Team 41

Multiple artifact removal from degraded images using CNN-ViT hybrid architecture

Tejas Dhopavkar • 7075870 | Panav Raina • 7075813

Problem Statement

Problem Statement

Identifying the key concern

- Real-world images are often affected by degradations through multiple artifacts.[4]
- Models have been trained to generate deblurred images based on either synthetic or natural data.
- Few models can tackle multiple degradations simultaneously and can generalize well for both (natural and synthetic) data.

Goals

Goals

We aim to achieve

- A model that can handle multiple degradations to produce a clear image.
- A model that generalizes well and gives good results on benchmark datasets.

Overview of Approach

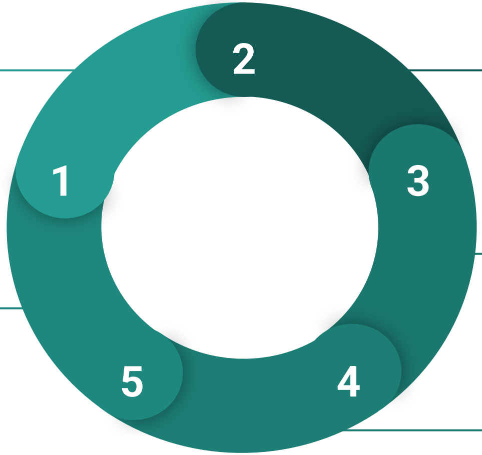
Approach

Dataset Collection

Datasets used are *GoPro* and *SIDD*.

Visualizations

Document loss, visualize attention and features.



Model Architecture

CNN-ViT Hybrid Architecture, ViT with image patching.

Model Training

Train the model over GPU's.

Model Evaluation

Evaluate performance of the model using metrics: PSNR, SSIM

Related Work

Related Work

Deep Image Deblurring: A Survey[2]

- The authors introduce deblurring as a classic low-level computer vision problem and provide a list of degradations that can affect an image.
- The authors also introduce us to various image deblurring and restoration architectures ranging from CNN, GAN, to Transformers while discussing their advantages and disadvantages.
- The authors also provide a brief description of the benchmark datasets like GoPro and SIDD.
- In addition, the authors also present the comparison between PSNR and SSIM of various models/architectures.

Related Work

DeblurDiff: Real-World Image Deblurring with Generative Diffusion Models. [1]

- Introduces a Latent Kernel Prediction Network with the pre-trained Stable Diffusion model.
- Element-wise adaptive convolution is applied to the kernel to preserve the structural information while deblurring.
- The kernels are iteratively refined using results from each diffusion step.
- The authors created their own dataset of 500,000 data pairs and used the GoPro dataset for evaluation.

Related Work

Restormer: Efficient Transformer for High-Resolution Image Restoration.[3]

- The authors propose a modified transformer architecture that focuses on attention across channels instead of spatial dimensions.
- They also utilize depth-wise convolutions. In regular CNNs, we apply the same filter across all the channels. However, in the Restormer, the authors apply a different filter to each of the channels before sending the features to compute cross attention.
- The Restormer is able to remove all degradations (noise, blur, de-raining, etc) successfully.
- The authors utilize the GoPro dataset along with many others for training and testing.

Proposed Approach and Experiments

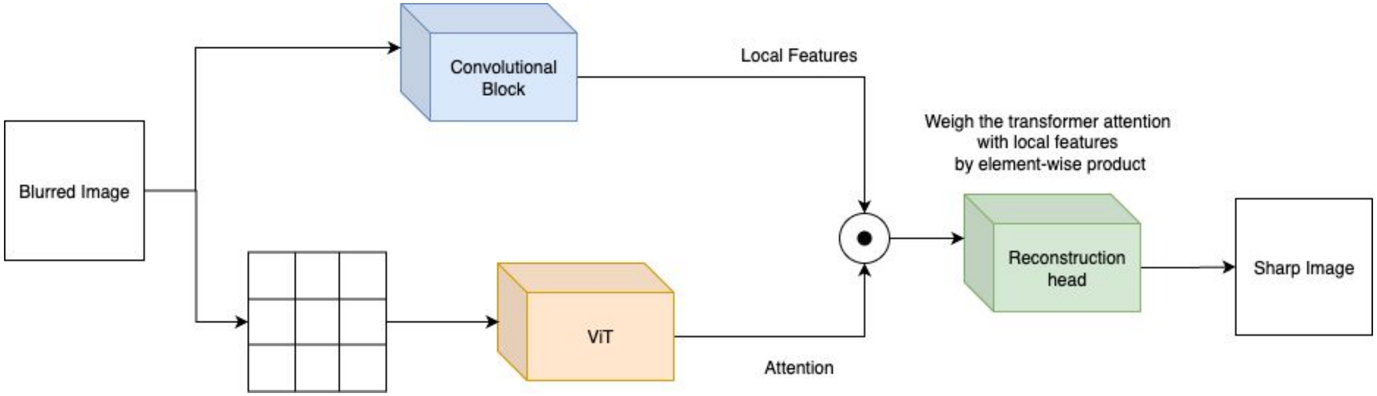
Dataset

Image Restoration Dataset

- GoPro[5] and SIDD[6] were combined.
- Synthetic noise, low resolution added to GoPro dataset.
- Synthetic blurring, low resolution added to SIDD dataset.
- Total samples: 2214 (Training) + 1160 (Testing) = 3374 images.

Experiment 1

CNN-ViT Hybrid Architecture



Experiment 1 - Details

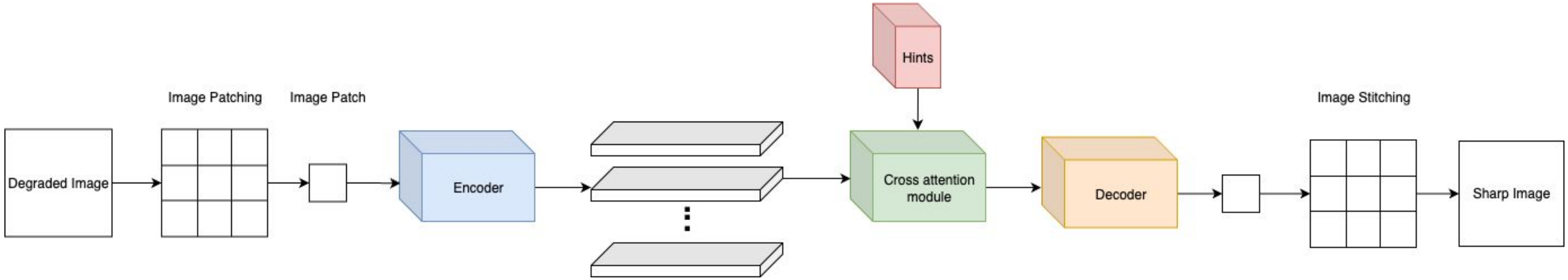
Parameter	Value
Epochs	100
Loss	MSE Loss
Optimizer	Adam Optimizer
Batch Normalization	NA
Hidden layers	5
Drop Probabilities	0.4
Transformer Weights	vit-base-patch16-224

Experiment 1 - Configuration

- High-Level Structure:
    1. Hybrid model combining Vision Transformer (ViT) and CNN. The CNN extracts the local features and ViT extracts the global features with it’s 12 attention heads. We compute the element-wise product of both features to get weighted-attention.
- Key Components:
    1. Custom Dataset: Image Restoration Dataset
    2. Model: CNN-ViT Hybrid Architecture
    3. Loss Function: MSE
    4. Training Pipeline: Extract the features using both branches, compute the product, reconstruct the image, and train the model on MSELoss.
    5. Evaluation & Visualization: Computes PSNR/SSIM, and visualizes input, predicted, and ground-truth images.

Experiment 2

ViT with Image Patching



Experiment 2 - Details

Parameter	Value
Epochs	100
Loss	MSE Loss
Optimizer	Adam Optimizer
Batch Normalization	Yes
Drop Probabilities	NA
Transformer Weights	Custom trained weights without initialization

Experiment 2 - Configuration

- High-Level Structure:
  1. Hybrid model combining Vision Transformer (ViT) for feature extraction and a convolutional decoder for image reconstruction, enhanced by cross-attention and hint integration.
- Key Components:
  1. Custom Dataset: Image Restoration Dataset
  2. Model: End-To-End ViT
  3. Loss Function: MSE
  4. Training Pipeline: Uses gradient accumulation, mixed-precision training, and a ReduceLROnPlateau scheduler for stable convergence.
  5. Evaluation & Visualization: Stitches patches, computes PSNR/SSIM, and visualizes input, predicted, and ground-truth images.

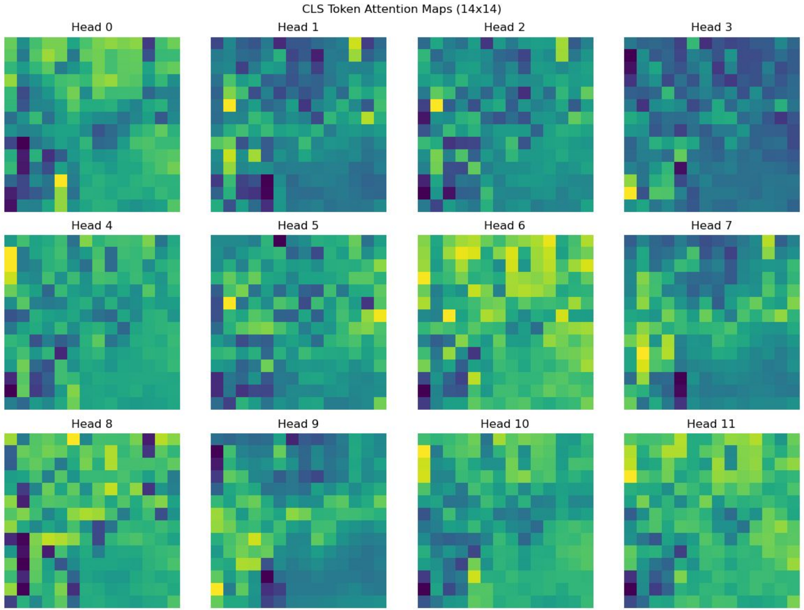
Results

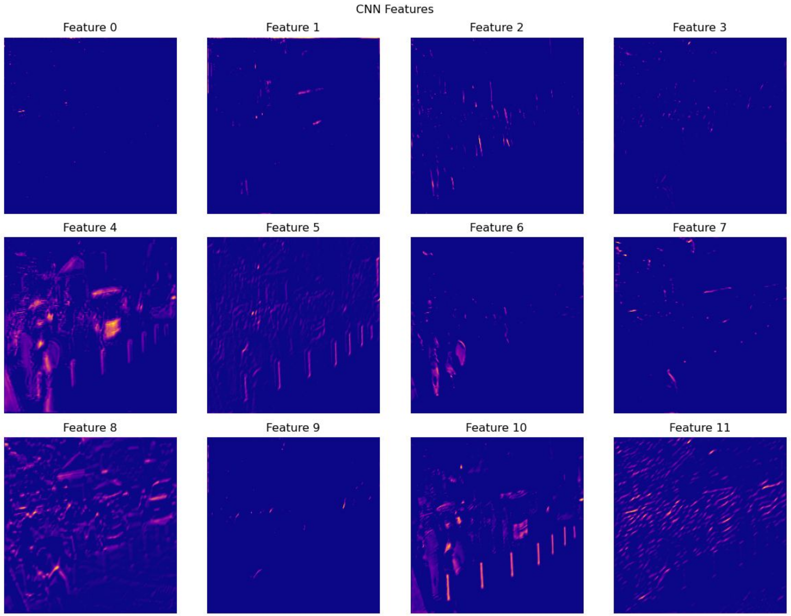
Experiment 1 - Results

Metric	Value
Average Loss	0.0019663786854552797
Average PSNR	34.09832070610481
Average SSIM	0.9226827049563671

Experiment 1 - Visualizations

Experiment 1 - Attention and CNN Features





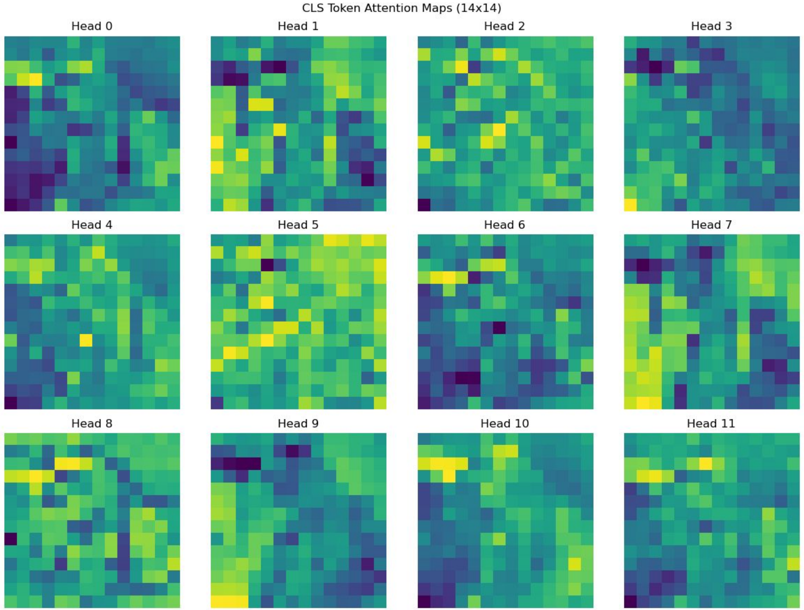
Visualizations - Experiment 1

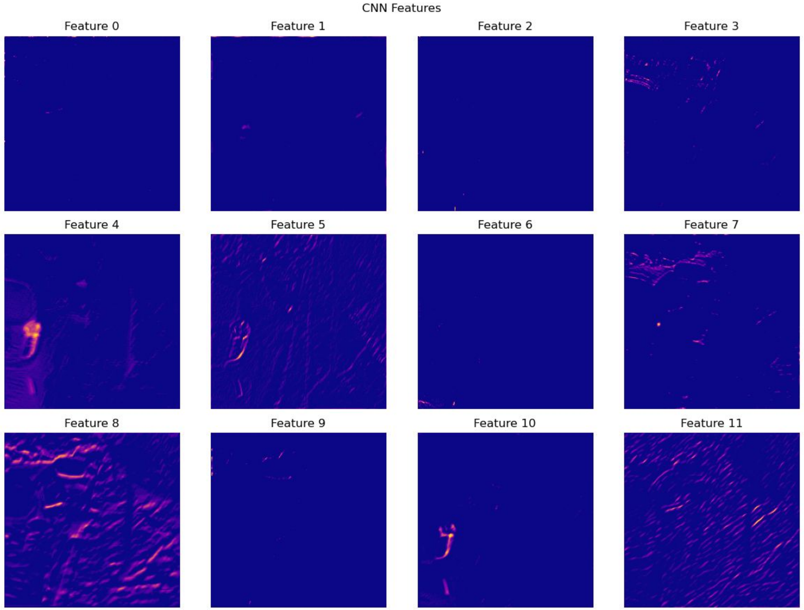
Blurred Image

Sharp Image

Output

Attention and CNN Features - Experiment 1

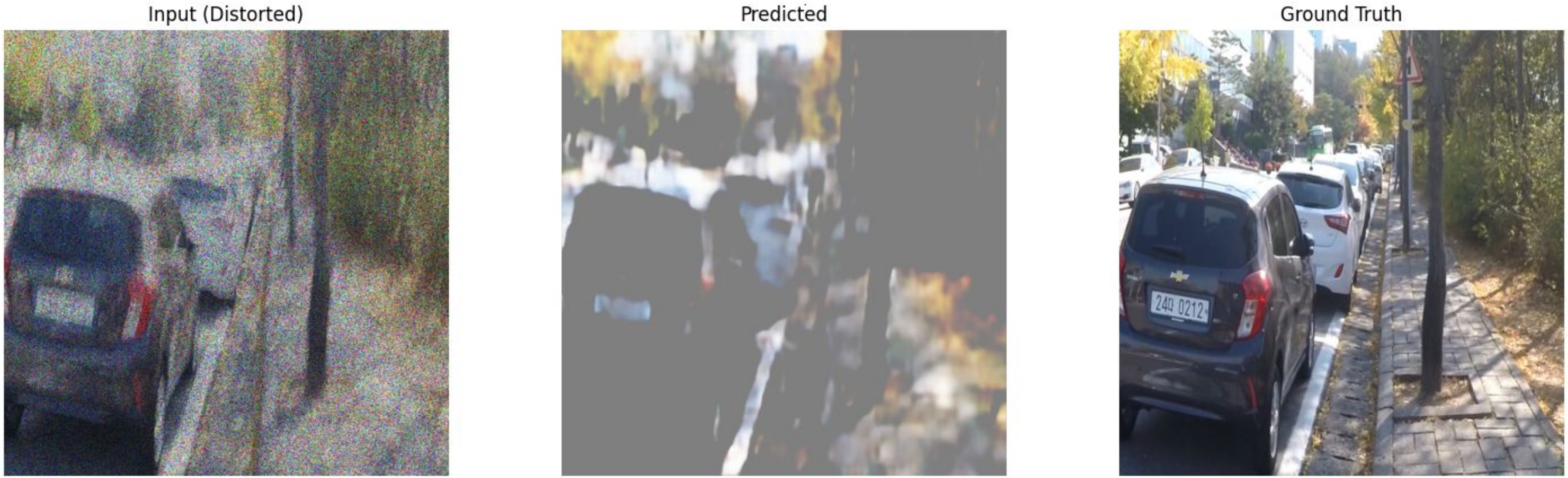




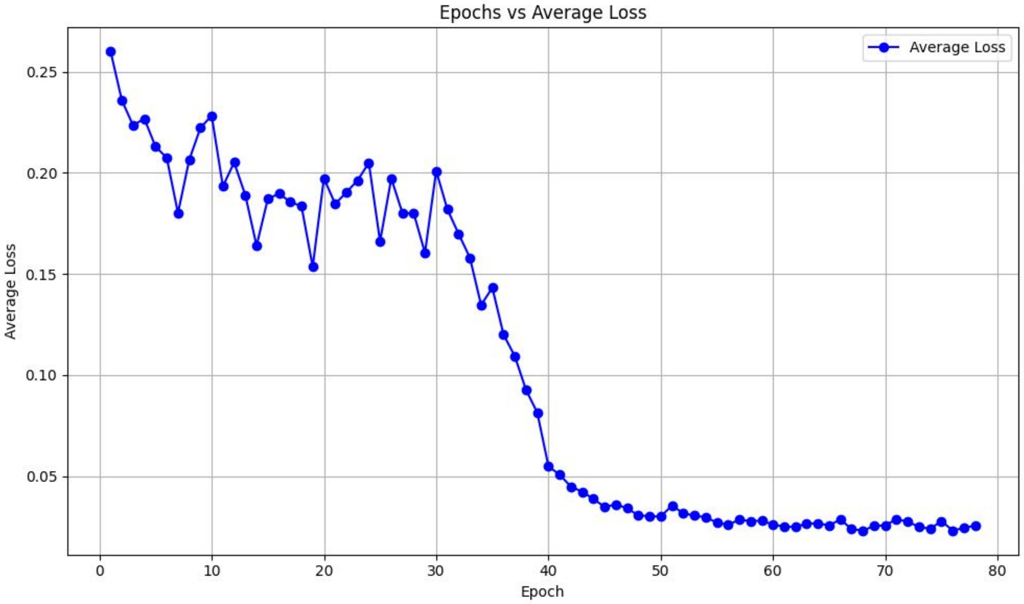
Results - Experiment 2

Metric	Value
Average Loss	0.002
Average PSNR	30.0
Average SSIM	0.90

Visualizations - Experiment 2



Visualizations - Experiment 2



Conclusion and Future Work

Conclusion

Remarks and summary

- Experiment 1 performs better and is able to resolve all three artifacts.
- “Fog” is observed in the generated output.
- Probable cause of the “fog” is a shallow decoder or inappropriate loss function.
- Proposed model has achieved good PSNR and SSIM values.

Future Work and Next Steps

Experiment 3

We aim to perform an ablation study, experimenting with individual branches of the proposed architecture, trying perceptual loss and including the results in the report.

Experiment 4

As an alternative approach to Experiment 2, we would compute the loss values on the complete image and not the patches by compromising the model complexity.

Report

Write the report and include all the findings.

References

1. Lingshun Kong, Jiawei Zhang, Dongqing Zou, Jimmy Ren, Xiaohe Wu, Jiangxin Dong, & Jinshan Pan. (2025). DeblurDiff: Real-World Image Deblurring with Generative Diffusion Models.
2. Kaihao Zhang, Wenqi Ren, Wenhan Luo, Wei-Sheng Lai, Bjorn Stenger, Ming-Hsuan Yang, & Hongdong Li. (2022). Deep Image Deblurring: A Survey.
3. Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, & Ming-Hsuan Yang. (2022). Restormer: Efficient Transformer for High-Resolution Image Restoration.
4. Kaihao Zhang, Wenqi Ren, Wenhan Luo, Wei-Sheng Lai, Bjorn Stenger, Ming-Hsuan Yang, & Hongdong Li. (2022). Deep Image Deblurring: A Survey.
5. Seungjun Nah, Tae Hyun Kim, & Kyoung Mu Lee. (2018). Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring.
6. A. Abdelhamed, S. Lin and M. S. Brown, "A High-Quality Denoising Dataset for Smartphone Cameras," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 1692-1700, doi: 10.1109/CVPR.2018.00182

Thank you

Questions and Feedback