Beyond Natural Images: A Benchmark for Cross-Domain Image Reconstruction

Anonymous CVPR submission

Paper ID *****

Abstract

This project will investigate how well state-of-the-art deep learning models for image reconstruction generalize across different visual domains. By evaluating CNN-, Transformer-, and Diffusion-based architectures on super-resolution, denoising, and inpainting tasks across diverse image types—such as natural scenes, text, astronomical, and stylized images—we aim to identify consistent strengths, weaknesses, and failure patterns. The findings will provide insights into cross-domain robustness and guide the development of more adaptable, content-aware reconstruction approaches.

1. Introduction

Every day, cameras capture billions of images—landscapes, selfies, night skies, documents, street scenes—and many of them are blurry, noisy, or low-resolution. Image reconstruction models promise to fix that: they take a damaged picture and make it clear again. But here's the catch—the same model that works beautifully on a photo of a mountain often fails on a picture of a starry sky, a page of text, or a sketch. You might get letters that melt together, stars that vanish, or cartoon lines that blur.

Right now, researchers and developers typically train separate models for each type of image. There are models just for faces, others for anime drawings, and others for natural scenes. This works in narrow settings but creates a messy ecosystem of one-off tools. If you want to enhance mixed content—say a photo of a person under the night sky with visible text—there's no reliable, all-purpose solution. Each model specializes but lacks general understanding of what it's looking at.

That's the gap we want to fill. Our goal is to first quantify that gap and test how well today's best reconstruction models actually generalize across different kinds of images—from skies and constellations to text, animals, and human portraits—and figure out why some architectures succeed where others fail. Instead of building yet another model from scratch, we'll take the most advanced existing ones (CNNs, Transformers, Diffusion models) and evaluate them on multiple visual domains using public datasets. By comparing their strengths and weaknesses, we'll learn what design choices make them robust and where they break.

If we succeed, we'll provide something useful for everyone—from photographers to scientists.

2. Planned Method

The central idea is to take State-of-The-Art models and evaluate their out-of-the-box generalization performance on unseen image domains without any fine-tuning. This approach directly tests their inherent robustness. By analyzing model designs that generalize best, we aim to contribute insights that could lead to a SoTA AI for improving image quality across diverse content. The process is broken down into three main stages.

2.1. Selection of Tasks and Architectures

We will focus on three core image reconstruction tasks: **Super-Resolution**, **Denoising**, and **In-Painting**. For each task, we will select representative, high-performing architectures that embody different design philosophies. Our selection will include CNN-based models like EDSR or ESRGAN, which excel at learning local features; Transformer-based models like SwinIR, which use attention to capture long-range dependencies and global structure; and Diffusion-based models like SR3, which iteratively reverse a noise process to synthesize photorealistic details.

2.2. Assembling a Cross-Domain Evaluation Suite

We will curate a benchmark suite composed of several challenging image domains, as detailed in Table 1. Models will be trained only on the Natural/Scenic domain, using standard benchmarks like DIV2K and Flickr2K, which will serve as our baseline. We will then test their generalization on several out-of-domain datasets. These include the Human/Animal domain (CelebA-HQ, AFHQ) to test reconstruction of fine facial features; the Text/Document domain (TextZoom) to evaluate the preservation of sharp edges and legibility; the Astronomy/Night Sky domain (Hubble, SDSS images) to assess performance on sparse, high-contrast data; and the Stylized/Line Art domain (anime/manga datasets) to test the handling of abstract content with sharp lines and flat colors.

Domain	Example Sources / Datasets	Purpose
Natural / Scenic	DIV2K, Flickr2K, COCO	Baseline domain
Human / Animal	CelebA-HQ, AFHQ	Faces, animals
Text / Document	TextZoom, scanned images	Structured symbols
Astronomy	Hubble, SDSS, astrophotos	Sparse, high-contrast
Stylized / Art	Anime/manga, Waifu2x	Strong edges, abstraction

Table 1. Cross-Domain Evaluation Datasets.

2.3. Evaluation Framework

For each task, we will use the official pre-trained weights for the selected architectures, which are typically trained on natural image datasets. We will then run inference with these models on the test sets from all domains listed above. The core analysis will focus on quantifying the performance degradation as models move from the familiar ”Natural” domain to the unfamiliar target domains.

3. Planned Experiments

Our experiments are designed to systematically evaluate how well existing deep learning models for image reconstruction generalize across different image domains. The evaluation will span three tasks (super-resolution, denoising, and inpainting) across six distinct image domains.

3.1. Experimental Setup

For each task, we will compare three representative architecture families: CNN-based models (ESRGAN, EDSR), a Transformer-based model (SwinIR), and a Diffusion-based model (SR3 or Stable Diffusion Upscaler). To ground our findings, we will include classical methods like bicubic interpolation and BM3D/DnCNN as baseline references. All deep learning models will be used in their pretrained form to ensure a fair comparison. Each model will be tested on a fixed set of 100 images per domain, with controlled degradations applied to simulate realistic conditions, such as downsampling, Gaussian noise, or random masking.

3.2. Metrics and Evaluation

To quantify performance, we will use a combination of metrics. For pixel-level fidelity, we will use **PSNR** and **SSIM**. For perceptual similarity, we will rely on **LPIPS**. Domain-specific metrics will also be employed, such as **OCR accuracy** for text domains and **star-count consistency** for astronomy images. We will also record runtime and GPU memory usage to assess computational efficiency. To specifically measure cross-domain robustness, we will compute a **Cross-Domain Drop (CDD)** metric, defined as the normalized performance difference between a model's in-domain and out-of-domain results.

3.3. Success Criteria

The project will be considered successful if our experiments demonstrate clear, measurable differences in how the model families behave across domains. A key indicator of success will be quantifying a significant generalization gap (a high CDD value) when models are applied to unfamiliar image types. We also aim to find consistent evidence that diffusion- and transformer-based models retain higher perceptual quality under domain shift.

4. Timeline and Feasibility

This project is designed to be feasible within a typical semester timeframe. We have broken down the work into several phases to ensure steady progress and timely completion.

Weeks 1-3: Setup and Baseline. This phase will involve a final literature review, setting up the computational environment, and downloading all datasets. We will also run all models on the baseline ”Natural/Scenic” domain to establish our in-domain performance metrics.

Weeks 4-8: Cross-Domain Experiments. This is the core experimental phase. We will systematically run the pre-trained models for all three tasks (super-resolution, denoising, inpainting) across the five out-of-domain test suites. Results and performance metrics will be carefully logged.

Weeks 9-11: Analysis and Interpretation. During this period, we will analyze the collected data, calculate the Cross-Domain Drop (CDD) for each model, and generate visualizations to compare performance. We will focus on identifying failure modes and patterns that explain why certain architectures generalize better than others.

Weeks 12-14: Final Report and Presentation. The final weeks will be dedicated to writing the final project report, structuring our findings into a coherent narrative, and preparing the final presentation.