Evaluating Out-of-Distribution Robustness in Image Inpainting

Sandra Elsa Sanjai
sandra.elsasanjai@studenti.unipd.it

Abstract

Image inpainting models are typically evaluated under in-distribution corruption settings, providing limited insight into their robustness under semantic distribution shift. In this work, we study out-of-distribution robustness in image inpainting through controlled perturbations that progressively increase spatial severity and semantic inconsistency. In addition to standard variations in mask size, blur, and semantic masking, we introduce a semantic shuffling experiment that inserts semantically incompatible facial components into fixed spatial locations. We propose a quantitative boundary measure based on perceptual similarity to capture the trade-off between learned priors and local visual evidence. Our results show that diffusion-based models exhibit greater robustness to semantic inconsistencies than classical approaches, highlighting the limitations of standard inpainting evaluations and motivating semantically grounded robustness benchmarks.

1. Introduction

Image inpainting aims to reconstruct missing or corrupted regions of an image in a way that is both visually plausible and semantically consistent. Recent advances in generative modeling, particularly diffusion-based approaches, have led to substantial improvements on standard benchmarks. However, most evaluations are conducted under tightly controlled or weakly perturbed settings, where the corruption process closely resembles the training distribution. In real-world scenarios, missing regions can vary widely in size, structure, and semantic content, making robustness to out-of-distribution (OOD) corruptions a critical aspect of inpainting performance.

In this work, we systematically evaluate the OOD robustness of image inpainting models (classical and diffusion-based) by perturbing the corruption process along multiple axes, including mask size, blur severity, semantic mask type, and semantic shuffling. Our main contributions are as follows:

- We introduce a structured experimental framework that evaluates image inpainting models under progressively more challenging out-of-distribution corruptions, spanning spatial severity, semantic masking, and semantic inconsistency.
- We propose a semantic shuffling experiment, providing a principled way to quantify the trade-off between learned priors and local visual evidence.
- We demonstrate that diffusion-based models, particularly those trained on diverse data, exhibit greater robustness to semantic distribution shift than deterministic reconstruction-based approaches.

2. Related Work

Classical Inpainting. Early learning-based inpainting methods focused on deterministic reconstruction by propagating visible structure and texture, beginning with Context Encoders [7]. CNN-based approaches such as Partial Convolutions [3] and EdgeConnect [6] further improved structural continuity by conditioning on valid pixels or predicted edges. More recently, transformer-based models such as MAE-FAR [1] adapt masked autoencoders for image reconstruction, achieving strong performance on large missing regions. However, these methods remain deterministic and often rely on local or token-level context, making them sensitive to misleading visible content.

Diffusion-based Inpainting. Generative approaches aim to model a distribution over plausible image completions rather than producing a single deterministic output. Diffusion-based methods, in particular, formulate inpainting as conditional sampling, enabling iterative refinement guided by a learned image prior. Techniques such as RePaint [5] and inpainting pipelines built on Stable Diffusion [8] have demonstrated strong global coherence and perceptual quality, especially for large or complex masks. These models outperform deterministic approaches but are typically evaluated under corruption patterns closely aligned with training.

Robustness under shift. Robustness in image inpainting is relatively underexplored compared to perceptual quality. Most evaluations focus on mask size or noise severity, which preserve semantic consistency and therefore do not constitute true distribution shift. Insights from robustness studies in vision suggest that models may exploit spurious cues under shift, yet this perspective is rarely applied to inpainting. Our work addresses this gap through semantic masking and semantic shuffling experiments that explicitly probe boundary reasoning under OOD conditions.

3. Dataset

All experiments are conducted on the CelebAMask-HQ dataset [2], a large-scale collection of aligned human face images with diverse identities, poses, and lighting conditions. This is an extension of the CelebA dataset [4], which is widely used in image inpainting research, additionally providing face segmentation masks. The dataset contains a consistent structure and has the presence of semantically meaningful regions such as eyes, nose, and mouth, making it well suited for evaluating semantic robustness under controlled corruptions.

From the testing set of 6000 data samples, we use a subset of 500 uniformly sampled images which are then normalized, center-cropped and resized to a fixed 256×256 resolution. Corresponding masks are generated with respect to each experiment's demands:

- **Varying Mask Size:** Rectangular masks are constructed by expanding the bounding box of the nose segmentation. The mask width and height are scaled by factors $(0.8, 1.0, 1.5, 2.0)$, where the final value covers the majority of the face. This yields normalized masked areas of 0.16, 0.25, 0.56, and 1.0, respectively. For a visual example, refer to Figure 1.

- **Varying Blur Size:** The ground truth image is corrupted using Gaussian blur with increasing kernel sizes $(0, 13, 25, 51)$ prior to masking. The rectangular mask is fixed with a scaling factor of 1.5 for all blur levels. For a visual example, refer to Figure 2.

- **Semantic Masks:** With the help of the available facial segmentations, individual semantic parts are masked (both-eye, left-eye, mouth, nose). For a visual example, refer to Figure 3.

- **Semantic Shuffle:** The mouth region is copied and overlaid onto the left-eye region using facial segmentation. This overlaid region is progressively masked, such that the ground-truth eye region is masked while the semantically misleading content is visible. For a visual example, refer to Figure 5.

Using the above tests, we progressively go from in-distribution to out-of-distribution robustness evaluation.

4. Method

4.1. Problem Statement

Let $x \in \mathbb{R}^{H \times W \times 3}$ denote a ground-truth image and $\mathcal{C}(\cdot)$ a corruption operator that removes or alters a subset of pixels. Image inpainting aims to reconstruct a completed image $\hat{x}$ given a corrupted input $\tilde{x} = \mathcal{C}(x)$, such that $\hat{x}$ is visually plausible and semantically consistent with the uncorrupted regions of x.

As detailed in section 3, we test robustness by evaluating multiple inpainting models under a shared corruption framework, progressively moving from standard in-distribution settings to semantically inconsistent out-of-distribution scenarios.

4.2. Corruption Operators

We define four families of corruption operators, each designed to probe a distinct aspect of robustness. Apart from the mask generation (Section 3), we discuss the problem that each corruption scheme inquires.

Varying Mask Size. To evaluate sensitivity to increasing occlusion, we construct rectangular masks anchored to the nose segmentation. This setting tests robustness to loss of spatial cues while preserving a fixed semantic reference.

Varying Blur Size. To isolate robustness to information degradation, we apply Gaussian blur with increasing kernel sizes prior to masking.

Semantic Masks. To probe semantic awareness, we selectively mask complete individual facial components using segmentation annotations. This setting would provide an idea of which semantic regions have stronger priors or rely on local features.

Semantic Shuffle. To explicitly test boundary reasoning under semantic inconsistency, we introduce a semantic shuffling operator. This creates corrupted inputs that contain misleading semantic content at a fixed spatial location. This setting provides a good way to test for the reliance on learned priors against local features.

The defined corruption operators span a continuum from in-distribution to out-of-distribution settings. Mask size and blur variations correspond to parametric perturbations that preserve semantic consistency, while semantic masking and semantic shuffling introduce distribution shifts at the semantic level. Evaluating all models across this continuum allows us to disentangle robustness to increased difficulty from robustness to novel and semantically inconsistent corruptions.

4.3. Models

We evaluate three inpainting models that represent complementary points in the design space of inpainting approaches.

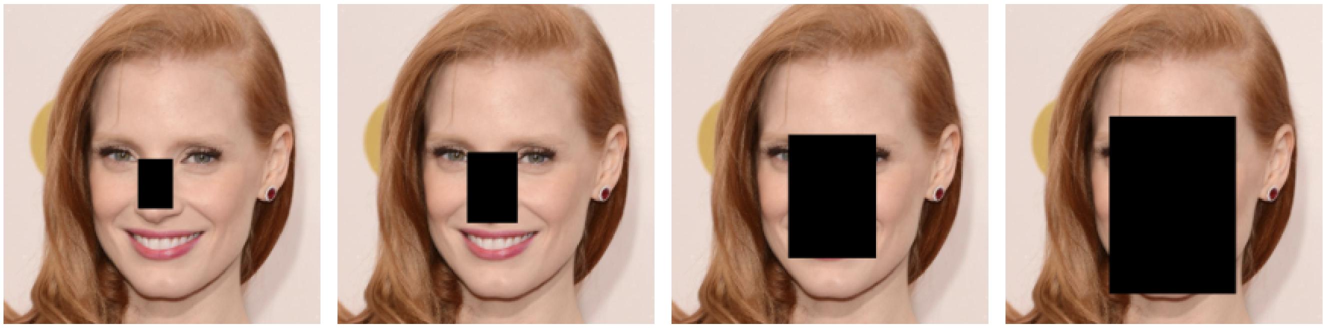

Figure 1: **Varying Mask Sizes** Rectangular masks, centered on the nose segmentation, are varied using scaling factors $0.8, 1.0, 1.5, 2.0$ from left to right.

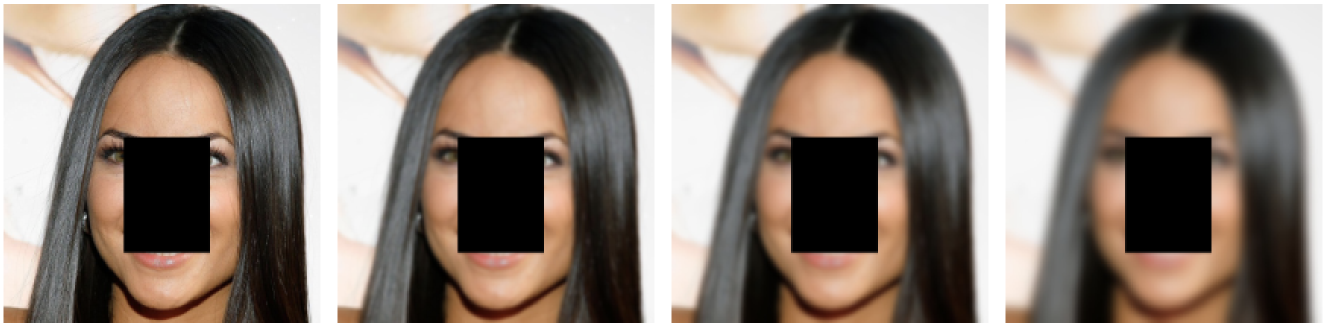

Figure 2: **Varying Blur Size** A fixed spatial rectangular mask (with scaling factor 1.5 is applied to gaussian blurred ground truths. The gaussian kernel size varies as $0, 13, 25, 51$ from left to right.

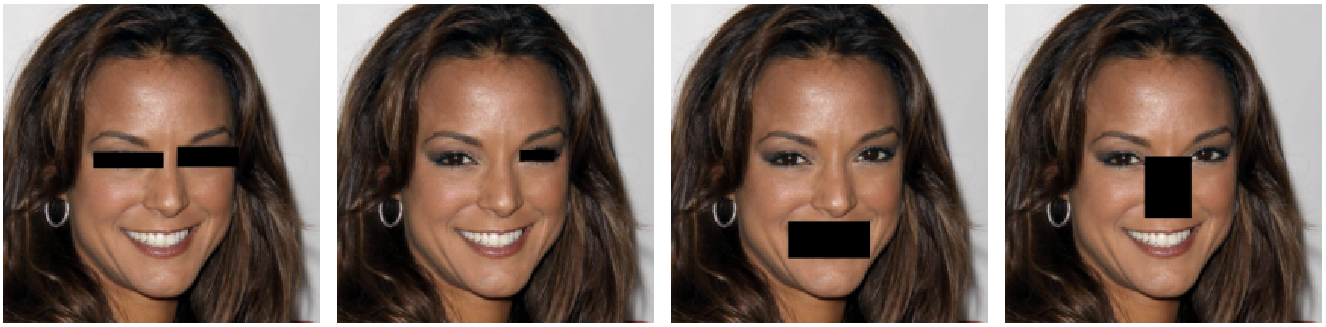

Figure 3: **Semantic Masks** Individual semantic regions in the face is masked using rectangular masks. From left to right, we mask both eyes, left eye, mouth, and nose.

MAE-FAR [1] is used as a deterministic baseline, using the checkpoint trained on the FFHQ dataset. It adapts masked autoencoders for image reconstruction and produces a single completion conditioned on visible tokens. As a non-generative method, MAE-FAR relies on learned reconstruction priors rather than sampling from a distribution, making it a strong representative of classical, deterministic inpainting approaches.

RePaint [5] represents diffusion-based inpainting models trained on domain-specific data (CelebA-HQ). It reformulates inpainting as a conditional diffusion process, repeatedly enforcing consistency with known pixels during sampling. Trained solely on facial data, RePaint serves as a specialized generative model with strong in-domain priors, making it well suited for evaluating robustness under facial-specific corruptions.

Stable Diffusion (Inpainting) [8] represents large-scale diffusion models trained on broad, heterogeneous datasets. Using a pretrained inpainting checkpoint [1], it performs conditional generation in a latent space. Although it supports both image and text conditions, we only rely on masked im-

[1] Checkpoint : runwayml/stable-diffusion-inpainting

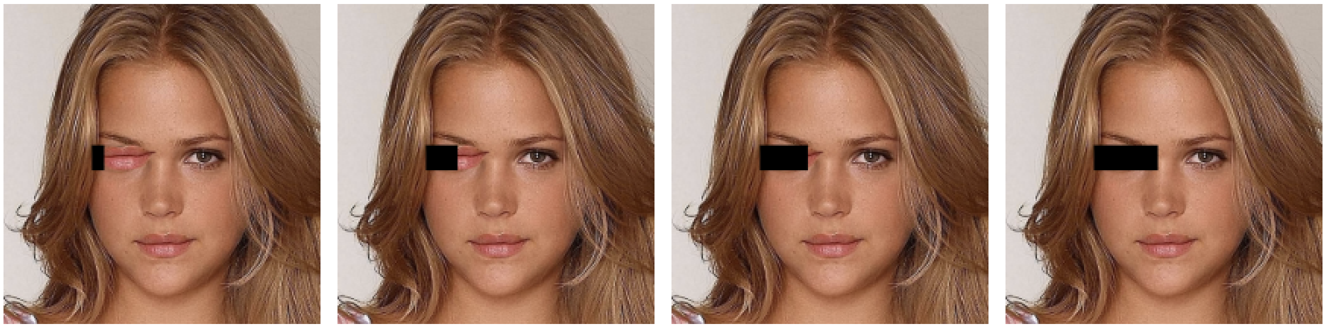

Figure 4: **Semantic shuffling** The mouth region is overlaid on top of the left-eye region, and masking is applied progressively. The above masks correspond to sizes $0.2, 0.5, 0.75, 1.0$.

age conditions (guidance scale is set to zero). Its general-purpose training makes it a useful contrast to RePaint, allowing us to examine how domain-specific versus general generative priors affect robustness under distribution shift.

4.4. Metrics

We evaluate inpainting quality using a combination of structural, perceptual, and distribution-level metrics to capture complementary aspects of reconstruction quality and robustness.

SSIM measures the similarity between reconstructed and ground-truth images based on luminance, contrast, and structural information. While SSIM provides insight into pixel-level fidelity, it is known to correlate weakly with perceptual realism, particularly for generative models.

LPIPS evaluates perceptual similarity using deep feature representations and is better aligned with human perception, making it suitable for assessing semantic plausibility.

KID measures the similarity between the distributions of reconstructed and ground-truth images in a learned feature space and provides a stable, unbiased estimate for small sample sizes. We use KID to assess how closely the distribution of inpainted outputs matches that of real images.

5. Experiments

We evaluate the robustness of three inpainting models—Stable Diffusion (inpainting) [8], RePaint [5], and MAE-FAR [1]—under progressively more challenging corruption settings. Experiments are designed to move from standard in-distribution perturbations to increasingly out-of-distribution and semantically inconsistent scenarios. Quantitative results are reported using SSIM, LPIPS, and KID, with arrows indicating the preferred direction for each metric.

5.1. Varying Mask Size

Table 1 reports performance as the size of the masked region increases. Across all models, increasing mask size leads to a consistent degradation in SSIM and LPIPS, indicating that larger occlusions pose a greater challenge regardless of model class. However, the rate and nature of degradation differ significantly.

Table 1: Varying mask sizes.

Method	**Mask Size**	KID ↓	SSIM ↑	LPIPS ↓
Stable-D	0.8	**4.344**	**0.8889**	**0.0364**
	1.0	4.466	0.8825	0.0385
	1.5	4.404	0.8563	0.0452
	2.0	4.880	0.8086	0.0607
RePaint	0.8	**4.213**	**0.9456**	**0.0245**
	1.0	4.558	0.9386	0.0267
	1.5	4.801	0.9106	0.0353
	2.0	5.415	0.8625	0.0535
MAE-FAR	0.8	**17.694**	**0.9372**	**0.0339**
	1.0	18.478	0.9165	0.0397
	1.5	21.529	0.8912	0.0548
	2.0	29.792	0.8445	0.0803

Diffusion-based models (Stable Diffusion [8] and RePaint [5]) exhibit relatively stable KID values across mask sizes, with only a moderate increase at the largest setting. In contrast, MAE-FAR [1] shows a sharp and monotonic increase in KID, indicating a growing mismatch between reconstructed and real image distributions as occlusion increases. While MAE-FAR [1] maintains competitive SSIM at smaller mask sizes, its performance degrades substantially for larger masks, suggesting limited robustness to large missing regions.

Among the diffusion models, RePaint [5] scores high in the perceptual metrics, showing that domain-specific priors are robust to spatial occlusion, compared to generally trained priors.

Overall, these results indicate that diffusion-based models are more robust to increasing spatial extent of occlusion, while deterministic reconstruction-based methods struggle as the task departs from their effective operating regime.

Table 2: Varying blur sizes.

Method	**Blur Size**	SSIM ↑	LPIPS ↓
Stable-D	0	0.9581	**0.0119**
	13	**0.9599**	0.0218
	25	0.9589	0.0329
	51	0.9533	0.0454
RePaint	0	**0.9532**	**0.0209**
	13	0.8833	0.1315
	25	0.8148	0.2407
	51	0.7299	0.3669
MAE-FAR	0	0.9479	**0.0224**
	13	**0.9524**	0.0278
	25	0.9520	0.0356
	51	0.9501	0.0459

Table 3: Semantic Masks

Method	**Mask Type**	KID ↓	SSIM ↑	LPIPS ↓
Stable-D	both-eye	4.261	0.9903	0.00443
	l-eye	4.237	**0.9983**	**0.00145**
	mouth	**4.119**	0.9907	0.00343
	nose	4.230	0.9900	0.00287
RePaint	both-eye	5.288	0.9487	0.02281
	l-eye	5.416	0.9485	0.02337
	mouth	4.903	0.9498	0.02248
	nose	**4.645**	**0.9499**	**0.02224**
MAE-FAR	both-eye	21.711	0.9901	0.00816
	l-eye	18.649	**0.9982**	**0.00244**
	mouth	19.439	0.9889	0.00627
	nose	**18.458**	0.9866	0.00515

5.2. Varying Blur Size

Table 2 evaluates robustness to low-frequency corruption by increasing Gaussian blur prior to masking. Here, we only consider perceptual metrics since we modify the ground truth, which can skew the distributional-metrics and make KID unreliable. Additionally, we also consider a masked ground truth and prediction, such that the perceptual metrics only attend to the masked inpainting region.

From the results, we see that Stable Diffusion [8] and MAE-FAR [1] maintain relatively stable SSIM values across blur levels, whereas RePaint exhibits a pronounced degradation as blur strength increases.

This behavior suggests that RePaint [5], despite being diffusion-based, is sensitive to deviations from the sharp facial statistics present in its training data. In contrast, Stable Diffusion [8] benefits from its large-scale and diverse training corpus, allowing it to better handle blurred inputs. MAE-FAR [1] also shows strong robustness to blur, likely because blur preserves coarse structure and does not introduce semantic ambiguity, which aligns well with deterministic reconstruction objectives.

These results highlight that robustness to blur depends not only on model class (generative vs deterministic) but also on training distribution and prior diversity.

5.3. Semantic Masking

Table 3 reports performance when masking specific semantic regions. Similar to section 5.2, we use masked inputs to calculate the perceptual metrics. From the results, we see that all models show variation across mask types, indicating that semantic importance of the missing region affects reconstruction difficulty.

For Stable Diffusion [8], performance remains consistently strong across all semantic masks, with minimal variation in SSIM, LPIPS, and KID. RePaint [5] shows moderate sensitivity to the masked region, with slightly worse performance for eye-related masks compared to the nose or mouth. This might possibly be due to its strong spatial priors favoring semantic regions with homogeneous textures.

MAE-FAR [1] achieves high SSIM values across all semantic masks but exhibits substantially higher KID compared to diffusion-based models, suggesting that high structural similarity does not necessarily correspond to realistic or well-distributed reconstructions. This might likely be due to reduced output diversity

5.4. Semantic Shuffling

To explicitly probe robustness under semantic inconsistency, we introduce a semantic shuffling experiment in which the mouth region is copied and overlaid onto the left-eye region, creating a corrupted input that contains semantically misleading content at a fixed spatial location. This experiment is designed to disentangle reliance on learned priors from sensitivity to local visual evidence.

We focus on comparing Stable Diffusion [8] and MAE-FAR [1] in this setting. RePaint [5] is not included, as preliminary experiments showed that its domain-specific training on facial data induces extremely strong spatial priors. In practice, this causes RePaint [5] to rely almost exclusively on these priors, largely ignoring local visual cues in the shuffled region, which makes it unsuitable for measuring the trade-off between learned semantic priors and local evidence that this experiment aims to capture.

To quantify robustness, we use LPIPS as a perceptual similarity measure and compute it against two different reference targets: the original ground-truth image and a modified ground truth in which the mouth fully replaces the left-eye region. As the mask size over the shuffled region increases, LPIPS with respect to the original ground truth decreases (the model is encouraged to reconstruct a plausible eye), while LPIPS with respect to the modified ground truth increases (the influence of the misleading mouth content di-

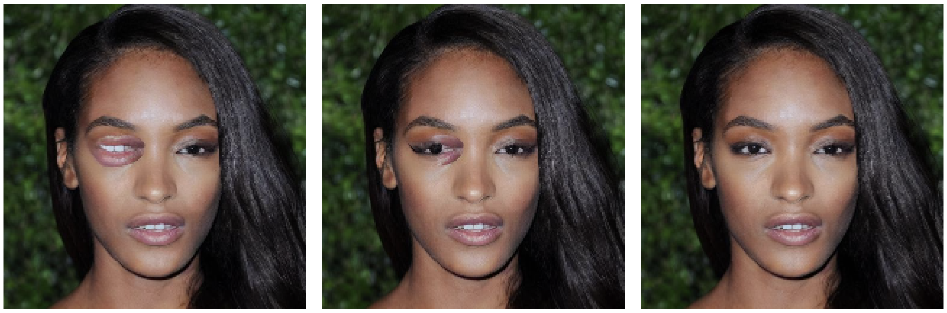

Figure 5: **Semantic shuffling qualitative results.** This shows the results with mask sizes corresponding to $0.2, 0.65, 1.0$ in the case of Stable Diffusion.

minishes). This produces two monotonic, opposing curves for each model.

The intersection point of these curves provides a quantitative boundary at which learned priors begin to dominate over local semantic evidence. As shown in Figure 6, this boundary occurs at a lower mask size for Stable Diffusion [8] (≈ 0.62) than for MAE-FAR [1] (≈ 0.80). This indicates that Stable Diffusion maintains its learned semantic priors more strongly and requires less removal of misleading local content to recover a semantically consistent reconstruction. In contrast, MAE-FAR [1] relies more heavily on local visual features and requires a larger masked region before overriding the inconsistent input.

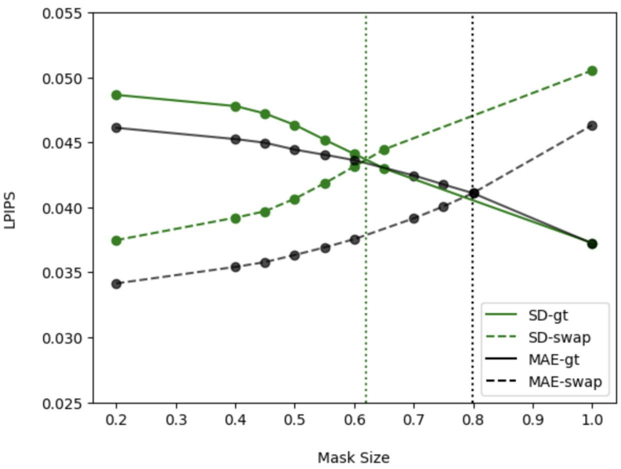

Figure 6: Semantic shuffling results. LPIPS is computed against the original ground truth and a semantically shuffled reference as the mask size over the inconsistent region increases. The intersection of the resulting curves defines a robustness boundary, occurring at 0.62 for Stable Diffusion and 0.799 for MAE-FAR.

5.5. Discussion

Deterministic and generative inpainting models exhibit different robustness behaviors as corruption severity and semantic complexity increase. While all models degrade under standard in-distribution perturbations, structural metrics alone can mask distributional failures.

Regarding out-of-distribution testing, the semantic shuffling experiment provides the clearest separation between models by explicitly testing how they resolve semantic inconsistency. By defining a robustness boundary based on the intersection of opposing LPIPS curves, we obtain a quantitative measure of the trade-off between learned priors and local evidence. The lower boundary observed for Stable Diffusion indicates stronger resistance to misleading local content and greater reliance on global semantic priors, whereas MAE-FAR requires substantially more masking before overriding inconsistent inputs. These findings highlight semantic robustness as a key axis along which diffusion-based models differ from classical reconstruction-based approaches.

6. Conclusion

In this work, we presented a systematic evaluation of out-of-distribution robustness in image inpainting models, with a particular focus on semantic consistency. Through controlled variations in mask size, blur, semantic masking, and semantic shuffling, we demonstrated that standard in-distribution evaluations are insufficient to fully characterize model robustness.

Our results show that diffusion-based models, especially those trained on diverse data, exhibit stronger robustness to semantic inconsistencies than deterministic reconstruction-based methods. By introducing a quantitative boundary measure based on perceptual similarity, we provide a principled way to assess the balance between learned priors and local visual evidence. These findings motivate the inclusion of semantically grounded OOD tests in inpainting benchmarks and evaluations.

References

[1] Chenjie Cao, Qiaole Dong, and Yanwei Fu. Learning prior feature and attention enhanced image inpainting, 2023.

[2] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[3] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions, 2018.

[4] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[5] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models, 2022.

[6] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. 2019.

[7] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting, 2016.

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.