Shadows Don't Lie and Lines Can't Bend! Generative Models don't know Projective Geometry...for now

Ayush Sarkar*[1] Hanlin Mai*[1] Amitabh Mahapatra*[1] Svetlana Lazebnik[1]
D.A. Forsyth[1] Anand Bhattad[2]
[1]University of Illinois Urbana-Champaign [2]Toyota Technological Institute at Chicago
https://projective-geometry.github.io/

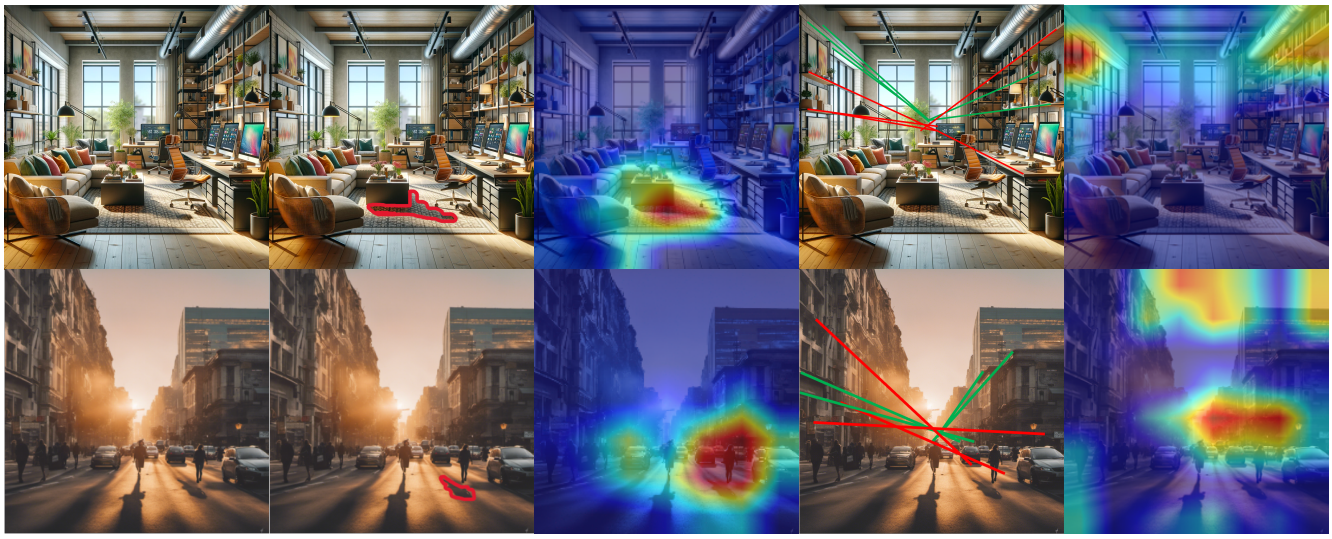


Figure 1. The first column presents visually compelling AI-generated images. However, a closer examination reveals fundamental inconsistencies, such as those in shadow alignment (second column) and vanishing point accuracy (fourth column). Our model's analysis, shown in the third and fifth columns, detects these shadow and perspective geometry errors. We show that these errors are systematic and can be used to identify generated images.

Abstract

Generative models can produce impressively realistic images. This paper demonstrates that generated images have geometric features different from those of real images. We build a set of collections of generated images, prequalified to fool simple, signal-based classifiers into believing they are real. We then show that prequalified generated images can be identified reliably by classifiers that only look at geometric properties. We use three such classifiers. All three classifiers are denied access to image pixels, and look only at derived geometric features. The first classifier looks at the perspective field of the image, the second looks at lines detected in the image, and the third looks at relations between detected objects and shadows. Our procedure detects generated images more reliably than SOTA local signal based detectors, for images from a number of distinct generators. Saliency maps suggest that the classifiers can identify geometric problems reliably. We conclude that current generators cannot reliably reproduce geometric properties of real images.

1. Introduction

Both StyleGAN [21–23] and diffusion models [35–37] are renowned for generating images that are strikingly similar to real-world photographs and consistently fool people. But, as we show, generated images have distinctive geometric features, likely from a failure to fully capture projective geometry.

Chen et al. [9], Zhan et al.[48], and Bhattad et al.[5] have shown generative models implicitly capture the complex scene properties, including normals, depth, albedo, and support relations. These works suggests these models "un-

*equal contribution

derstand" geometry, which would be useful for rendering 3D scenes. Our detailed, population-level analysis of generated images suggests generative models [1, 3, 4, 11, 31] cannot fully translate this "understanding" into accurate geometry. Specifically, we demonstrate that generative models produce images with lines that differ from those of real images (likely due to problems aligning vanishing points); that generative models produce images with perspective fields that are unlike those of real images; and that object-shadow relations in generated images differ reliably from those in real images. We use advanced pretrained models (Line Segment Detection [30]; Perspective Fields [20]; and PointNet [33]) that inspect geometric representations to distinguish between real and generated images.

To ensure our findings' integrity and accuracy, we adopt a rigorously designed data curation process. This critical step involves meticulously filtering out any biases related to color, texture, and local features within our test set. Such precision in data selection is crucial to isolate and accurately assess the subtle, yet significant, inconsistencies in projective geometry and illumination present in generated images. This careful approach ensures that our results are not obscured by common artifacts typically found in generated images, thereby enhancing the reliability of our conclusions. Our contributions are:

- **Unearthing Geometric Discrepancies:** We present a comprehensive analysis that goes beyond existing literature to both demonstrate and quantify geometric discrepancies produced by current generative models.
- **Data curation:** We offer a data curation process that can be used to hone in on generative model errors.
- **Broadening the Scope of Model Assessment:** Our approach offers an alternative method for evaluating models: do they get projective geometry right?

2. Related Work

Generative Models: The advancement of generative models, particularly in creating visually realistic images, marks a significant milestone in computer vision. Pioneering efforts by Karras et al. [21–23] with StyleGAN, and the emergence of diffusion models [35–37], have set new benchmarks in realism. These models, used in diverse fields from art to data augmentation, have yet to fully grasp the nuances of projective geometry, which is the focus of our analysis, primarily using open Stable Diffusion models.

Geometric Understanding in Generative Models: While studies like Chen et al. [9], Zhan et al. [48], and Bhattad et al. [5] demonstrate these models' potential in understanding scene geometry, our work diverges by scrutinizing the generated images themselves, examining their adherence to the principles of projective geometry and illumination, rather than analyzing learned features.

Detecting Generated Images: The realism of modern generative models has made image forensics increasingly challenging. Traditional methods focused on detecting synthetic images using signals like resampling artifacts [32] and JPEG quantization [2]. Kee et al. [24] introduced a geometric technique for detecting shadow inconsistencies, paralleling our pursuit of physical realism. However, our work extends beyond identifying photo manipulation to evaluating the overall perspective geometry and illumination consistency in images from generative models.

Zhang et al.'s work [12] focuses on detecting AI-generated images using diverse generative models and online training for future model adaptation. Our research, in contrast, assesses the projective geometry in these images, examining their ability to render scenes with accurate perspective and illumination. Boháček et al. [6], while detecting geometric inconsistencies related to shadows, align with our interest in physical realism. However, we delve deeper, thoroughly evaluating perspective geometry and illumination in generative models for a more comprehensive understanding of their geometric accuracy.

The rise of deep generative methods has steered image forensics towards using discriminative methods to detect synthetic content [7, 13, 18, 41, 42, 47, 49]. These advancements align with our objective of analyzing the physical and geometrical congruence of generated images. However, our work goes a step further by critically assessing whether generative models fundamentally understand and accurately replicate projective geometry, rather than simply distinguishing between real and synthetic images. This deeper level of analysis aims to unveil the intricacies and limitations of current models in faithfully rendering geometrically coherent images.

Evaluation Metrics: Traditional metrics like the Inception Score (IS) [38] and Fréchet Inception Distance (FID) [17] focus on pixel-level fidelity. The emergence of CLIP-based scores [16, 34] and DIRE [44] offers a semantic perspective. In contrast, our approach, distinct in its focus on perspective geometry and illumination consistency, seeks to ensure comprehensive realism, bridging the gap between visual and physical authenticity.

Recent studies like Davidsonian Scene Graph [10] and ImagenHub [25] address fine-grained evaluation inconsistencies, while the HEIM benchmark [26] assesses models across multiple aspects. Our work complements these by providing an in-depth evaluation of the physical and geometric realism of images generated by state-of-the-art models.

3. Background on Projective Geometry

Projective geometry is a mathematical framework that enables the accurate representation of three-dimensional spaces in two-dimensional images. It provides the rules for perspective, which are crucial for creating realistic scenes with depth and spatial orientation [14]. In this section, we will examine

the common inconsistencies that may arise during image synthesis according to projective geometry. Our evaluation framework is intended to detect and measure these discrepancies, which are essential for evaluating the realism and physical plausibility of generated images.

Inconsistent Vanishing Points. Vanishing points are fundamental to capturing the essence of perspective in images. They should align with the direction of parallel lines converging at a distance. Generated images often exhibit inconsistencies where these lines do not meet at the correct vanishing points, leading to a distorted sense of perspective.

Lighting and Shadow Inconsistencies. Accurate shadows are essential for reinforcing the position and shape of objects within a scene. Discrepancies in shadow direction, length, and softness can indicate a misalignment with the scene's light sources, disrupting the image's three-dimensionality.

Scale Discrepancies. The principle of size constancy dictates that objects of the same size should appear smaller as their distance from the observer increases. Generated images sometimes fail to maintain this scaling, resulting in a compromised depth perception.

Distortion of Geometric Figures. Geometric figures should maintain their shape when projected onto the image plane, barring intentional perspective distortion. Errors in this projection can result in circles appearing as ellipses or squares as trapezoids, indicating a flawed perspective rendering.

Depth Cues. Depth perception in images is conveyed through cues such as overlapping, texture gradients, and relative size. Misrepresentation of these cues can lead to an unnatural spatial arrangement that the human eye can readily detect as artificial.

Our evaluation framework, detailed in the subsequent sections, is designed to rigorously test generated images against these projective geometry principles. While a comprehensive evaluation of projective geometry would consider all the aforementioned inconsistencies, our framework prioritizes the detection of inconsistent vanishing points and lighting and shadow inconsistencies. These elements are particularly telling indicators of an image's projective geometry realism and are often the most challenging for generative models to replicate accurately.

4. Dataset Curation via Multi-Stage Filtering

Our data curation process begins with the selection of real images, subsequently captioned using a recent state-of-the-art (SOTA) method, where we utilize the ViT-bigG-14/laion2b_s39b_b160k model [19] along with the BLIP-2 [27] model. These captioned images are then processed through the Stable Diffusion-XL model [31]. This initial phase establishes a robust dataset where real and generated images are aligned based on common captions. Such alignment is crucial for a thorough and equitable evaluation of projective geometry nuances in generated images, ensuring

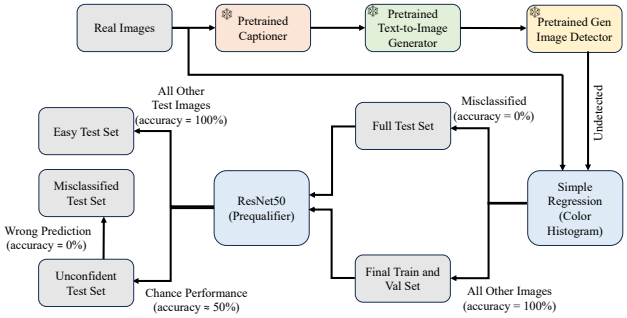


Figure 2. The diagram illustrates the sequential flow of our data curation methodology. Initially, images undergo captioning [27] [19], followed by processing through Stable Diffusion-XL [31]. Post-generation, these images are subjected to CNN Detection [42] and then filtered using a Histogram-Based Logistic Regression model. Subsequently, a ResNet50 [15] Prequalifier refines the selection, resulting in the final curated dataset.

that our dataset is ideally suited for examining the subtle aspects of image generation.

Transitioning from this initial phase, the integrity of our evaluation framework becomes paramount. To guarantee its robustness, we employ a multi-stage data curation process. This meticulous procedure is tailored to eliminate biases that may arise from small signal artifacts, color distribution, or textural features—elements often exploited by conventional CNN detectors. Systematically filtering out images based on these criteria allows us to isolate those that exhibit subtle yet essential inconsistencies in projective geometry and illumination. This rigorous approach ensures that our evaluation is not skewed by superficial artifacts, but rather, it becomes a true test of a generative model's grasp of complex physical and geometric principles. The following sub-sections detail each stage of this multi-stage filtering process and our data curation process is summarized in Figure 2.

4.1. Off-the-shelf CNNDetector Validation

Our process begins with the generation of scenes using the Stable Diffusion XL model, each scene crafted from real images paired with descriptive captions. These generated images are subjected to the CNNDetector [42], a CNN-based state-of-the-art detector. Surprisingly, the CNNDetector could not distinguish any of the generated scenes from Stable Diffusion-XL, indicating a level of visual fidelity that surpasses conventional detection capabilities.

4.2. Filtering Using Color Histograms

The next step in refining our dataset involves a logistic regression model that leverages color histogram statistics to differentiate between real and synthetic images. This model analyzes the color distribution within each image and assigns a probability of it being generated based on learned patterns in these distributions.

Table 1. Statistical overview of the Data Curation and Filtering Process: We present the distribution of real and generated images for indoor and outdoor datasets through consecutive stages of our data curation pipeline. Starting with large sets, we apply CNN detection and LR histogram filtering to refine the datasets, significantly reducing the number of generated images and highlighting the effectiveness of these preliminary filters. The ResNet50 Prequalifier further narrows down the datasets, creating an 'Unconfident Set' for images with low classifier certainty and a 'Misclassified Subset' for images incorrectly labeled by the classifier. These rigorously curated sets are instrumental for our subsequent analysis, concentrating on projective geometry while mitigating the influence of signal cues. The datasets for each prequalifier—indoor, outdoor, and combined—are prepared separately to tailor the models to their specific contexts.

	Indoor Real	**Indoor Generated**	**Outdoor Real**	**Outdoor Generated**	**Combined Real**	**Combined Generated**
Total Images	400,000	400,000	125,000	125,000	525,000	525,000
Post CNN Detection & LR Histogram Filtering						
Remaining Images	53,974	53,974	8,316	8,316	62,290	62,290
Training and Test Sets						
Training Set Size	280,000	280,000	95,000	95,000	375,000	375,000
Validation Set Size	48,000	48,000	17,000	17,000	65,000	65,000
Test Set Size	53,974	53,974	8,316	8,316	62,290	62,290
Post ResNet50 Prequalifier						
Unconfident Set	10,078	19,588	6,399	5,249	10,511	16,739
Misclassified Subset	3,366	10,928	3,420	1,910	3,535	8,132

This stage effectively filters out approximately 90% of the dataset, highlighting the predictive power of color histograms in identifying synthetic content that a CNN detector was otherwise unable to detect. We pool the misclassified set at this stage as our full test set and all other detected images are pooled into our training and validation test set.

4.3. Texture Consistency Examination

As we progress through our data curation process, we integrate a ResNet50 [15] classifier to serve as a prequalifier, drawing on methods established in CNNDetection [42] and the recent online detection of AI-generated images [13]. This prequalifier, though analyzing images in their entirety, demonstrates a fine-tuned sensitivity to both local distortions and textural inconsistencies—attributes attributable to the architectural depth and sophistication of ResNet50.

The ResNet50 classifier is shown to be good at distinguishing the intricate textural features that differentiate real images from generated ones, capitalizing on the local and global discrepancies introduced by generative processes. It evaluates the overall texture consistency and coherency of the image, offering a holistic yet detail-oriented perspective.

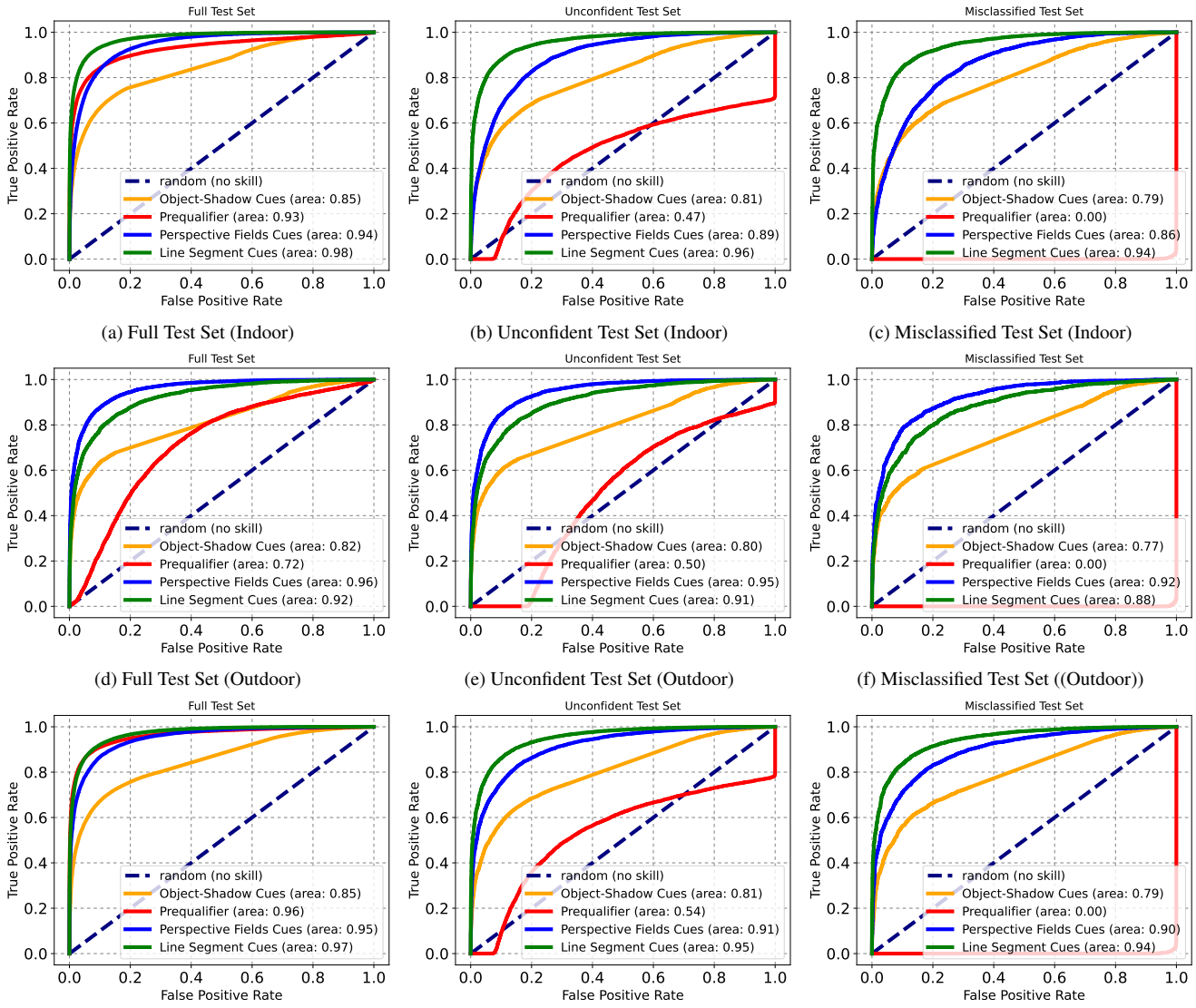
5. Analyzing Projective Geometry

After curating the dataset, we meticulously analyze the unfiltered images, consisting of 10,928 indoor and 1,910 outdoor scenes. These images have successfully passed prior filtering stages; that is they are detected as real images. We then subject them to a thorough evaluation for geometric and shadow inconsistencies, focusing on their conformity to projective geometry principles. This process ensures that our dataset tests the models rigorously, challenging them to replicate not only surface details (color or texture inconsistencies) but also the underlying geometric correctness and photometric accuracy of generated images.

It must be noted that projective geometry inconsistencies are prevalent in most of the generated images, yet they often escape detection by conventional analysis. Our approach focuses on the "hard set"—challenging scenarios where the prequalifying classifier, trained directly on images, either operates at the chance (the **unconfident test set**) or inversely misclassifies real and generated images (the **misclassified test set**). This distinction is critical, as it allows us to rigorously test our models, which, unlike the prequalifier, do not have direct access to the images. Trained solely on geometric abstractions and projective cues extracted from the images, our models are competent at detecting subtle but decisive inaccuracies that simple texture artifacts cannot account for. This level of abstraction in training guarantees that our models focus on the fundamental aspects of projective geometry, identifying errors that could significantly impact the practical applicability of generated visuals in real-world contexts. Below, we detail different models that comprise our approach.

5.1. Line Segment Cues

Our method for assessing the projective geometry in generated images starts by identifying key structural lines within each image using a Deep Learning-based Line Segment Detector (Deep LSD) [30]. These lines are crucial for our analysis as they indicate how well the generated images adhere to the rules of perspective. To classify the images based on these line segments, we train a PointNet-like architecture [33] known for its ability to handle unordered data sets like the ones we encounter with line segments. This model is trained to recognize patterns that help differentiate real images from generated ones by understanding the arrangement



(a) Full Test Set (Indoor) (b) Unconfident Test Set (Indoor) (c) Misclassified Test Set (Indoor)

(d) Full Test Set (Outdoor) (e) Unconfident Test Set (Outdoor) (f) Misclassified Test Set ((Outdoor))

(g) Full Test Set (Indoor + Outdoor) (h) Unconfident Test Set (Indoor + Outdoor) (i) Misclassified Test Set (Indoor + Outdoor)

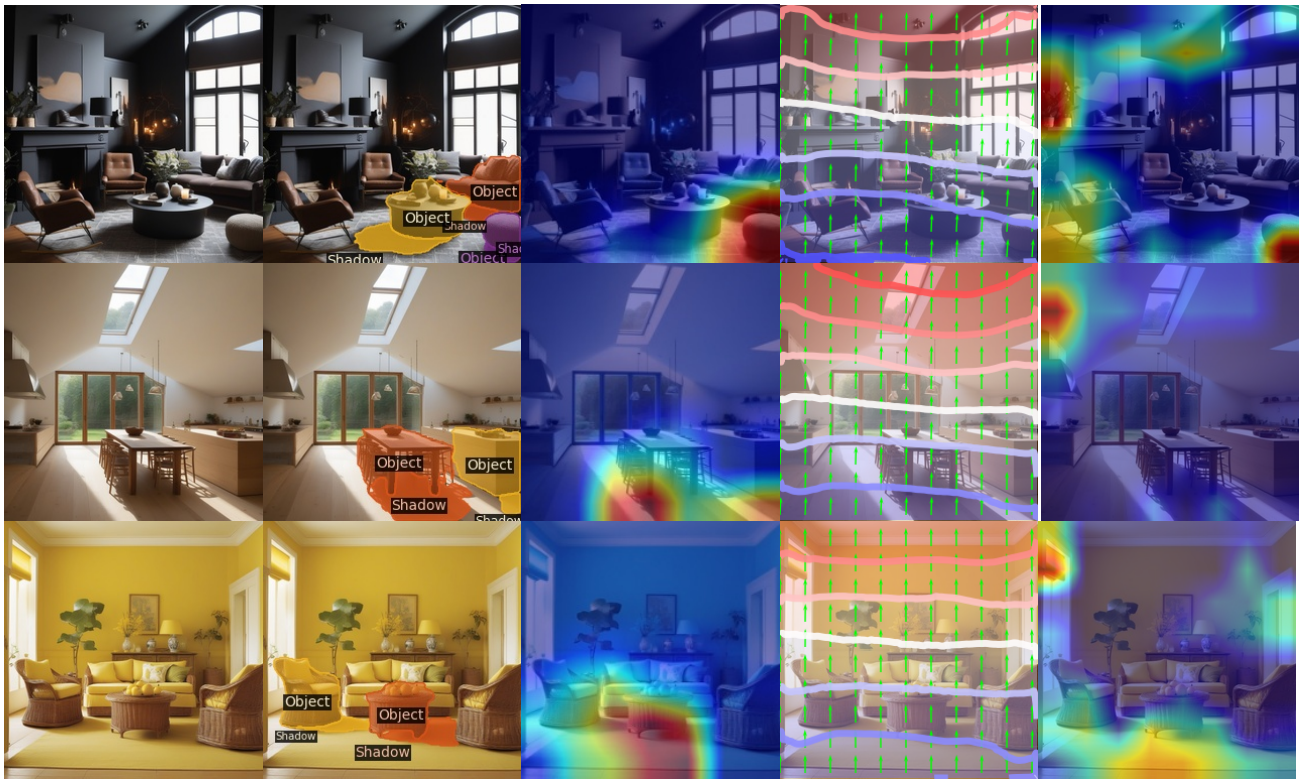
Figure 3. ROC Curves Assessing Projective Geometry Cues in Generated Images by Stable Diffusion XL. We trained separate models for indoor scenes, outdoor scenes, and a combination of indoor and outdoor scenes. In the full test set (a), (d), (g), our line segment classifier was found to be more accurate (with AUCs of 0.98, 0.92, and 0.97) compared to the prequalifier (with AUCs of 0.93, 0.72, and 0.96). Prequalification helps to determine whether the methods are using signal cues or not. Our results show that they are not, as demonstrated by (b), (c), (e), (f), (h), and (i). For the unconfident test set, where the prequalifier has an AUC of 0.47 (b), 0.50 (e), and 0.54 (h) for indoor, outdoor, and combined partition, our classifiers can still accurately identify the generated images with high AUCs. Similarly, for the misclassified test set, where the prequalifier has an AUC of 0.00, our classifiers remain reliable. We conclude that generated images contain geometric structures not seen in real images, and these structures very reliably identify generated images.

and consistency of these lines.

Unlike traditional models that may require the data to be in a specific format, PointNet is flexible and considers each line segment without the need for pre-sorting, making it particularly suited for our geometric analysis. It assigns a score representing the likelihood that an image is real based on the spatial arrangement of its lines. By analyzing the scores from PointNet, we can determine the model's proficiency in detecting subtle discrepancies in line arrangements that often indicate a generated image.

5.2. Perspective Field Cues

Our framework's second model utilizes Perspective Fields [20], vector fields that encode the spatial orientation of pixels in relation to vanishing points and the horizon. These dense fields could be instrumental in assessing the



Generated Image | Object-Shadow (OS) | OS GradCam | Perspective Fields | Perspective Fields GradCam

Figure 4. Grad-CAM can be applied to our Object-Shadow and Perspective Field classifiers. Doing so suggests that the high AUCs of Figure 3 are vested in real geometric errors. Here we show indoor scenes. The first column displays images generated by Stable Diffusion-XL. The second column overlays detected object-shadow pairs from [43], highlighting the model's ability to identify these features. The third column applies Grad-CAM to our Object-Shadow classifier. This shows areas most diagnostic of synthetic generation. Note: in the first row, the shadow cast by the ottoman is in the wrong direction and Grad-CAM identifies this error as diagnostic for our classifier; in the second row, the Grad-CAM weights suggest a shadow problem at the left side chair, which is difficult to check but plausible; in the third row, the shadow cast by the coffee table is in the wrong direction and Grad-CAM identifies this error as diagnostic. The fourth column shows the Perspective Fields of [20], and the fifth column shows Grad-CAM when applied to our Perspective Fields classifier. Note: in the first row, Grad-CAM weights identify an oddly oriented line in the top left corner; in the second row, Grad-CAM weights identify a problem with the top of the cupboard on the left, which is difficult to confirm but plausible; in the third row, Grad-CAM weights identify a visible problem with the blind on the left. Best viewed on screen.

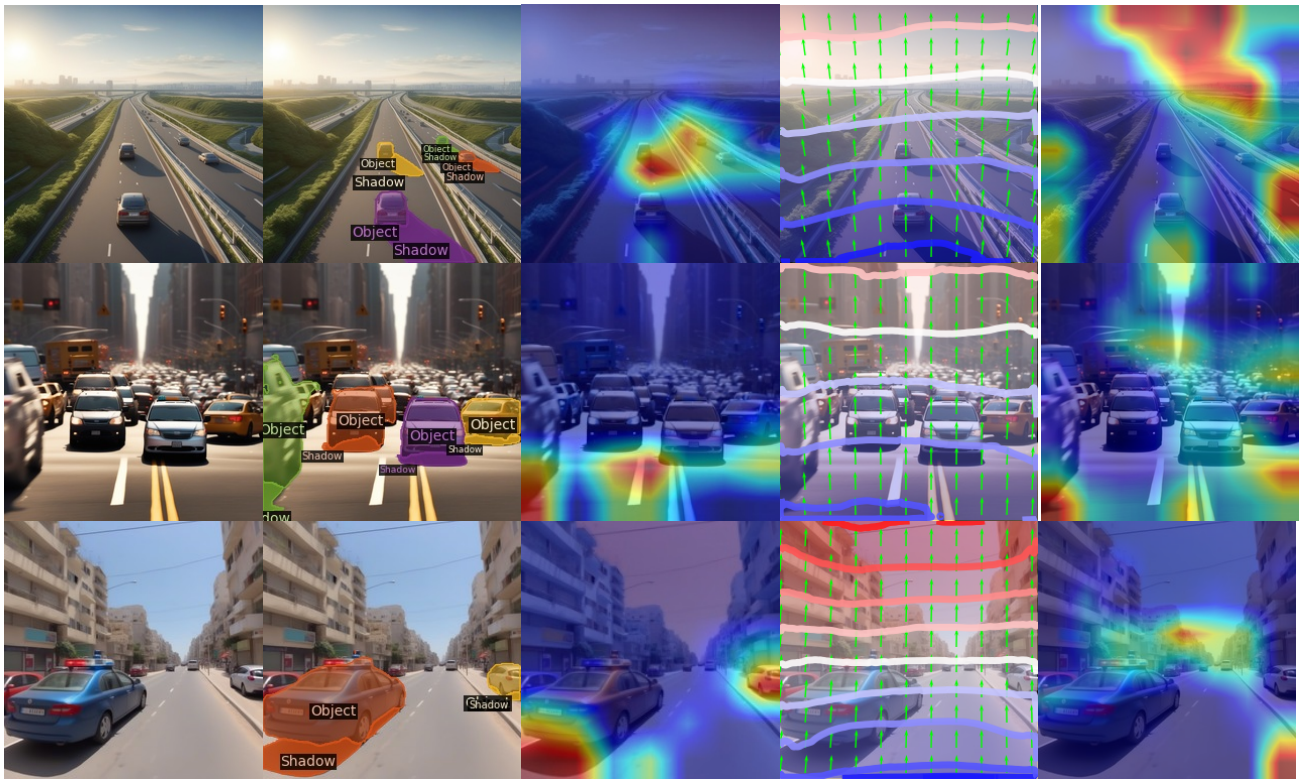
projective geometry of images. We use a pretrained model to generate Perspective Fields from single images, which serve as a basis for understanding the scene's geometric structure.

We then train a ResNet50 classifier on these fields to differentiate between real and generated images, focusing on anomalies in projective geometry. The classifier evaluates the consistency of Perspective Fields with projective geometry principles, scoring images on their geometric plausibility. This method allows for a precise and focused evaluation of projective geometry in generated images, enhancing the detection of subtle inconsistencies.

5.3. Object-Shadow Cues

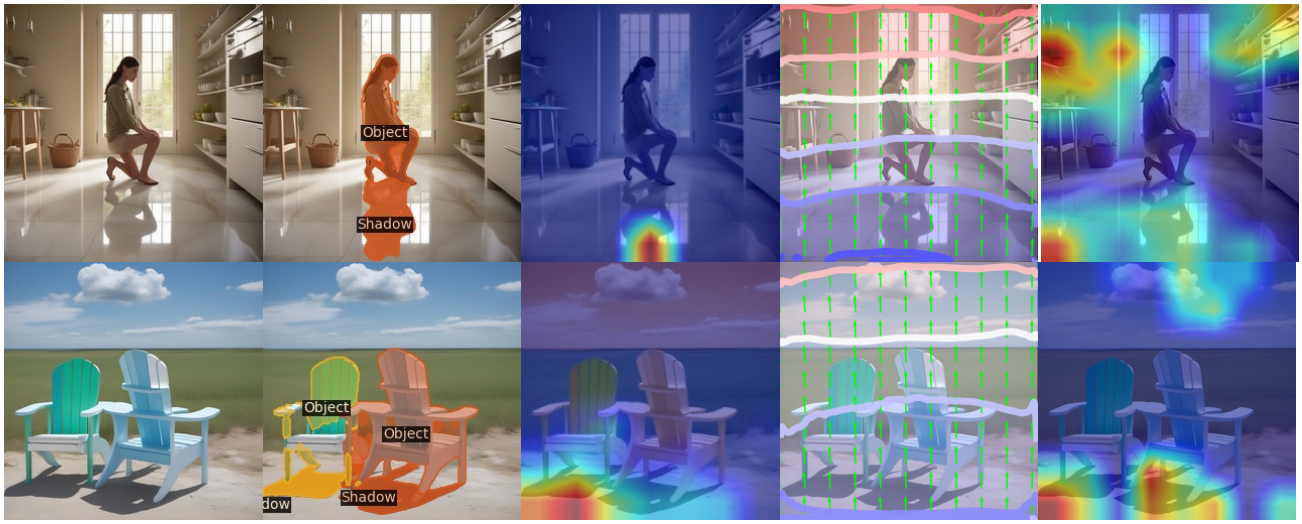
The third model in our framework addresses the illumination aspect by examining object-shadow relationships. Shadows are inherently tied to the shapes that cast them, following the principles of projective geometry that dictate how three-dimensional forms are translated onto a two-dimensional plane. The direction, length, and shape of a shadow should be consistent with the light source's position and the geometry of the casting object. Any inconsistency in this alignment can reveal the synthetic nature of an image.

To detect such inconsistencies, we analyze the shadows in relation to their corresponding objects and the presumed light source direction. We employ an object-shadow instance detection algorithm [43] to identify shadows and then use geometric heuristics to evaluate their plausibility given the objects and their orientation in the scene. This is accomplished by training a ResNet50 classifier on binary masks of object and shadow instances. The consistency of these shadows with the objects is scored, and images with implausible shadows are marked as likely generated.



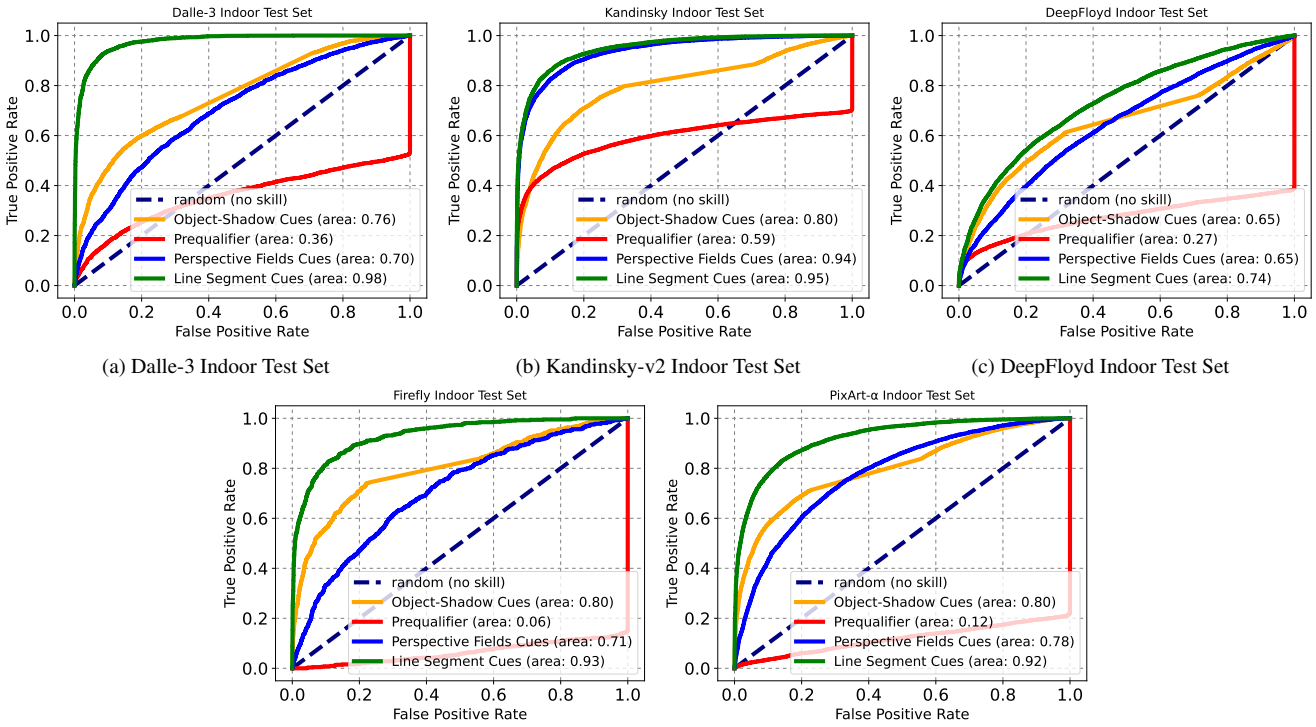
Generated Image | Object-Shadow (OS) | OS GradCam | Perspective Fields | Perspective Fields GradCam

Figure 5. Grad-CAM results for outdoor scenes, after the model of Figure 4. The first column displays images generated by Stable Diffusion-XL. Note: in the first row, the cars cast shadows in different directions and Grad-CAM identifies this error as diagnostic; in the second row, two cars in front cast shadows in different directions and Grad-CAM identifies this error as diagnostic; in the third row, Grad-CAM identifies the (very odd) structure of the buildings near the vanishing point as a problem, based on perspective field distortion.



Generated Image | Object-Shadow (OS) | OS GradCam | Perspective Fields | Perspective Fields GradCam

Figure 6. Our projective geometry classifiers identify distinct types of problems. The top row shows an example that was classified real by the Object-Shadow classifier, but correctly identified as generated by the Perspective Fields classifier. The shadow cast by the person appears realistic; but, as the Perspective Fields GradCAM identifies, the shelf on the top left has problematic geometry. The bottom row shows an example that was correctly identified as generated by the Object-Shadow classifier but was classified as real by the Perspective Fields classifier. Here the perspective effects in the image appear inoffensive, but the two chairs are casting shadows from different light sources, as the Grad-CAM weights correctly show.



(a) Dalle-3 Indoor Test Set (b) Kandinsky-v2 Indoor Test Set (c) DeepFloyd Indoor Test Set

(d) Adobe Firefly Indoor Test Set (e) PixArt-α Indoor Test Set

Figure 7. Evaluating the effectiveness of our classifiers in distinguishing projective geometry cues in indoor images generated by various models. We use our classifiers trained on the Stable Diffusion XL dataset and evaluate their performances on test sets generated by Dall-E 3 (a), Kandinsky-v2 (b), and DeepFloyd (c), using the same text prompts from the "unconfident" Stable Diffusion XL generated test set as described in Table 1. The Prequalifier's AUC scores were notably lower across all test sets, registering AUCs of 0.36 for Dall-E 3, 0.59 for Kandinsky-v2, and 0.27 for DeepFloyd. In the Dall-E 3 Test Set, Line Segment Cues classifier showed the highest accuracy with an AUC of 0.98, while Object-Shadow Cues and Perspective Fields Cues also showed decent accuracy with AUCs of 0.76 and 0.70, respectively. In the Kandinsky-v2 Test Set, Line Segment Cues and Perspective Fields Cues demonstrated robust detection with AUCs of 0.95 and 0.94, while Object-Shadow cues were detected with an AUC of 0.80. The DeepFloyd Test Set seems to have smaller geometric distortion, with Line Segment Cues, Object-Shadow Cues, and Perspective Fields Cues achieving AUCs of 0.74, 0.65, and 0.65, respectively, outperforming the Prequalifier (AUC of 0.27). In the Adobe Firefly test set, the Line Segment Cues classifier demonstrates the highest discrimination ability with an AUC of 0.93, closely followed by Object-Shadow Cues with an AUC of 0.80, while the Prequalifier lags behind with an AUC of 0.06. Similarly, in the PixArt-α Indoor Test Set, Line Segment Cues lead with an AUC of 0.92, indicating robust performance across different generative models. Based on these findings, we can conclude that current generative models exhibit a fundamental gap in replicating projective geometry, and our derived geometry cues can reliably distinguish between real and synthetically generated images.

6. Evaluation

6.1. Dataset

Our evaluation consists of a diverse set of images, including:

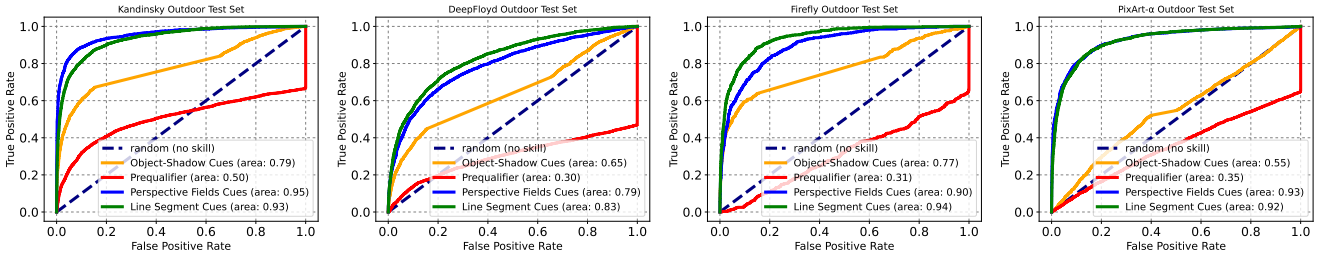
Indoor Scenes: A collection of interior images featuring a variety of furniture arrangements and lighting conditions. These were sourced from LSUN (specifically Bedroom, Dining Room, Kitchen, and Living Room) [45].

Outdoor Scenes: A dataset of outdoor environments with varying landscapes and urban settings. These were sourced from Berkeley Deep Drive 100K [46] and Mapillary Vistas[29].

6.2. Classifiers Results

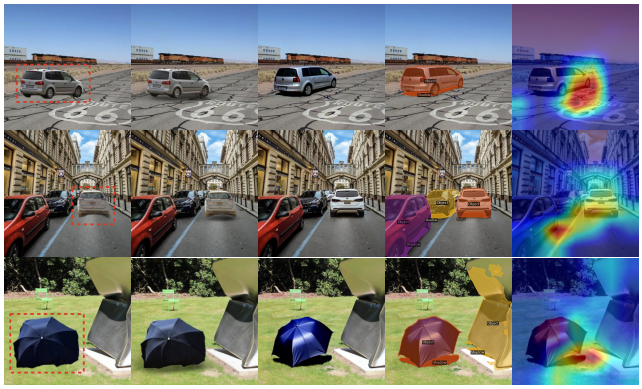
Figure 3 shows ROC curves for each method on indoor, outdoor and combined (indoor+outdoor) scenes. In each case, classifiers are trained on images that are *not* prequalified and tested on prequalified scenes, meaning that performance estimates are biased *low* — likely training on prequalified data would lead to even more accurate classification. Each classifier is effective, with AUCs ranging from 0.72 to 0.97. Recall these classifiers see *only* derived geometric features and do not see the image itself.

Qualitative examples using Grad-CAM [39] appear in Figures 4 and 5. Notice how images that might be acceptable to a line analysis often fail a shadow analysis. Figure 6 shows examples to emphasize this point.



(a) Kandinsky-v2 Outdoor Test Set (b) DeepFloyd Outdoor Test Set (c) Adobe Firefly Outdoor Test Set (d) PixArt-α Outdoor Test Set

Figure 8. ROC Curves for Outdoor Generated Scenes: The performance of our classifiers on outdoor test sets from Kandinsky-v2 (a), DeepFloyd (b), Adobe Firefly (c), and PixArt-α (d). Across the datasets, our derived geometry cues consistently outperform the Prequalifier, showcasing their robustness in distinguishing generated images with high accuracy (AUC > 0.90). Based on these findings, we can conclude that current generative models exhibit a fundamental gap in replicating projective geometry, and our derived geometry cues can reliably distinguish between real and synthetically generated images.



Copy-Paste | Traditional Composites | ObjectStitch Composites | Object-Shadow (OS) | OS GradCam

Figure 9. Detecting Composite Errors with Object-Shadow (OS) Cues. We show images directly taken from Figure 1 (teaser) of [40]. Our OS cues effectively identify composite images, such as those created by the recent SOTA insertion method [40], by pinpointing inaccuracies in shadow orientation. The bottom row provides a clear example where, despite the sun being positioned behind the camera, the shadows are mistakenly cast to the right. Also see the shadow of the adjacent object (marked in yellow), which is pointing upward in the opposite direction. Similarly, in the top row, shadows are cast in an implausible direction. The OS GradCam visualizations on the right successfully highlight these misdirected shadows.

7. Other Generators Evaluated

Our investigation primarily leverages the widely utilized and open-sourced Stable Diffusion XL (SDXL) model [31] as the training data for our classifiers. Our classifiers do not see pixels, but derived geometric features. This means that one could expect a form of generalization across generators. We illustrate that this generalization occurs - ROC curves in Figure 7 demonstrate that classifiers trained to distinguish Stable Diffusion XL images from real images can also reliably distinguish Kandinsky-v2 [3], DeepFloyd [11] and PixArt-α[8] from the open-source domain. Additionally, we assess the efficacy of our models against images from proprietary generators such as OpenAI's Dalle-3[4] and Adobe's Firefly [1], representing some of the most advanced tools in image generation. Finally, we show we can detect composite made by a recent SOTA method [28] by looking at Object-Shadow cues in Figure 9.

8. Discussion

We have shown that generated images can be reliably distinguished from real images by looking only at derived geometric cues. This is likely because image generators do not fully implement the geometry one observes in real images. Producing accurate perspective geometry or accurate shadow geometry requires very tight coordination of detailed information over very long spatial scales. Our results, together with the notorious tendency of face image generators to award subjects' left and right earlobes of different shapes, suggest that doing so is beyond the capacity of current generators. We speculate that fixing this difficulty requires structural innovation in the generator, rather than simply exposing the generator to more data.

References

[1] Adobe. Firefly. https://www.adobe.com/sensei/generative-ai/firefly.html, 2023. Accessed: 2023-11. 2, 9

[2] Shruti Agarwal and Hany Farid. Photo forensics from jpeg dimples. In *2017 IEEE workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2017. 2

[3] AI Forever. Kandinsky-2. https://github.com/ai-forever/Kandinsky-2, 2023. Accessed: 2023-11. 2, 9

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions, 2023. 2, 9

[5] Anand Bhattad, Daniel McKee, Derek Hoiem, and DA Forsyth. Stylegan knows normal, depth, albedo, and more. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2

[6] Matyáš Boháček and Hany Farid. A geometric and photometric exploration of gan and diffusion synthesized faces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 874–883, 2023. 2

[7] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *ECCV*, 2020. 2

[8] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 9

[9] Yida Chen, Fernanda Viégas, and Martin Wattenberg. Beyond surface statistics: Scene representations in a latent diffusion model. *arXiv preprint arXiv:2306.05720*, 2023. 1, 2

[10] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023. 2

[11] Deep Floyd. Iterative filter. https://github.com/deep-floyd/IF, 2023. Accessed: 2023-11. 2, 9

[12] David C. Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. Online detection of ai-generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 382–392, 2023. 2

[13] David C Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. Online detection of ai-generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 382–392, 2023. 2, 4

[14] David A Forsyth and Jean Ponce. *Computer vision: a modern approach*. prentice hall professional technical reference, 2002. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 4

[16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 2

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2

[18] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *ECCV*, 2018. 2

[19] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 3

[20] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F Fouhey. Perspective fields for single image camera calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17307–17316, 2023. 2, 5, 6

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[23] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2

[24] Eric Kee, James F O'Brien, and Hany Farid. Exposing photo manipulation with inconsistent shadows. *ACM Transactions on Graphics (ToG)*, 32(3):1–12, 2013. 2

[25] Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. Imagenhub: Standardizing the evaluation of conditional image generation models. *arXiv preprint arXiv:2310.01596*, 2023. 2

[26] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon, and Percy Liang. Holistic evaluation of text-to-image models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 3

[28] Oscar Michel, Anand Bhattad, Eli VanderBilt, Ranjay Krishna, Aniruddha Kembhavi, and Tanmay Gupta. Object 3dit: Language-guided 3d-aware image editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 9

[29] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5000–5009, 2017. 8

[30] Rémi Pautrat, Daniel Barath, Viktor Larsson, Martin R. Oswald, and Marc Pollefeys. Deeplsd: Line segment detection and refinement with deep image gradients. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 4

[31] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach.

Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3, 9

[32] Alin C Popescu and Hany Farid. Exposing digital forgeries by detecting traces of resampling. *IEEE Transactions on signal processing*, 53(2):758–767, 2005. 2

[33] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2, 4

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 2

[38] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. 2

[39] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2019. 8

[40] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18310–18319, 2023. 9

[41] Sheng-Yu Wang, Oliver Wang, Andrew Owens, Richard Zhang, and Alexei A Efros. Detecting photoshopped faces by scripting photoshop. In *ICCV*, 2019. 2

[42] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020. 2, 3, 4

[43] Tianyu Wang, Xiaowei Hu, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection with a single-stage detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3259–3273, 2022. 6

[44] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22445–22455, 2023. 2

[45] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 8

[46] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 8

[47] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7556–7566, 2019. 2

[48] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. What does stable diffusion know about the 3d scene? *arXiv preprint arXiv:2310.06836*, 2023. 1, 2

[49] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *CVPR*, 2018. 2

Shadows Don't Lie and Lines Can't Bend!
Generative Models don't know Projective Geometry...for now

Supplementary Material

9. Additional Analysis

In Table 2, we provide quantitative analysis that our Line Segment cues and Perspective Field cues are correlated and look at similar geometric cues while Object-Shadow cues look for different geometric cues to identify if an image is generated or real.

We also provide statistical distributions of geometry cues leveraged for detecting projective geometry distortion. These include Object-Shadow pairs, Perspective Fields, and line segment distributions obtained from DeepLSD. The distributions are in Figures 13, 14, 15, 16, and 17.

An ROC plot in Figure 18 shows that while using statistical biases helps detect generated images over chance, ResNet classifiers trained directly on these cues still outperform them.

Table 2. We quantify the distribution of detection agreement among three types of cues: Line Segment (LS), Perspective Fields (PF), and Object-Shadow (OS), for the images processed by Stable Diffusion-XL. The output indicates whether each method can accurately identify generated images as either real or generated. The "Yes" indicates that the method has correctly detected generated images, whereas "No" indicates that the method has identified generated images as real. We have also provided the absolute and percentage values of images for both indoor and outdoor domains' unconfident test set in the last two columns. The table reveals a statistically significant correlation between Line Segment and Perspective field cues (p-value $\approx 2e^{-16}$), suggesting they are not independent in their detection of generated images. Conversely, Object-Shadow Cues demonstrate a different pattern of detection, with the probability of identifying an image as generated being lower than that of Line Segment Cues. This shows that they are complementary and look at distinct discrepancies in the images. A qualitative figure demonstrating a complementary capability is in Figure 6 of the main text.

LS cues	PF cues	OS cues	Indoor	Outdoor
Yes	Yes	Yes	10520 (53.71%)	2382 (45.38%)
Yes	Yes	No	4844 (24.73%)	1314 (25.03%)
Yes	No	Yes	1033 (5.27%)	287 (5.47%)
Yes	No	No	725 (3.70%)	260 (4.95%)
No	Yes	Yes	872 (4.45%)	322 (6.13%)
No	Yes	No	874 (4.46%)	423 (8.06%)
No	No	Yes	285 (1.45%)	102 (1.94%)
No	No	No	435 (2.22%)	159 (3.03%)

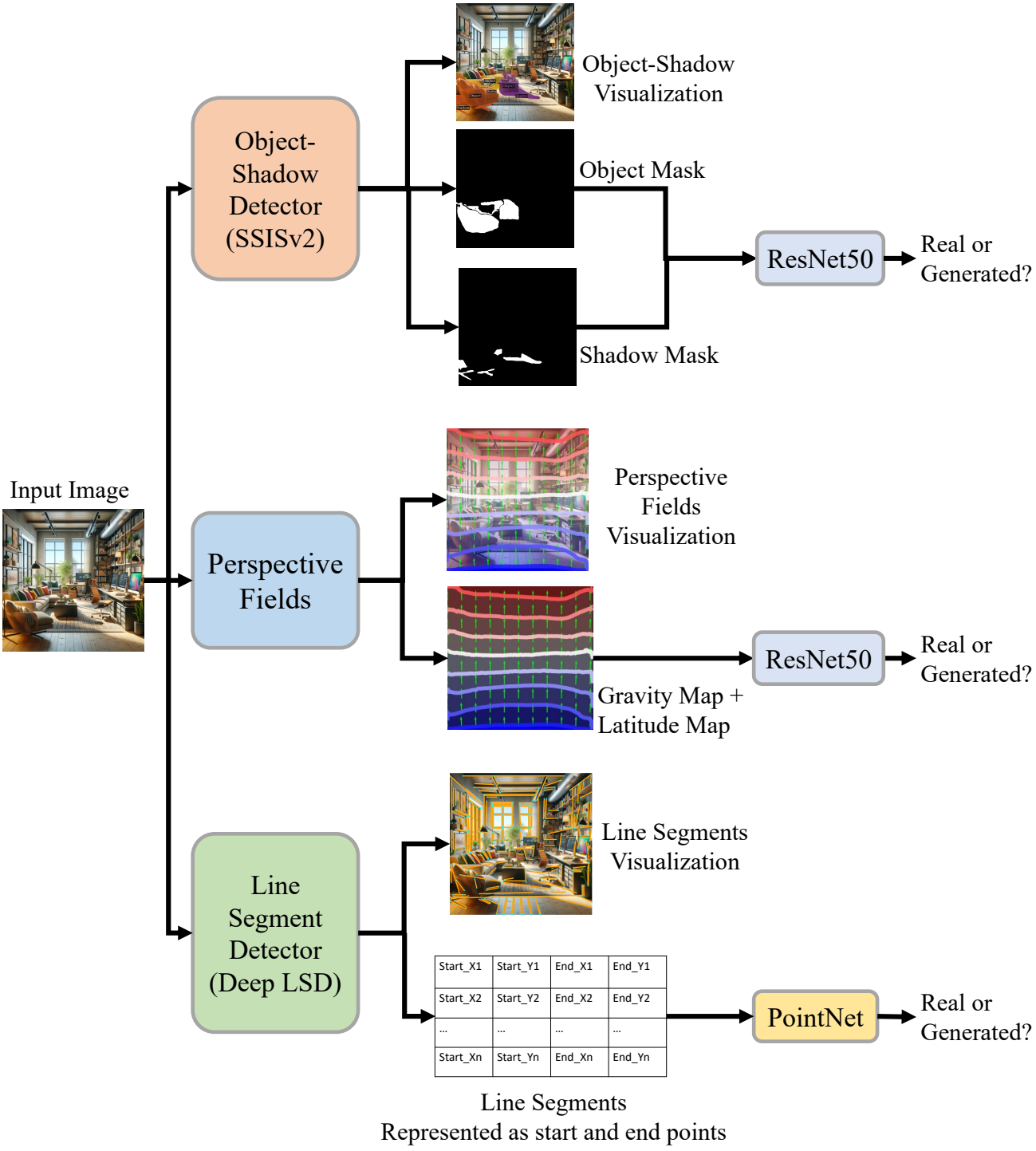
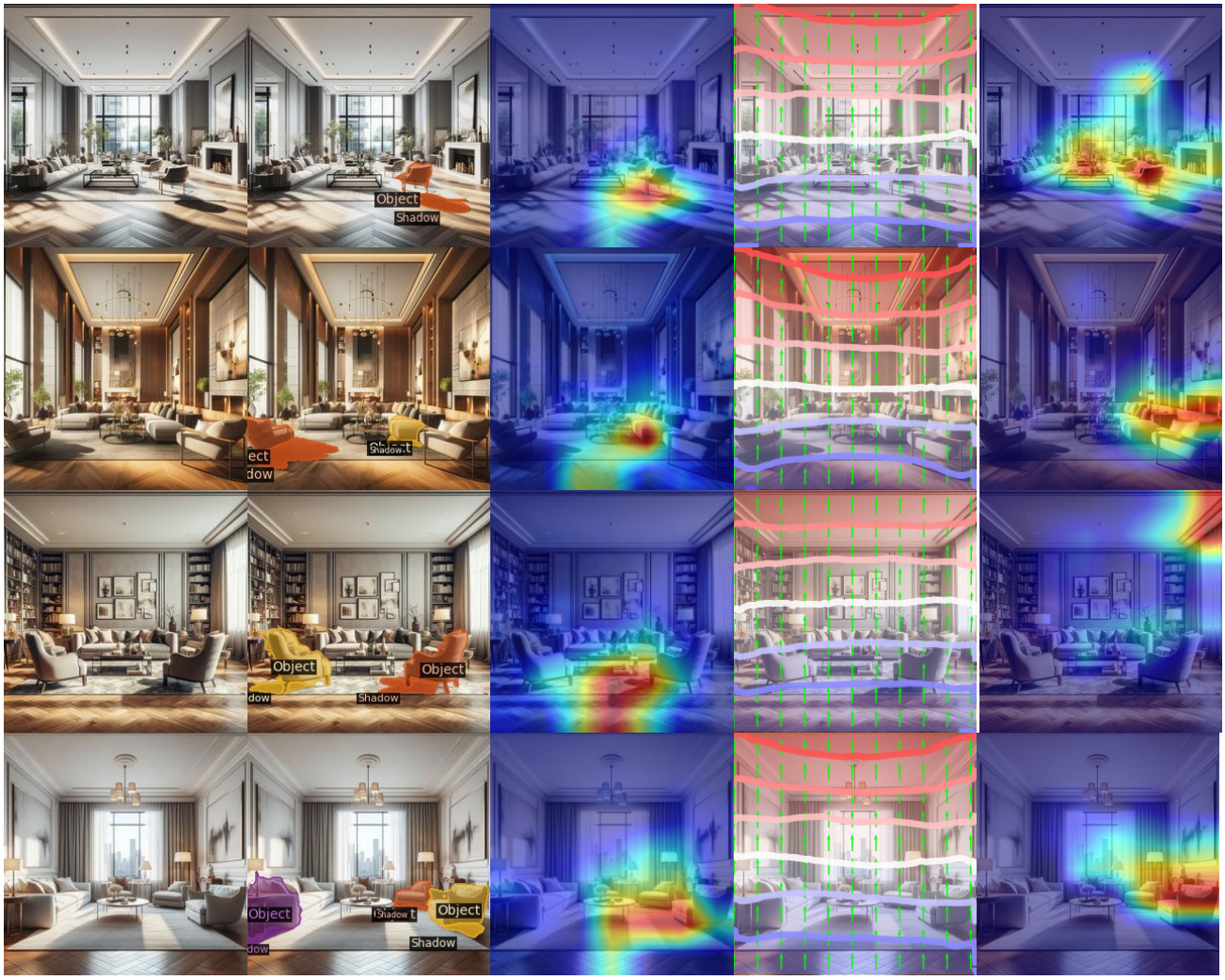
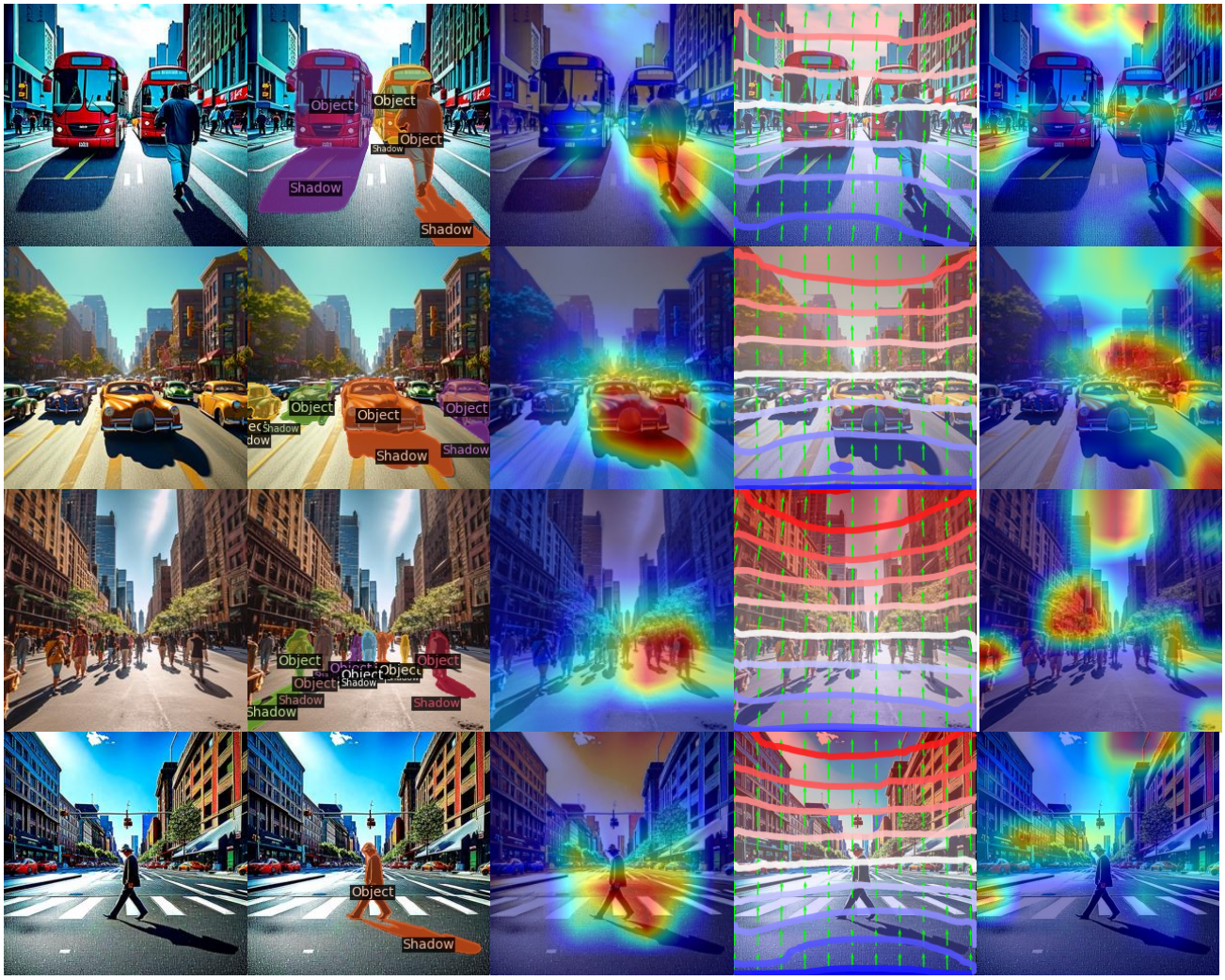


Figure 10. The schematic represents our pipeline, which begins by extracting geometric cues from images (left), such as object-shadow associations (top), perspective fields (middle), and line segments (bottom). These cues serve as the data describing geometry of images for training our classifiers. We utilize a ResNet architecture for Object-Shadow and Perspective Fields and PointNet for Line Segments in processing unordered data sets (Right).



Generated Image | Object-Shadow (OS) | OS GradCam | Perspective Fields | Perspective Fields GradCam

Figure 11. All interior scenes generated using Dalle-3. We analyze them using Object-Shadow (OS) cues and Perspective Fields (PF), along with their respective GradCam visualizations. The OS GradCam highlights areas where shadow directions or lengths don't appear to match the scene's lighting. For example, in the first and third rows, the shadows beneath the furniture don't seem to fit the objects casting them. The second row's OS GradCam shows an unnatural shadow on the sofa that's difficult to spot. Meanwhile, the PF analysis exposes inaccuracies in line alignment and vanishing points. In the top and third rows, the PF GradCam highlights inconsistencies along the room's ceiling lines and window frames that don't match the rest of the scene's perspective geometry. In the second and fourth rows, it detects inconsistencies on the side wall beneath the painting region.



Generated Image | Object-Shadow (OS) | OS GradCam | Perspective Fields | Perspective Fields GradCam

Figure 12. The generated street scenes in Adobe's Firefly have inconsistencies in projective geometry. We show Object-Shadow (OS) and Perspective Fields (PF) analyses and have presented each generated image alongside the results. In the first row, the shadow of the bus on the left is in one direction, while the shadow of the bus on the right and the pedestrian point is in opposite directions. The second row shows the OS GradCam pinpointing a car's shadow that is unrealistically elongated on one side. In the third and fourth rows, we observe pedestrians with shadows that are inconsistent with the lighting. The Perspective Fields analysis in rows two and three detects line inconsistencies deep in the scene and near vanishing points, while in the first and last rows, it captures discrepancies on the road markings and building facades.

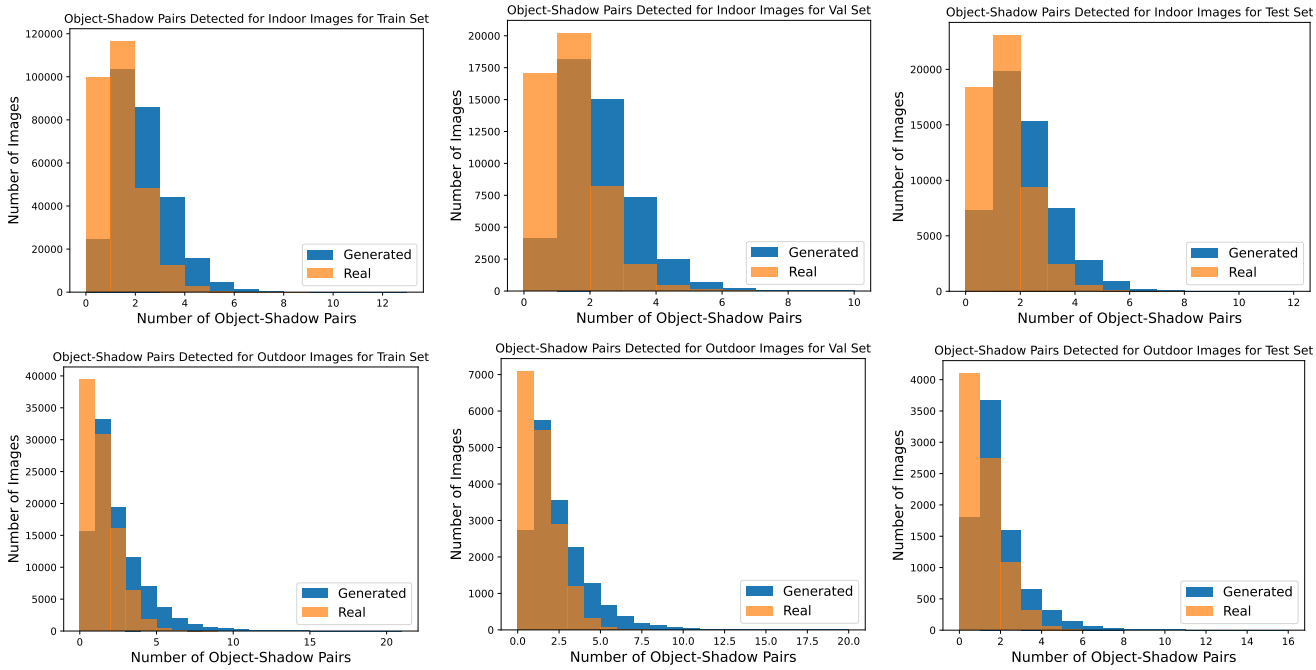


Figure 13. A statistical distribution analysis of a number of object-shadow pairs for both indoor and outdoor datasets. A classifier could exploit some of the statistical biases to distinguish between generated and real images. However, we found that our derived geometry cues perform much better than a classifier trained to look at such statistical signals, as shown in Figure 18. Furthermore, the GradCam analysis indicates that these derived object-shadow cues correctly identify erroneous regions.

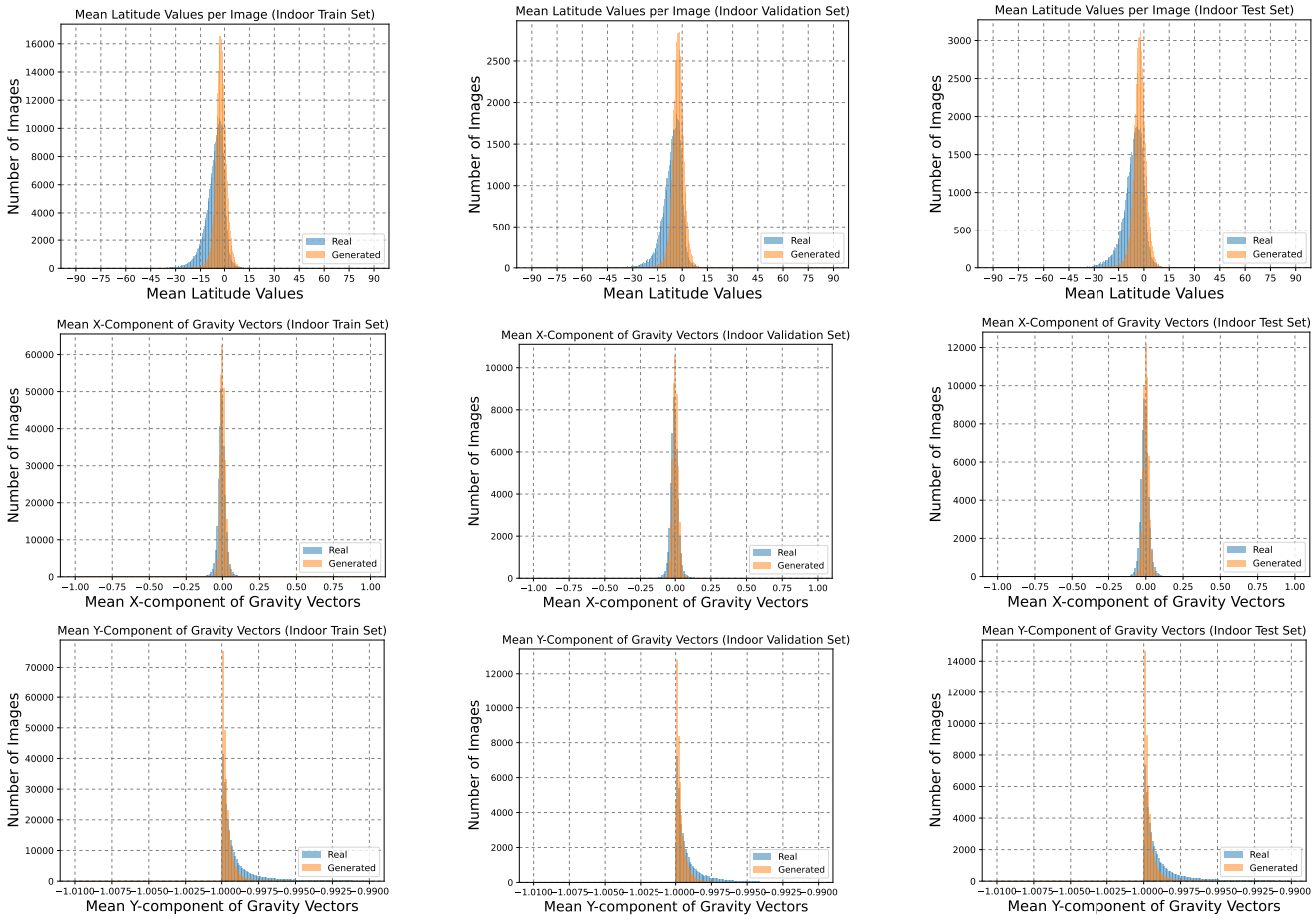


Figure 14. This set of histogram plots displays the statistical distribution of perspective field metrics in indoor scenes, comparing the training, validation, and test sets. The top row histograms reveal a significant difference in the distribution of latitude angles between real and generated images. The middle and bottom row plots illustrate the mean X and Y components of gravity vectors in the images, showing a clear separation between the real and generated images. These metrics indicate minor spatial inconsistencies between the real and generated images. Although these basic statistical differences provide some discriminative power, they are less effective than our ResNet classifier trained on Perspective Fields, which efficiently detects and focuses on critical geometric inconsistencies. This is validated by our comprehensive ROC analysis in Figure 18.

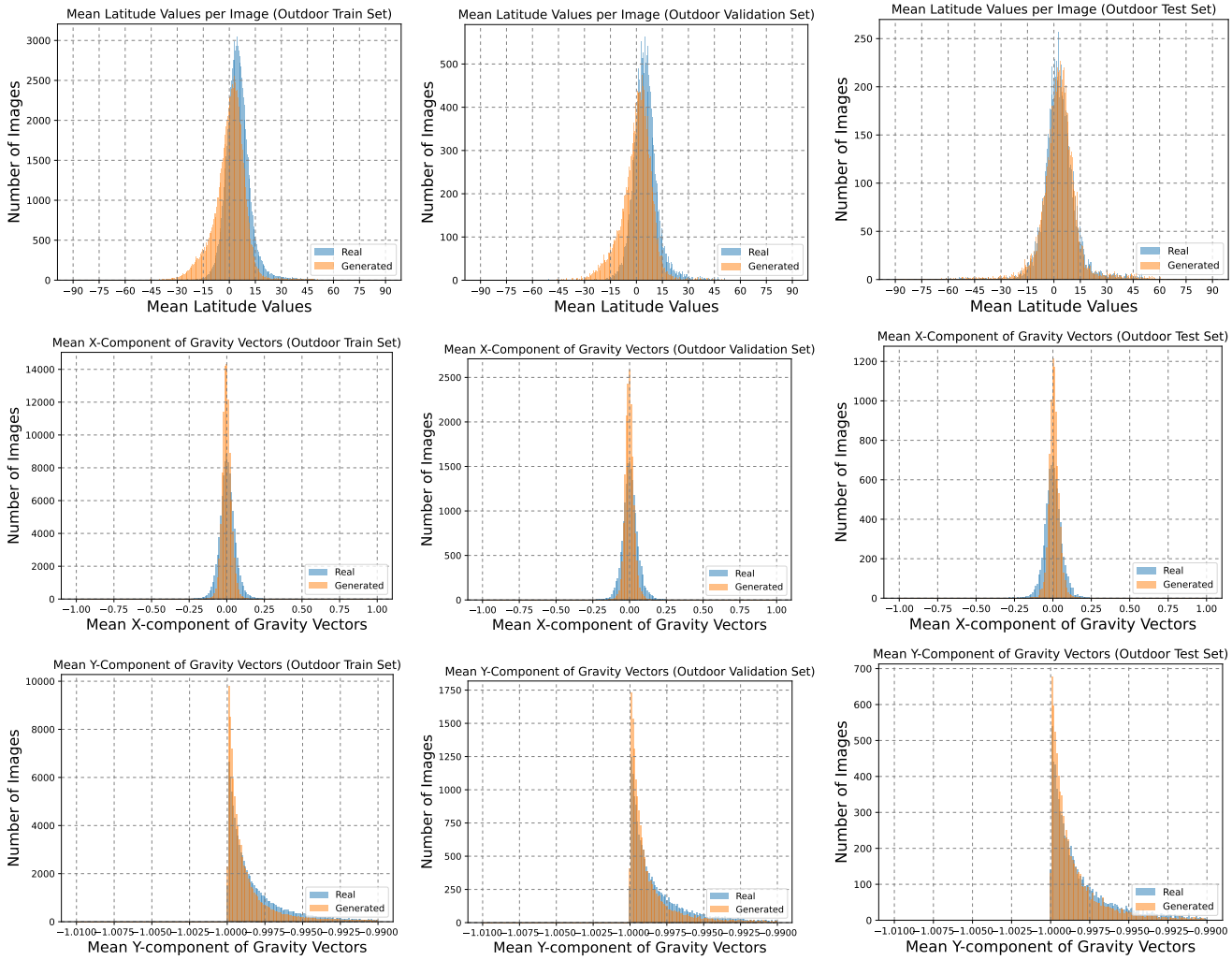


Figure 15. This set of histogram plots displays the statistical distribution of perspective field metrics in outdoor scenes, comparing the training, validation, and test sets. The top row histograms reveal a significant difference in the distribution of latitude angles between real and generated images. The middle and bottom row plots illustrate the mean X and Y components of gravity vectors in the images, showing a clear separation between the real and generated images. These metrics indicate minor spatial inconsistencies between the real and generated images. Although these basic statistical differences provide some discriminative power, they are less effective than our ResNet classifier trained on Perspective Fields, which efficiently detects and focuses on critical geometric inconsistencies. This is validated by our comprehensive ROC analysis in Figure 18.

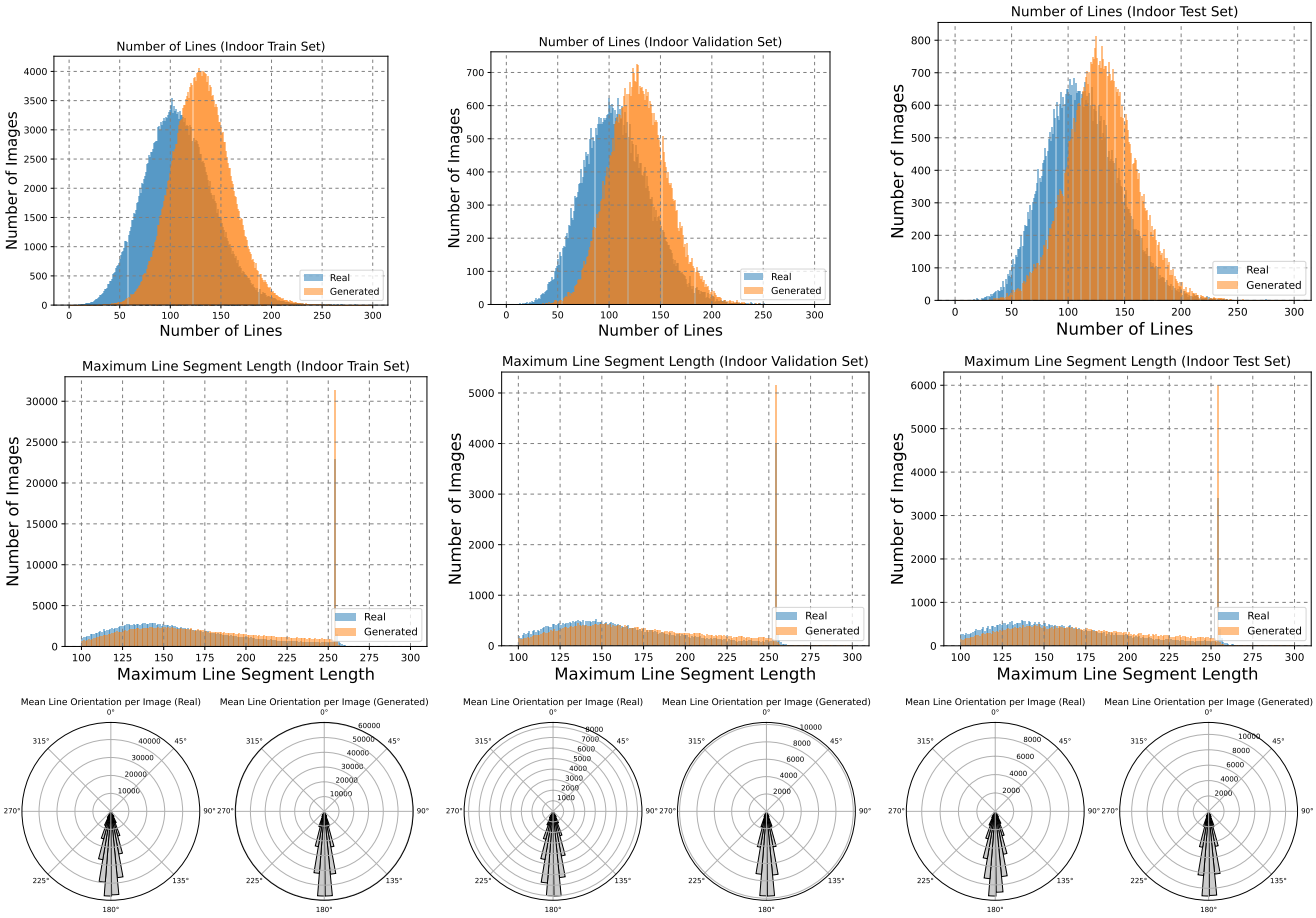


Figure 16. Line Segment Distribution in Indoor Scenes: We show the distribution of line segment counts and lengths in indoor scenes across training, validation, and test sets. The histograms (top row) compare the number of line segments detected in real versus generated images, with generated images generally exhibiting a different distribution, suggesting a discrepancy in line segment occurrence. The line segment length plots (middle row) show the maximum length of line segments. The polar plots (bottom row) illustrate the mean line orientation per image. While these basic statistical differences provide some discriminative power, they are notably less effective than our PointNet classifiers, which demonstrate a profound ability to detect and focus on critical geometric inconsistencies, as validated by our comprehensive ROC analysis in Figure 18.

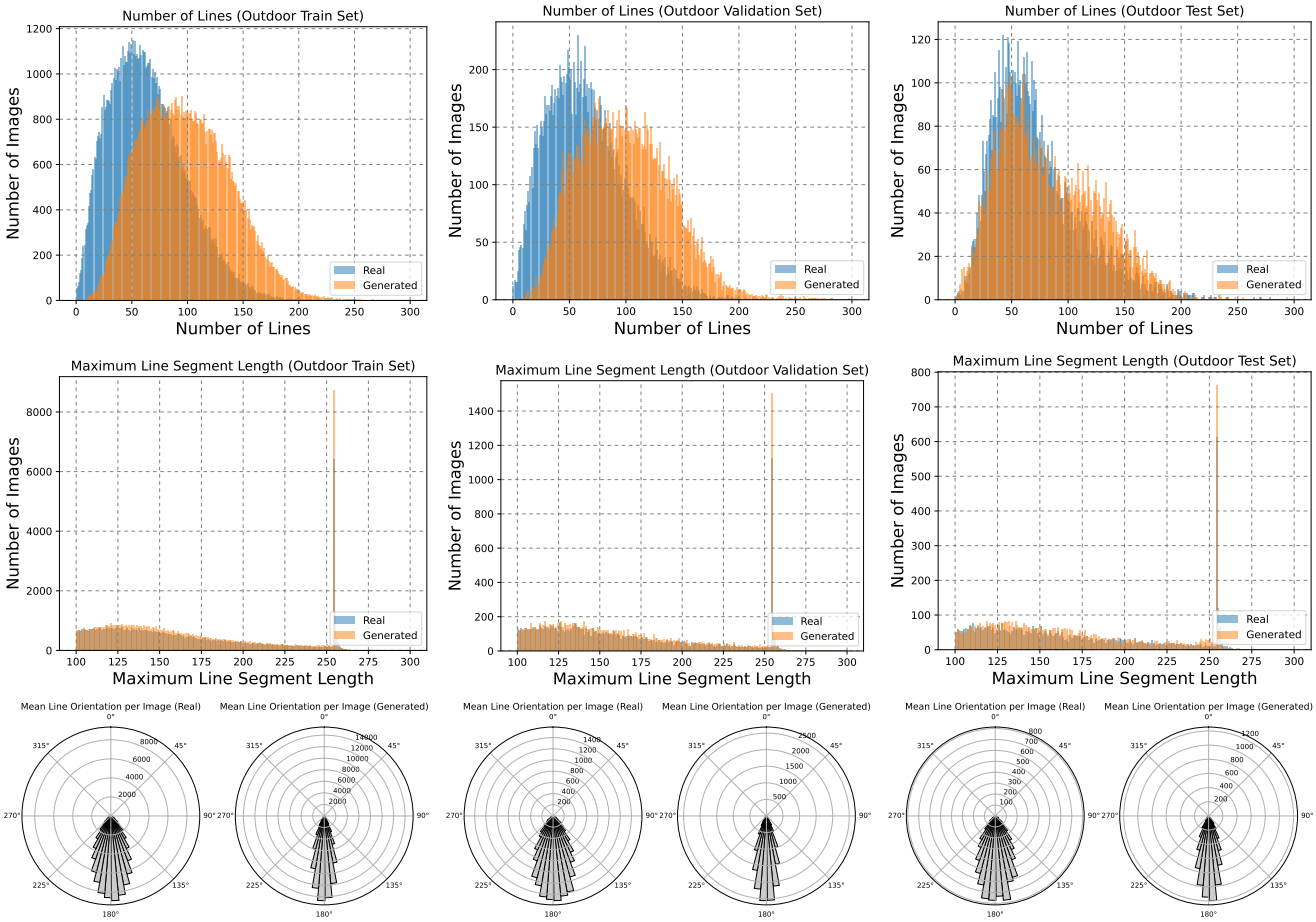


Figure 17. Line Segment Distribution in Outdoor Scenes: We show the distribution of line segment counts and lengths in indoor scenes across training, validation, and test sets. The histograms (top row) compare the number of line segments detected in real versus generated images, with generated images generally exhibiting a different distribution, suggesting a discrepancy in line segment occurrence. The line segment length plots (middle row) show the maximum length of line segments. The polar plots (bottom row) illustrate the mean line orientation per image. While these basic statistical differences provide some discriminative power, they are notably less effective than our PointNet classifiers, which demonstrate a profound ability to detect and focus on critical geometric inconsistencies, as validated by our comprehensive ROC analysis in Figure 18.

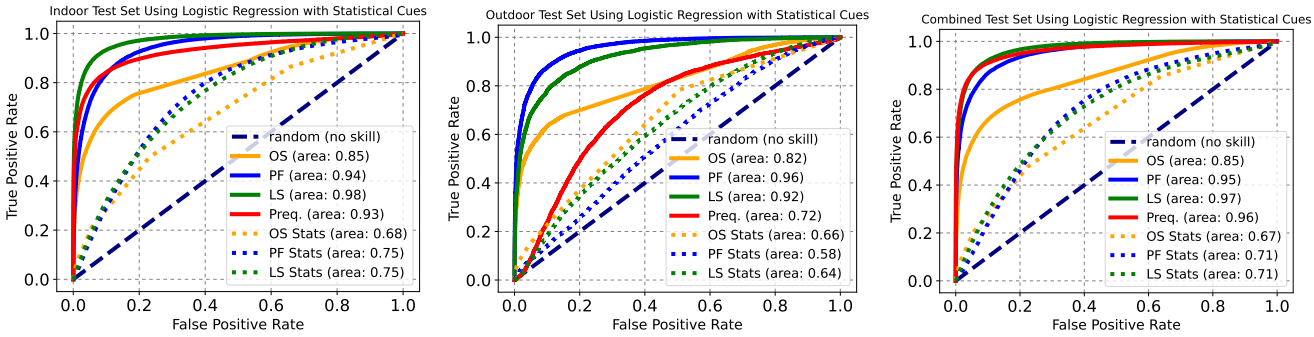


Figure 18. ROC analysis comparing our classifiers against basic statistical cues on our full test set. We compare the performance of our sophisticated classifiers – Object-Shadow (OS), Perspective Fields (PF) ResNet, and Line Segments (LS) PointNet classifiers – with basic statistical measures applied via logistic regression (LR) on indoor, outdoor, and combined test sets shown in dotted lines. While basic statistical cues like the count and mean lengths of line segments, the number of object shadows, and gravity changes per pixel indicate better-than-chance performance (AUCs ranging from 0.58 to 0.75), they are eclipsed by the more robust classifiers we developed and also the ResNet prequalifier. Our classifiers excel in identifying generated images with incorrect projective geometry by focusing on incorrect regions, not just statistical cues, as demonstrated by GradCAM visualizations.