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 [22–24] and diffusion models [36–38] are renowned for generating images that are strikingly similar to real-world photos and consistently fool people. But, as we show, generated images have distinctive geometric features, likely from a failure to fully capture projective geometry.

Bhattad et al. [5, 6], Chen et al. [10], Du et al. [13], Zhan et al. [49], and have shown generative models implicitly capture the complex scene properties, including normals, depth, albedo, and support relations. These works suggest these models "understand" geometry, which would

*equal contribution

be useful for rendering 3D scenes. Our detailed, population-level analysis of generated images suggests generative models [1, 3, 4, 12, 32] 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 [31]; Perspective Fields [21]; and PointNet [34]) that inspect geometric representations to distinguish between real and generated images. We rely on derived geometry cues from SOTA methods as manual image analysis and explicit geometry-based rules like drawing lines perspective lines are not scalable or automatable.

To guarantee the accuracy of our findings, we follow a strict process of data curation. This involves using a controlled pixel-level classifier to filter out any biases related to color, texture, and local features within our test set. This precision in data selection is crucial to isolate and accurately assess the inconsistencies in projective geometry and illumination present in generated images. By carefully screening the data, we can ensure that our results are not obscured by common artifacts that are usually found in generated images. This enhances 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.
- **Scalable, Data-Driven Approach:** We introduce a scalable, data-driven approach by utilizing three distinct projective geometry cues to detect geometric inaccuracies.
- **Robust and Generalizable:** We demonstrate robustness and generalizability by effectively identifying geometric errors across a wide array of images from recent generative models, including those with the latest time stamps.
- **Broadening the Scope of Model Assessment:** Our approach can be used as 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. [22–24] with StyleGAN, and the emergence of diffusion models [36–38], 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 diffusion models.

Geometric Understanding in Generative Models: While studies like Bhattad et al. [6], Du et al.[13], Chen et al. [10] and Zhan et al. [49] 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 [33] and JPEG quantization [2]. Kee et al. [25] 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 [14] 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. [7], 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 generative methods has steered image forensics towards using discriminative methods to detect synthetic content [8, 15, 19, 42, 43, 48, 50]. 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) [39] and Fréchet Inception Distance (FID) [18] focus on pixel-level fidelity. The emergence of CLIP-based scores [17, 35] and DIRE [45] 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 [11] and ImagenHub [26] address fine-grained evaluation inconsistencies, while the HEIM benchmark [27] 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 [16]. 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

Our data curation process is carefully designed to distinguish between real images and generated images from several generative models. This process includes models that the classifier has not seen during its training. Another important goal is to ensure our prequalifier effectively identifies images with recent timestamps. This helps prevent the classifier from relying on cues specific to the dataset, such as whether the image is from the training distribution or not, rather than determining whether the image is real or generated.

4.1. Real Images

We experiment with a diverse set of images, including:
(a) Indoor Scenes: A collection of 400,000 interior images featuring a variety of furniture arrangements and lighting conditions, sourced from LSUN [46]. Specifically, we used 100,000 images each from Bedroom, Dining Room, Kitchen, and Living Room categories.
(b) Outdoor Scenes: A dataset of 125,000 outdoor scenes with varying landscapes and urban settings, sourced from Berkeley Deep Drive 100K [47] and Mapillary Vistas [30]. We selected images that represent a wide range of weather conditions and times of day.
(c) Combination of Indoor and Outdoor scenes: We also analyzed a combination of above indoor and outdoor scenes to assess the performance on a more diverse dataset.
(d) Recent Timestamp Images: A curated test set of 500 indoor and 500 outdoor images with timestamps ranging from May 2023 to March 2024. These images were collected from various social media platforms and online sources to ensure our classifier's ability to handle recent real-world data and that the models are not obscured because of any data-source-specific biases.

4.2. Image Captions

We use the ViT-bigG-14/laion2b_s39b_b160k model [20] and the BLIP model [28] in succession, to generate refined captions for real images. These models were chosen for their state-of-the-art performance in image captioning tasks. The ViT model uses a Vision Transformer architecture, while BLIP employs a multimodal pre-training approach. By using common captions, we ensure a fair evaluation of the projective geometry in generated images.

4.3. Generated Images

We generate images from Stable Diffusion XL v1.0 [32], Kandinsky-v3 [3], DeepFloyd IF v1.0 [12], and PixArt-α v1.0 [9]. We use the same caption from the real images to generate these images with the default settings of each generative model.

4.4. Robust Prequalifier

Our goal is to identify challenging images that a signal-based classifier may struggle to differentiate. Therefore, we need to develop a robust prequalifier. This is to ensure that our results are not affected by any false data signals. Thus, we aim to eliminate all potential factors that may influence

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. The ResNet50 Prequalifier helps us curate 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
Training and Test Sets						
Training Set Size	75,000	75,000	25,000	25,000	100,000	100,000
Validation Set Size	10,000	10,000	5,000	5,000	15,000	15,000
Test Set Size	315,000	315,000	95,000	95,000	410,000	410,000
Post ResNet50 Prequalifier						
Unconfident Set	23213	13840	1444	1018	44392	2540
Misclassified Subset	10399	5756	609	443	23800	825

the geometry cues we obtain because of signal weirdness in the generated images. To accomplish this, we begin by training signal-based classifiers using a vanilla ResNet-50 and primarily concentrate on identifying instances that prove challenging for these signal-based classifiers.

We trained prequalifiers on three distinct settings - indoor scenes, outdoor scenes and a combination of the two. Each prequalifier was trained on a dataset consisting of real images and images generated by one of the four models: Stable Diffusion XL, Kandinsky-v3, DeepFloyd IF, and PixArt-α. While all four types of prequalifiers performed well on their respective test sets, we found that Kandinsky exhibited superior generalization capabilities when evaluated on images generated by other models or on recent timestamp images. The fundamental experiment conducted to assess the generalizability of different generators is covered in Figure 9 in the Supplementary Material.

For indoor scenes, our Kandinsky-based prequalifier achieves an Area Under the Curve (AUC) of 0.99 on its test set with an accuracy of 97.43 and maintained high performance on images generated by other models such as Stable Diffusion XL and PixArt-α, achieving AUCs of 0.97 and 0.98 respectively. Furthermore, our Kandinsky-trained prequalifier meant for both indoor and outdoor settings combined showed robust performance on recent timestamp generated indoor and outdoor images, achieving AUC scores of 0.90, 0.95, and 0.72 on images generated by Stable Diffusion XL, PixArt-α, and DeepFloyd IF, respectively.

Given their robustness and strong generalizability, we selected the Kandinsky-trained prequalifiers as the basis for training our derived geometry classifiers on Kandinsky-generated images. This choice ensures that our classifiers effectively look at geometric discrepancies and can handle various generated images. This enhances the reliability of our approach and supports our conclusion that the classifiers are not affected by spurious signal artifacts.

4.5. Final Test Sets

We categorize our test sets into three groups based on the accuracy of the prequalifier: easy, unconfident, and misclassified. The **easy test set** includes images with 100% accuracy, indicating that the prequalifier can reliably distinguish between real and generated images. The **unconfident test set** includes images where the prequalifier performs at chance level, suggesting that it struggles to make confident predictions. Finally, the **misclassified test set** includes images where the prequalifier makes completely wrong predictions, either classifying real images as generated or vice versa. A summary of this split for Kandinsky-v3 is provided in Tab. 1.

Our approach primarily evaluates the "hard set" (unconfident and misclassified) for geometric and shadow inconsistencies, assessing adherence to projective geometry principles. This ensures rigorous testing of generative models' ability to reproduce geometric correctness and photometric accuracy, beyond surface-level or signal details.

It is important to note that projective geometry inconsistencies are prevalent in generated images but often go undetected by conventional methods. Our models, trained on geometric abstractions and projective cues, can identify subtle but critical inaccuracies that texture artifacts cannot explain. 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.

We employ a suite of models to capture different facets of projective geometry, trained on datasets emphasizing geometric consistency and photometric accuracy. By combining their strengths, we comprehensively assess the quality of generated images and provide insights into the limitations and potential improvements of generative models from a projective geometry perspective.

5. Analyzing Projective Geometry

For our analysis, we rely on three geometry cues – object shadow, line segments, and perspective fields. We train

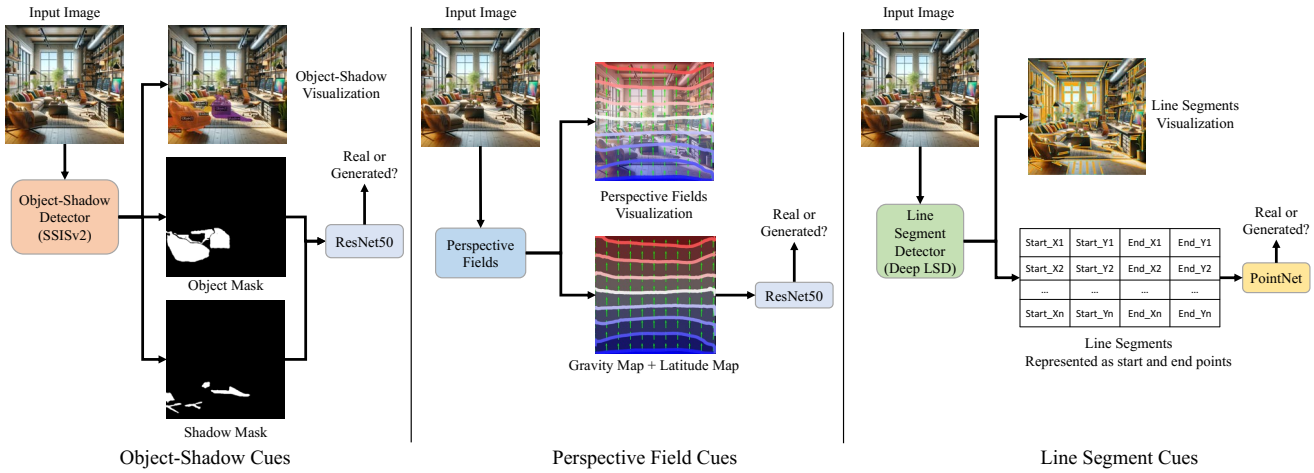


Figure 2. We train three classifiers to identify discrepancies in projective geometry. These classifiers are trained using derived geometry cues such as object-shadow associations (left), perspective fields (middle), and line segments (right) without looking at image intensity. We use the ResNet architecture for Object-Shadow and Perspective Fields, and PointNet for Line Segments to process unordered data sets.

three separate classifiers solely based on these derived geometry cues, without utilizing any pixel information. The training process for each classifier is described below, and a schematic pipeline of these classifiers can be found in Fig. 2.

5.1. Object-Shadow Cues

Our first model examines object-shadow relationships to address illumination. Shadows follow projective geometry principles, and inconsistencies can reveal generated images.

We use a pretrained object-shadow instance detection model [44] to identify shadows and geometric heuristics to evaluate their plausibility given the objects and their orientation. A ResNet50 classifier is trained on binary masks of object and shadow instances to score the consistency of shadows with objects. Images with implausible shadows are marked as likely generated.

5.2. Perspective Field Cues

Our framework's second model uses Perspective Fields [21], vector fields encoding the spatial orientation of pixels relative to vanishing points and the horizon, to assess projective geometry. We generate these fields from single images using a pretrained model.

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

5.3. Line Segment Cues

Our method also assesses projective geometry in generated images by identifying key structural lines using Deep LSD [31]. These lines indicate adherence to perspective rules. We then train a PointNet-like architecture [34] to classify images based on line segment patterns, differentiating real from generated images.

PointNet's flexibility in handling unordered data makes it suitable for analyzing line segments without pre-sorting. The model assigns scores representing the likelihood of an image being real based on the spatial arrangement of its lines. Analyzing these scores reveals the model's ability to detect subtle discrepancies in line arrangements, which often indicate a generated image.

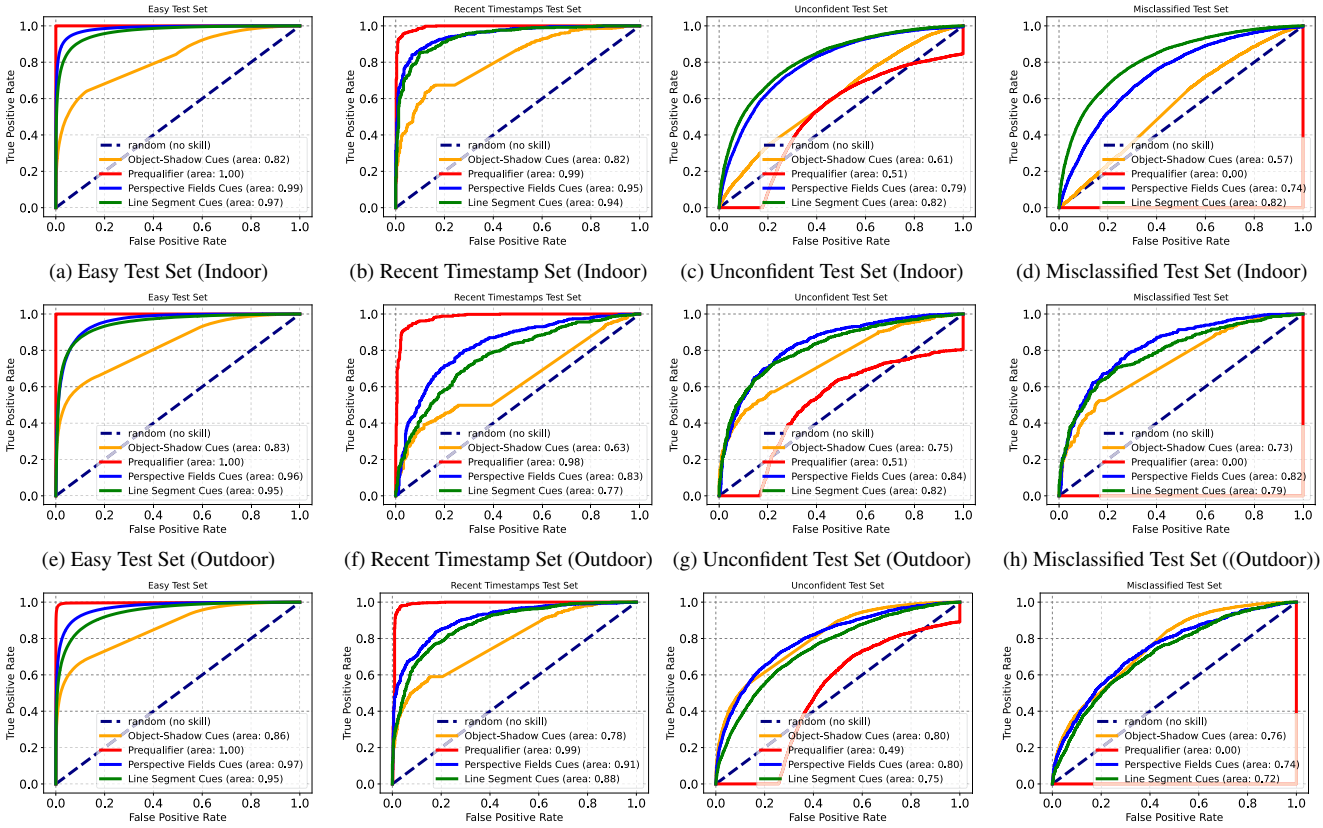
6. Evaluation

Based on our prequalification analysis, we use Kandinsky as the primary source for our generated data to train our classifiers. We evaluated our three geometry-derived cues classifiers and analyzed their ability to generalize to other generated images that were not used during training. In addition, we performed a GradCam analysis [40] to identify any potential geometric discrepancies in images.

6.1. Classifiers Results

Fig. 3 shows ROC curves for each of our methods 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 appear in Fig. 5 and Fig. 4. Notice how images that might be acceptable to a line analysis often fail a shadow analysis. Fig. 6 shows



(a) Easy Test Set (Indoor) (b) Recent Timestamp Set (Indoor) (c) Unconfident Test Set (Indoor) (d) Misclassified Test Set (Indoor)

(e) Easy Test Set (Outdoor) (f) Recent Timestamp Set (Outdoor) (g) Unconfident Test Set (Outdoor) (h) Misclassified Test Set ((Outdoor))

(i) Easy Test Set (Combined) (j) Recent Timestamp Set (Combined) (k) Unconfident Test Set (Combined) (l) Misclassified Test Set (Combined)

Figure 3. ROC Curves Assessing Projective Geometry Cues in Generated Images trained on Kandinsky-v3. We trained separate models for indoor scenes (top row), outdoor scenes (middle row), and a combination of indoor and outdoor scenes (last row). All our derived geometry cues classifiers are trained without looking at image intensity information and can reliably detect projective geometry errors. The recent timestamp test set (second column) confirms that these models are robust. We find hard examples using a prequalifier trained on image pixels. Our derived geometry cues consistently show high AUC for finding projective geometry errors on hard test sets – the last two columns – unconfident and misclassified test sets. For the unconfident test set, where the prequalifier has an AUC of 0.51 (c), 0.51 (g), and 0.49 (h) for indoor, outdoor, and combined partition, our classifiers can still accurately identify the generated images with high AUCs – 0.82 from line segments in the indoor set, 0.84 from perspective fields in the outdoor set, and 0.80 from perspective field cues and object shadows in the combined set. Similarly, for the misclassified test set, where the prequalifier has an AUC of 0.00, as it should, our classifiers remain reliable with AUC up to 0.82. We conclude that generated images contain geometric structures not seen in real images, and these structures very reliably identify generated images by only looking at derived geometry cues.

examples to emphasize this point.

7. Other Generators Evaluated

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 Fig. 7 demonstrate that classifiers trained to distinguish Kandinsky images from real images can also reliably distinguish Stable DIffusion XL [3], DeepFloyd [12] and PixArt-α[9] 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 [29] by looking at Object-Shadow cues in the supplement.

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. Our findings have significant implications for the development of image generation models, as the inability to

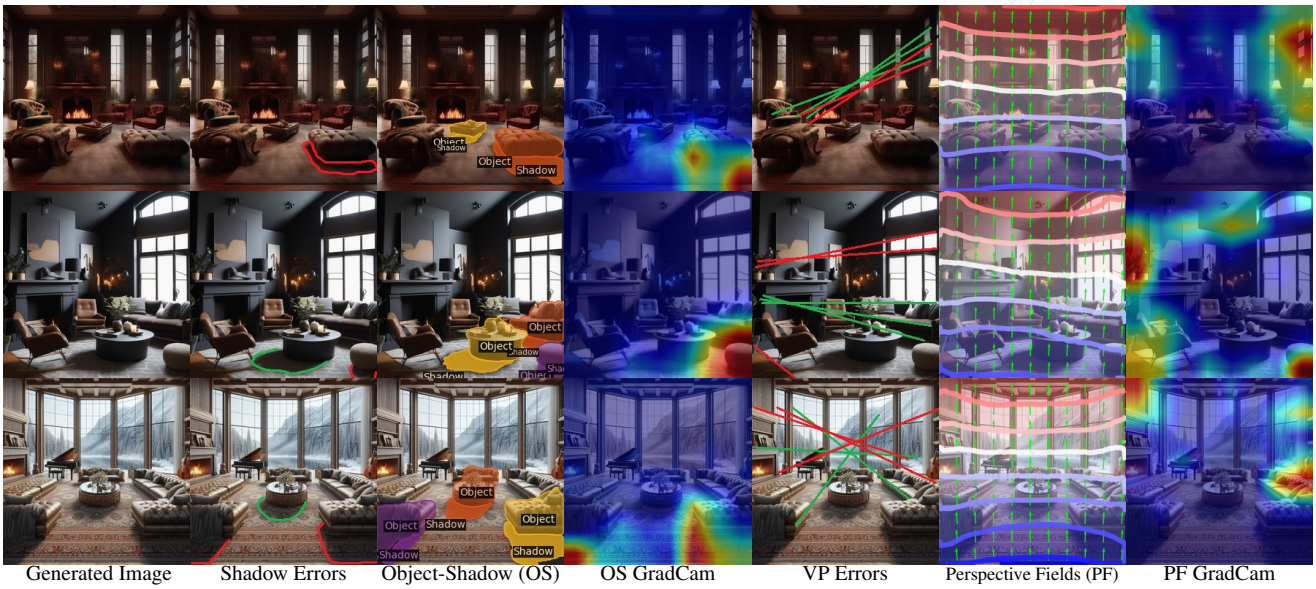


Figure 4. Grad-CAM applied to our Object-Shadow and Perspective Field classifiers reveals that the high AUCs in Fig. 3 are based on real geometric errors in indoor scenes generated by Kandinsky, Stable Diffusion XL and Dalle-3 shown in each row respectively. The second and fifth columns highlight shadow and vanishing point errors, respectively. The third column overlays detected object-shadow pairs [44]. Grad-CAM applied to our Object-Shadow classifier (fourth column) identifies diagnostic areas for synthetic generation, such as inconsistent shadow directions (in all three rows), mismatched shadow lengths (second row). The sixth column shows Perspective Fields [21], and Grad-CAM applied to our Perspective Fields classifier (last column) reveals geometric errors in all three rows, particularly at ceilings and side walls, with noticeable errors also present in window grills in the first and second rows.

Figure 5. Grad-CAM results for outdoor scenes generated by Kandinsky, Stable Diffusion XL, and Adobe Firefly, shown in each row respectively. The second column highlights shadow errors, while the third column overlays detected object-shadow pairs [44]. Grad-CAM applied to our Object-Shadow classifier (fourth column) reveals incorrect shadow shapes in the first and second rows, with shadows on the right-side pedestrians pointing in a different direction than those on the left. The fifth column shows vanishing point errors, and the sixth column presents Perspective Fields [21]. Grad-CAM applied to our Perspective Fields classifier (last column) confirms large perspective distortions on building facades and road markings, corroborating the vanishing point errors in the fifth column.

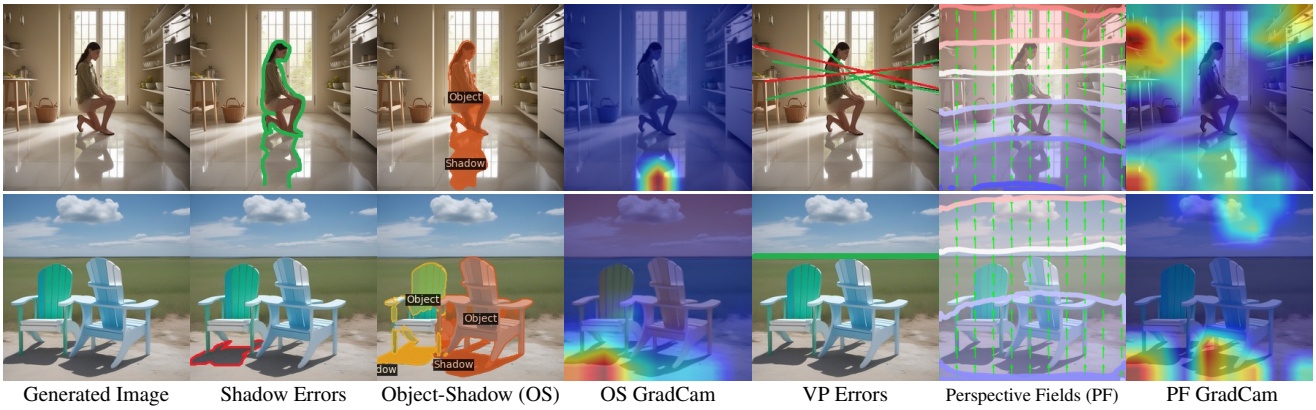
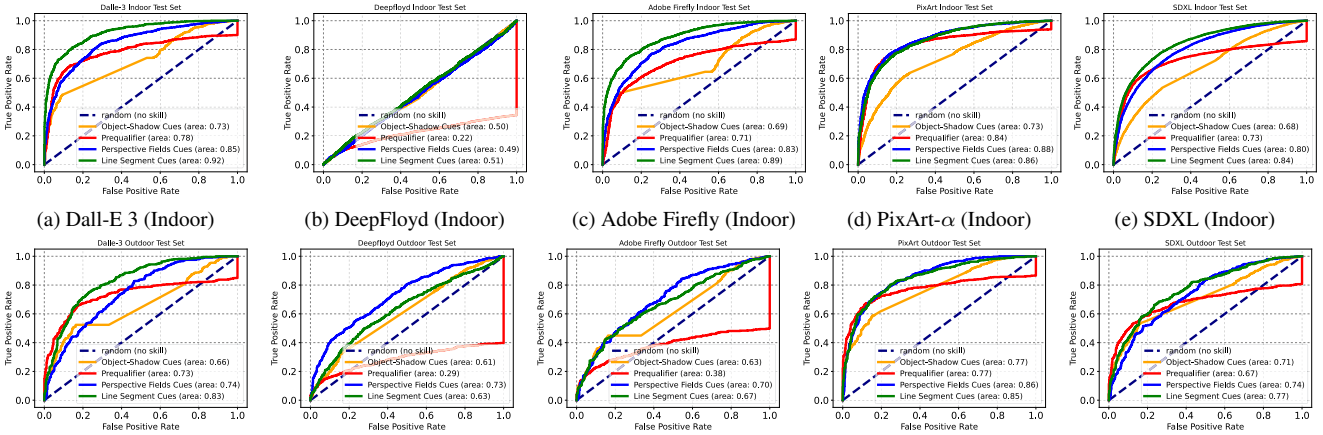


Figure 6. Our projective geometry classifiers identify distinct types of problems in generated images. The top row presents an example that was classified as real by the Object-Shadow classifier but correctly identified as generated by the Perspective Fields classifier. While the shadow cast by the person appears realistic, the Perspective Fields Grad-CAM highlights the problematic geometry of the shelf on the top left. In contrast, the bottom row shows an example that was correctly identified as generated by the Object-Shadow classifier but misclassified as real by the Perspective Fields classifier. Although the perspective effects in the image appear plausible, the Grad-CAM weights correctly reveal that the two chairs are casting shadows from different light sources, indicating inconsistency in scene's illumination.



(a) Dall-E 3 (Indoor) (b) DeepFloyd (Indoor) (c) Adobe Firefly (Indoor) (d) PixArt-α (Indoor) (e) SDXL (Indoor)

(f) Dall-E 3 (Outdoor) (g) DeepFloyd (Outdoor) (h) Adobe Firefly (Outdoor) (i) PixArt-α (Outdoor) (j) SDXL (Outdoor)

Figure 7. Our model trained on Kandinsky-v3 generalizes in detecting projective geometry errors from images generated by various unseen models. We evaluate the model's performance on test sets generated by Dall-E 3 (a), DeepFloyd (b), Adobe Firefly (c), PixArt-α (d), and Stable Diffusion XL or SDXL (e) using the same text prompts from the 'misclassified' Kandinsky-v3 generated test set. The top row shows indoor scenes and the last row shows outdoor scenes. Our model generalizes to all evaluated generators except DeepFloyd. However, DeepFloyd-generated images can be reliably detected when the model is trained on a DeepFloyd-generated training set, but it shows poor generalization capabilities to other generators compared to Kandinsky.

accurately replicate projective geometry extends across various state-of-the-art models, indicating a widespread issue rather than a problem specific to a particular generator.

We speculate that fixing this difficulty requires structural innovation in the generator, rather than simply exposing the generator to more data. The complex interplay of light, shadows, and perspective in real-world scenes may necessitate novel approaches to modeling and encoding spatial relationships within the generator's architecture. Potential avenues for improvement could include incorporating explicit geometric reasoning or developing new loss functions that prioritize the consistency of projective geometry. Furthermore, our work highlights the importance of developing robust evaluation metrics for image generation models that assess the geometric coherence of generated images. By tackling these challenges, we could create image-generation models that more faithfully capture the complex geometric relationships in real-world scenes.

Acknowledgment

This paper is based on work supported in part by the National Science Foundation under Grant No. 2106825 and a gift from Boeing Corporation.

References

[1] Adobe. Firefly. https://www.adobe.com/sensei/generative-ai/firefly.html, 2023. Accessed: 2023-11. 2, 6

[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] Vladimir Arkhipkin, Andrei Filatov, Viacheslav Vasilev, Anastasia Maltseva, Said Azizov, Igor Pavlov, Julia Agafonova, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky 3.0 technical report, 2023. 2, 3, 6

[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, 6

[5] Anand Bhattad and D.A. Forsyth. Stylitgan: Prompting stylegan to generate new illumination conditions. In *arXiv*, 2023. 1

[6] 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

[7] 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

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

[9] 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. 3, 6

[10] 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

[11] 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

[12] Deep Floyd. Iterative filter. https://github.com/deep-floyd/IF, 2023. Accessed: 2023-11. 2, 3, 6

[13] Xiaodan Du, Nicholas Kolkin, Greg Shakhnarovich, and Anand Bhattad. Generative models: What do they know? do they know things? let's find out! *arXiv preprint arXiv:2311.17137*, 2023. 1, 2

[14] 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

[15] 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

[16] David A Forsyth and Jean Ponce. *Computer vision: a modern approach*. prentice hall professional technical reference, 2002. 3

[17] 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

[18] 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

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

[20] 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

[21] 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, 7, 16, 17

[22] 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

[23] 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.

[24] 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

[25] 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

[26] 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

[27] 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

[28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 3

[29] 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. 6

[30] 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. 3

[31] 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, 5

[32] 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

[33] 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

[34] 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, 5

[35] 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

[36] 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

[37] 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.

[38] 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

[39] 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

[40] 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. 5

[41] 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. 11

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

[43] 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, 12

[44] 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. 5, 7, 16, 17

[45] 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

[46] 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. 3

[47] 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. 3

[48] 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

[49] 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

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