose the use of a joint loss function that incorporates a strict contrastive constraint with both positive and negative examples. *Last,* Performer [29] further reduces the computational complexity of the architecture making our network more efficient.

The key contributions of our work can be summarized as follows:

- We design an end-to-end trainable Densely Connected Convolutional Transformer (DCCT) architecture; the first work combining local features captured by convolutions and global Performer attention in an encoder–decoder architecture for image dehazing.
- We preserve information flow across the network by using the proposed densely connected Learnable Connection Layer (LCL) across the entire architecture. The proposed LCL is composed of a simple convolutional kernels that enable a weighted combination of features at different hierarchies.
- The proposed *DCCT* directly restores the clear image without independently estimating the transmission map and atmospheric light and is guided by two strong generalizable constraints; namely, Supervised Contrastive constraint and Total Variational (TV) Loss. The contrastive constraint pushes the features of the output image to a more conducive space by considering its distance from multiple positive and negative examples. Further, TV Loss controls the level of noise in the output image by considering the sum of pixel values in a neighborhood.
- We do extensive analysis and ablation to empirically validate the choice of components of the proposed *DCCT* and test it on a variety of benchmarks to show that it is highly competitive with the current state of single image dehazing methods.

The remainder of this paper is organized as follows. In Section 2 we present an overview of existing literature related to the proposed method, in Section 3 we discuss the notation and the key components of *DCCT*. Following, in Section 4 we share ablation studies and comparison of the proposed *DCCT* with existing literature. Lastly, in Section 5 we present the conclusion of our study.

2. Related works

In this section, we discuss the literature related to the proposed DCCT based single image dehazing. Single Image Dehazing is a challenging problem that requires to restore a clean haze-free image from the given haze degraded input. Two broad approaches are used to tackle this problem; prior based and learning based (data driven).

2.1. Prior based image dehazing

These techniques [13,14,30–33] are centered around prior-information based on observations about characteristics of the haze degraded images to perform image dehazing. He et al. [13] observed that hazy images tend to have maximum information in at least one color channel with low intensity and proposed the Dark Channel Prior technique for image dehazing. [14] attempted to restore clear images through contrast enhancement by contracting Markov Random Fields. Fattal [30] formulated the optical transmission for hazy scenes and recovered the clear image considering the local statistical correlation of surface shading and transmission map. Zhu et al. [32] estimated the depth map of the image using a color attenuation prior and a supervised linear model. Berman et al. [31] highlighted the limitations of local patch-based priors. Further, [31] proposed a non-trainable estimation algorithm that used a haze-line prior for haze-free image estimation. Priors are designed with focus on the target image that makes it less robust to a complex set of scenes with different haze statistics. [13] under-performs for images with entities that resemble haze conditions. An example of this would be the sky region with clouds and white objects. Similarly, [14] is also unable to handle non-homogeneous haze conditions.

2.2. Learning based image dehazing

The recent growth of deep learning techniques and large scale training datasets encouraged its use in various domains like image dehazing [18–20,20–23,26,27,34–40], low-light image enhancement [41,42], object detection [4–6] and many others. Cai et al. [18] proposed a CNN-based technique that estimates a transmission map and uses the scattering model to recover the clear image. However, Cai et al. [18] assumes that atmospheric light remains locally constant which leads to halo effects. Ren et al. [19] proposed an improved multi-scale CNN for better estimation of the transmission map. Li et al. [20] proposed a lightweight CNN for image dehazing that uses a *K-estimation* module that effectively combines the atmospheric light and transmission map into one. However, the independent estimation of atmospheric light and transmission map aggregates the error in the disjoint estimation and also requires twice the supervision as shown by Liu et al. [22]. Consequently, [22] used end-to-end GridNet [43] as an attention based architecture to show its effectiveness in the dehazing task. Qin et al. [23] built upon GridDehazenet [22] using an independent pixel and channel attention mechanism with a deeper convolutional network for image dehazing to achieve superior performance. Deng et al. [21] draw inspiration from image restoration by modeling the single image dehazing task as a layer separation problem and fusing the output of the layer separation function with the atmospheric model using attention modules. Liu et al. [27] leverages real world hazy images to extenuate the prevailing issue of domain shift. Liu et al. [27] train a consistency (teacher) network to force learning more general representation. Yang et al. [24], Das and Dutta [44] extend the idea of a traditional U-Net [45] and aggregate multi-scale features for direct clear image estimation. Dong et al. [46] proposed a multi-scale U-Net that combines the boosting strategy and back-projection for single image dehazing. Dong et al. [46] also design a dense fusion module to prevent information loss. Different from Dong et al. [46], we augment our architecture with dense *learnable* skip connection. Researchers [26,34,36,38,47,48] attempted to perform image dehazing using an adversarial strategy to guide the training of neural networks. Adversarial training is useful for domain adaptation and related applications, but suffers from severe over-enhancement issues. Unlike the most representative existing deep learning based dehazing techniques, the proposed method is able to do end-to-end image dehazing without using the scattering model, and considers both local features using convolutions and global self-attention. Similar to [26,49], we use dense connections for the entire architecture, but use Learnable Connection Layers instead of skip connections. Lastly, most methods only consider spatial information preserving losses that are not sufficient to model the features of the input, we propose to use a supervised metric learning similar to Khosla et al. [50] that extends the existing perceptual loss to cater to multiple positive and negative pairs for each anchor.

2.3. Vision transformers

Recently, transformers have dominated both natural language processing [51–55] and computer vision [56]. The key idea behind using transformers is self-attention that captures global interactions and attends to relevant areas of the input sequence. Dosovitskiy et al. [56]

provided a method to process images as sequences. In [56], the RGB input image $I \in \mathbb{R}^{C\times H\times W}$ is first divided into N patches of constant size $\mathbb{R}^{C\times p\times p}$ with the height, width and channels denoted by H, W, C respectively. Then the resulting $N = (H\times W)/(p\times p)$ patches are linearly projected to a constant size D. To maintain positional information about each patch a positional embedding $E_i \in \mathbb{R}^D$ is added to each patch feature f_i, the same as [25]. The series of tokens $E_i + f_i$ are passed to the transformer blocks where $i \in [1, \dots N]$.

Other recent methods explore ways to do more challenging tasks like object detection and segmentation [57–59] with transformers. Conformer [60] is another variant of ViT that uses a convolutional module similar to [61] for Automatic Speech Recognition. Wu et al. [62] introduced a convolutional token embedding instead of a linear projection to make ViT more efficient. Since transformers form an attention matrix $(N \times N)$ for each input token with the rest of the sequence, they tend to scale quadratically. A number of derivative works [29,61,63] have provided more optimized techniques for better scaling. Performer Attention [29] approximates the softmax attention kernel using a linear space and time complexity alternative called Fast Attention via positive Orthogonal Random (FAVOR+) features approach. The key intuition behind FAVOR+ is the decomposition of the softmax operation which forces the transformer to produce a $D\times D$ matrix where $D \ll N$ and does not grow with input size.

Different from the existing methods, our convolutional-performer is combined with two additional convolutional modules for downsampling and upsampling; and is densely connected with the proposed learnable connection layers.

3. Method

In this section, we first discuss the motivation for our approach, followed basic notations and then present the proposed Densely Connected Convolutional Transformer (*DCCT*). The primary components of the proposed architecture are: (1) *Convolutional Downsampling Module* (CDSM:- Section 3.3.1), (2) *Convolutional Performer Encoder and Decoder* (Section 3.3.2), (3) *Learnable Connection Layer* (LCL:- Section 3.3.4), (4) *Convolutional Upsampling Module* (CUSM:- Section 3.3.3) and (5) *Combined Loss Function* (Section 3.3.5).

3.1. Motivation

The encoder–decoder networks in the traditional U-Net [45] architectures perform downsampling and upsampling respectively. Where the encoder captures the relevant distributions and the decoder is responsible for restoring the image. In our work, the *CDSM* allows the network to downsample the input to primarily reduce the number of tokens processed by the subsequent performer model. Additionally, it also learns low-level image features that are pertinent for image restoration task. Different from the traditional U-Net however, the *CDSM* is densely connected with the upsampling block using learnable skip connections (*LCL*).

The predominant CNN-based techniques (Section 2) do not learn global haze-relevant relationships. To that end, we propose the first convolutional performer encoder–decoder for image dehazing. A symmetrical encoder–decoder model is composed of sequential performer blocks and convolutions. Each performer block captures the global relationship between the input tokens and the convolution blocks explore any local biases in the feature space. This combination allows the encoder to efficiently construct rich haze-relevant embeddings and the decoder is tasked with reconstructing the haze-removed tokens. These reconstructed tokens are then augmented with important low level information (like edges) from the *CDSM* to finally restore the clear image. Further, to prevent any loss of information, we connect the entire architecture using the *LCL*.

The design of a U-Net like architecture (well established in image restoration) with the proposed convolution-performer blocks (effective representation learning), learnable connection layers (better feature fusion) and simple *CDSM, CUSM* (low-level information preservation) allow our network to effectively restore clear images.

3.2. Notation

Single Image Dehazing aims to recover the clear output J from a given input hazy image or the direct attenuation I using a Neural Network function f. Some methods employ a two stage strategy to estimate the transmission map and atmospheric light first, followed by clear image estimation using the atmospheric scattering equation. In this paper, we approach the problem in a single step as follows:

$$\hat{J} = f(I, \theta) \tag{1}$$

where neural network f is parameterized by θ and the estimated clear image is $\hat{J}$. While training, an input of size $B\times C\times H\times W$ is fed to the proposed network and an output of size $B\times 3\times H\times W$ is restored, where B denotes the size of the input batch, $C\times H\times W$ are the channels, height and width of the input images respectively. We use 3 channel RGB images of dimension 512×512 as inputs while training.

3.3. Densely connected convolutional transformer

We summarize *DCCT* and show the outline of the proposed architecture in Fig. 1.

3.3.1. Convolutional downsampling module

We extract low-level features from the input I using a Convolutional Downsampling Module (*CDSM*). The *CDSM* transforms the input I of shape $B\times C\times H\times W$ to $I_{1/4}$ of shape $B\times(4\times 32)\times H/4\times W/4$ using a small CNN. The architectural details are illustrated in Fig. 2.

3.3.2. Convolutional performer encoder-decoder

Projection: The downsampled feature map $(I_{1/4})$ is reshaped to a vector (I_R) of size $B\times(HW/16p^2)$. Further, I_R is linearly projected into $N = (HW)/16p^2$ tokens, each of size D (transformer dimension) as done in [56]. Let this linearly projected vector be $Y_{1/4}$.

Encoder: As shown in Fig. 1, the projected patches are fed to the encoder of the *DCCT* composed of three successive transformer and convolution layers. Each convolution progressively downsamples its input. Like the U-Net [45] architecture, encoder learns a haze-relevant features by compressing the input to a bottleneck representation. Mathematically, the ith layer of the encoder can be shown as follows:

$$Y_{1/(4\times 2^i)} = PL_i(Y_{1/(4\times 2^i)}) \in \mathbb{R}^{B\times N\times D} \tag{2}$$

$$Y_{1/(4\times 2^{i+1})} = R^T(Glu(Conv(R(Y_{1/(4\times 2^i)})))) \in \mathbb{R}^{B\times N\times D/2} \tag{3}$$

$$Y_{1/(4\times 2^{i+1})} = LCL_1(Y_{1/(4\times 2^{i+1})}, Y_{1/(4\times 2^j)})\forall j \in [0, i) \tag{4}$$

Here, each transformer layer denoted by PL is a Performer block performing MSA, LN followed by FFN as in Vaswani et al. [25]. Note that before each convolution, we reshape the input to the spatial dimension. The reshaped tensor is passed through a single convolution layer $Conv$ and conditioned to Glu [64] non-linearity. Each convolution has a stride of 2 and a kernel size of $5, 3, 1$ for the 0th, 1st, 2nd encoder layer. The convolved tensor is finally reshaped back (R^T) to a 3D tensor such that it is compatible with the transformer. The Learnable Connection Layer *LCL* (Section 3.3.4) combines the features $Y_{1/(4\times 2^j)}$ from the all the previous encoder layers. Here j represents previous layer number. Note that we interpolate the previous features to the size of the current features before combining the same.

Decoder: The final bottleneck representation $Y'_{1/32}$ is fed to the symmetrical decoder layers with transposed convolution for upsampling. We represent the ith decoder layer as follows:

$$Y'_{2^i/32} = PL_i(Y'_{2^i/32}) \in \mathbb{R}^{B\times N\times D/8} \tag{5}$$

$$Y'_{2^{i+1}/32} = R^T(Glu(ConvT(R(Y_{2^i/(4)})))) \in \mathbb{R}^{B\times N\times D/4} \tag{6}$$

$$Y'_{2^{i+1}/32} = LCL_1(Y'_{2^{i+1}/32}, Y'_{2^j/32})\forall j \in [0, i) \tag{7}$$

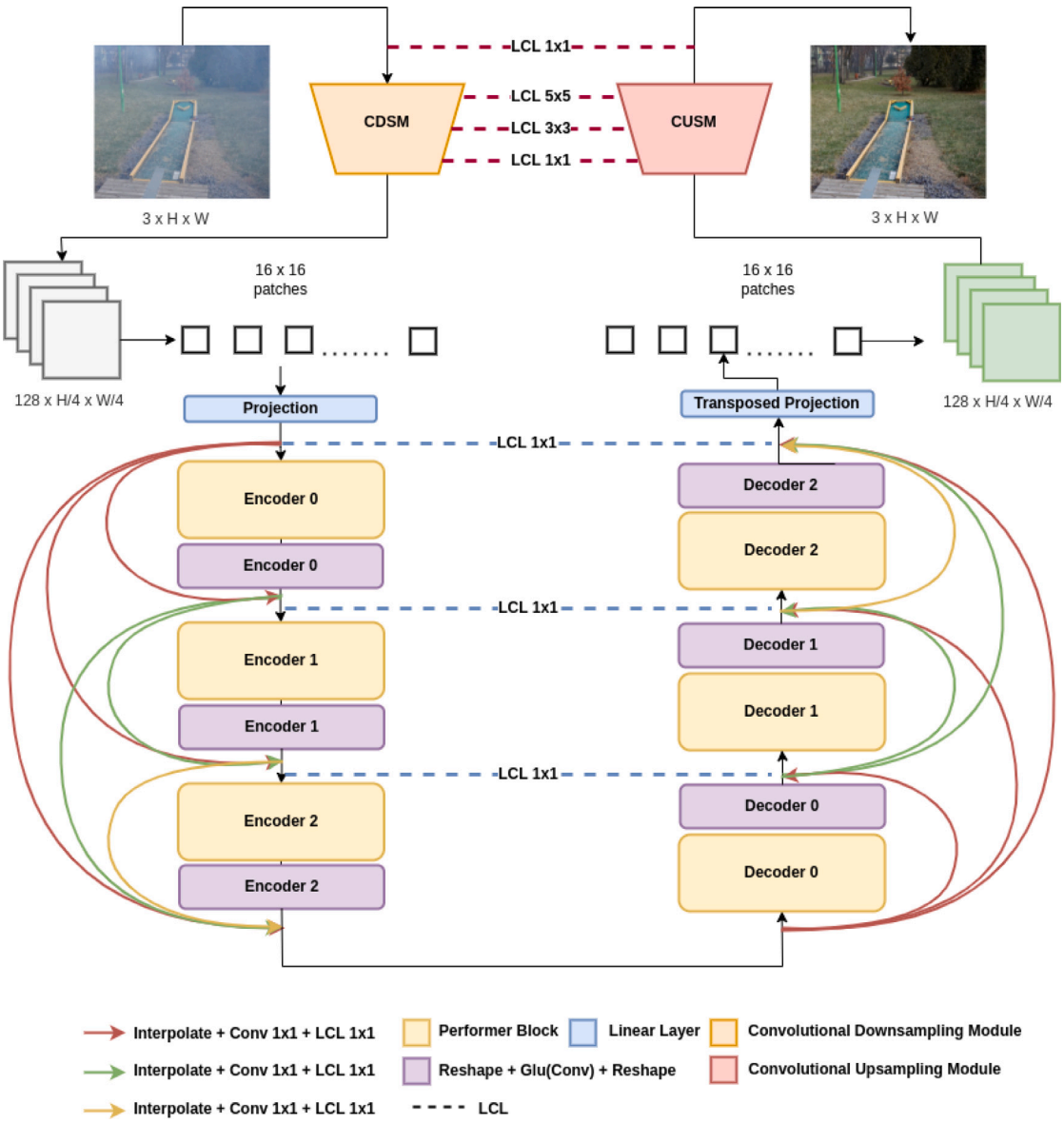


Fig. 1. An overview of the proposed *DCCT*: the architecture is densely connected using *LCL* and contains three primary components: (1) *First,* we extract low level features from the input hazy image using the proposed *CDSM*. (2) *Second,* we linearly project the feature maps and feed it along with positional embeddings [56] through three encoder–decoder layers of the Conv-Performer (3) *Third,* we reproject the output and use the proposed *CUSM* to restore the clear output.

$$Y'_{2^{i+1}/32} = LCL_1(Y'_{2^{i+1}/32}, Y'_{1/(4\times 2^{k+1})}) \tag{8}$$

The performer layers and transposed convolutional layer in the decoder are similar to the encoder. The transposed convolution layer $ConvT$ upsamples the input feature map by a factor of two, and is symmetric to the encoder. The output of these two layers is combined using *LCL* with: (a) All previous feature maps produced by the decoder, and (b) Symmetric encoder feature maps as in U-Net [45]. Note that $Y'_{2^{i+1}/32}, Y'_{1/(4\times 2^{k+1})}$ are of the same shape with i denoting the ith decoder layer and k denoting the corresponding encoder feature maps. Concretely, the output of the first encoder layer is combined with the output of last decoder layer and the output of the second encoder layer is combined with the output of the second decoder layer.

Transposed Projection: The decoder restores $Y'_{1/4}$ that is of the same shape as the input $Y_{1/4}$. The output of the final decoder layer $Y'_{1/4}$ is projected back from $\mathbb{R}^{B\times N\times D}$ to $\mathbb{R}^{B\times(HW/16p^2)\times(4\times 32\times p^2)}$. The projected vector is then reshaped back to the tensor $I'_{1/4} \in \mathbb{R}^{B\times(4\times 32)\times H/4\times W/4}$ which is the of the same shape as $I_{1/4}$.

3.3.3. Convolutional upsampling module

The reshaped output of the final decoder layer $I'_{1/4}$ is progressively upsampled using transposed convolutions, symmetric to the initial downsampling stage. The output of the Convolutional Upsampling Module (*CUSM*) is the estimated clear image $\hat{J}$. The architecture of the proposed *CUSM* is presented in Fig. 3.

Concretely, feature map produced by the $ConvT$ layer of the *CUSM* is fused (using *LCL*) with the corresponding feature map of same size produced by the $Conv$ layer in the *CDSM*. Finally, the restored mask I' is fused (using *LCL*) with the input hazy image I to produce the estimated clear image $\hat{J}$.

3.3.4. Learnable connection layer

Skip connections are an essential part of Neural Network architectures since they provide a simple way to build deeper networks [65]. The key idea behind using skip connections is to facilitate low level feature preservation which might get diminished in a deep network. The use of skip connections in segmentation networks like U-Nets [45]

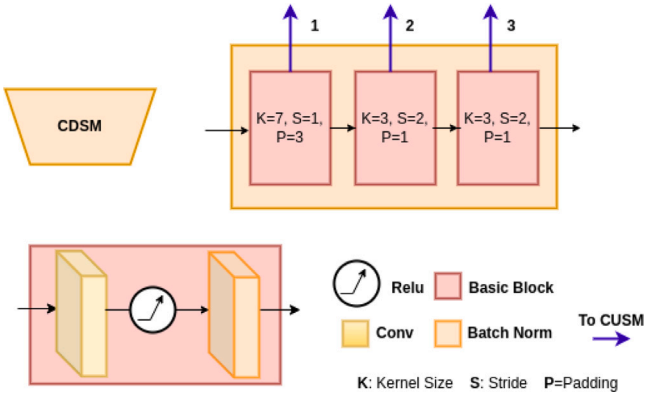


Fig. 2. An overview of the Convolutional Downsampling Module (CDSM): The CDSM contains three successive convolution layers with ReLU activation and BatchNorm. The outputs produced by each layer is connected with the CUSM as shown in Fig. 3.

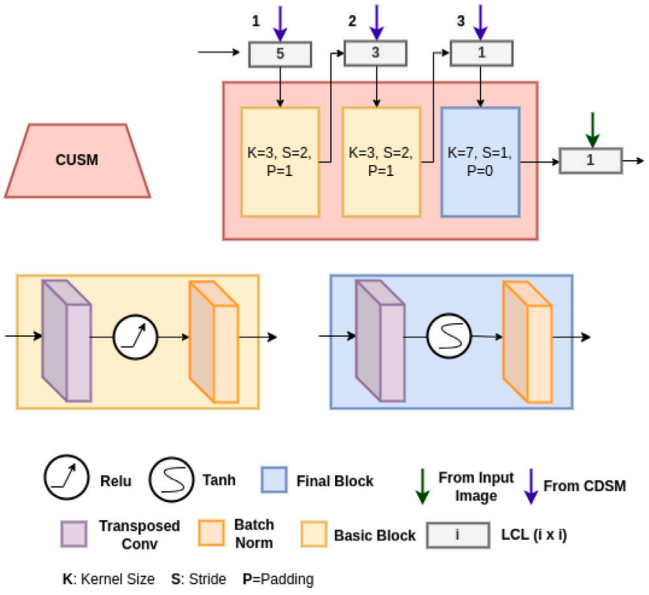


Fig. 3. An overview of the Convolutional Upsampling Module: The CUSM combines inputs from the CDSM and the outputs produced by basic blocks composed of transposed convolution operation. The activation at the final layer is tanh. The final mask is then combined using LCL with the input hazy image to produce the estimated clear image.

have shown that they are critical to preventing vanishing gradients during training. Similarly, skip connections have been extensively used in Image Dehazing as well [22–24,26,44] due to their obvious advantages in image restoration. [66] showed that replacing naive skip connections with discriminate learnable alternatives can boost the performance of segmentation networks. Gathering inspiration from [66], we propose a convolutional weighted fusion of features at different levels. This is done so that the network discriminates between higher and lower level features on its own instead of strict additive or multiplicative skip connections. Let G, G' be lower and higher level features in the encoder–decoder architecture respectively. Then *LCL* operation can be shown as:

$$G_{combined} = Tanh(Conv(G)) + Tanh(Conv(G')) \tag{9}$$

where $Tanh$ is the hyperbolic tangent activation function and $Conv$ is the convolution operation. For all intermediate layers we learn a 1×1 convolution, whereas for connecting *CDSM* and *CUSM*, we choose a larger learnable kernel for a more adaptive combination of features. LCL_i denotes the *LCL* operation with convolutional kernel size i. In Table 1 (row 3–4), we show that using *LCL* over simple dense skip connections increases the *PSNR* from 18.73 to 19.47.

Dense Connections: Huang et al. [49] presented a CNN architecture that uses dense connections between the layers of the CNN through Dense Blocks and Transition Blocks, achieving superior performance over prior art with less parameter count and simpler optimization. [49] argue that the addition of a large number of skip connections through the architecture achieves a higher supervision from the loss. Similarly, Zhang and Patel [26] show that fusing information from different levels of the encoder and decoder can maximize information flow for better transmission map estimation. We also maximize information flow between the layers of the proposed *DCCT* by densely connecting the encoder and decoder. Different from [26,49] that concatenate the feature maps using a Dense Block, we use the *LCL* that allows us to use significantly fewer parameters, yet achieve the desired performance gain. We fuse feature maps of different spatial dimensions by interpolating to the desired size before feeding the transformed feature maps to the *LCL* (Eq. (9)).

3.3.5. Combined loss

We condition the training of f using two primary objective functions. The first objective function, $L_{spatial}$ backpropagates gradients that capture the spatial information loss between the estimated output $\hat{J}$ and the clear ground truth J. We empirically verify that training the proposed *DCCT* yields the best validation results with Smooth L1 loss over $L2$ or $L1$ loss in Table 1.

$$L_{spatial} = SL1(\hat{J}, J) \tag{10}$$

where $SL1$ is the Smooth L1 Loss [67]. The second loss term constraints the solution space by acting as a regularizer. Let $I_{k,1}, I_{k,2} \ldots, I_{k,m}$ be a set of hazy images and $J_{k,1}, J_{k,2} \ldots, J_{k,m}$ be the corresponding clear images in the kth training batch. Then the multi image perceptual loss with anchor $J'_{k,1}$ is given by

$$L_{mip} = \frac{1}{N} \sum_{i=1}^{N} \max(\mathcal{D}^2(\gamma(J'_{k,1}), \gamma(J_{k,N})) - \mathcal{D}^2(\gamma(J'_{k,1}), \gamma(I_{k,N})), 0) \tag{11}$$

where $J_{k,N}, I_{k,N}$ are positive and negative samples respectively, and $\mathcal{D}$ is the distance function in the embedding space. We do not explicitly mine positive and negative samples in our work. γ is the VGG19 [68] feature extractor, and is used to produce features of the given inputs at different levels.

We visualize grainy noise on the output conditioned to these two losses alone and mitigate the noise issue using the total variational [69] loss denoted by L_{tv}

We combine the above losses into a joint loss with a tradeoff parameter α as follows:

$$L = \alpha(L_{spatial}) + (1 - \alpha)(\frac{L_{mip}}{2} + \frac{L_{tv}}{2}) \tag{12}$$

The hyperparameter $\alpha = 0.8$ for all experiments based on empirical evaluation.

4. Experiments

In this section, we present the evaluation of performance of the proposed *DCCT* on the Single Image Dehazing task. We provide detailed ablation studies that establish the effectiveness of each component of the proposed model and compare our method with the predominant works in literature.

4.1. Implementation details

DCCT is trained in an end-to-end fashion using the proposed joint objective function. We empirically verify the value of the tradeoff parameter α in the combined loss to be 0.8. For training, we use the Adam solver [70] with a learning rate of $1e-4$ and default values of β_1 and β_2. Cosine Scheduler with warmup is used to condition the learning rate during training. We use early stopping with a patience threshold of 5 to monitor over-fitting and train our pipeline for minimum 15 epochs with a batch size of 12 for all experiments and ablations. We use Gradient Clipping with a threshold of 1.0 to prevent exploding gradients. Further, we apply Stochastic Weight Averaging while training [71]. We train the transformer with the following hyperparameters: (1) number of heads = 4, (2) input dimension $D = 128$, (3) patch size $p = 16$ and (4) head dimension = 32. All experiments were implemented using PyTorch 1.1 [72] and conducted on two Nvidia Titan RTX GPUs and Intel Xeon Silver 4114 CPU. The proposed *DCCT* has 12.6 million parameters and takes approximately 0.081 s to process a 512 × 512 image on an Nvidia Titan RTX GPU.

4.2. Dataset

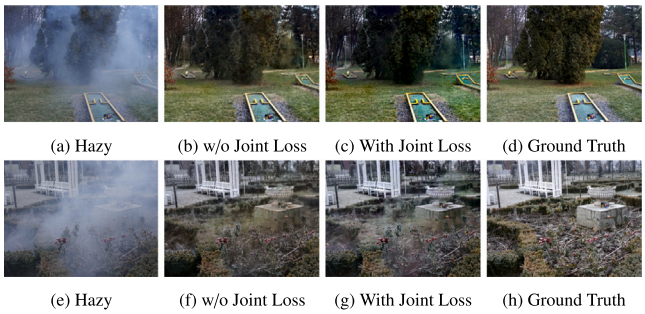
We evaluate the proposed *DCCT* on both synthetic and real world datasets. We leverage RESIDE (REalistic Single Image DEhazing) [17] which is a popular training and testing benchmark composed of both real world and synthetic subsets. For training and validation we use 512 × 512 randomly cropped 3 channel images from the RESIDE synthetic subset OTS (Outdoor Training Set) composed of 313 950 hazy images generated from 8500 ground truth images. After training the proposed network, we evaluate its performance on RESIDE SOTS outdoor (500 test images). We demonstrate the effectiveness of the proposed method on real world datasets like, NH-Haze [73], and Dense-Haze [74]. We train the model on NH-Haze and Dense-Haze for a fair comparison with other methods. NH-haze is a dataset with non-homogeneous hazy images with 45 training, 5 validation and 5 test images. Dense Haze is a benchmark consisting of 45 training, 5 validation and 5 test samples with homogeneous dense haze. For subjective evaluation, we leverage the RESIDE HSTS dataset and O-Haze [75] as test datasets respectively.

4.2.1. Data augmentation

1. *RESIDE outdoor:* We apply data augmentations on this dataset only during training [76]. All hazy training samples and their corresponding ground truth images are randomly cropped to a constant size of 512 × 512 and flipped vertically and horizontally with a probability of 0.5. We rotate the inputs with a probability of 0.5 through randomly sampled angles between −90 deg, +90 deg.
2. *NH-Haze [73] and Dense Haze [74]:* Since the dataset size is only 45 images, we apply static augmentations to increase the size of the dataset first. The training images are augmented by randomly resizing, cropping, flipping and rotating the inputs to create a larger training dataset of 100 000 images. While training, we cut random patches of 512 × 512 from the augmented input samples. Empirical evidence suggested that using a larger dataset (statically augmented) performs significantly better than using the same augmentations dynamically.

4.3. Evaluation metrics

We use Peak Signal to Noise Ratio (*PSNR*) and Structural Similarity Index (*SSIM*) [77] as the full reference Image Quality Analysis metrics that have been used in the existing dehazing literature task. *PSNR* provides a measure of the ratio of the maximum possible signal power and the power of noise in decibels. *SSIM* measures the structural similarity between the query image and corresponding ground truth. Higher values of *SSIM* and *PSNR* indicate better quality output.



(a) Hazy (b) w/o Joint Loss (c) With Joint Loss (d) Ground Truth

(e) Hazy (f) w/o Joint Loss (g) With Joint Loss (h) Ground Truth

Fig. 4. DCCT with and without joint loss. The color consistency and perceptual quality of outputs produced by *DCCT* trained with joint loss is significantly better (zoom in for better view).

Table 1
Ablation studies.

Model	SSIM	PSNR
Vanilla (base) Transformer	0.6929	18.55
Performer (base)	0.6823	18.33
Performer + Dense	0.7120	18.73
Performer + Dense + LCL	0.7121	19.47
DCCT with L1 Loss	0.7029	18.95
DCCT with L2 Loss	0.6995	18.02
Performer + Dense + LCL + MIP Loss + TV Loss + SL1 Loss (**DCCT**)	**0.7250**	**20.18**

4.4. Ablation studies

In this section, we present ablation studies and related discussions that validate our design choices. A summary of ablations conducted are given in Table 1.

4.4.1. Effect of performer blocks

The primary reason for using Performer [29] over basic transformer blocks is training time and memory consumption. We observe that a standard transformer takes 2 GPU days to train for 30 000 steps compared to only 1.5 GPU days with less memory consumption (↓ 4000 MiB) while using performer. The vanilla transformer model allows for slight performance gains (0.01 higher *SSIM* and 0.2 higher *PSNR*). Despite the marginal gains, we select the performer-based architecture due to clear train-time benefits.

4.4.2. Effect of dense connections

We get significant gains upon combining features from different levels within the encoder and decoder via simple interpolation, followed by addition. We validate our hypothesis that presence of existing skip-connections in the performer blocks are insufficient by showing clear empirical superiority (*SSIM* ↑ 0.0398 and *PSNR* ↑ 0.4), by densely connecting the entire architecture. Further, we see a quick saturation in performance after 17 epochs in vanilla architectures, however the addition of Dense Connections allows the network to train for longer without overfitting.

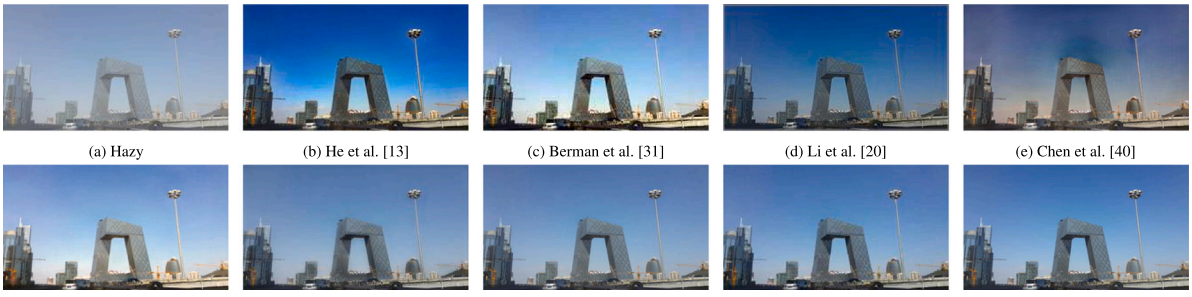
4.4.3. Effect of LCL

Single Image Dehazing is an ill posed problem wherein several possible outputs can be considered approximations of the clear image. This motivates the need to search a larger solution space in order to find best approximations of the clear image. The *LCL* allows the same by incorporating a learnable connection that weighs the lower and higher level features using a simple convolution operation before fusion. We show that using the *LCL* over additive skip connections can provide a quantitative boost (*PSNR* ↑ 0.74).

(a) Hazy (b) He et al. [13] (c) Berman et al. [31] (d) Li et al. [20] (e) Chen et al. [40]

(f) Shao et al. [36] (g) Yang et al. [24] (h) Qin et al. [23] (i) Ours (j) Ground Truth

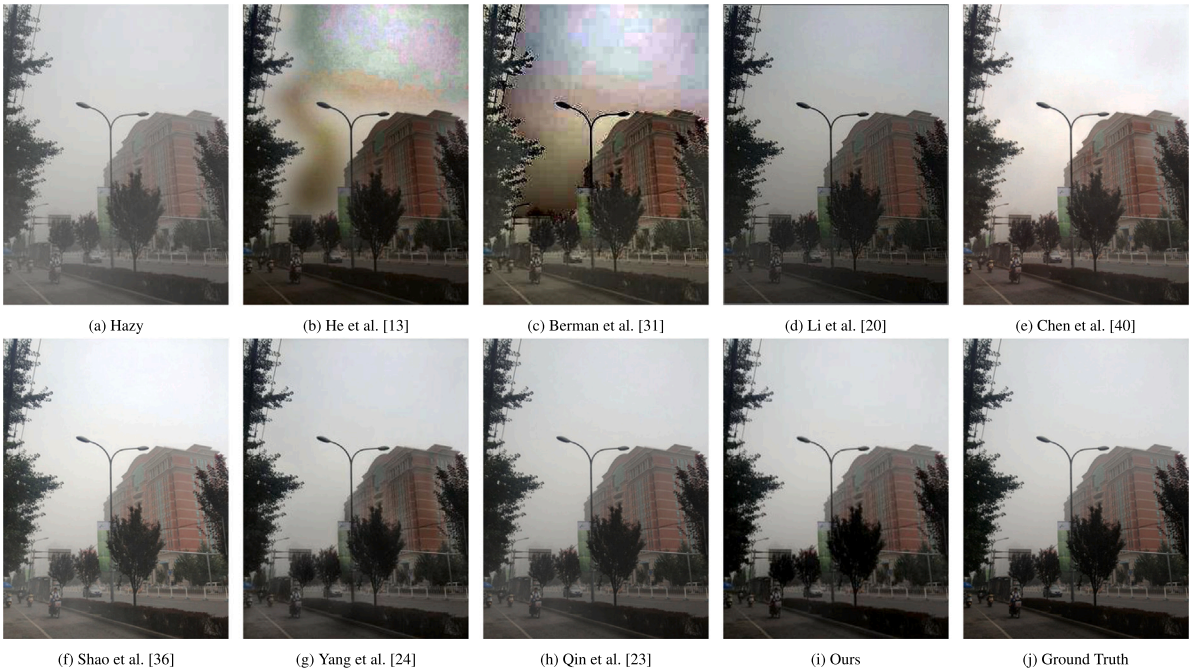
Fig. 5. Visual comparison of image 1 from RESIDE SOTS dataset.



(a) Hazy (b) He et al. [13] (c) Berman et al. [31] (d) Li et al. [20] (e) Chen et al. [40]

(f) Shao et al. [36] (g) Yang et al. [24] (h) Qin et al. [23] (i) Ours (j) Ground Truth

Fig. 6. Visual comparison of image 2 from RESIDE SOTS dataset.



(a) Hazy (b) He et al. [13] (c) Berman et al. [31] (d) Li et al. [20] (e) Chen et al. [40]

(f) Shao et al. [36] (g) Yang et al. [24] (h) Qin et al. [23] (i) Ours (j) Ground Truth

Fig. 7. Visual comparison of image 3 from RESIDE SOTS dataset.

4.4.4. *Effect of joint loss*

The joint loss not only boosts performance of our network, but largely improves the subjective quality of the outputs. Fig. 4 shows the outputs with and without the joint loss on NH-Haze dataset. Clearly, the joint loss ensures color consistency and improves perceptual quality of the outputs. For instance, the color of the round artifact in the garden

(a) Hazy (b) He et al. [13] (c) Berman et al. [31] (d) Li et al. [20] (e) Chen et al. [40]

(f) Shao et al. [36] (g) Yang et al. [24] (h) Qin et al. [23] (i) Ours (j) Ground Truth

Fig. 8. Visual comparison on validation image from O-Haze Dataset [75].

Table 2
Quantitative Comparison between the proposed *DCCT* and existing Single Image Dehazing algorithms. The best and second to best method is represented with **bold** and underline respectively.

Method	SOTS (outdoor)		NH-Haze		Dense Haze	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
He et al. [13]	0.8148	19.13	0.5231	10.57	0.2460	12.12
Berman et al. [31]	0.7500	17.27	–	–	–	–
Ren et al. [19]	0.8100	17.57	–	–	–	–
Cai et al. [18]	0.8514	22.46	0.5472	17.01	0.4225	13.84
Li et al. [20]	0.8765	20.29	0.5670	15.41	0.4140	13.14
Liu et al. [22]	0.9819	30.86	0.5370	13.80	0.3681	13.31
Chen et al. [40]	0.9450	28.13	0.5839	14.73	0.3615	10.71
Yang et al. [24]	0.9470	26.61	–	–	–	–
Qin et al. [23]	0.9840	33.52	0.6915	19.87	**0.4524**	14.39
Shao et al. [36]	0.9300	27.76	–	–	–	–
Deng et al. [21]	0.9844	**34.29**	–	–	–	–
Dong et al. [46]	0.9840	33.97	–	–	–	–
Liu et al. [27]	0.9700	29.42	–	–	–	–
Ours (DCCT)	**0.9863**	34.22	**0.7250**	**20.18**	0.3895	**14.94**

in Fig. 4(g) is closer to the ground truth compared to Fig. 4(f). Further the checkerboard is much clearer in outputs produced with joint loss. Additionally, we also validate the effectiveness of Smooth L1 Loss over: (a) L1 loss (*SSIM* ↓ 0.0221, *PSNR* ↓ 1.23) and (b) L2 loss (*SSIM* ↓ 0.0255, *PSNR* ↓ 2.16). Note that the comparison of different spatial information preservation losses is done keeping the rest of the architecture same as the proposed *DCCT*.

4.4.5. Effect of CDSM and CUSM

The *CDSM* and *CUSM* allow feasible training of a transformer with input size of 512 × 512 and a patch size of 16. A smaller patch size is essential for more fine grained reconstruction. Using a patch size higher than 16 results in unstable optimization and poor qualitative outputs. Specifically, using a patch size of 32 leads to a significant drop in *SSIM* (↓ 0.046) and *PSNR* (↓ 1.01) on the RESIDE dataset.

4.5. Comparison

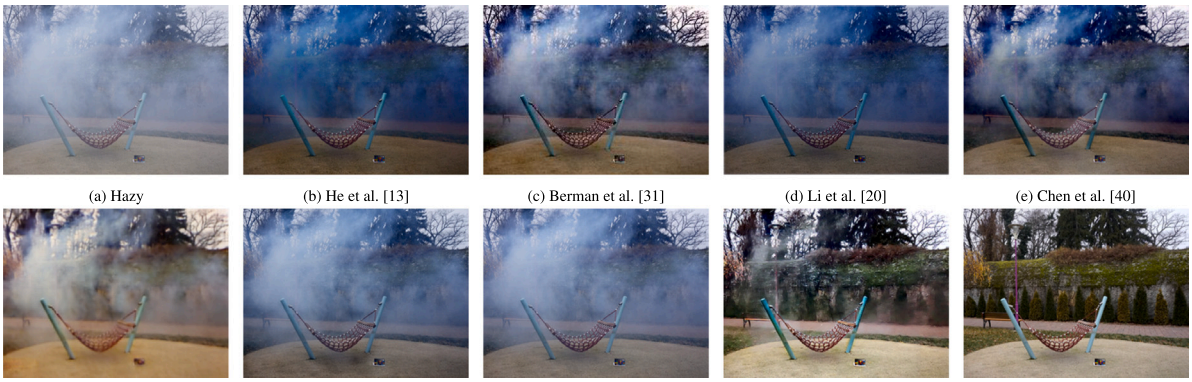
We compare the proposed *DCCT* with techniques proposed in seminal dehazing literature He et al. [13], Berman et al. [31], initial deep learning based techniques Ren et al. [19], Cai et al. [18] and the recent state of the art algorithms Li et al. [20], Liu et al. [22], Chen et al. [40], Qin et al. [23], Yang et al. [24], Shao et al. [36]. The quantitative comparison is done on RESIDE SOTS [17], NH-Haze [73], and Dense-Haze [74] datasets.

4.5.1. Quantitative results

The quantitative comparison has been shown in Table 2. The proposed *DCCT* achieves outperforms all existing methods in terms of *PSNR* and most methods in terms of *SSIM* for all three datasets. We beat all existing techniques on RESIDE SOTS synthetic benchmark by a margin at least of 0.70 PSNR (other than [21]) and 0.0023 *SSIM*. On recent NTIRE dehazing challenge datasets [73], we achieve significantly higher metrics compared to the next best method [23]. This establishes that our method can generalize to non-homogeneous and dense haze conditions in addition to the synthetic haze conditions.

4.5.2. Qualitative results

We present the visual outputs on Ancuti et al. [73], Li et al. [17], Ancuti et al. [75] datasets to demonstrate the effectiveness of the proposed method on both synthetic and real dehazing datasets in Figs. 5–11. We observe that prior based methods He et al. [13], Berman et al. [31] suffer from common problems like over-enhancement and failure to preserve original colors. For example, over-enhanced sky region is evident in Figs. 5(b) and 6(b). Li et al. [20] are either unable to remove haze effectively in dense haze conditions or produces low-brightness outputs compared to the original. In Fig. 6, none of the existing methods is able to remove haze from the "two-wheeler" (bottom-left of the image) other than Chen et al. [40] and the proposed method. However, the output produced by Chen et al. [40] in Fig. 6(e) has patches of haze in the image and artifacts around the street light (patches on the building and sky-region artifacts are visible upon zooming in), whereas the proposed method yields a more realistic image. For real world dataset examples presented in Figs. 8–9, the proposed *DCCT* is able to remove larger amount of haze compared to the predominant techniques in a diverse set of environments. In Figs. 10 and 11, we show a visual comparison on the RESIDE HSTS dataset, wherein, the proposed *DCCT* is able to produce outputs that are most true to the ground truth. Most existing methods Yang et al. [24], Chen et al. [40], Li et al. [20] are unable to remove haze from objects located farther away from the camera as shown in Fig. 11.



(a) Hazy (b) He et al. [13] (c) Berman et al. [31] (d) Li et al. [20] (e) Chen et al. [40]

(f) Shao et al. [36] (g) Yang et al. [24] (h) Qin et al. [23] (i) Ours (j) Ground Truth

Fig. 9. Visual comparison on validation image from NH-Haze Dataset.

(a) Hazy (b) He et al. [13] (c) Berman et al. [31] (d) Li et al. [20] (e) Chen et al. [40]

(f) Shao et al. [36] (g) Yang et al. [24] (h) Qin et al. [23] (i) Ours (j) Ground Truth

Fig. 10. Visual comparison on image 1 RESIDE HSTS Dataset.

(a) Hazy (b) He et al. [13] (c) Berman et al. [31] (d) Li et al. [20] (e) Chen et al. [40]

(f) Shao et al. [36] (g) Yang et al. [24] (h) Qin et al. [23] (i) Ours (j) Ground Truth

Fig. 11. Visual comparison on image 2 from RESIDE HSTS Dataset.

5. Conclusion

In this paper, we introduce a novel dense convolutional transformer network for Single Image Dehazing, viz Densely Connected Convolutional Transformer (*DCCT*). The three primary components of the proposed *DCCT* include a Convolutional Downsampling Module (*CDSM*), Convolutional-Performer Encoders and Decoders, and a Convolutional Upsampling Module (*CUSM*). Initially, the *CDSM* captures low level features and downsamples the input, which is converted to a sequence of small patches. The patches are then processed through the Convolutional Performer blocks that efficiently apply global self-attention to capture haze relevant features. Finally, we reconstruct the clear image using the *CUSM*. We also show that using the proposed dense learnable skip connections (*LCL*) across the architecture signif-

icantly boosts performance. We train *DCCT* in an end-to-end fashion with a joint objective function that allows for better perceptual quality through contrastive regularization. Through extensive experiments and ablations we establish the design choices of our architecture and demonstrate its effectiveness over the current state of the art. We also show that *DCCT* is able to remove haze from a diverse set of conditions and produces more realistic outputs, encouraging its use in other areas of low-level image processing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] C. Wang, Y. Qian, W. Gong, J. Cheng, Y. Wang, Y. Wang, Cross-layer progressive attention bilinear fusion method for fine-grained visual classification, J. Vis. Commun. Image Represent. 82 (2022) 103414.

[2] Z. Yang, Z. Wang, L. Luo, H. Gan, T. Zhang, SWS-DAN: Subtler WS-DAN for fine-grained image classification, J. Vis. Commun. Image Represent. 79 (2021) 103245.

[3] A.A. Baffour, Z. Qin, Y. Wang, Z. Qin, K.-K.R. Choo, Spatial self-attention network with self-attention distillation for fine-grained image recognition, J. Vis. Commun. Image Represent. 81 (2021) 103368.

[4] Q. Wang, L. Zhou, Y. Yao, Y. Wang, J. Li, W. Yang, An interconnected feature pyramid networks for object detection, J. Vis. Commun. Image Represent. 79 (2021) 103260.

[5] Q. Zhang, L. Zhang, D. Wang, Y. Shi, J. Lin, Global and local information aggregation network for edge-aware salient object detection, J. Vis. Commun. Image Represent. 81 (2021) 103350.

[6] S. Xie, C. Liu, J. Gao, X. Li, J. Luo, B. Fan, J. Chen, H. Pu, Y. Peng, Diverse receptive field network with context aggregation for fast object detection, J. Vis. Commun. Image Represent. 70 (2020) 102770.

[7] Y. Zhan, W.-L. Zhao, Instance search via instance level segmentation and feature representation, J. Vis. Commun. Image Represent. 79 (2021) 103253.

[8] J. Zhang, Z. Li, C. Zhang, H. Ma, Stable self-attention adversarial learning for semi-supervised semantic image segmentation, J. Vis. Commun. Image Represent. 78 (2021) 103170.

[9] Y. Wang, J. Choi, Y. Chen, S. Li, Q. Huang, K. Zhang, M.-S. Lee, C.-C.J. Kuo, Unsupervised video object segmentation with distractor-aware online adaptation, J. Vis. Commun. Image Represent. 74 (2021) 102953.

[10] H. Zhou, S. Zhou, Scene categorization towards urban tunnel traffic by image quality assessment, J. Vis. Commun. Image Represent. 65 (2019) 102655.

[11] D. Xiao, L. Zhuo, J. Li, J. Li, Structure-prior deep neural network for lane detection, J. Vis. Commun. Image Represent. 81 (2021) 103373.

[12] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A review of video surveillance systems, J. Vis. Commun. Image Represent. 77 (2021) 103116.

[13] K. He, J. Sun, X. Tang, Single image haze removal using dark channel prior, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 2341–2353.

[14] R.T. Tan, Visibility in bad weather from a single image, 2008, http://dx.doi.org/10.1109/CVPR.2008.4587643.

[15] S. Nayar, S. Narasimhan, Vision in bad weather, 2 (1999) 820–827. vol.2. http://dx.doi.org/10.1109/ICCV.1999.790306.

[16] E. McCartney, Optics of the Atmosphere: Scattering by Molecules and Particles, Wiley, New York, 1976.

[17] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, Z. Wang, Benchmarking single-image dehazing and beyond, IEEE Trans. Image Process. 28 (1) (2018) 492–505.

[18] B. Cai, X. Xu, K. Jia, C. Qing, D. Tao, Dehazenet: An end-to-end system for single image haze removal, IEEE Trans. Image Process. 25 (11) (2016) 5187–5198.

[19] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, M. Yang, Single image dehazing via multi-scale convolutional neural networks, in: B. Leibe, N. Sebe, M. Welling, J. Matas (Eds.), Computer Vision - 14th European Conference, ECCV 2016, Proceedings, Springer Verlag, Germany, 2016, pp. 154–169, http://dx.doi.org/10.1007/978-3-319-46475-6_10.

[20] B. Li, X. Peng, Z. Wang, J. Xu, D. Feng, AOD-net: All-in-one dehazing network, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4780–4788, http://dx.doi.org/10.1109/ICCV.2017.511.

[21] Z. Deng, L. Zhu, X. Hu, C.-W. Fu, X. Xu, Q. Zhang, J. Qin, P.-A. Heng, Deep multi-model fusion for single-image dehazing, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2453–2462.

[22] X. Liu, Y. Ma, Z. Shi, J. Chen, Griddehazenet: Attention-based multi-scale network for image dehazing, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7314–7323.

[23] X. Qin, Z. Wang, Y. Bai, X. Xie, H. Jia, FFA-net: Feature fusion attention network for single image dehazing, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 11908–11915.

[24] H. Yang, C.H. Yang, Y. James Tsai, Y-net: Multi-scale feature aggregation network with wavelet structure similarity loss function for single image dehazing, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 2628–2632.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[26] H. Zhang, V.M. Patel, Densely connected pyramid dehazing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3194–3203.

[27] Y. Liu, L. Zhu, S. Pei, H. Fu, J. Qin, Q. Zhang, L. Wan, W. Feng, From synthetic to real: Image dehazing collaborating with unlabeled real data, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 50–58.

[28] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701–1708.

[29] K.M. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J.Q. Davis, A. Mohiuddin, L. Kaiser, D.B. Belanger, L.J. Colwell, A. Weller, Rethinking attention with performers, in: International Conference on Learning Representations, 2021, URL: https://openreview.net/forum?id=Ua6zuk0WRH.

[30] R. Fattal, Single image dehazing, ACM Trans. Graph. 27 (3) (2008) 1–9.

[31] D. Berman, T. Treibitz, S. Avidan, Non-local image dehazing, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1674–1682, http://dx.doi.org/10.1109/CVPR.2016.185.

[32] Q. Zhu, J. Mai, L. Shao, A fast single image haze removal algorithm using color attenuation prior, IEEE Trans. Image Process. 24 (11) (2015) 3522–3533.

[33] F. Yuan, Y. Zhou, X. Xia, X. Qian, J. Huang, A confidence prior for image dehazing, Pattern Recognit. 119 (2021) 108076.

[34] Y. Liu, H. Al-Shehari, H. Zhang, Attention mechanism enhancement algorithm based on cycle consistent generative adversarial networks for single image dehazing, J. Vis. Commun. Image Represent. (2022) 103434.

[35] K. Singh, A.S. Parihar, Variational optimization based single image dehazing, J. Vis. Commun. Image Represent. 79 (2021) 103241.

[36] Y. Shao, L. Li, W. Ren, C. Gao, N. Sang, Domain adaptation for image dehazing, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2805–2814, http://dx.doi.org/10.1109/CVPR42600.2020.00288.

[37] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, M. Yang, Gated fusion network for single image dehazing, 2018, CoRR abs/1804.00213.

[38] D. Engin, A. Genç, H.K. Ekenel, Cycle-dehaze: Enhanced CycleGAN for single image dehazing, 2018, CoRR abs/1805.05308.

[39] A. Mehra, P. Narang, M. Mandal, TheiaNet: Towards fast and inexpensive CNN design choices for image dehazing, J. Vis. Commun. Image Represent. 77 (2021) 103137.

[40] D. Chen, M. He, Q. Fan, J. Liao, L. Zhang, D. Hou, L. Yuan, G. Hua, Gated context aggregation network for image dehazing and deraining, 2018, CoRR abs/1811.08747.

[41] H. Rohilla, G. Asnani, K. Singh, A.S. Parihar, Low-light image enhancement using multi-exposure sequence generation and image fusion, 29 (2020) 4481–4490.

[42] W. Kim, Low-light image enhancement by diffusion pyramid with residuals, J. Vis. Commun. Image Represent. 81 (2021) 103364.

[43] D. Fourure, R. Emonet, É. Fromont, D. Muselet, A. Tremeau, C. Wolf, Residual conv-deconv grid network for semantic segmentation, 2017, http://dx.doi.org/10.5244/C.31.181.

[44] S.D. Das, S. Dutta, Fast deep multi-patch hierarchical network for nonhomogeneous image dehazing, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 1994–2001, http://dx.doi.org/10.1109/CVPRW50498.2020.00249.

[45] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: MICCAI, 2015.

[46] H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, M.-H. Yang, Multi-scale boosted dehazing network with dense feature fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2157–2167.

[47] Y. Dong, Y. Liu, H. Zhang, S. Chen, Y. Qiao, FD-GAN: Generative adversarial networks with fusion-discriminator for single image dehazing, in: AAAI, 2020.

[48] R. Malav, A. Kim, S.R. Sahoo, G. Pandey, DHSGAN: An end to end dehazing network for fog and smoke, in: C. Jawahar, H. Li, G. Mori, K. Schindler (Eds.), Computer Vision – ACCV 2018, Springer International Publishing, Cham, 2019, pp. 593–608.

[49] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[50] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, Adv. Neural Inf. Process. Syst. 33 (2020) 18661–18673.
[51] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423, URL: https://aclanthology.org/N19-1423.
[52] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2020, arXiv:1910.01108.
[53] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019.
[54] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019, URL: http://arxiv.org/abs/1907.11692. cite arXiv:1907.11692.
[55] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, in: ICLR, OpenReview.net, 2020, URL: http://dblp.uni-trier.de/db/conf/iclr/iclr2020.html#LanCGGSS20.
[56] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16 × 16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021, URL: https://openreview.net/forum?id=YicbFdNTTy.
[57] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer, 2020, pp. 213–229.
[58] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable {DETR}: Deformable transformers for end-to-end object detection, in: International Conference on Learning Representations, 2021, URL: https://openreview.net/forum?id=gZ9hCDWe6ke.
[59] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, PMLR, 2021, pp. 10347–10357.
[60] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, R. Pang, Conformer: Convolution-augmented transformer for speech recognition, 2020, pp. 5036–5040, http://dx.doi.org/10.21437/Interspeech.2020-3015.
[61] Z. Wu, Z. Liu, J. Lin, Y. Lin, S. Han, Lite transformer with long-short range attention, in: ICLR, 2020, URL: https://openreview.net/forum?id=ByeMPlHKPH.
[62] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, CvT: Introducing convolutions to vision transformers, 2021, pp. 22–31, http://dx.doi.org/10.1109/ICCV48922.2021.00009.
[63] S. Wang, B. Li, M. Khabsa, H. Fang, H. Ma, Linformer: Self-attention with linear complexity, 2020, URL: http://arxiv.org/abs/2006.04768. cite arXiv:2006.04768.
[64] Y.N. Dauphin, A. Fan, M. Auli, D. Grangier, Language modeling with gated convolutional networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 933–941.
[65] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015, CoRR abs/1512.03385 (2015).
[66] S.A. Taghanaki, A. Bentaieb, A. Sharma, S.K. Zhou, Y. Zheng, B. Georgescu, P. Sharma, Z. Xu, D. Comaniciu, G. Hamarneh, Select, attend, and transfer: Light, learnable skip connections, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2019, pp. 417–425.
[67] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[68] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015, URL: http://arxiv.org/abs/1409.1556.
[69] D. Strong, T. Chan, Edge-preserving and scale-dependent properties of total variation regularization, Inverse Problems 19 (6) (2003) S165.
[70] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015, URL: http://arxiv.org/abs/1412.6980.
[71] P. Izmailov, D. Podoprikhin, T. Garipov, D.P. Vetrov, A.G. Wilson, Averaging weights leads to wider optima and better generalization, in: A. Globerson, R. Silva (Eds.), Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018, AUAI Press, 2018, pp. 876–885, URL: http://auai.org/uai2018/proceedings/papers/313.pdf.
[72] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 32, Curran Associates, Inc., 2019, pp. 8024–8035, URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.
[73] C.O. Ancuti, C. Ancuti, F.-A. Vasluianu, R. Timofte, et al., NTIRE 2020 challenge on NonHomogeneous dehazing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE CVPR 2020, 2020.
[74] C.O. Ancuti, C. Ancuti, M. Sbert, R. Timofte, Dense haze: A benchmark for image dehazing with dense-haze and haze-free images, in: IEEE International Conference on Image Processing (ICIP), IEEE ICIP 2019, 2019.
[75] C.O. Ancuti, C. Ancuti, R. Timofte, C.D. Vleeschouwer, O-HAZE: a dehazing benchmark with real hazy and haze-free outdoor images, in: IEEE Conference on Computer Vision and Pattern Recognition, NTIRE Workshop, NTIRE CVPR'18, 2018.
[76] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, L. Beyer, How to train your vit? Data, augmentation, and regularization in vision transformers, 2021, CoRR abs/2106.10270.
[77] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.