Team 41: Multiple distortion artifact removal using CNN-Vision Transformer hybrid architecture

Tejas Dhopavkar
7075870

Panav Raina
7075813

1. Task and Motivation

1.1. Task Statement and Definitions

We aim to develop a unified deep learning model capable of handling three common types of image degradation: blur, noise, and low resolution. The model should accept a degraded image (which may exhibit one or more of these artifacts) and restore it to a high-quality version by performing image deblurring, denoising, and super-resolution simultaneously.

1.2. Motivation

Real-world images are often affected by multiple degradations due to limitations in imaging systems, environmental conditions, or transmission. Existing solutions typically address each degradation individually, limiting their practical utility. A unified model would be significantly more efficient and deployable in real-world applications such as surveillance, medical imaging, and autonomous systems. Furthermore, combining these tasks allows the model to learn shared representations and leverage complementary features.

While diffusion-based methods [12] have demonstrated excellent performance across multiple restoration tasks, they are computationally heavy and slow at inference. In contrast, we propose a more efficient and lightweight Vision Transformer (ViT)-based model enhanced with task-specific conditional hints. This allows us to dynamically adapt the model's focus to present artifacts while achieving real-time performance.

1.3. Related Work

Joint image restoration has seen progress through several recent models. JIRSR [7] presents a joint image super-resolution network capable of handling multiple degradations using a single architecture. D3Net [5] proposes a dual-domain denoising network tailored for real noise. JR-CNN [2] is a cascade CNN model targeting joint restoration. Diffusion-based approaches like Zhang et al. [12] explore the generative capability of denoising diffusion probabilistic models for restoration tasks. The restormer model proposed by Zamir et al. [10] introduces a transformer based model than can work for large as well as small images with various types of artifacts. However, very few models tackle all three degradation types jointly while incorporating both synthetic and naturally occurring artifacts.

1.4. Challenges

The primary challenges include:

1. Multi-degradation generalization: Designing a model that can handle combinations of blur, noise, and resolution loss effectively.

2. Natural vs. synthetic artifacts: Ensuring generalization across both naturally occurring and artificially introduced distortions.

3. Efficient architecture: Balancing performance with computational feasibility.

4. Conditional inference: Incorporating external hints indicating which degradations are present to guide restoration.

2. Goals

2.1. Challenges Addressed

This project addresses the lack of unified models capable of performing deblurring, denoising, and super-resolution together. We aim to bridge the gap between synthetic benchmarks and real-world applicability by training on both natural and augmented degradation types.

Our model will consist of 2 branches. One with a denoising and convolutional layers to suppress high-frequency noise and extract features, while the second branch will be a Vision Transformer (ViT) that incorporates task-specific hints (blur, noise, low-resolution indicators) into the attention mechanism.

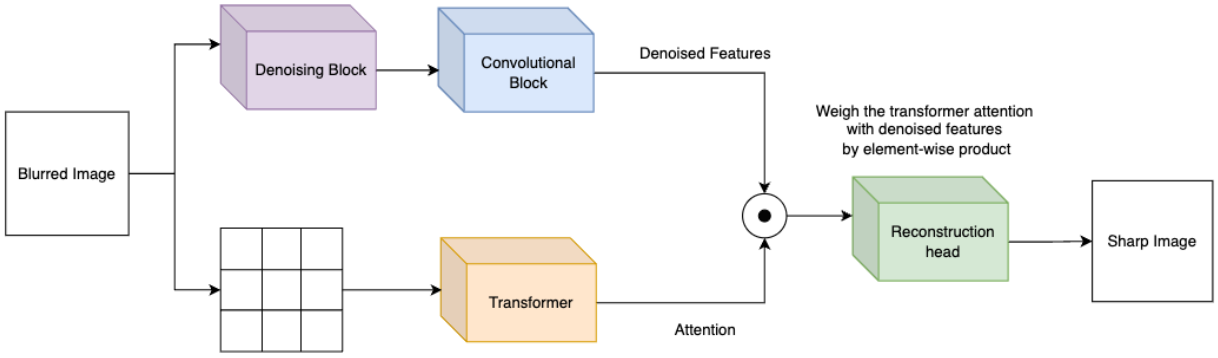

Figure 1. Proposed architecture.

2.2. Mid-Term Goals

By the mid-term checkpoint, we aim to:

1. Generate the synthetic data we require.
2. Finalize the initial convolutional stack of layers to be used.

3. Methods

3.1. Models and frameworks/libraries

We aim to utilize convolution operations for noise reduction and the vision transformer [3] for the image deblurring task. We propose a branched architecture where the first branch consists of de-noising and convolution operations. These operations will help reduce the noise in the image and extract the local features. The second branch consists of the vision transformer. The vision transformer will be trained to remove the blur in the image. The outputs of both branches will undergo an element-wise multiplication. As a result, we obtain a weighted embedding of the image. This weighted embedding is then passed to the image-reconstruction head to obtain the sharpened image.

To implement the proposed architecture, we plan to use *PyTorch* and *HuggingFace*. *PyTorch* and *HuggingFace* provide implementations of various state-of-the-art models, including the Vision Transformer. The *VisionTransformer* base class can be utilized to implement a vision transformer that takes image tokens as input [3].

3.2. What makes the proposed architecture be effective?

The authors of [11] mention that an image can experience distortions from various sources, in addition to blurring. The distortions can be in the form of noise, low resolution, low/high contrast among others. With the proposed architecture, our aim is to tackle the induction of noise and low resolution, along with blurring.

The first branch will provide us with relatively noiseless features. We also aim to utilize appropriate padding and kernel size to maintain the dimensions of the image. The vision transformer will be trained to de-blur the image by generating contextual features. The hypothesis is that by performing a point-wise multiplication of the outputs of both branches, we will be able to obtain an embedding weighted by prominent local features, thus helping in the sharpening process.

By using vision transformers, we aim to have a lightweight architecture as compared to diffusion models.

3.3. Comparison with other models

Existing methods use synthetic data for model training. As a result, they perform well on synthetic data but performs poorly on real-world data [11]. Kong et al. [6] have utilized diffusion models for de-blurring images. However, their model is heavy and requires more than 500,000 data points. Hence, model training is resource heavy and time consuming.

3.4. Computational budget

We plan to train our own model. We have access to a GPU on a local system. However, we might need access to the GPU provided by the university for model training purposes.

4. Datasets

We intend to use the following datasets:

1. GoPro Dataset [8], RealBlur dataset [4]: These are widely-used and are often considered as benchmark datasets for image deblurring.

2. DND dataset [9], SIDD dataset [1]: Many datasets utilize images induced with synthetic noise. Models tend to perform better on synthetic noise as compared to real-world noise [11]. Hence, we chose these datasets as they have images with real-world noise.

5. Evaluation

5.1. Learning paradigm and loss functions

We aim to follow the supervised learning paradigm. To optimize the model, Zhang et al. [11] have listed various loss functions like pixel loss, perceptual loss, adversarial loss and relativistic loss.

5.2. Metrics

To evaluate the performance of the model, our preferred choice of metric would be either PSNR or SSIM [11] as they are widely used for image reconstruction tasks like deblurring. We do not intend to use any other custom metric.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A high-quality denoising dataset for smartphone cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[2] Gadipudi Amaranageswarao and S. Deivalakshmi. Joint restoration convolutional neural network for low-quality image super resolution. 11 2020. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 2

[4] Jucheol Won Sunghyun Cho Jaesung Rim, Haeyun Lee. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[5] Tomáš Kerepecký and Filip Sroubek. D3net: Joint demosaicking, deblurring and deringing. pages 1–8, 01 2021. 1

[6] Lingshun Kong, Jiawei Zhang, Dongqing Zou, Jimmy Ren, Xiaohe Wu, Jiangxin Dong, and Jinshan Pan. Deblurdiff: Real-world image deblurring with generative diffusion models, 2025. 2

[7] Guoping Li, Zhenting Zhou, and Guozhong Wang. A joint image super-resolution network for multiple degradations removal via complementary transformer and convolutional neural network. *IET Image Processing*, 18:n/a–n/a, 01 2024. 1

[8] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, July 2017. 2

[9] Tobias Plötz and Stefan Roth. Benchmarking denoising algorithms with real photographs, 2017. 3

[10] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming–Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5718–5729, 2022. 1

[11] Kaihao Zhang, Wenqi Ren, Wenhan Luo, Wei-Sheng Lai, Björn Stenger, Ming-Hsuan Yang, and Hongdong Li. Deep image deblurring: A survey. *CoRR*, abs/2201.10700, 2022. 2, 3

[12] Yi Zhang, Xiaoyu Shi, Dasong Li, Xiaogang Wang, Jian Wang, and Hongsheng Li. A unified conditional framework for diffusion-based image restoration. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc. 1