# AI May Secretly Rank Images by Device Brand, Not Content

Originally published at https://www.unite.ai/ai-may-secretly-rank-images-by-device-brand-not-content/

*New research finds that popular image-centric AI systems don't just look at what's in a photo, they also pick up on how the photo was taken. Hidden details like camera type or image quality can quietly affect what the AI thinks it sees, leading to wrong results – just because the photo came from a different device.*

In 2012 it was revealed that a travel website was routinely showing higher prices to users that it could deduce were browsing on Apple devices, equating the Apple AAPL 0.36% brand with higher spending power. Later investigation concluded that this device-focused 'wallet-sniffing' had become almost routine for ecommerce sites.

Similarly, which smartphone or capture device took a particular photograph can be deduced by forensic means, based on the known characteristics of a limited number of lenses in the models. In such cases, the model of a capture device is usually estimated by *visual* traces; and, as with the 2012 incident, knowing what type of camera took an image is a potentially exploitable characteristic

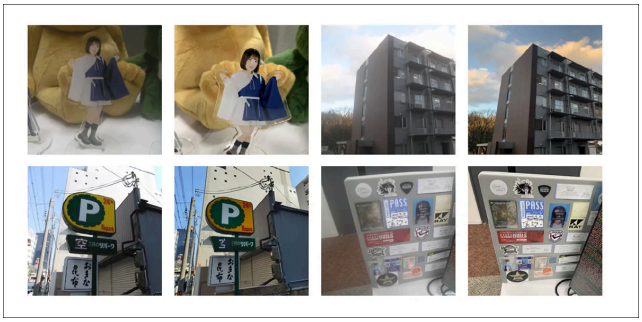
Though capture devices tend to embed significant metadata into an image, this feature can often be turned off by users; even where it is left on, distribution platforms such as social media networks may strip out some or all of the metadata, either for logistical or privacy purposes, or both.

Nonetheless, metadata in user-uploaded images is frequently either re-written/interpreted (rather than deleted) or else left intact, as a secondary source of information not about what is in the picture, but how the picture was taken. As the 2012 case revealed, information of this kind can be valuable – not only to commercial platforms, but also, potentially, to hackers and bad actors.

## Twin Viewpoints

A new research collaboration between Japan and the Czech Republic has found that the traces left by camera hardware and image processing (such as JPEG quality or lens sharpening) are not only detectable by forensic tools but are also silently encoded in the 'global understanding' of leading AI vision models.

This includes CLIP and other large-scale visual encoders, which are widely used in everything from search engines to content moderation. The new work demonstrates that these models do not merely interpret what is *in* a photo, but can also learn how the photo was *made*; and this hidden signal can sometimes overpower the visible content.



*Example image pairs from the authors' PairCams dataset, created to test how camera type affects AI image models. Each pair shows the same object or scene photographed at the same moment using a non-smartphone (left) and a smartphone (right).* Source: https://arxiv.org/pdf/2508.10637

The study asserts that even when AI models are given heavily masked or cropped versions of the image, they can still guess the make and model of the camera with surprising accuracy. This means that the representation space these systems use to judge image similarity can become entangled with irrelevant factors, such as the user's device, with unpredictable consequences.

For instance, in downstream tasks such as classification or image retrieval, this undesirable 'weighting' can cause the system to favor certain camera types, regardless of what the image actually shows.

The paper states:

> *'Metadata labels leaving traces in visual encoders to the point of overshadowing semantic information can lead to unpredictable outcomes, compromising generalizability, robustness, and potentially undermining the trustworthiness of the models.*
>
> *'More critically, this effect could be exploited maliciously; for instance, an adversarial attack may manipulate metadata to intentionally mislead or deceive a model, posing risks in sensitive domains like healthcare, surveillance, or autonomous systems.'*

The paper finds that Contrastive Visual-Language (CVL) systems such as CLIP, now one of the most influential encoders in computer vision, are particularly likely to obtain such inferences from the data:

*Search results for a query image, showing how foundation models rank similar images based not only on visual content but also on hidden metadata such as JPEG compression or camera model.*

The new paper is titled *Processing and acquisition traces in visual encoders: What does CLIP know about your camera?*, and comes from six researchers across The University of Osaka and Czech Technical University in Prague.

## Method and Data*

To test the influence of hidden metadata on visual encoders such as CLIP, the authors worked with two categories of metadata: image processing parameters (such as JPEG compression or sharpening) and acquisition parameters (such as camera model or exposure settings).

Rather than train new models, the researchers evaluated 47 widely used visual encoders in their frozen, pretrained state, including contrastive vision-language models such as CLIP, self-supervised models such as DINO, and conventionally supervised networks.

For processing parameters, the researchers applied controlled transformations to the ImageNet and iNaturalist 2018 datasets, including six levels of JPEG compression, three sharpening settings, three resizing scales, and four interpolation methods.

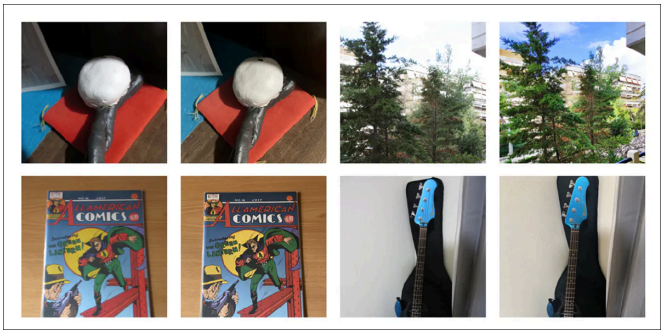
Examples of images and associated annotations from the iNaturalist dataset. Source: https://arxiv.org/pdf/1707.06642

The models were tested on their ability to recover each transformation setting using only the image content, with successful predictions indicating that the encoder retains information about these processing choices in its internal representation.

To examine acquisition parameters, the researchers compiled a 356,459-image dataset called *FlickrExif*, containing preserved Exif metadata, and constructed a second dataset called *PairCams*, made up of 730 image pairs captured simultaneously with a smartphone and a non-smartphone camera.

The FlickrExif dataset was built using the Flickr API to download images with accompanying Exif metadata. Between 2,000 and 4,000 safe-for-work images were collected per month, dated from early 2000 through mid-2024, and filtered to include only those with permissive licenses. To prevent overrepresentation by prolific users, each individual contributor was limited to ten images per month for any given year.

For the PairCams dataset, every photo was taken using automatic settings and no flash, allowing for a comparison of the way that visual encoders respond to differences in camera hardware alone, regardless of the image content:



*Further examples from the PairCams dataset curated by the authors.*

The authors tested for two sets of parameters: image processing parameters, such as compression and color transformations; and image acquisition parameters, such as camera make or model:

| parameter | type | class | description |
|---|---|---|---|
| JPEG compression | proc. | 6 | amount of JPEG compression |
| sharpening | proc. | 3 | amount of sharpening |
| resizing | proc. | 3 | amount of resizing |
| interpolation | proc. | 4 | type of interpolation during resize |
| make | acq. | 9 | manufacturer of the camera |
| model (all) | acq. | 88 | specific camera model used |
| model (smart) | acq. | 12 | specific smartphone used |
| model (smart vs non-smart) | acq. | 2 | whether camera is a smartphone |
| exposure | acq. | 16 | amount of light captured by sensor |
| aperture | acq. | 17 | size of the opening in the lens |
| ISO speed | acq. | 16 | camera sensor's sensitivity to light |
| focal length | acq. | 13 | distance from lens to sensor |

*Image processing and acquisition parameters analyzed, with number of classes for each.*

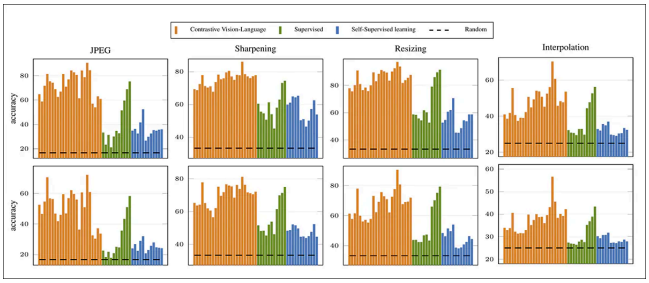
## Tests

To determine whether information about image processing and camera type is preserved inside visual encoder embeddings, the authors trained a classifier to predict metadata labels directly from those embeddings. If the classifier performed no better than random guessing, it would suggest that details about processing or device are not captured by the model.

However, any performance above chance would indicate that these technical traces are indeed being encoded, and could influence downstream tasks.

To test for processing traces, the authors assigned each training image a random processing setting, such as a particular JPEG compression level, while all test images in a batch shared the same setting.

Averaged classification accuracy across all settings was then combined with repeated trials under different random seeds, so that it could be determined whether technical details of image processing were consistently captured in the model's internal representation:




*Classification accuracy for predicting image-processing parameters from encoder embeddings, using a linear classifier applied to frozen models. Results are shown for JPEG compression, sharpening, resizing, and interpolation, with three model categories, contrastive vision-language (orange), supervised (green), and self-supervised (blue), evaluated on ImageNet (top row) and iNaturalist 2018 (bottom row). Random-guessing baselines are marked with dashed lines.*

Across all four processing parameters, contrastive vision-language models showed the highest ability to recognize hidden image manipulations. Some of the models achieved more than 80% accuracy when predicting JPEG compression, sharpening, and resizing settings from ImageNet embeddings.

Supervised encoders, particularly those based on ConvNeXt, also performed strongly, whereas self-supervised models were consistently weaker.

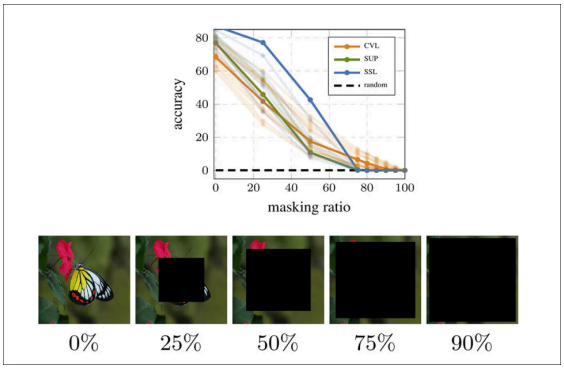
Interpolation was the most difficult parameter to detect, yet the top CVL and supervised models still achieved results well above the random baseline of 25% on both datasets.

Next, to test whether camera-related information is embedded in model representations, the authors created separate training and test sets for each acquisition parameter (such as camera make, camera model, exposure, aperture, ISO, and focal length).

For most parameters, only classes with at least 5,000 examples were used; 500 images were randomly set aside for testing, and the remaining examples were downsampled so that every class had 200 training samples. For the 'model (all)' and 'model (smart)' parameters, which had less data per class, the authors instead used classes with at least 500 images, and split each class into *train* and *test* subsets at a four-to-one ratio.

Photographers were kept separate across training, validation, and test sets, and a simple classifier was trained to predict camera information based on the image features.

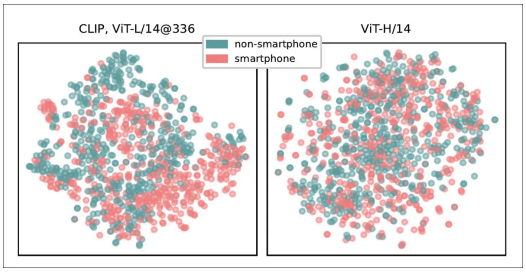
To ensure that the classifier was not influenced by the semantic content of the images, 90% of each image was center-masked (see examples below). The authors assert that at this level of masking, all visual encoders perform close to random on ImageNet, indicating that the semantic signal has been effectively suppressed:




*ImageNet validation accuracy as a function of masking ratio. At 90% masking, all models drop to near-random performance on semantic label prediction, indicating that semantic cues have been effectively removed. The example images at the bottom illustrate the masking levels.*

Even with 90% of each image masked, most contrastive vision-language models and the supervised ConvNeXt encoders still predicted camera-related labels at well above chance levels. Many CVL models exceeded 70% accuracy in distinguishing smartphone from non-smartphone images.

Other supervised encoders, SigLIP, and all self-supervised models performed much worse. When no masking was applied, CVL models again showed the strongest clustering by camera type, confirming that these models embed acquisition information more deeply than the others:

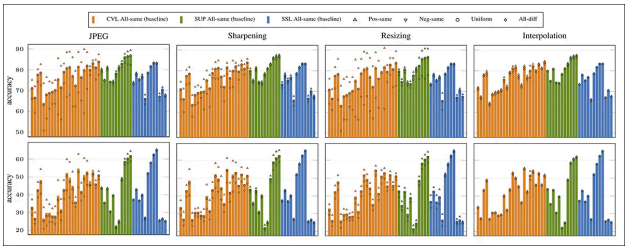



*t-SNE visualizations for two visual encoders, with colors indicating whether each image was captured by a smartphone or a non-smartphone camera.*

## Downstream Significance

Having established that metadata influences the models in this way, the propensity for hidden processing traces to interfere with image interpretation was then evaluated.

When two versions of the same image were processed differently, embeddings were often organized according to the *processing style* rather than the content. In several cases, a heavily compressed photo of a dog was treated as more similar to an unrelated image with the same compression setting than to its own uncompressed version:




*Impact of processing parameters on semantic prediction, featuring semantic classification accuracy for ImageNet (top) and iNaturalist (bottom) under five processing setups. In the baseline, all training and test images share the same processing label; in the all-diff setting, the test image uses a processing value not present in the training set; in pos-same and neg-same, the processing label is aligned either with semantically similar or dissimilar images; in the uniform setting, processing labels are randomly assigned across the training set. Results are reported using k = 10 for ImageNet, and k = 1 for iNaturalist.*

The strongest distortions were caused by JPEG compression, followed by sharpening and resizing, whereas interpolation produced only a minor effect. The authors assert that these results demonstrate that processing traces can override semantic information and dictate how an image is understood.

In conclusion, they warn:

> *'While we have identified that metadata labels are encoded in foundational visual encoders and provided hints about potential causes, we cannot definitively pinpoint the source of the problem. Investigating this further is challenging due to the cost of retraining such models and the frequent use of private datasets and undisclosed implementation details.*
>
> *'Although we do not propose specific mitigation techniques, we highlight the issue as an important area for future research.'*

## Conclusion

In the literature there is a growing forensic interest with regard to the traces and signs of 'method over content'; the easier it is to identify a framing domain or a specific dataset, the easier it is to leverage this information in the form of – for instance – deepfake detectors, or systems designed to categorize the provenance or age of data and models.

This all goes against the core intent of training AI models, which is that central distilled concepts should be curated independent of the means of production, and should bear no trace of them. In fact, datasets and capture devices have characteristics and domain traits that are effectively impossible to separate from content, because in themselves, they also represent a 'historical perspective'.

* *The paper is laid out unconventionally, and we will adapt as best we can to its unusual formatting and presentation. A great deal of material that should have been in a (non-existent) 'Method' section has been shunted to diverse parts of the appendix, presumably to restrict the main paper to eight pages – though at the considerable expense of clarity. If we have missed any opportunity to improve this, due to lack of time, we apologize.*

*First published Wednesday, August 20, 2025*

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**