A Framework for Detecting and Quantifying Hallucinations in Remote Sensing Image Inpainting

Konstantinos Zafeirakis and Grigorios Tsagkatakis

***Abstract*—Remote sensing images are often degraded due to sensor malfunctions or adverse atmospheric conditions and require inpainting to estimate missing observations. Despite the advancements, deep learning models for inpainting face multiple challenges including hallucinations, where the model incorrectly introduces non-existent elements in the image. This study introduces a novel framework for detecting hallucinations using an image inpainting generator coupled with a two-class discriminator and a class activation mapping (Grad-CAM) model. The experimental setup involves diverse masking techniques and analyzes the inpainting results across different image classes. Our findings reveal significant impacts of mask type and size on hallucination metrics, with rectangular masks generally yielding better results than irregular and random masks. Additionally, each class-specific generator exhibited unique inpainting behaviors, influenced by mask size. The study identifies the in-distribution Dice metric and out-of-distribution prediction value as effective measures for hallucination detection, with the FID metric proving optimal for reconstruction quality.**

***Index Terms*—Image Inpainting; Hallucination Detection; Generative Adversarial Networks; Explainable AI.**

I. Introduction

Remote sensing images are often compromised by sensor malfunctions, atmospheric conditions like cloud cover, or data transmission errors, resulting in missing or corrupted regions [1]. Image inpainting, the task of filling in these missing regions, is a critical preprocessing step to ensure data completeness for subsequent analysis [2]. In remote sensing, deep learning-based inpainting methods are the gold standard [3] due to their ability to handle the high resolution and geographical variations [4]. Generative Adversarial Networks (GANs) [5], [6], which learn the underlying distribution of natural images to generate highly realistic completions and transformer-based models like the Mask-Aware Transformer (MAT) [7], have pushed performance further, leveraging self-attention mechanisms to capture long-range dependencies and excel at inpainting large, irregular holes.

These models excel at producing plausible textures and structures, but their generative nature introduces a significant risk, *hallucinations* []. Hallucination occurs when a model generates content that is plausible in appearance but factually incorrect or inconsistent with the ground truth. In remote sensing, this could manifest as fabricating a building in a field or inpainting a river with forest texture, thereby corrupting the dataset for downstream applications. Despite the critical importance of data integrity in remote sensing, there is a lack of systematic methods to detect, localize, and quantify such hallucinations. Most evaluations of inpainting models focus on pixel-wise reconstruction error (e.g., MSE) or perceptual similarity (e.g., LPIPS, FID), which do not explicitly measure factual correctness.

This paper addresses this gap by proposing a dedicated framework for hallucination detection, conceptually illustrated in Fig. 1. The main contributions of this work include:

- A novel framework combining an inpainting generator, a two-class discriminator, and an explainability model to systematically detect and quantify hallucinations.
- A thorough investigation into how mask type (rectangular, random, irregular) and size influence the emergence and severity of hallucinations in remote sensing imagery.
- An extensive experimental evaluation that validates our proposed metrics—the out-of-distribution (OOD) prediction score and an in-distribution (ID) Dice score—as effective indicators of hallucinatory content.

The remainder of this paper is organized as follows: Section II reviews related work. Section III details our proposed framework. Section IV describes the experimental setup, and Section V presents and analyzes the results. Finally, Section VI concludes the paper.

II. Methodology

A. Image Inpainting Generator

The core of our system is a deep neural network acting as the generator, G, responsible for filling masked image regions. We employ the Mask-Aware Transformer (MAT) [7], a state-of-the-art transformer-based model for large-hole inpainting. Given a ground truth image I and a binary mask M (where 0 denotes missing pixels), the generator takes the masked image $I_M = I \odot M$ as input and produces a reconstructed image I_G. The MAT architecture consists of a convolutional head, a transformer body, a convolutional tail, a style manipulation module, and a Conv-U-Net for refinement.

The generator is trained on a single class of images, which we define as the in-distribution (ID) class. The training objective is to minimize a composite loss function that balances realism and perceptual quality:

$$L = L_G + \gamma R_1 + \lambda L_p \tag{1}$$

where L_G is the non-saturating adversarial loss, $R_1 = \mathbb{E}_x[\|\nabla D(x)\|]$ is an R1 gradient penalty for stabilizing training, and $L_p = \sum_i \eta_i \|\phi_i(x) - \phi_i(\hat{x})\|_1$ is the perceptual loss

K. Zafeirakis is with the Faculty of Science, University of Amsterdam, Amsterdam, The Netherlands (e-mail: konstantinos.zafeirakis@student.uva.nl).

G. Tsagkatakis is with the Institute of Computer Science, Foundation for Research and Technology - Hellas (FORTH), Heraklion, Greece & the Computer Science Department of the University of Crete (e-mail: greg@ics.forth.gr).

Manuscript received [Date of Submission]; revised [Date of Revision].

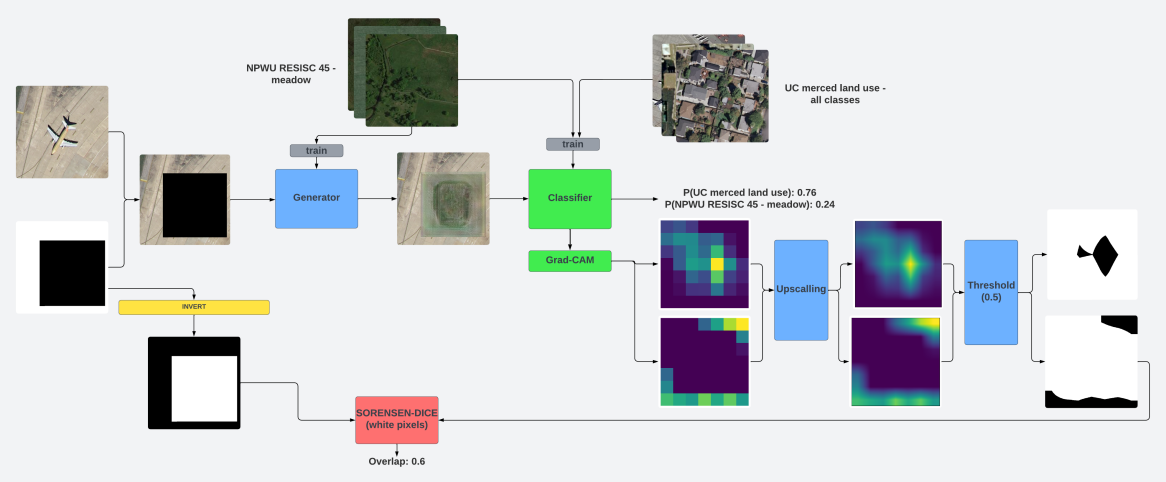

Fig. 1: Overview of the Proposed Hallucination Detection Framework. An out-of-distribution ground truth image (e.g., 'airport') is masked and fed to an inpainting generator trained on an in-distribution class (e.g., 'meadow'). The completed image is passed to a two-class discriminator, which outputs a prediction score (hallucination likelihood) and a Grad-CAM heatmap highlighting the regions responsible for the OOD classification.

calculated from layer activations ϕ_i of a pre-trained VGG-19 network. Following [7], we set $\gamma = 10$ and $\lambda = 0.1$.

B. Hallucination Discriminator

To detect hallucinations, we use a two-class supervised deep neural network, C, acting as the discriminator. Its task is to determine if an inpainted image I_G belongs to the generator's ID training class (class 0) or an out-of-distribution (OOD) class (class 1). We build the classifier via transfer learning using a MobileNetV2 [8] model pre-trained on ImageNet. The original classification head is replaced by a Global Average Pooling (GAP) 2D layer, a Dropout layer ($p = 0.5$), and a final Dense layer with two outputs.

The classifier is trained on two sets: images from the generator's ID class and images from all other classes in the dataset, which form the OOD set. This setup trains the classifier to be an expert at recognizing the generator's expected output style versus anything else.

The primary hallucination metric is the Out-of-Distribution Prediction (OODP) value, derived from the classifier's logits. The softmax function converts the classifier's raw logit outputs for the ID class (z_0) and OOD class (z_1) into probabilities. The OODP is the probability assigned to the OOD class:

$$\text{OODP} = \frac{e^{z_1}}{e^{z_0} + e^{z_1}} \tag{2}$$

A high OODP value indicates high confidence from the classifier that the inpainted image contains features inconsistent with the generator's training data, thus signaling a hallucination.

C. Class Activation-Based Detection

To localize hallucinations and derive a spatial metric, we use Gradient-weighted Class Activation Mapping (Grad-CAM) [9]. Grad-CAM leverages the gradients flowing into the final convolutional layer of the classifier to produce a heatmap highlighting the image regions most influential for a given class prediction. The neuron importance weights α_k^c for a class c and feature map k are computed by global average pooling the gradients of the class score y^c with respect to the feature map activations A^k:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3}$$

The final Grad-CAM heatmap is a weighted combination of the feature maps, passed through a ReLU function to isolate positive contributions:

$$L_{Grad-CAM}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right) \tag{4}$$

We generate a heatmap for the in-distribution (ID) class, which highlights regions the classifier recognizes as belonging to the generator's training domain. This heatmap is upscaled and binarized with a threshold of 0.5 to create a mask, ID_{THmap}. Our spatial metric measures the overlap between this identified ID region and the actual unmasked region of the image, using the Sørensen-Dice coefficient.

III. Experimental Setup

A. Datasets and Implementation

We use the **NWPU-RESISC45** dataset [10] for our primary experiments, with additional testing on the **UC Merced Land Use** dataset [11]. For each experiment, a MAT generator was trained on 300 images (512x512 pixels) from a single class (e.g., 'meadow', 'dense residential'). Generator training was monitored and stopped when the Fréchet Inception Distance (FID) score, evaluated on a validation set, converged. The MobileNetV2 discriminator was trained using images resized to 224x224, with the convolutional base frozen. All models were implemented using PyTorch.

To simulate varying data loss scenarios, we used the three mask types namely: (i) a single, randomly placed rectangular

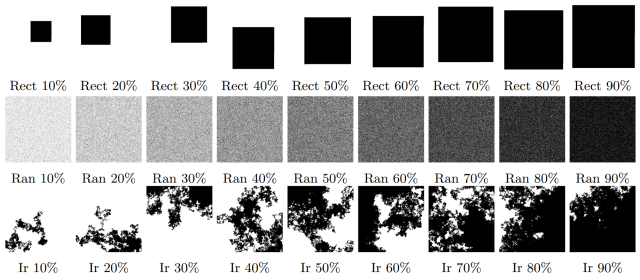

Fig. 2: Examples of the three mask types used in our experiments: Rectangular, Random, and Irregular, shown at increasing coverage percentages.

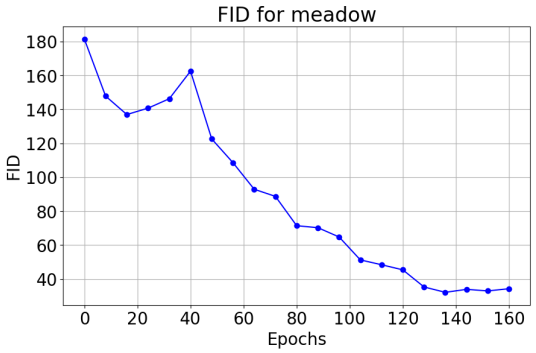

Fig. 3: FID score of the 'meadow' generator during training. The score converges, indicating successful model training.

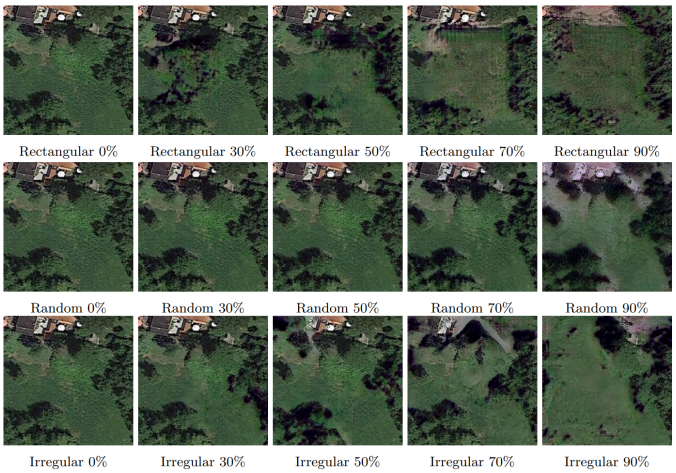

Fig. 4: Qualitative baseline results for 'meadow' inpainting. The inpainted images show the generator's performance under ideal, in-distribution conditions with different mask types and sizes.

patch **Rectangular**, (ii) Randomly sampled individual pixels **Random**, (iii) A contiguous mask grown from a random seed **Irregular**, illustrated in Fig. 2: For each type, we varied the masked area from 10% to 90% of the total image area.

In addition to our proposed hallucination metrics, we evaluate the generator's reconstruction quality using three standard perceptual metrics: **Mean Squared Error (MSE)**: Calculated between grayscale versions of the ground truth and the inpainted images. **Fréchet Inception Distance (FID)** [12]: Measures perceptual quality by comparing feature distributions. **Learned Perceptual Image Patch Similarity (LPIPS)** [13]: Quantifies perceptual similarity using deep features.

IV. Results and Analysis

We conducted a series of experiments to validate our framework. We first establish the baseline performance of our inpainting generator in an ideal, in-distribution (ID) scenario. We then analyze hallucination detection across scenarios of varying difficulty, defined by the visual similarity between the generator's training class and the out-of-distribution (OOD) target image class.

A. Baseline Inpainting Performance

Before evaluating hallucinations, it is essential to confirm that our inpainting generator is well-trained and capable of high-quality reconstruction under normal conditions.

Training Convergence: Fig. 3 shows the FID score of the 'meadow' generator during training. The score steadily decreases and converges after approximately 160 epochs, indicating that the model has learned a stable representation of the target class.

Reconstruction Quality: Table II presents the evaluation of MSE, FID, and LPIPS metrics for the 'meadow' generator on in-distribution 'meadow' test images. As expected, all reconstruction metrics degrade as the masked area increases (Table II). Notably, Random masks consistently yield the best reconstruction quality (lowest error), as they preserve distributed contextual guidance. Conversely, rectangular masks perform the worst, particularly regarding the FID metric, as they create a large, context-free void that forces the generator to hallucinate heavily. Fig. 4 visually confirms this: the random mask inpainting results in a slightly blurry but structurally coherent meadow, while the rectangular mask struggles to recreate fine details. These combined quantitative and qualitative results confirm the generator's competence and establish a robust performance benchmark.

B. Hallucination Detection: High Dissimilarity Scenario

Having established the generator's baseline, we begin our hallucination analysis with the most straightforward case: applying the 'meadow'-trained generator to the highly dissimilar 'airport' class.

Quantitative Analysis: The core hallucination metrics for this scenario (OOD Prediction Score and ID Dice Score) are detailed in Table I. Rectangular masks consistently yield the highest OOD prediction scores (Rct column under 'Airport'), increasing sharply with masked area to exceed 40% confidence at high coverage. This provides a strong, clear signal that the discriminator is confidently identifying the inpainted content as out-of-distribution. The ID Dice Score (Table I) confirms a stable increase for rectangular masks, indicating the discriminator correctly localizes the hallucinated region from the known ground truth context. In contrast, random and irregular masks produce a much weaker and less consistent

TABLE I: Hallucination Detection Metrics (OODP and ID Dice Score) for Meadow Generator

Masked %	OOD Prediction Value									ID Dice Score								
	Airport			River			Forest			Airport			River			Forest		
	Rnd	Rct	Irr	Rnd	Rct	Irr	Rnd	Rct	Irr	Rnd	Rct	Irr	Rnd	Rct	Irr	Rnd	Rct	Irr
0	.012	.012	.012	.008	.008	.010	.075	.075	.075	0	0	0	0	0	0	0	0	0
10	.010	.015	.012	.016	.038	.019	.090	.116	.075	.355	.315	.290	.512	.394	.562	.592	.392	.453
20	.005	.035	.008	.018	.021	.024	.044	.092	.066	.345	.435	.435	.506	.572	.694	.562	.414	.578
30	.008	.067	.010	.036	.044	.041	.080	.144	.112	.402	.422	.495	.526	.462	.608	.572	.436	.528
40	.015	.090	.010	.019	.089	.039	.048	.146	.102	.396	.545	.468	.472	.668	.548	.524	.532	.512
50	.028	.135	.030	.031	.145	.078	.082	.142	.076	.428	.548	.512	.536	.662	.646	.591	.466	.542
60	.028	.166	.028	.044	.165	.053	.062	.154	.070	.422	.635	.532	.568	.658	.592	.574	.484	.528
70	.050	.210	.035	.072	.372	.061	.034	.196	.070	.448	.642	.518	.602	.824	.590	.502	.492	.478
80	.140	.242	.075	.091	.542	.096	.160	.214	.066	.525	.705	.562	.674	.898	.636	.528	.542	.525
90	.195	.433	.108	.169	.758	.264	.246	.364	.088	.655	.772	.622	.706	.964	.774	.665	.762	.512

TABLE II: Reconstruction Quality Metrics (MSE, FID, LPIPS) for In-Distribution (Meadow) Inpainting

Masked Area (%)	MSE			FID			LPIPS		
	Rnd	Rct	Irr	Rnd	Rct	Irr	Rnd	Rct	Irr
10	3	6	5	22	43	35	.102	.045	.052
20	8	13	11	41	78	60	.194	.082	.095
30	9	20	18	53	127	93	.195	.142	.145
40	15	25	21	73	130	111	.262	.175	.185
50	16	35	29	91	175	155	.264	.224	.228
60	21	42	39	124	188	187	.328	.264	.265
70	33	49	45	141	238	205	.415	.315	.320
80	59	55	54	215	237	233	.558	.356	.372
90	67	64	68	245	250	262	.482	.394	.422

signal for both metrics, suggesting their resulting artifacts are more subtle.

The visual results in Fig. 5 confirm the quantitative findings. The inpainted images clearly show the generator filling the masked areas with amorphous green 'meadow' texture, which is visually jarring against the structured airport background. The hallucination is most blatant in the case of large rectangular masks, corresponding to the highest OODP scores.

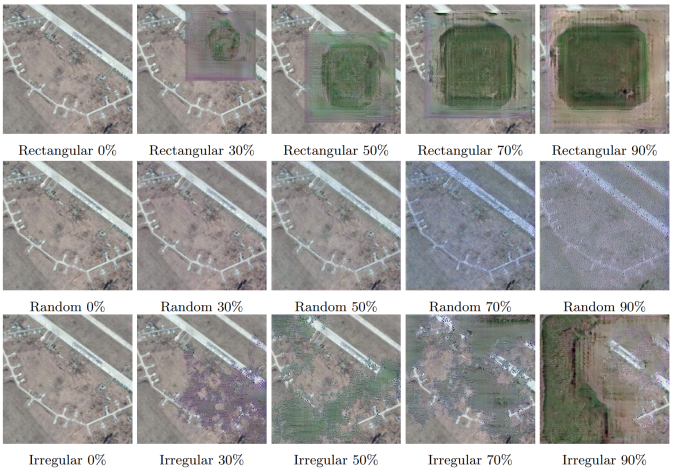

Fig. 5: Qualitative results for 'meadow' generator on 'airport' images. The generator fills the masked areas with out-of-context 'meadow' texture.

C. Hallucination Detection: Moderate and Low Dissimilarity

To test the framework's sensitivity, we analyze its performance on moderately dissimilar ('river') and highly similar ('forest') classes.

Quantitative Analysis: As shown in Table I, the OODP score for rectangular masks in the river scenario rises sharply, reaching 0.76 at 90% coverage, demonstrating the framework's ability to detect even subtle hallucinations. In the forest scenario, OODP scores remain low across mask types (max 0.364), underscoring the challenge of identifying near-distribution hallucinations. The ID Dice scores (Table I) likewise show robust spatial detection for rectangular masks in the river case, but reduced performance when hallucinated content closely resembles the ground truth (forest case).

Qualitative results for 'river' (Fig. 6) parallel the 'airport' case, though artifacts are subtler due to closer semantic similarity.

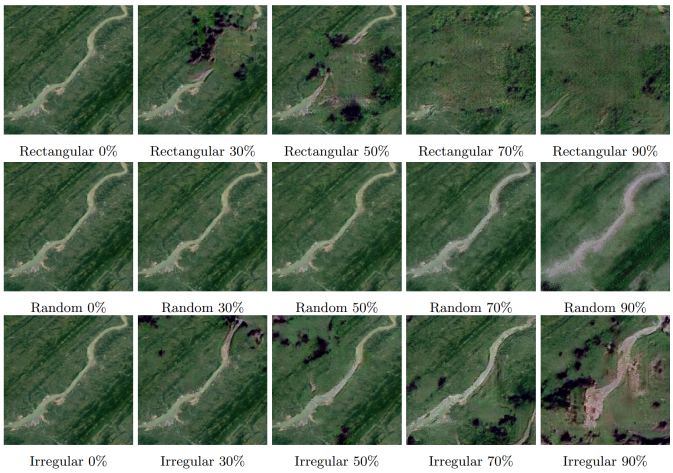

Fig. 6: Qualitative River

D. Analysis of Generator-Specific Behavior

To demonstrate that hallucinations are predictable, generator-specific artifacts, we analyze an alternative scenario using a 'dense residential' generator on an 'airplane' image. The qualitative progression is shown in Fig. 7. At a small mask size, the generator attempts a plausible reconstruction. As the mask grows, it begins to hallucinate building-like textures. By 90% coverage, it fabricates a full high-altitude cityscape, which is accurately localized by our framework. This behavior is entirely different from the 'meadow' generator's output, confirming that hallucinations are structured, class-conditional artifacts, not random noise.

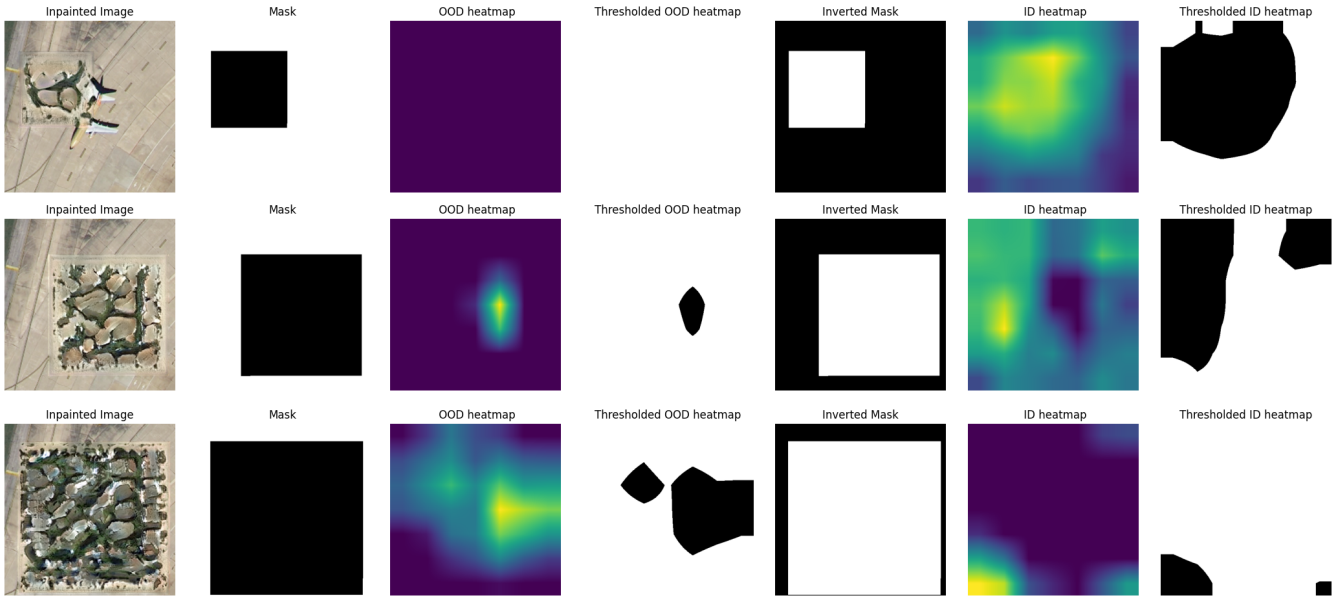

Fig. 7: Qualitative progression of hallucination from a 'dense residential' generator on an 'airplane' image. Hallucination severity and detection accuracy increase with mask size.

V. Discussion

Our results provide critical insights into the nature and detection of hallucinations in remote sensing inpainting. The primary implication is that the risk of data corruption is non-uniform: hallucination severity strongly correlates with the semantic distance between the generator's training class and the target content. The predictable, class-conditional nature of hallucinations (e.g., 'dense residential' generator consistently producing cityscapes) confirms they are a systematic failure mode rooted in the learned prior, not random noise. While this predictability makes detection feasible, the generated artifacts can be highly realistic, presenting a major vulnerability to automated analysis if left undetected. Our finding that rectangular masks induce the most severe and detectable hallucinations has direct practical consequences. Large, contiguous data loss, typical of cloud cover or sensor striping, represents the highest-risk scenario for generating factually incorrect content. Conversely, sparse data loss (random masks) is less likely to cause severe semantic hallucinations. Therefore, practitioners must exercise the most caution when applying inpainting models to repair large, continuous gaps in critical imagery. The framework faces limitations in near-distribution scenarios, as seen in the 'meadow' vs. 'forest' experiment, where the discriminator struggles to draw a clear semantic boundary.

Acknowledgment

This research has been supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "3rd Call for H.F.R.I.'s Research Projects to Support Faculty Members & Researchers" (H.F.R.I. Project Number: 26302).

References

[1] W. Huang, Y. Deng, S. Hui, and J. Wang, "Image inpainting with bilateral convolution," *Remote Sensing*, vol. 14, no. 23, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/23/6140

[2] Z. Qin, Z. Wang, C. Chen, Z. Wang, and Y.-G. Wang, "Image inpainting based on deep learning: A review," *Displays*, vol. 69, p. 102028, 2021.

[3] A. Pondaven, M. Bakler, D. Guo, H. Hashim, M. Ignatov, and H. Zhu, "Convolutional neural processes for inpainting satellite images," *arXiv preprint arXiv:2205.12407*, 2022.

[4] A. Kumar, D. Tamboli, S. Pande, and B. Banerjee, "Rsinet: Inpainting remotely sensed images using triple gan framework," *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 143–146, 2022.

[5] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2536–2544.

[6] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, jul 2017. [Online]. Available: https://doi.org/10.1145/3072959.3073659

[7] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[8] M. Z. A. Z. L.-C. C. Mark Sandler, Andrew Howard, "Mobilenetv2: Inverted Residuals and Linear Bottlenecks," 2019.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, p. 336–359, Oct. 2019. [Online]. Available: http://dx.doi.org/10.1007/s11263-019-01228-7

[10] Y. Liu, Y. Zhong, S. Shi, and L. Zhang, "Scale-aware deep reinforcement learning for high resolution remote sensing imagery classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 209, pp. 296–311, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271624000224

[11] V. R. Jakkula, "Tutorial on support vector machine (svm)," 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:15115403

[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," 2018.

[13] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," 2018.