



Learning A Sparse Transformer Network for Effective Image Deraining



- 트랜스포머 기반 방법은 **비지역적인 정보**를 모델링할 수 있음
- 가장 유용한 self-attention 값만을 유지하여 통합된 특징을 제공하는 **Sparse Transformer(DRSformer)**를 제안
- 기존 트랜스포머 모델의 한계 극복(모든 토큰의 유사성 학습)

1. Introduction

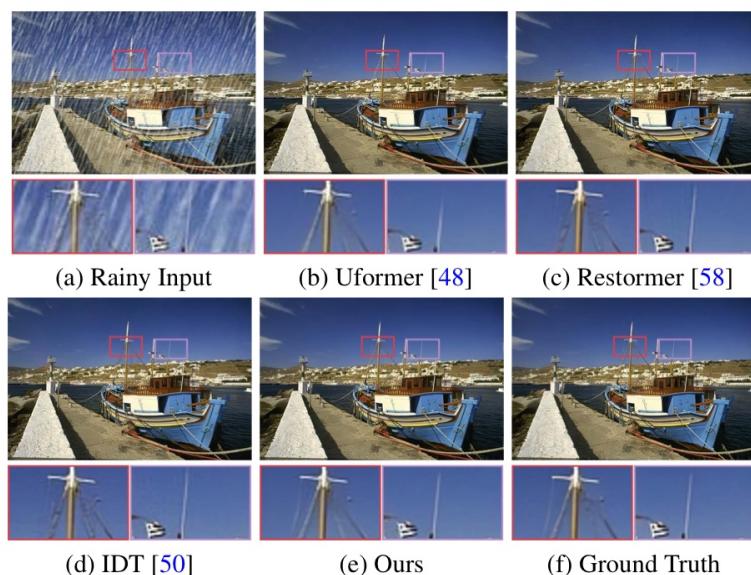


Figure 1. Image deraining results between our method and recent Transformer-based methods [48, 50, 58]. Our method can generate high-quality image with more accurate detail and texture recovery.

- 해당 논문은 deraining(비오는 날/흐린 날 등에 외부에서 촬영된 이미지에서 객체만을 남겨 깨끗한 이미지로 복원하는 task) 문제에 대한 새로운 접근 방식으로 Sparse Transformer 네트워크인 **DRSformer**를 제안

- CNN은 지역적인 특징을 잘 파악하지만 장거리 deraining을 처리하는 데는 제한이 있음
- 반면, 트랜스포머는 비-로컬 정보를 잘 모델링하여 고품질 이미지 복원에 도움이 되지만, 이미지의 지역적인 세부 사항을 적절히 처리하지 못함
⇒ 기존의 CNN과 트랜스포머 모델의 한계를 극복하고 deraining 성능을 향상
- DRSformer는 희소 트랜스포머 블록과 혼합 스케일 피드 포워드 네트워크를 결합
 - 희소 트랜스포머 블록은 self-attention 메커니즘을 사용하여 중요한 정보에 집중하고 불필요한 정보를 제거하여 특성 집계를 개선
 - 혼합 스케일 피드 포워드 네트워크는 다중 스케일 정보를 활용하여 이미지 디레이닝 성능을 향상
- DRSformer의 구조는 세 가지 이점을 가짐
 1. 불필요한 특성 간섭에 민감도가 낮음 → 견고한 모델
 2. 지역성을 향상시키고 전역적인 특성의 활용 능력을 향상시킴
 3. 데이터와 콘텐츠의 희소성을 동시에 탐색하여 디레이닝 성능을 향상시킴

2. Related Work

Single image deraining

- 전통적인 이미지 디레이닝 방법은 수작업 이미지 선험을 개발하여 추가 제약 조건을 제공
 - 그러나 이러한 선험은 주로 경험적이며 깨끗한 이미지의 본질적인 특성을 잘 모델링하지 못함
- 이에 대한 대안으로 CNN 기반의 다양한 프레임워크가 개발되었지만, 합성곱의 제한으로 인해 장거리 종속성을 캡처하는 데 어려움이 있음
 - 이에 반해, 트랜스포머를 네트워크 백본으로 사용하는 방법은 비-로컬 정보를 모델링하여 이미지 디레이닝을 향상시킬 수 있음

Vision Transformers

- 트랜스포머는 이전의 CNN 기반 방법보다 이미지 복원 작업에서 우수한 성능을 보임
 - 이미지 deraining 작업에서는 트랜스포머의 self-attention을 배경 복구 네트워크와 결합하여 동적 연관 deraining 네트워크를 설계하는 작업 등에 트랜스포머가 활용되었음

- 최근에는 이미지 디레이닝 트랜스포머(IDT)를 개발하여 탁월한 결과를 얻었음
- 그러나 기존의 dot-product self-attention 방식은 중복된 특성이나 더 작은 가중치의 특성이 attention map에 간섭할 수 있음
→ 잠재적
 잡음 발생 가능성

⇒ 이 연구에서는 트랜스포머에서 **sparse attention**을 도입

Sparse representation

- 심층 신경망의 표현이 희소해지는 것이 시각 및 자연어 처리 작업에서 유용하다는 개념이 각광받고 있음
 - 희소 표현은 저수준 비전 문제를 다루는 데 중요한 역할을 하며, 데이터 기반과 콘텐츠 기반 희소 주의로 나눌 수 있음
- 해당 논문에서는 상위 k개의 선택을 기반으로 한 self-attention의 간단하면서도 효과적인 근사 방법을 도입하여 **sparse attention**을 달성하는 방법을 제안

Top-k selection

- Zhao 등은 NLP 작업에서 상위 k 메커니즘을 기반으로 한 명시적 선택 방법을 처음으로 제안하였음
 - 그들의 성공을 토대로, k-NN 주의가 시각 트랜스포머의 성능을 향상시키기 위해 도입되었음
- 공간 차원에서 상위 k 선택을 수행하는 것과는 달리, 해당 논문에서는 효율적인 상위 k 개의 유용한 채널 선택 연산자를 설계하였음

3. Proposed Method

3-1. Overall pipeline

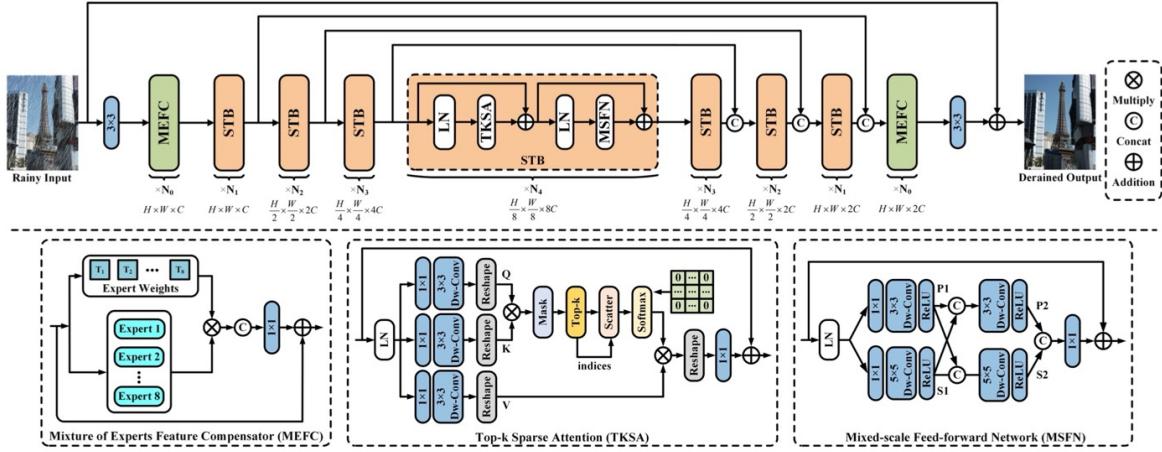


Figure 2. The overall architecture of the proposed sparse Transformer network for image deraining (DRSformer), which mainly contains sparse Transformer block (STB) with top- k sparse attention (TKSA) and mixed-scale feed-forward network (MSFN), and mixture of experts feature compensator (MEFC). LN refers to the layer normalization and DW-Conv refers to the depth-wise convolution.

- 계층적 encoder-decoder 기반 구조

- $H \times W$: feature map의 공간 해상도
- 3*3 convolution을 활용하여 겹치는 이미지 패치를 임베딩
- 네트워크 backbone에서는 $N_{i \in [1,2,3,4]}$ 개의 STB 를 쌓아 공간적으로 다양한 비 분포를 추출

⇒ 고유한 공간 해상도와 채널 차원을 다룸

- 특징 다운샘플링 및 업샘플링

- pixel-unshuffle과 pixel-shuffle 작업을 적용
- 안정적인 훈련을 위한 연속 중간 특징 연결을 위해 skip-connections을 추가

- 각 STB

- 특징 희소성을 달성하기 위해 TKSA 를 개발하여 특징 집계 과정을 더 효과적으로 강제
- MSFN 이 도입되어 다중 스케일 로컬 정보를 풍부하게 하고 이미지 복원에 도움

- 모델 학습의 초기와 최종 단계에서 N_0 개의 MEFC 를 도입하여 보충적인 특성 정제를 제공하여 높은 품질의 명확한 출력을 최종적으로 재구성

- DRSformer가 적응 콘텐츠와 비오는 이미지의 고유한 속성을 모두 활용하여 원치 않는 비 흔적과 잠재적인 명확한 배경을 분리할 수 있게 되었음

- 최종 재구성 결과

The final reconstructed result is obtained by: $I_{derain} = \mathcal{F}(I_{rain}) + I_{rain}$, where $\mathcal{F}(\cdot)$ is the overall network and it is trained by minimizing the following loss function:

$$\mathcal{L} = \|I_{derain} - I_{gt}\|_1, \quad (1)$$

where I_{gt} denotes the ground-truth image, and $\|\cdot\|_1$ denotes the L_1 -norm.

3-2. Sparse Transformer block

- 표준 transformer는 전역 범위에서 모든 token을 활용하여 self-attention을 계산
→ 노이즈가 많이 섞일 수 있는 문제

⇒ **Sparse Transformer Block(STB)**를 도입

that emerged in neural networks [64]. Formally, given the input features at the $(l-1)$ -th block \mathbf{X}_{l-1} , the encoding procedures of STB can be defined as:

$$\underbrace{\mathbf{X}'_l}_{\text{output}} = \underbrace{\mathbf{X}_{l-1}}_{\text{input features}} + \text{TKSA} \left(\underbrace{\text{LN}(\mathbf{X}_{l-1})}_{\substack{\uparrow \text{layer normalization}}} \right), \quad (2)$$

$$\mathbf{X}_l = \mathbf{X}'_l + \text{MSFN} (\text{LN}(\mathbf{X}'_l)), \quad (3)$$

where LN denotes the layer normalization; \mathbf{X}'_l and \mathbf{X}_l denote the outputs from the top- k sparse attention (TKSA) and mixed-scale feed-forward network (MSFN), which are described below.

Top-k sparse attention(TKSA)

Q , key K and value V with the dimension of $\mathbb{R}^{L \times d}$, the output of dot-product attention is generally formulated as:

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \left(\frac{\mathbf{Q} \mathbf{K}^\top}{\lambda} \right) \mathbf{V}, \quad (4)$$

* dimension: $\mathbb{R}^{L \times d}$
↑ query ↑ key ↑ value
↑ temperature factor (optional)

where \mathbf{Q} , \mathbf{K} , and \mathbf{V} denote the matrix forms of Q , K , and V , respectively. λ is an optional temperature factor defined by $\lambda = \sqrt{d}$. Generally, multi-head attention is implemented to each of the k new Q , K and V , yielding $d = C/k$ channel dimensional outputs which are concatenated and then got the final result for all heads via the linear projection.

- top-k 점수보다 작은 요소들의 경우 분산 함수를 통해 주어진 인덱스에서의 해당 확률을 0으로 대체
 - 이러한 동적 선택은 dense → sparse로 attention을 이동

function. This dynamic selection makes the attention from *dense* to *sparse*, which is derived by:

$$\text{SparseAtt}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \left(\mathcal{T}_k \left(\frac{\mathbf{Q}\mathbf{K}^\top}{\lambda} \right) \right) \mathbf{V}, \quad (5)$$

where $\mathcal{T}_k(\cdot)$ is the learnable top- k selection operator:

$$[\mathcal{T}_k(\mathbf{S})]_{ij} = \begin{cases} S_{ij} & S_{ij} \in \underset{\substack{\downarrow \text{top-}k \text{ value} \\ \rightarrow \text{부록을 보자}}}{\text{top-}k(\text{row } j)} \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

Finally, we multiply the softmax and value by matrix multiplication. As we use the multi-head strategy, we concatenate all the outputs of multi-head attention, and then get the final result by the linear projection.

Mixed-scale feed-forward network(MSFN)

- 기존 연구에서는 일반적으로 단일 규모의 심층 컨볼루션을 일반 feed-forward network에 도입하여 지역성을 개선
 - 이러한 방식의 경우 다중 스케일 비줄기의 상관 관계가 무시됨
- 해당 논문에서는 전송 프로세스에 두 개의 multi-scale depth-wise convolution 경로를 삽입하여 **MSFN**을 설계하였음
 - input tensor: $X_{t-1} \in \Re^{H*W*C}$
 - layer normalization 후 일단 1x1 convolution을 활용하여 채널 차원을 확장 → 이후 2개의 평행 브랜치에 전송(feed)
- MSFN 전개 과정

$$\begin{aligned}
 \hat{\mathbf{X}}_l &= f_{1 \times 1}^c(\text{LN}(\hat{\mathbf{X}}_{l-1})), \\
 \mathbf{X}_l^{p_1} &= \sigma(f_{3 \times 3}^{dwc}(\hat{\mathbf{X}}_l)), \mathbf{X}_l^{s_1} = \sigma(f_{5 \times 5}^{dwc}(\hat{\mathbf{X}}_l)), \# \text{ 두 개의 } \\
 &\quad \text{파행 branchall feed} \\
 \mathbf{X}_l^{p_2} &= \sigma(f_{3 \times 3}^{dwc}[\mathbf{X}_l^{p_1}, \mathbf{X}_l^{s_1}]), \mathbf{X}_l^{s_2} = \sigma(f_{5 \times 5}^{dwc}[\mathbf{X}_l^{s_1}, \mathbf{X}_l^{p_1}]), \# \text{ feature } \\
 &\quad \text{branch local 특성} \\
 \mathbf{X}_l &= f_{1 \times 1}^c[\mathbf{X}_l^{p_2}, \mathbf{X}_l^{s_2}] + \mathbf{X}_{l-1},
 \end{aligned} \tag{7}$$

where $\sigma(\cdot)$ is a ReLU activation, $f_{1 \times 1}^c$ represents 1×1 convolution, $f_{3 \times 3}^{dwc}$ and $f_{5 \times 5}^{dwc}$ denote 3×3 and 5×5 depth-wise convolutions, and $[\cdot]$ is the channel-wise concatenation.

3-3. Mixture of Experts Feature Compensator

- MEFC는 **DRSformer**에 희소성을 통합하기 위한 핵심 요소로, 병렬 레이어인 "전문가"를 통해 다양한 희소 CNN 작업을 적용
 - 각 전문가의 가중치를 조절하는 self-attention을 기반으로 하며, 각 전문가의 출력은 1×1 컨볼루션과 ReLU 함수를 통해 생성됨
 - 이를 통해 MEFC는 다양한 비오는 효과를 적응적으로 제거할 수 있음
- 전체 과정

tions depending on the inputs. Given an input feature map $\mathbf{X}_{l-1} \in \mathbb{R}^{H \times W \times C}$, we first apply the channel-wise average to generate C -dimensional channel descriptor $\mathbf{z}_c \in \mathbb{R}^C$:

$$\mathbf{z}_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W \mathbf{X}_{l-1}(i, j), \quad (8)$$

channel descriptor
↑ channel-wise average
input feature map
position

where $\mathbf{X}_{l-1}(i, j)$ is the (y, x) position of the feature \mathbf{X}_{l-1} . Then, the coefficient vector of each expert is allocated corresponding to the learnable weight matrices $\mathbf{W}_1 \in \mathbb{R}^{T \times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{O \times T}$. Here, T is the dimension of the weight matrices. To avoid altering the sizes of its inputs and outputs, we zero pad the input feature maps computed by each expert. Finally, the output of the l -th MEFC is calculated by:

$$\begin{aligned} \mathbf{T}_l &= \mathbf{W}_2 \sigma(\mathbf{W}_1 \mathbf{z}_c), \\ \mathbf{X}_l &= f_{1 \times 1}^c \left[\sum_{i=1}^O f_{exp}(\mathbf{X}_l, \mathbf{T}_l) \right] + \mathbf{X}_{l-1}, \end{aligned} \quad (9)$$

weight matrix
↑ dimension
ReLU
of experts
expert operation
convolution

where f_{exp} and O represent the expert operations and the number of experts, respectively. $f_{1 \times 1}^c$ represents 1×1 convolution, $\sigma(\cdot)$ is a ReLU function, and $[.]$ is the channel-wise concatenation. With this design, MEFC is now closely linked to the main STBs so that is able to adaptively remove the rainy effects of diverse appearances.

4. Experiments and Analysis

4-1. Experimental settings

Datasets

- Rain200L/H, DID Data 및 DDN-Data와 같은 다중 공개 벤치마크에서 디레이닝 실험을 구현
- SPA-Data와 같은 대규모 실제 데이터셋을 사용하여 평가하였음

Comparison methods

- 다양한 모델들과의 비교 실험 진행

- DSC 및 GMM과 같은 이전 기반 모델
- DDN, RESCAN, PReNet, MSPFN, RCDNet, MPRNet, DualGCN 및 SPDNet과 같은 CNN 기반 모델
- Transformer 기반 방법인 Uformer, Restormer 및 IDT

Evaluation metrics

- 위에서 언급한 벤치마크들에 대해 **PSNR** 및 **SSIM**을 평가 지표로 채택
 - 이전의 디레이닝 방법을 따라, YCbCr 공간의 Y 채널에서 PSNR 및 SSIM 지표를 계산
 - 또한, 참조 이미지가 없는 비오는 이미지의 경우에는 NIQE와 BRISQUE와 같은 비참조 메트릭을 사용

Training details

- $\{N_0, N_1, N_2, N_3, N_4\} = \{4, 4, 6, 6, 8\}$ 로 설정
 - 동일한 레벨의 네 개의 STB에 대한 어텐션 헤드의 수는 $\{1, 2, 4, 8\}$ 로 설정
- 초기 채널: $C = 48$, 확장 비율: 2
- MEFC
 - 전문가 수 $O = 8$
 - 가중치 행렬의 차원 $T = 32$
 - Rain200L 및 SPA-Data의 경우 복잡하지 않고 학습하기 쉬운 비오는 효과 때문에 MEFC를 사용 x
- STB
 - TKSA의 희소성 $[\text{delta_1}; \text{delta_2}] = [1/2, 4/5]$
 - MSFN의 채널 확장 계수: $r = 2.66$
- iteration = 300K
- 배치 크기 = 8인 AdamW 옵티마이저를 사용
- 학습률
 - 초기: 92K 번의 반복 동안 $1 * 10^{-4}$ 로 고정
 - 이후 코사인 단절 스키마에 따라 208K 번의 반복 동안 $1 * 10^{-6}$ 으로 감소
- 데이터 증강을 위해 수직 및 수평 반전이 무작위로 적용

- 전체 프레임워크는 4개의 NVIDIA GeForce RTX 3090 GPU에서 PyTorch에서 수행되며, 대규모 사전 훈련 없이 end-to-end 학습 방식으로 작동

4-2. Comparisons with the state-of-the-arts

Synthetic datasets

- 제안된 방법은 PSNR을 특히 향상시키며 다양한 벤치마크에서 우수한 성능을 보임

Table 1. Comparison of quantitative results on synthetic and real datasets. **Bold** and underline indicate the best and second-best results

Datasets		Rain200L		Rain200H		DID-Data		DDN-Data		SPA-Data	
Metrics		PSNR	SSIM								
Prior-based methods	DSC [30]	27.16	0.8663	14.73	0.3815	24.24	0.8279	27.31	0.8373	34.95	0.9416
	GMM [24]	28.66	0.8652	14.50	0.4164	25.81	0.8344	27.55	0.8479	34.30	0.9428
CNN-based methods	DDN [8]	34.68	0.9671	26.05	0.8056	30.97	0.9116	30.00	0.9041	36.16	0.9457
	RESCAN [23]	36.09	0.9697	26.75	0.8353	33.38	0.9417	31.94	0.9345	38.11	0.9707
	PReNet [36]	37.80	0.9814	29.04	0.8991	33.17	0.9481	32.60	0.9459	40.16	0.9816
	MSPFN [19]	38.58	0.9827	29.36	0.9034	33.72	0.9550	32.99	0.9333	43.43	0.9843
	RCDNet [43]	39.17	0.9885	30.24	0.9048	34.08	0.9532	33.04	0.9472	43.36	0.9831
	MPRNet [59]	39.47	0.9825	30.67	0.9110	33.99	0.9590	33.10	0.9347	43.64	0.9844
	DualGCN [9]	40.73	0.9886	31.15	0.9125	34.37	0.9620	33.01	0.9489	44.18	0.9902
	SPDNet [56]	40.50	0.9875	31.28	0.9207	34.57	0.9560	33.15	0.9457	43.20	0.9871
Transformer-based methods	Uformer [48]	40.20	0.9860	30.80	0.9105	35.02	0.9621	33.95	0.9545	46.13	0.9913
	Restormer [58]	40.99	<u>0.9890</u>	32.00	0.9329	35.29	<u>0.9641</u>	34.20	<u>0.9571</u>	47.98	<u>0.9921</u>
	IDT [50]	40.74	0.9884	32.10	0.9344	34.89	0.9623	33.84	0.9549	47.35	0.9930
	DRSformer	41.23	0.9894	32.18	0.9330	35.38	0.9647	34.36	0.9590	48.53	0.9924

- 특히, DRSformer는 평균적으로 IDT보다 0.4 dB 높은 성능을 나타냄

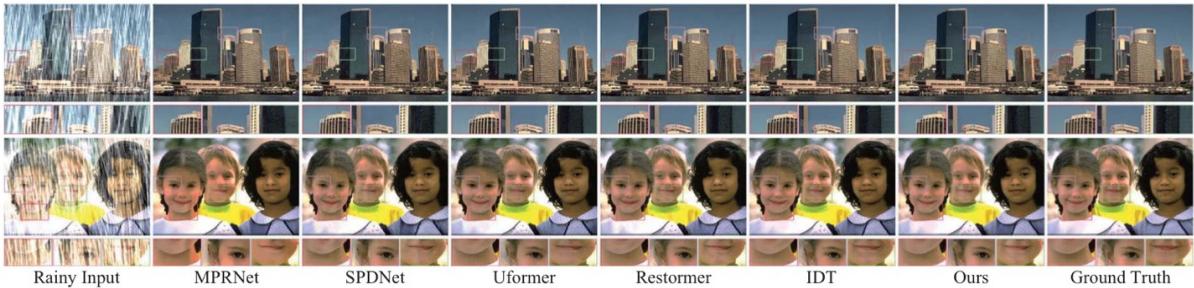


Figure 3. Visual quality comparison on the Rain200H dataset. Zooming in the figures offers a better view at the deraining capability.

- CNN 기반 모델과 비교하면 우리의 방법은 더 나은 결과를 보임
 - DID-Data 및 DDN-Data 벤치마크에서도 우수한 결과를 보여주며, 시각적으로도 우수한 결과를 제시
- 기존의 Transformer 기반 방법과 비교하여 세부 사항과 질감 복구에서 우수한 결과를 얻음

Real-world datasets

- SPA-Data 벤치마크에서 실험을 추가로 진행하였음
 - DRSformer는 여전히 최고의 PSNR/SSIM 값을 유지하며, 우수성을 입증
- 시각적 비교

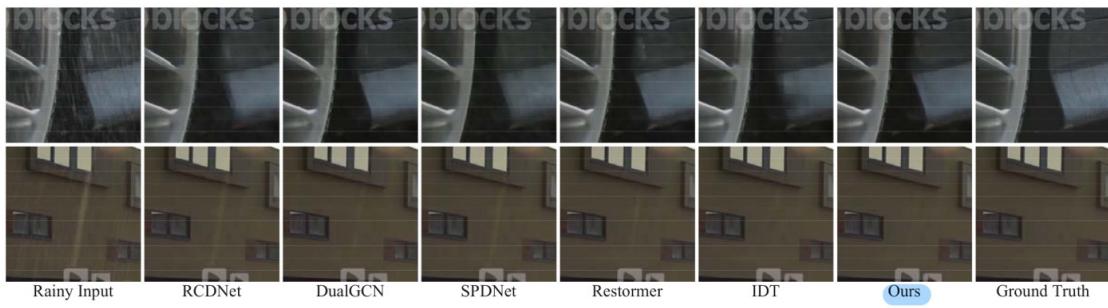


Figure 4. Visual quality comparison on the SPA-Data dataset. Zooming in the figures offers a better view at the deraining capability.



Figure 5. Visual quality comparison on real-world rainy images. Zooming in the figures offers a better view at the deraining capability.

- 또한, 인터넷 데이터에서 무작위로 선택한 20개의 실제 비오는 이미지에 대한 평가를 수행하여 DRSformer의 효과를 더 검증했음

Table 2. Comparison of quantitative results on real-world rainy images, and note that lower scores indicate better image quality.

Methods	Rainy Input	MPRNet [59]	SPDNet [56]	Uformer [48]	Restormer [58]	IDT [50]	Ours
NIQE ↓ / BRISQUE ↓	5.829 / 33.129	4.740 / 32.018	4.422 / 26.173	4.833 / 28.106	5.005 / 34.036	4.238 / 25.573	4.095 / 22.730

⇒ DRSformer이 실제 세계 데이터 유형에도 잘 일반화되는 것을 확인

4-3. Ablation studies

Effective of Top-k selection

- Top-k selection을 사용하는 경우 PSNR 값이 더 커짐

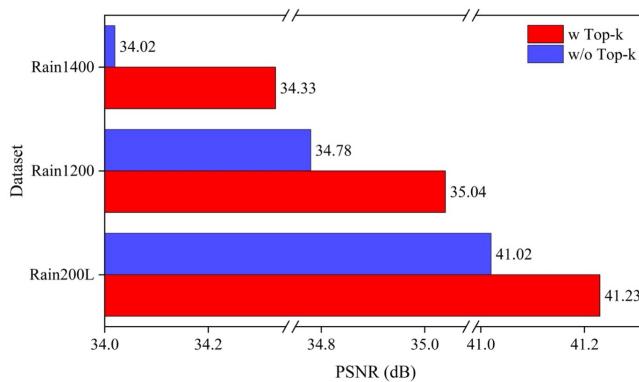


Figure 6. Ablation analysis for top-k selection on the benchmarks.

▼ cf> PSNR(Peak Signal-to-noise ratio)

- 영상 화질 손실양을 평가하기 위해 사용되는 지표

- 이미지 저장, 전송, 압축, 영상 처리 등에서 영상 화질이 바뀌었을 때 사용
- 클수록 좋음
- self-attention matrix softmax를 적용하는 경우 고주기성의 정보를 제거하는 경향이 있음 ⇒ 과도하게 스무딩된 결과가 도출됨
- 이러한 top-k 선택의 효과를 이해하기 위해, 학습된 특징을 시각화하기 위한 고주파 필터링 (HPF)를 사용

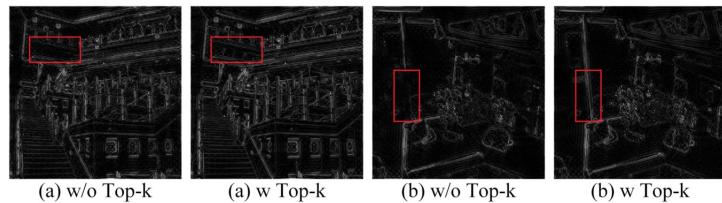


Figure 8. Visualization of feature maps. Our proposed top- k selection can effectively leverage pixel-dependent properties of image structure and generate more precise high-frequency details.

⇒ 더 미세한 세부 특징들까지 재구성하고 잠재적인 복원 품질을 향상시킬 수 있음

Effect of the number of k

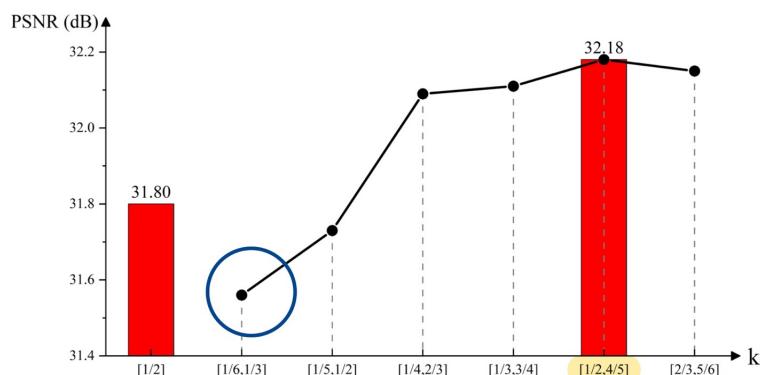


Figure 7. Ablation analysis for different number k in the TKSA.

- 모델에서 제안하는 TKSA의 주요 매개변수는 k 임
 - 최적의 k 선택이 희소율의 경계 제어를 결정한다는 점에서 유의
- 과도한 검색을 피하기 위해, k 에 대해 조절 가능한 간격 범위를 설정하여 가장 기여하는 점수를 동적으로 학습
- 작은 k 값은 전역 정보 집계가 충분하지 않아 성능이 감소
 - 최적의 결과는 TKSA에서 $[\delta_1, \delta_2]$ 가 $[1/2, 4/5]$ 로 할당될 때 얻어졌으며, k 값이 증가함에 따라 불필요하고 쓸모 없는 특징이 도입되어 디레이닝 성능이

점차 감소

Effectiveness of MSFN

- 제안된 MSFN의 효과를 평가하기 위해 세 가지 기준선과 비교
 - 전통적인 피드포워드 네트워크(FN)
 - Dconv 피드포워드 네트워크(DFN)
 - 게이트 Dconv 피드포워드 네트워크(GDFN)
- Rain200H에서의 정량적 분석 결과
 - GDFN은 성능 이점을 가져 오기 위해 두 개의 동일한 스케일 깊이별 합성곱 스트림에 게이팅 메커니즘을 도입하지만, 여전히 디레이닝을 위한 다중 스케일 지식을 무시
 - 서로 다른 스케일에서의 로컬 특징 추출과 융합을 추가함으로써, MSFN은 실제로 성능을 더 향상시킬 수 있으며, GDFN 대비 0.21 dB의 PSNR 이득을 달성

Effectiveness of MEFC

- MEFC의 효과를 평가하기 위해, 여러 가지 다른 모델 변형을 기반으로 실험을 수행하였음

Table 4. Ablation study for different variants of our DRSformer.
MEFC-1 and MEFC-2 denote MEFC in early and final stages.

Models	MEFC-1	STBs	MEFC-2	Experts	PSNR / SSIM
(a)		✓		0	32.03 / 0.9308
(b)	✓	✓		8	32.01 / 0.9311
(c)		✓	✓	8	32.07 / 0.9328
(d)	✓	✓	✓	1	32.06 / 0.9316
(e)	✓	✓	✓	4	32.14 / 0.9325
(f)	✓	✓	✓	8	32.18 / 0.9330

⇒ DRSformer의 최종 성능에 각 설계 전략이 각각의 고유한 기여를 함을 확인할 수 있음

4-4. Closely-related methods

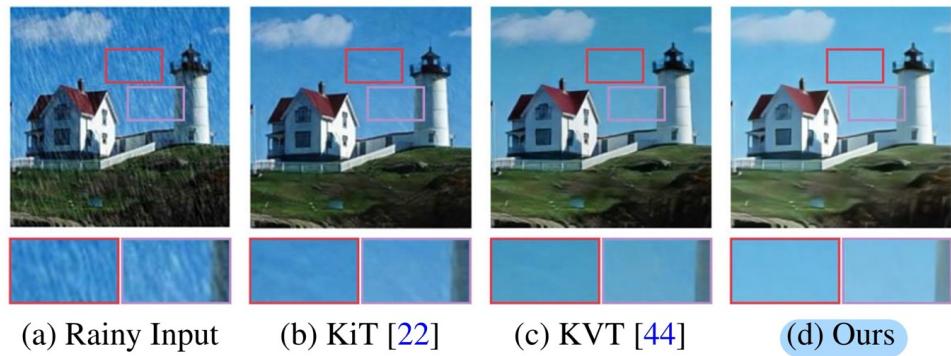


Figure 10. Comparison results with closely-related methods.

- 최근에는 이미지 복원을 해결하기 위해 k-NN 이미지 트랜스포머 (KiT)를 제안함
 - 쌍별 로컬 어텐션으로 k개의 유사 패치를 집계하는 복잡한 지역 민감 해싱을 사용
→ 충분한 전역 상호 작용을 보장할 수 없음
 - 반면 top-k 선택 메커니즘은 지역성을 즐기면서도 전역 관계 추론 능력을 강화
- KiT는 내용을 흐리게 만들고 색상 왜곡을 일으키는 경향이 있음
 - 반면, DRSformer는 더 나은 deraining 결과를 도출
- KVT와 달리, DRSformer는 채널 간의 희소 어텐션을 계산하는 데 더 효율적

5. Concluding Remarks

- 이미지 디레이닝을 해결하기 위한 효과적인 희소 트랜스포머 네트워크인 **DRSformer**를 제안
 - Transformer의 기본 자기-주의 메커니즘이 관련 없는 정보의 전역 상호 작용으로 인해 고통을 겪을 수 있다는 관찰을 기반으로, 더 나은 특성 집합을 위해 최우선 자기-주의 값을 유지하는 top-k 희소 어텐션을 개발하였음
 - 비에 대한 집계된 특성을 용이하게 하기 위해, 우리는 다중 스케일 표현을 더 잘 탐색하기 위해 혼합 스케일 피드포워드 네트워크를 개발하였음
 - 또한, 희소 트랜스포머 백본에 협력적인 정제를 제공하기 위해 전문가 특성 보상기를 도입하여 재구성된 이미지의 미묘한 세부 정보를 보존

⇒ 실험 결과는 DRSformer가 최신 기술과 유리하게 경쟁한다는 것을 입증하였음
- 제한사항) 모델 효율성
 - 모델은 크기가 256x256인 이미지에 대해 33.7백만 개의 매개변수가 필요하며 242.9G FLOP를 요구

- 신뢰할 수 있는 모델 압축을 달성하면서 원래의 디레이닝 성능을 유지하기 위해 모델에 가지치기(pruning) 또는 증류(distillation) 방법을 적용할 예정