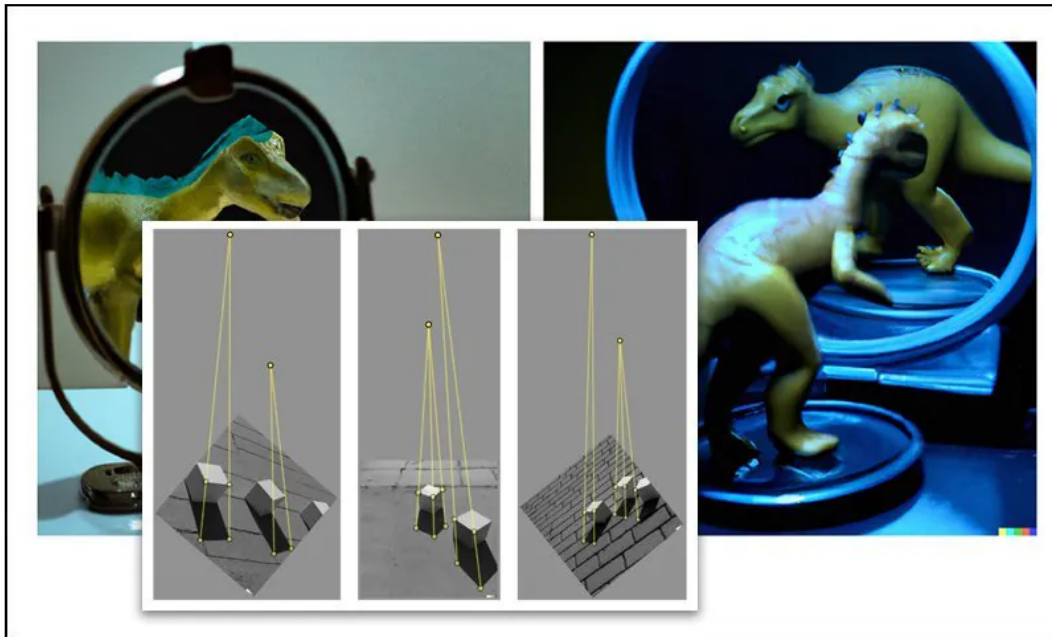


A Detection System for Pure Image Synthesis Frameworks Like DALL-E 2

Originally published at <https://www.unite.ai/a-detection-system-for-pure-image-synthesis-frameworks-like-dall-e-2/>



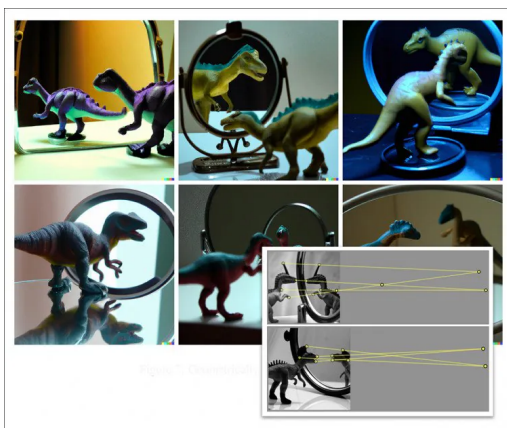
New [research](#) from the University of California at Berkeley offers a method to determine whether output from the new generation of image synthesis frameworks – such as Open AI's [DALL-E 2](#), and Google's [Imagen](#) and [Parti](#) – can be detected as 'non-real', by studying geometry, shadows and reflections that appear in the synthesized images.

Studying images generated by text prompts in DALL-E 2, the researchers have found that in spite of the impressive realism of which the architecture is capable, some persistent inconsistencies occur related to the rendering of global perspective, the creation and disposition of shadows, and especially regarding the rendering of reflected objects.

The paper states:

'[Geometric] structures, cast shadows, and reflections in mirrored surfaces are not fully consistent with the expected perspective geometry of natural scenes. Geometric structures and shadows are, in general, locally consistent, but globally inconsistent.'

'Reflections, on the other hand, are often rendered implausibly, presumably because they are less common in the training image data set.'



A lack of consistent intersections between the rendered object and the rendering of its reflection is currently a reliable way to detect a DALL-E 2 image, according to the new study. Source: <https://arxiv.org/pdf/2206.14617.pdf>

The paper represents an early foray into what may eventually become a noteworthy strand in the computer vision research community – Image Synthesis detection.

Since the advent of deepfakes in 2017, [deepfake detection](#) (primarily of [autoencoder](#) output from packages such as [DeepFaceLab](#) and [FaceSwap](#)) has become an [active and competitive](#) academic strand, with various papers and methodologies targeting the evolving 'tells' of synthesized faces in real video footage.

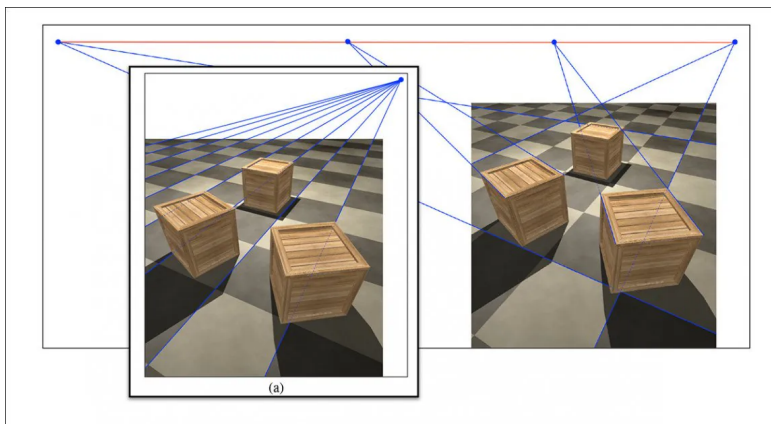
However, until the very recent emergence of hyperscale-trained image generations systems, the output from text-prompt systems such as [CLIP](#) posed no threat to the status quo of 'photoreality'. The authors of the new paper believe that this is about to change, and that even the inconsistencies that they have discovered in DALL-E 2 output may not make much difference to output images' potential to deceive viewers.

The authors state*:

'[Such] failures may not matter much to the human visual system which has been found to be surprisingly inept at certain geometric judgments including inconsistencies in [lighting](#), [shadows](#), [reflections](#), [viewing position](#), and [perspective distortion](#).'

Vanishing Credibility

The authors' first forensic examination of DALL-E 2 output relates to perspective projection – the way that the positioning of straight edges in nearby objects and textures should resolve uniformly to a 'vanishing point'.



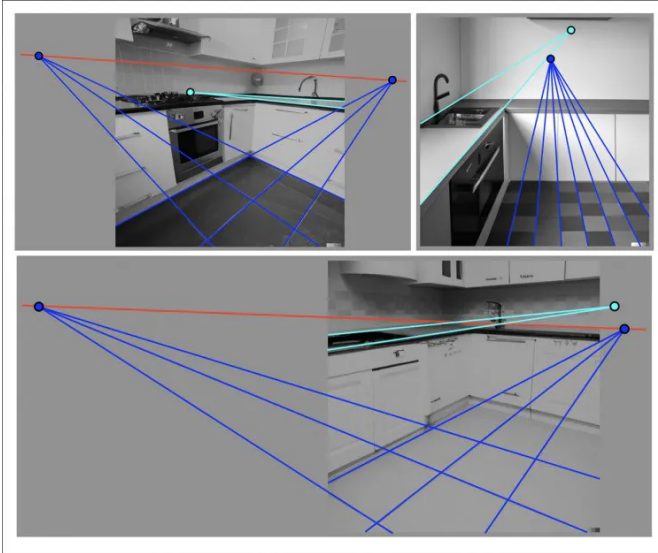
Left, parallel lines on the same plane resolve to a common vanishing point; right, multiple vanishing points on the same and parallel planes define a vanishing line

To test DALL-E 2's consistency in this regard, the authors used DALL-E 2 to generate 25 synthesized images of kitchens – a familiar space that, even in well-appointed dwellings, is usually confined enough to provide multiple possible vanishing points for a range of objects and textures.

Examining output from the prompt 'a photo of a kitchen with a tiled floor', the researchers found that in spite of a generally convincing representation in each case (bar some strange, smaller artifacts unrelated to perspective), the objects depicted never seem to converge correctly.



The authors note that while each set of parallel lines from the tile pattern are consistent and intersect at a sole vanishing point (blue in the image below), the vanishing point for the counter-top (cyan) disagrees with both the vanishing lines (red) and the vanishing point derived from the tiles.



The authors observe that even if the counter-top was not parallel to the tiles, the cyan vanishing point should resolve to the (red) vanishing line defined by the vanishing points of the floor tiles.

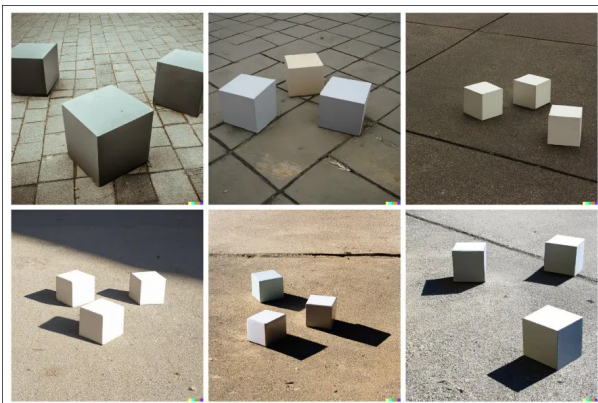
The paper states:

‘While the perspective in these images is – impressively – locally consistent, it is not globally consistent. This same pattern was found in each of 25 synthesized kitchen images.’

Shadow Forensics

As anyone who has ever dealt with ray-tracing knows, shadows also have potential vanishing points, indicating single or multi-source illumination. For exterior shadows in harsh sunlight, one would expect shadows across all the facets of an image to resolve consistently to the single source of light (the sun).

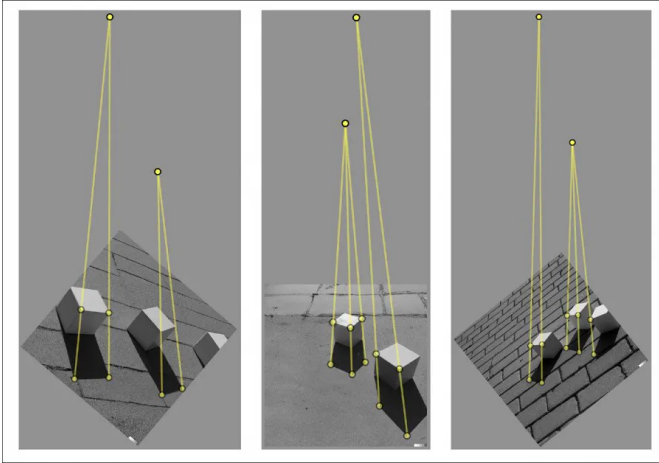
As with the previous experiment, the researchers created 25 DALL-E 2 images with the prompt ‘three cubes on a sidewalk photographed on a sunny day’, as well as a further 25 with the prompt ‘three cubes on a sidewalk photographed on a cloudy day’.



In the top row, images created from the researchers’ prompt ‘three cubes on a sidewalk photographed on a cloudy day’; in the lower row, images created from the prompt ‘three cubes on a sidewalk photographed on a sunny day’.

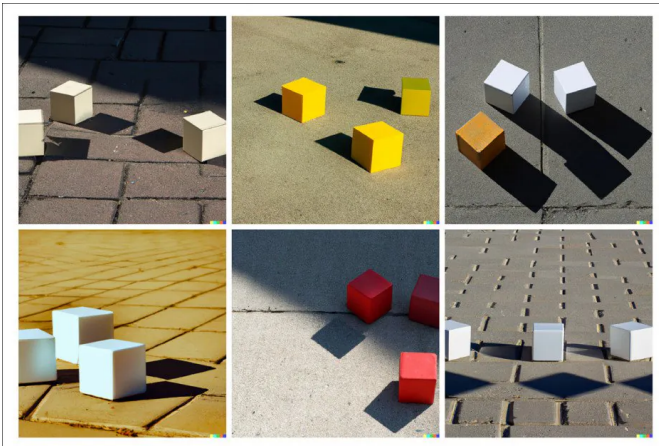
The researchers note that when representing cloudy conditions, DALL-E 2 is able to render the more diffuse associated shadows in a convincing and plausible manner, perhaps not least because this type of shadow is likely to be more prevalent in the dataset images on which the framework was trained.

However, some of the ‘sunny’ photos, the authors found, were inconsistent with a scene illuminated from a single light source.



For the above image, the generations have been converted to grayscale for clarity, and show each object with its own dedicated 'sun'.

Though the average viewer may not spot such anomalies, some of the generated images had more manifest examples of 'shadow failure':



While some of the shadows are simply in the wrong place, many of them, interestingly, correspond to the kind of visual discrepancy produced in CGI modelling when the sample rate for a virtual light is too low.

Reflections in DALL-E 2

The most damning results in terms of forensic analysis came when the authors tested DALL-E 2's ability to create highly reflective surfaces, which is a burdensome calculation also in CGI ray-tracing and other traditional rendering algorithms.

For this experiment, the authors produced 25 DALL-E 2 images with the prompt 'a photo of a toy dinosaur and its reflection in a vanity mirror'.

In all cases, the authors report, the mirror image of the rendered toy was in some way disconnected from the 'real' toy dinosaur's aspect and disposition. The authors state that the problem was resistant to variations in the text prompt, and it seems to be a fundamental weakness in the system.



There seems to be a logic in some of the errors – the first and third examples in the top row appears to show a dinosaur that is *duplicated* very well, but not mirrored.

The authors comment:

'Unlike the cast shadows and geometric structures in the previous sections, DALL·E-2 struggles to synthesize plausible reflections, presumably because such reflections are less common in its training image data set.'

Glitches like these may be ironed out in future text-to-image models that are able to review more effectively the overall semantic logic of their output, and which will be able to impose abstract physical rules on scenes which have, to an extent, been assembled from word-pertinent features in the system's latent space.

In the light of a growing trend towards ever-larger synthesis architectures, the authors conclude:

'[It] may just be a matter of time before paint-by-text synthesis engines learn to render images with full-blown perspective consistency. Until that time, however, geometric forensic analyses may prove useful in analyzing these images.'

** My conversion of the authors' inline citations to hyperlinks.*

First published 30th June 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.

Portfolio site: martinanderson.ai

Contact: martin@martinanderson.ai



Discover More