Dissecting Image Crops

Basile Van Hoorick
Columbia University
basile@cs.columbia.edu

Carl Vondrick
Columbia University
vondrick@cs.columbia.edu

Abstract

The elementary operation of cropping underpins nearly every computer vision system, ranging from data augmentation and translation invariance to computational photography and representation learning. This paper investigates the subtle traces introduced by this operation. For example, despite refinements to camera optics, lenses will leave behind certain clues, notably chromatic aberration and vignetting. Photographers also leave behind other clues relating to image aesthetics and scene composition. We study how to detect these traces, and investigate the impact that cropping has on the image distribution. While our aim is to dissect the fundamental impact of spatial crops, there are also a number of practical implications to our work, such as revealing faulty photojournalism and equipping neural network researchers with a better understanding of shortcut learning. Code is available at https://github.com/basilevh/dissecting-image-crops.

1. Introduction

The basic operation of cropping an image underpins nearly every computer vision paper that you will be reading this week. Within the first few lectures of most introductory computer vision courses, convolutions are motivated as enabling feature invariance to spatial shifts and cropping [52, 31, 2]. Neural networks rely on image crops as a form of data augmentation [28, 50, 21]. Computational photography applications will automatically crop photos in order to improve their aesthetics [47, 12, 60]. Predictive models extrapolate pixels out from crops [51, 57, 55]. Even the latest self-supervised efforts depend on crops for contrastive learning to induce rich visual representations [13, 20, 45, 49].

This core visual operation can have a significant impact on photographs. As Oliva and Torralba told us twenty years ago, scene context drives perception [44]. Recently, image cropping has been at the heart of media disinformation. Figure 1 shows two popular photographs where the photographer or media organization spatially cropped out part of the context, altering the message of the image. Twitter's autocrop feature relied on a saliency prediction network that was racially biased [10].

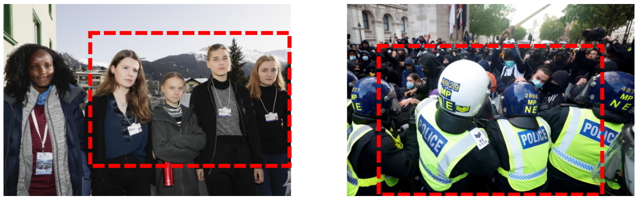


Figure 1: We show two infamous image crops, visualized by the red box. (**left**) An Ugandan climate activist had been cropped out of the photo before it was posted in an online news article, the discovery of which sparked controversy [16]. (**right**) A news network had cropped out a large stick being held by a demonstrator during a protest [14]. Cropping dramatically alters the message of the photographs.

The guiding question of this paper is to understand the traces left behind from this fundamental operation. What impact does image cropping have on the visual distribution? Can we determine when and how a photo has been cropped?

Despite extensive refinements to the manufacturing process of camera optics and sensors, nearly every modern camera pipeline will leave behind subtle lens artefacts onto the photos that it captures. For example, vignetting is caused by a lens focusing more light at the center of the sensor, creating images that are slightly brighter in the middle than near its borders [36]. Chromatic aberration, also known as purple fringing, is caused by the lens focusing each wave length differently [5]. Since these artefacts are correlated with their spatial position in the image plane, they cause image crops to have trace signatures.

Physical aberrations are not the only traces left behind during the operation. Photographers will prefer to take photos of interesting objects and in canonical poses [53, 4, 22]. Aesthetically pleasing shots will have sensible compositions that respect symmetry and certain ratios in the scene. Violating these principles leaves behind another trace of the cropping operation.

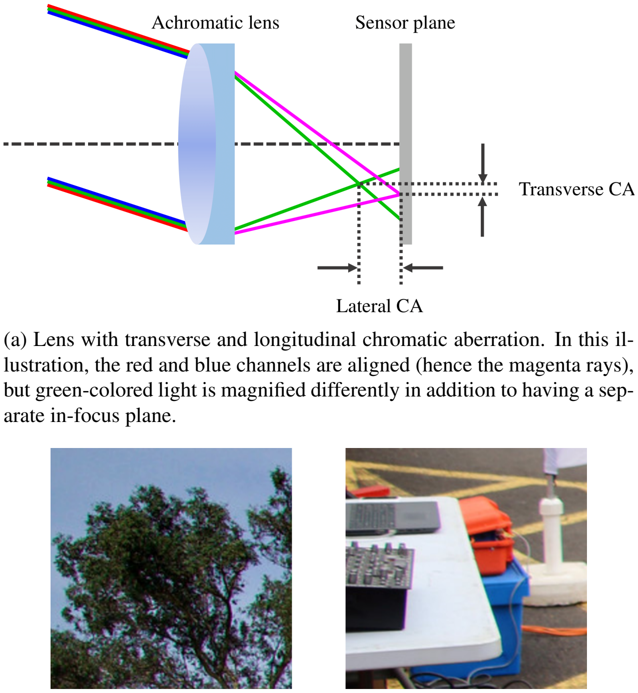
These traces are very subtle, and the human eye often cannot detect them, which makes studying and characterizing them challenging. However, neural networks are excellent at identifying these patterns. Indeed, extensive effort goes into preventing neural networks from learning such shortcuts enabled by image crops [15, 43].

In this paper, we flip this around and declare that these shortcuts are not bugs, but instead an opportunity to dissect and understand the subtle clues left behind from image cropping. Capitalizing on a large, high-quality collection of natural images, we train a convolutional neural network to predict the absolute spatial location of a patch within an image. This is only possible if there exist visual features that are *not* spatially invariant. Our experiments analyze the types of features that this model learns, and we show that it is possible to detect traces of the cropping operation. We can also use the discovered artefacts, along with semantic information, to recover where the crop was positioned in the original sensor plane.

While the aim of this paper is to analyze the fundamental traces of image cropping in order to question conventional assumptions about translational invariance and the crucial role of data augmentation pervading the field, we believe our investigation could have a large practical impact as well. Historically, asking fundamental questions has spurred significant insight into core computer vision problems, such as invariances to scale [9], asymmetries in time [46], the speediness of videos [6], and visual chirality [34]. For example, insight into image crops could enable detection of soft tampering operations, or spur developments to mitigate shortcut learning.

2. Background and Related Work

Optical aberrations. No imaging device is perfect, and every step in the imaging formation pipeline will leave traces behind onto the final picture. The origins of these signatures range from the physics of light in relation to the camera hardware, to the digital demosaicing and compression algorithms used to store and reconstruct the image. Lenses typically suffer from several aberrations, including chromatic aberration, vignetting, coma, and radial distortion [26, 5, 33, 24]. As shown in Figure 2a, chromatic aberration is manifested in two ways: *transverse* (or *lateral*) chromatic aberration (TCA) refers to the spatial discrepancies in focus points across color channels perpendicular to the optical axis, while *longitudinal* chromatic aberration (LCA) refers to shifts in focus along the optical axis instead [23, 24]. TCA gives rise to color channels that appear to be scaled slightly differently relative to each other, while LCA causes the distance between the focal surface and the lens to be frequency-dependent, such that the degree of blurring varies among color channels. Chromatic aberration can be leveraged to extract depth maps from defocus blur [19, 54, 24], although the spatial sensitivity of these cues is often undesired [15, 42, 43, 40]. TCA is leveraged by [59] to measure the angle of an image region relative to the lens as a means to detect cropped images. We instead present a learning-based approach that discovers additional clues without the need for carefully tailored algorithms.



(a) Lens with transverse and longitudinal chromatic aberration. In this illustration, the red and blue channels are aligned (hence the magenta rays), but green-colored light is magnified differently in addition to having a separate in-focus plane.

(b) Close-up of two photos, revealing visible transverse chromatic aberration (TCA) artefacts.

Figure 2: The origin behind, and examples of, chromatic aberration.

Patch localization. While one of the first major works in self-supervised representation learning focused on predicting the *relative* location of two patches among eight possible configurations [15], it was also discovered that the ability to perform *absolute* localization seemed to arise out of chromatic aberration. For the best-performing 10% of images, the mean Euclidean distance between the ground truth and predicted positions of single patches is 31% lower than chance, and this gap narrowed to 13% if every image was pre-processed to remove color information along the green-magenta axis. Although there are reasons to believe that modern network architectures might perform better, these rather modest performance figures suggest a priori that the attempted task is a difficult one. Note that the learnability of absolute location is often regarded as a bug; treatments used in practice include random color channel dropping [15], projection [15], grayscale conversion [42, 43], jittering [42], and chroma blurring [40].

Visual crop detection. In the context of forensics, almost all existing research has centered around 'hard' tampering such as splicing and copy-move operations. We argue that some forms of 'soft' tampering, notably cropping, are also worth investigating. While a few papers have addressed image crops [59, 39, 17], they are typically tai-

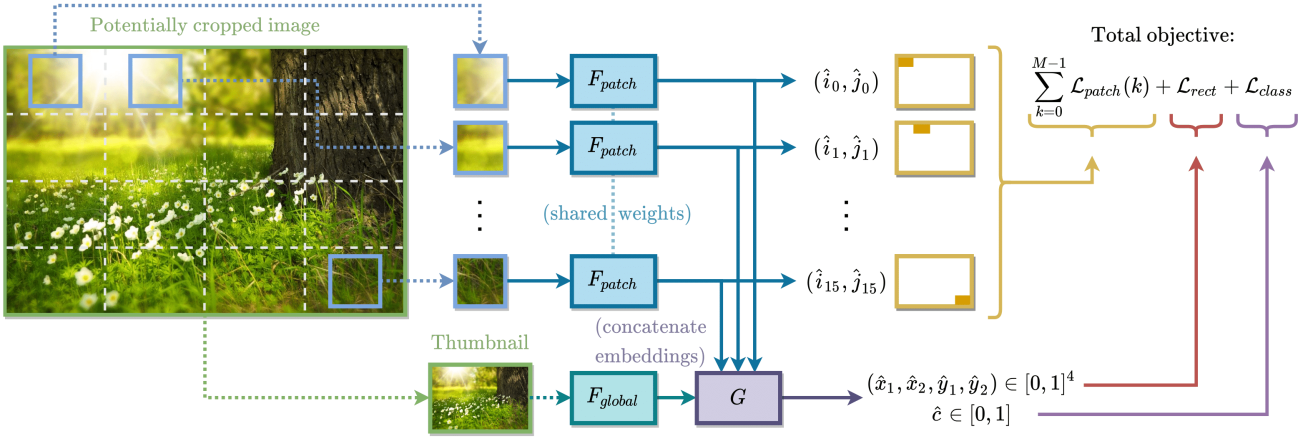

Potentially cropped image
F_{patch}
F_{patch}
(shared weights)
F_{patch}
(concatenate embeddings)
Thumbnail
F_{global}
G
$(\hat{i}_0, \hat{j}_0)$
$(\hat{i}_1, \hat{j}_1)$
$(\hat{i}_{15}, \hat{j}_{15})$
Total objective:
$\sum_{k=0}^{M-1} \mathcal{L}_{patch}(k) + \mathcal{L}_{rect} + \mathcal{L}_{class}$
$(\hat{x}_1, \hat{x}_2, \hat{y}_1, \hat{y}_2) \in [0,1]^4$
$\hat{c} \in [0,1]$

Figure 3: **Full architecture of our crop detection model.** We first extract $M = 16$ patches from the centers of a regularly spaced grid within the source image, a priori not knowing whether it is cropped or not. The patch-based network F_{patch} looks at each patch and classifies its absolute position into one out of 16 possibilities, whereby the estimation is mostly guided by low-level lens artefacts. The global image-based network, F_{global} instead operates on the downscaled source image, and tends to pick up semantic signals, such as objects deviating from their canonical pose (*e.g.* a face is cut in half). Since these two networks complement each other's strengths and weaknesses, we integrate their outputs into one pipeline via the multi-layer perceptron G. Note that F_{patch} is supervised by all three loss terms, while F_{global} only controls the crop rectangle $(\hat{x}_1, \hat{x}_2, \hat{y}_1, \hat{y}_2)$ and the final score $\hat{c}$.

lored toward specific types of pictures only. For example, both [39] and [17] rely heavily on structured image content in the form of vanishing points and lines, which works only if many straight lines (*e.g.* man-made buildings or rooms) are prominently visible in the scene. Various previous works have also explored JPEG compression, and some have found that it may help reveal crops under specific circumstances, mostly by characterizing the regularity and alignment of blocking artefacts [32, 8, 41]. In contrast, our analysis focuses on camera pipeline artefacts and photography patterns that exist independently of digital post-processing algorithms.

3. Dataset

The natural clues for detecting crops are subtle, and we need to be careful to preserve them when constructing a dataset. Our underlying dataset has around $700,000$ high-resolution photos from Flickr, which were scraped during the fall of 2019. We impose several constraints on the training images, most importantly that they should not already have been cropped and that they must maintain a constant, fixed aspect ratio and resolution. Appendix A describes this selection and collection process in detail.

We generate image crops by first defining the *crop rectangle* $(x_1, x_2, y_1, y_2) \in [0, 1]^4$ as the relative boundaries of a cropped image within its original camera sensor plane, such that $(x_1, x_2, y_1, y_2) = (0, 1, 0, 1)$ for unmodified images. We always maintain the aspect ratio and pick a random size factor f uniformly in $[0.5, 0.9]$, representing the relative length in pixels of any of the four sides compared to the original photo: $f = x_2 - x_1 = y_2 - y_1$.

After randomly cropping exactly half of all incoming photos, we give our model access to small image patches as well as global context. We select square patches of size 96×96 (*i.e.* around 5% of the horizontal image dimension), which is sufficiently large to allow the network to get a good idea of the local texture profile, while also being small enough to ensure that neighbouring patches never overlap. In addition, we downscale the whole image to a 224×149 thumbnail, such that it remains accessible to the model in terms of its receptive field and computational efficiency.[1]

Interrelating contextual, semantic information to its spatial position within an image might turn out to be crucial for spotting crops. We therefore add coordinates as two extra channels to the thumbnail, similarly to [35]. Note that the model does not know a priori whether its input had been cropped or not. Lastly, several shortcut fuzzing procedures had to be used to ensure that the learned features are generalizable; see Appendix B for an extensive description.

4. Approach

We describe our methodology and the challenges associated with revealing whether and how a variably-sized single image has been cropped. First, we construct a neural network that can trace image patches back to their original position relative to the center of the lens. Then, we use this

[1] The reason we care about receptive field is because, even though high-resolution images are preferable when analyzing subtle lens artefacts, a ResNet-L with $L \leq 50$ has a receptive field of only ≤ 483 pixels [3], which pushes us to prefer *lower* resolutions instead.

novel network to expose and analyze possibly incomplete images using an end-to-end trained crop detection model, which also incorporates the global semantic context of an image in a way that can easily be visualized and understood. Figure 3 illustrates our method.

4.1. Predicting absolute patch location

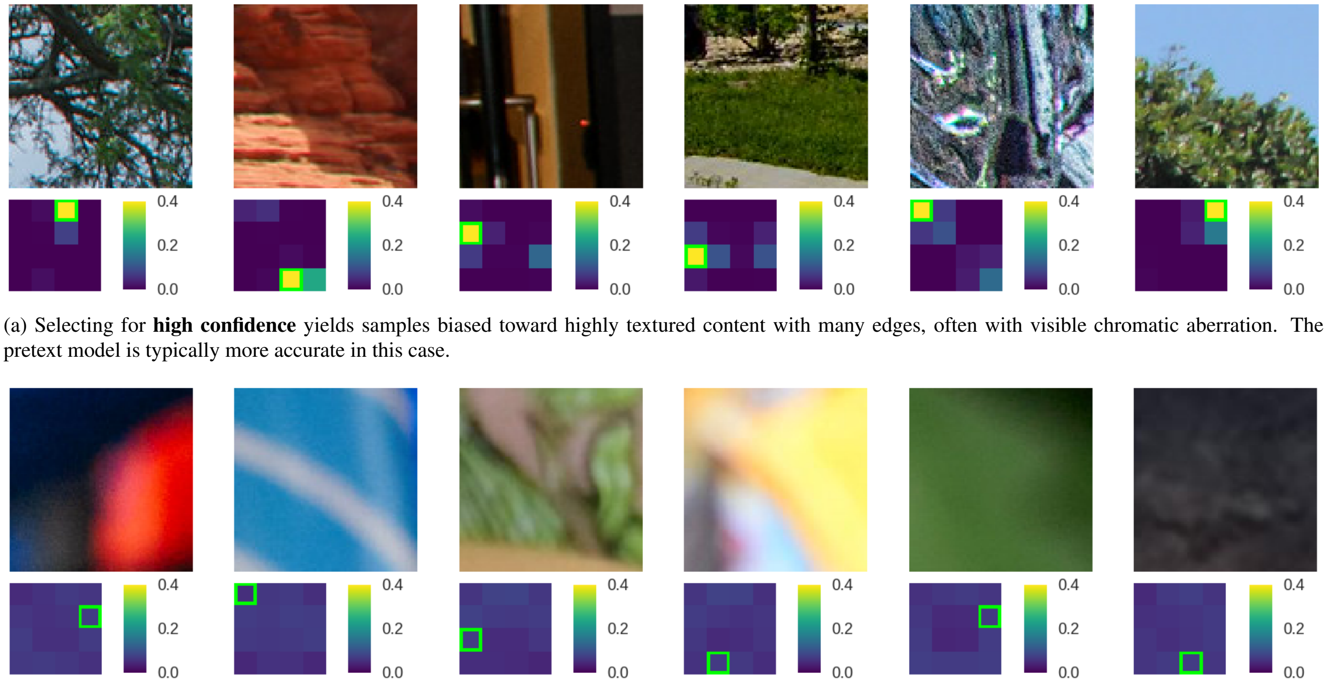
One piece of the puzzle towards analyzing image crops is a neural network called F_{patch}, which discriminates the original position of a small image patch with respect to the center of the lens. We frame this as a classification problem for practical purposes, and divide every image into a grid of 4×4 evenly sized cells, each of which represents a group of possible patch positions. Since this pretext task can be considered to be a form of self-supervised representation learning, with crop detection being the eventual downstream task, we call F_{patch} the *pretext model*.

But before embarking on an end-to-end crop detection journey that simply integrates this module into a larger system right from the beginning, it is worth asking the following questions: When exactly does absolute patch localization work well in the first place, and how could it help in distinguishing cropped images in an interpretable manner? To this end, we trained F_{patch} in isolation by discarding F_{global} and forcing the network to decide based on information from patches only. The 16-way classification loss term $\mathcal{L}_{patch}$ is responsible for pretext supervision, and is applied onto every patch individually.

Intriguing patterns emerge when discriminating between different levels of confidence in the predictions produced by F_{patch}. Although the accuracy of this localization network is not that high ($\sim$21% versus $\sim$6% for chance) due to the inherent difficulty of the task, Figure 4 shows that it works quite well for some images, particularly those with a high degree of detail coupled with apparent lens artefacts. On the flip side, blurry photos taken with high-end cameras tend to make the model uncertain. This observation suggests that chromatic aberration has strong predictive power for the original locations of patches within pictures. Hence, it is reasonable to expect that incorporating patch-wise, pixel-level cues into a deep learning-based crop detection framework will improve its capabilities.

4.2. Architecture and objective

Guided by the design considerations laid out so far, Figure 3 shows our main model architecture. F_{patch} is a ResNet-18 [21] that converts any patch into a length-64 embedding, which then gets converted by a single linear layer



(a) Selecting for **high confidence** yields samples biased toward highly textured content with many edges, often with visible chromatic aberration. The pretext model is typically more accurate in this case.

(b) Selecting for **low confidence** yields blurry or smooth samples, where the lack of detail makes it difficult to expose physical imperfections of the lens. The pretext model tends to be inaccurate in this case.

Figure 4: **Absolute patch localization performance.** By leveraging classification, an uncertainty metric emerges for free. Here, we display examples where the pretext model F_{patch} performs either exceptionally well or badly at recovering the patches' absolute position within the full image. The output probability distribution generated by the network is also plotted as a spatial heatmap (□ = ground truth).

on top to a length-16 probability distribution describing the estimated location $(\hat{i}_k, \hat{j}_k) \in \{0 \dots 3\}^2$ of that patch. F_{global} is a ResNet-34 [21] that converts the downscaled global image into another length-64 embedding. Finally, G is a 3-layer perceptron that accepts a 1088-dimensional concatenation of all previous embeddings, and produces 5 values describing (1) the crop rectangle $(\hat{x}_1, \hat{x}_2, \hat{y}_1, \hat{y}_2) \in [0,1]^4$, and (2) the actual probability $\hat{c}$ that the input image had been cropped. By simultaneously processing and combining aggregated patch-wise information with global context, we allow the network to draw a complete picture of the input, revealing both low-level lens aberrations and high-level semantic cues. The total, weighted loss function is as follows (with $M = 16$):

$$\mathcal{L} = \frac{\lambda_1}{M} \sum_{k=0}^{M-1} \mathcal{L}_{patch}(k) + \frac{\lambda_2}{4} \mathcal{L}_{rect} + \lambda_3 \mathcal{L}_{class} \tag{1}$$

Here, $\mathcal{L}_{patch}(k)$ is a 16-way cross-entropy classification loss between the predicted location distribution $\hat{\boldsymbol{l}}(k)$ of patch k and its ground truth location $\boldsymbol{l}(k)$. For an uncropped image, $\boldsymbol{l}(k) = k$ and $(i_k, j_k) = (k \mod 4, \lfloor k/4 \rfloor)$, although this equality obviously does not necessarily hold for cropped images. Second, the loss term $\mathcal{L}_{rect}$ encourages the estimated crop rectangle to be near the ground truth in a mean squared error sense. Third, $\mathcal{L}_{class}$ is a binary cross-entropy classification loss that trains $\hat{c}$ to state whether or not the photo had been cropped. More formally:

$$\mathcal{L}_{patch}(k) = \mathcal{L}_{CE}(\hat{\boldsymbol{l}}(k), \boldsymbol{l}(k)) \tag{2}$$

$$\begin{aligned} \mathcal{L}_{rect} = [&(\hat{x}_1 - x_1)^2 + (\hat{x}_2 - x_2)^2 \\ &+ (\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2] \end{aligned} \tag{3}$$

$$\mathcal{L}_{class} = \mathcal{L}_{BCE}(\hat{c}, c) \tag{4}$$

Note that the intermediate outputs $(\hat{i}_k, \hat{j}_k)$ and $(\hat{x}_1, \hat{x}_2, \hat{y}_1, \hat{y}_2)$ exist mainly to encourage a degree of interpretability of the internal representation, rather than to improve the accuracy of the final score $\hat{c}$. Specifically, the linear projection of F_{patch} to $(\hat{i}_k, \hat{j}_k)$ should make the embedding more sensitive to positional information, thus helping the crop rectangle estimation.

4.3. Training details

In our experiments, all datasets are generated by cropping exactly 50% of the photos with a random crop factor in $[0.5, 0.9]$. After that, we resize every example to a uniformly random width in $[1024, 2048]$ both during training and testing, such that the image size cannot have any predictive power. We train for up to 25 epochs using an Adam optimizer [27], with a learning rate that drops exponentially from $5 \cdot 10^{-3}$ to $1.5 \cdot 10^{-3}$ at respectively the first and last epoch. The weights of the loss terms are: $\lambda_1 = 2.4$, $\lambda_2 = 3$, and $\lambda_3 = 1$.

5. Analysis and Clues

We quantitatively investigate the model in order to dissect and characterize visual crops. We are interested in conducting a careful analysis of what factors the network might be looking at within every image. For ablation study purposes, we distinguish three variants of our model:

- **Joint** is the complete patch- and global-based model from Figure 3 central to this work;
- **Global** is a naive classifier that just operates on the thumbnail, *i.e.* the whole input downscaled to 224 × 149, using F_{global};
- **Patch** only sees 16 small patches extracted from consistent positions within the image, using F_{patch}.

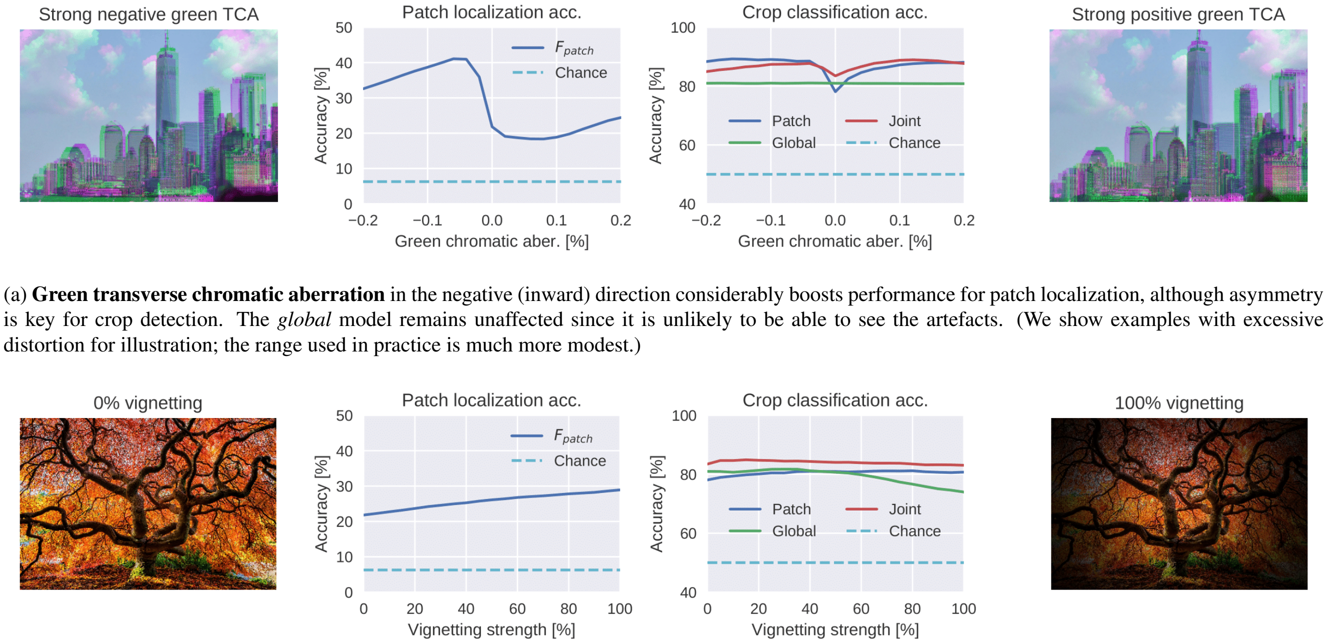
We classify the information that a model uses as evidence for its decision into two broad categories: **(1) characteristics of the camera or lens system**, and **(2) object priors**. While (1) is largely invariant of semantic image content, (2) could mean that the network has learned to leverage certain rules in photography, *e.g.* the sky is usually on top, and a person's face is usually centered.

To gain insight into what exactly our model has discovered, we first investigate the network's response to several known lens characteristics by artificially inflating their corresponding optical aberrations on the test set, and computing the resulting performance metrics. Next, we measure the changes in accuracy when the model is applied on datasets that were crafted specifically as to have divergent distributions over object semantics and image structure. We expect both lens flaws and photographic conventions to play different but interesting roles in our model.

A discussion of chromatic aberration expressed along the green channel, vignetting, and photography patterns follows; see Appendix C for the effect of color saturation, radial lens distortion, and chromatic aberration of the red and blue channels. Note that all discussed image modifications are applied *prior* to cropping, as a means of simulating a real lens that exhibits certain controllable defects.

5.1. Effect of chromatic aberration

A common lens correction to counter the frequency-dependence of the refractive index of glass is to use a so-called *achromatic doublet*. This modification ensures that the light rays of two different frequencies, such as the red and blue color channels, are aligned [26]. Because the remaining green channel still undergoes TCA and will therefore be slightly downscaled around the optical center, this artefact is often visible as green or purple fringes near edges and other regions with contrast or texture [7]. Figure 2b depicts real examples of what chromatic aberration looks like. Note that the optical center around which radial magnification occurs does not necessarily coincide with the image center due to the complexity of multi-lens systems [58],



(a) **Green transverse chromatic aberration** in the negative (inward) direction considerably boosts performance for patch localization, although asymmetry is key for crop detection. The *global* model remains unaffected since it is unlikely to be able to see the artefacts. (We show examples with excessive distortion for illustration; the range used in practice is much more modest.)

(b) **Vignetting** also contributes positively to the pretext model's accuracy. Interestingly, the crop detection performance initially increases but then drops slightly for strong vignetting, presumably because the distorted images are moving out-of-distribution.

Figure 5: **Breakdown of image attributes that contribute to features relevant for crop detection.** In these experiments, we manually exaggerate two characteristics of the lens on 3,500 photos of the test set, and subsequently measure the resulting shift in performance.

although both points have been found to be very close in practice [23]. Furthermore, chromatic aberration can vary strongly from device to device, and is not even present in all camera systems. Many high-end, modern lenses and/or post-processing algorithms tend to accurately correct for them, to the point that it becomes virtually imperceptible.

Nonetheless, our model still finds this spectral discrepancy in focus points to be a distinctive feature of crops and patch positions: Figure 5a (left plot) demonstrates that artificially downscaling the green channel significantly improves the pretext model's performance. This is because the angle and magnitude of texture shifts across color channels can give away the location of a patch relative to the center of the lens. Consequently, the downstream task of crop detection (right plot) becomes easier when TCA is introduced in either direction. Horizontally mirrored plots were obtained upon examining the red and blue channels, confirming that the green channel suffers an inward deviation most commonly of all in our dataset. It turns out that the optimal configuration from the perspective of F_{patch} is to add a little distortion, but not too much — otherwise we risk hurting the realism of the test set.

5.2. Effect of vignetting

A typical imperfection of multi-lens systems is the radial brightness fall-off as we move away from the center of the image, seen in Figure 5b. Vignetting can arise due to mechanical and natural reasons [36], but its dependence on the position within a photo is the most important aspect in this context. We simulate vignetting by multiplying every pixel value with $\frac{1}{g(r)}$, where:

$$g(r) = 1 + ar^2 + br^4 + cr^6 \tag{5}$$

$$(a, b, c) = (2.0625, 8.75, 0.0313) \tag{6}$$

$g(r)$ is a sixth-grade polynomial gain function, the parameters a, b, c are assigned typical values taken from [36], and r represents the radius from the image center with $r = 1$ at every corner. The degree of vignetting is smoothly varied by simply interpolating every pixel between its original (0%) and fully modified (100%) state.

Figure 5b shows that enhanced vignetting has a positive impact on absolute patch localization ability, but this does not appear to translate into noticeably better crop detection accuracy. While the gradient direction of the brightness across a patch is a clear indicator of the angle that it makes with respect to the optical center of the image, modern cameras appear to correct for vignetting well enough such that the lack of realism of the perturbed images hurts F_{global}'s performance more so than it helps.

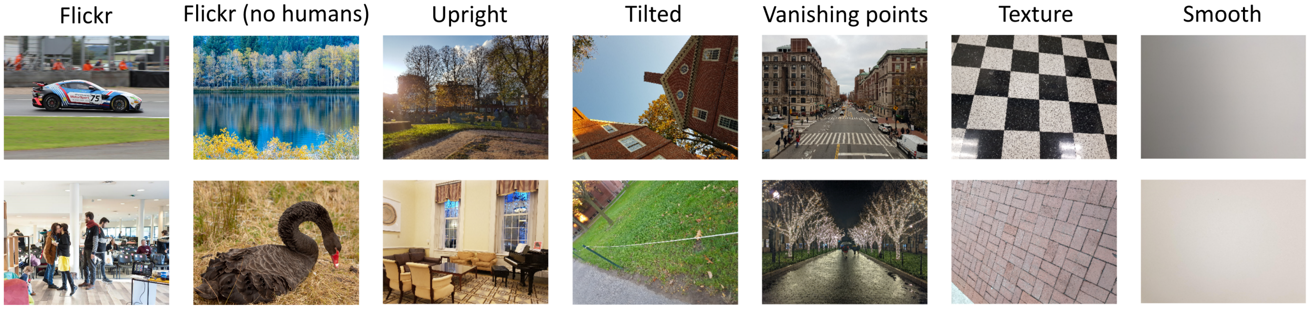


Figure 6: **Representative examples of the seven test sets.** The first two are variants of Flickr, one unfiltered and one without humans or faces, and the remaining five are custom photo collections we intend to measure various other kinds of photographic patterns or biases with. These were taken in New York, Boston, and SF Bay Area, and every category contains between 15 and 127 pictures.

Dataset	**Joint**	**Global**	**Patch**	**Human**
Flickr	86%	79%	77%	67%
Flickr (no humans)	81%	75%	73%	-
Upright	80%	72%	76%	-
Tilted	71%	58%	70%	-
Vanish	82%	75%	79%	-
Texture	66%	54%	67%	-
Smooth	50%	51%	55%	-

Table 1: **Accuracy comparison between three different crop detection models on various datasets.** All models are trained on Flickr, and appear to have discovered common rules in photography to varying degrees.

5.3. Effect of photography patterns and perspective

The desire to capture meaningful content implies that not all images are created equal. Interesting objects, persons, or animals will often intentionally be centered within a photo, and cameras are generally oriented upright when taking pictures. Some conventions, *e.g.* grass is usually at the bottom, are confounded to some extent by the random rotations during training, although there remain many facts to be learned as to what constitutes an appealing or sensible photograph. One clear example of these so-called *photography patterns* in the context of our model is that when a person's face that is cut in half, this might reveal that the image had been cropped. This is because, intuitively speaking, it does not conform to how photographers typically organize their visual environment and constituents of the scene.

The structure of the world around us not only provides high-level knowledge on where and how objects typically exist within pictures, but also gives rise to perspective cues, for example the angle that horizontal lines make with vertical lines upon projection of a 3D scene onto the 2D sensor, coupled with the apparent normal vector of a wall or other surface. Measuring the exact extent to which all of these aspects play a role is difficult, as no suitable dataset exists. The ideal baseline would consist of photos without any adherence to photography rules whatsoever, taken in uniformly random orientations at arbitrary, mostly uninteresting locations around the world.

We constructed and categorized a small-scale collection of such photos ourselves, using the Samsung Galaxy S8 and Google Pixel 4 smartphones, spanning the 5 right-most columns in Figure 6. Columns 3 and 5 depict photos that are taken with the camera in an upright, biased orientation. Column 5 specifically encompasses vanishing line-heavy content, where perspective clues may provide clear pointers. Columns 4, 6, and 7 contain pictures that are unlikely to be taken by a normal photographer, but whose purpose is instead to measure the response of our system on photos with compositions that make less sense.

Quantitative results are shown in Table 1. On the Flickr test set, the crop classification accuracy is 79% for the thumbnail-based model, 77% for the patch-based model, and 86% for the joint model. For comparison purposes, we also asked 16 people to classify 100 random Flickr photos into whether they look cropped or not, resulting in a human accuracy of 67%. This demonstrates that integrating information across multiple scales results in a better model than a network that only sees either patches or thumbnails independently, in addition to having a significant performance margin over humans.

Our measurements also indicate that the model tends to consistently perform better on sensible, upright photos. Analogous to what makes many datasets curated [11, 49], Flickr in particular seems to exhibit a high degree of photographic conventions involving people, so we also tested a manually filtered subset of 100 photos that do not contain humans or faces, resulting in a modest drop in accuracy. Interestingly, the patch-based network comes very close to the joint network on *tilted* and *texture*, suggesting that global context can sometimes confuse the model if the photo is taken in an abnormal way. Fully smooth, white-wall images appear to be even more out-of-distribution. However, most natural imagery predominantly contains canonical and appealing arrangements, where our model displays a stronger ability to distinguish crops.

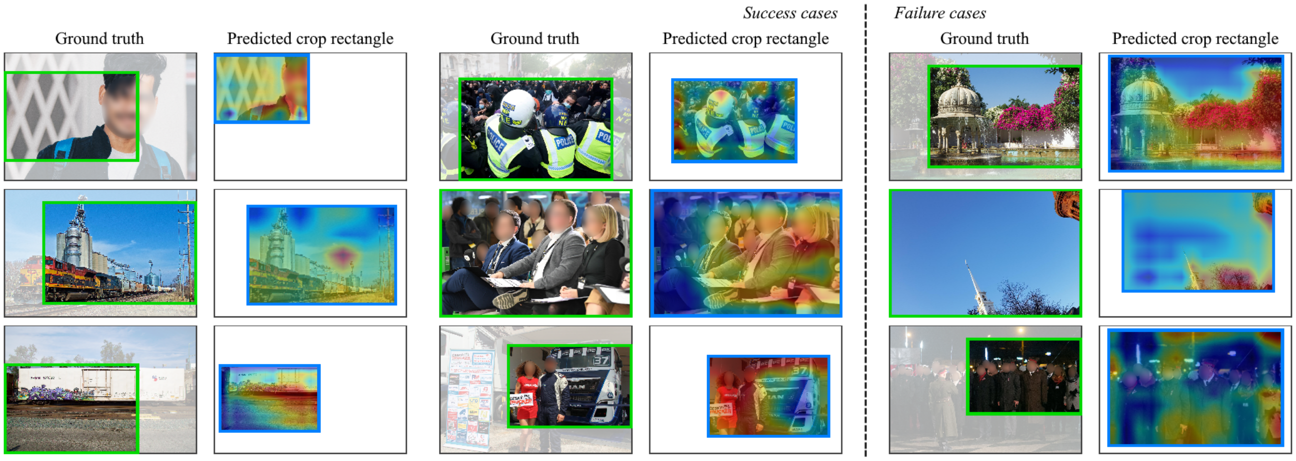


Figure 7: **Qualitative examples and interpretation of our crop detection system.** High-level cues such as persons and faces appear to considerably affect the model's decisions. Note that images don't always *look* cropped, but in that case, patches can act as the giveaway whenever they express lens artefacts. Regardless, certain scene compositions are more difficult to get right, such as in the failure cases shown on the right. (Faced blurred here for privacy protection.)

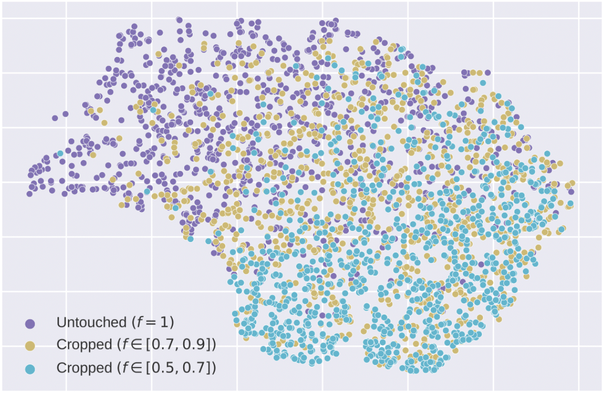


Figure 8: **Dimensionality-reduced embeddings generated by F_{global} on Flickr.** Here, the size factor f stands for the fraction of one cropped image dimension relative to the original photo. The model is clearly able to separate untampered from strongly cropped images, although lightly cropped images can land almost anywhere across the spectrum as the semantic signals might be less pronounced and/or less frequently present.

6. Visualizing Image Crops

In order to depict the changing visual distribution as images are cropped to an increasingly stronger extent, we look at the output embeddings produced by the thumbnail network F_{global}. In Figure 8, we first apply Principal Component Analysis (PCA) to transform the data points from 64 to 24 dimensions, and subsequently apply t-SNE [37] to further reduce the dimensionality from 24 to 2.

As discussed in the previous sections, there could be many reasons as to why the model predicts that a certain photo appears or does not appear to be cropped. However, to explain results obtained from any given single input, we can also apply the Grad-CAM technique [48] onto the global image. This procedure allows us to construct a heatmap that attributes decisions made by F_{global} and G back to the input regions that contributed to them.

Figure 7 showcases a few examples, where we crop untouched images by the green ground truth rectangle and subsequently feed them into the network to visualize its prediction. The model is often able to *uncrop* the image, using semantic and/or patch-based clues, and produce a reasonable estimate of which spatial regions are missing (if any). For example, the top left image clearly violates routine principles in photography. The top or bottom images are a little harder to judge by the same measure, though we can still recover the crop frame thanks to the absolute patch localization functionality.

7. Discussion

We found that image regions contain information about their spatial position relative to the lens, refining established assumptions about translational invariance [30]. Our network has automatically discovered various relevant clues, ranging from subtle lens flaws to photographic priors. These features are likely to be acquired to some extent by many self-supervised representation learning methods, such as contrastive learning, where cropping is an important form of data augmentation [13, 49]. Although they are often treated as a bug, there are also compelling cases where the clues could prove to be useful. For example, our crop detection and analysis framework has implications for revealing misleading photojournalism. We also hope that our work inspires further research into how the traces left behind by image cropping, and the altered visual distributions that it gives rise to, can be leveraged in other interesting ways.

References

[1] Opencv: Geometric image transformations. 12

[2] Alexander Amini and Ava Soleimany. Mit 6.s191: Introduction to deep learning, spring 2020. 1

[3] André Araujo, Wade Norris, and Jack Sim. Computing receptive fields of convolutional neural networks. *Distill*, 2019. https://distill.pub/2019/computing-receptive-fields. 3

[4] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, pages 9453–9463, 2019. 1

[5] Steven Beeson and James W Mayer. *Patterns of light: chasing the spectrum from Aristotle to LEDs*. Springer Science & Business Media, 2007. 1, 2

[6] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9922–9931, 2020. 2

[7] David Brewster and Alexander Dallas Bache. *A Treatise on Optics...: First American Edition, with an Appendix, Containing an Elementary View of the Application of Analysis to Reflexion and Refraction*. Carey, Lea, & Blanchard, 1833. 5

[8] AR Bruna, Giuseppe Messina, and Sebastiano Battiato. Crop detection through blocking artefacts analysis. In *International Conference on Image Analysis and Processing*, pages 650–659. Springer, 2011. 3

[9] Peter Burt and Edward Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on communications*, 31(4):532–540, 1983. 2

[10] Katie Canales. Twitter is making changes to its photo software after people online found it was automatically cropping out black faces and focusing on white ones, Oct 2020. 1

[11] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2959–2968, 2019. 7

[12] Jiansheng Chen, Gaocheng Bai, Shaoheng Liang, and Zhengqin Li. Automatic image cropping: A computational complexity study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 507–515, 2016. 1

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 1, 8

[14] London Broadcasting Company. Bbc criticised for cropping out weapon in black lives matter protest photo, Jun 2020. 1

[15] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 2, 12

[16] Sahar Esfandiari and Will Martin. Greta thunberg slammed the associated press for cropping a black activist out of a photo of her at davos, Jan 2020. 1

[17] Marco Fanfani, Massimo Iuliani, Fabio Bellavia, Carlo Colombo, and Alessandro Piva. A vision-based fully automated approach to robust image cropping detection. *Signal Processing: Image Communication*, 80:115629, 2020. 2, 3

[18] Alex Franz and Thorsten Brants. All our n-gram are belong to you, Aug 2006. 12

[19] Josep Garcia, Juan Maria Sanchez, Xavier Orriols, and Xavier Binefa. Chromatic aberration and depth extraction. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 1, pages 762–765. IEEE, 2000. 2

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1

[21] 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. 1, 4, 5, 13

[22] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019. 1

[23] Sing Bing Kang. Automatic removal of chromatic aberration from a single image. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2, 6

[24] Masako Kashiwagi, Nao Mishima, Tatsuo Kozakaya, and Shinsaku Hiura. Deep depth from aberration map. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4070–4079, 2019. 2

[25] Josh Kaufman. Github: first20hours/google-10000-english, Aug 2019. 12

[26] Michael J Kidger. Fundamental optical design. In *Fundamental optical design*. SPIE Bellingham, 2001. 2, 5

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1

[29] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018. 11

[30] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 8

[31] Fei-Fei Li, Ranjay Krishna, and Danfei Xu. Stanford cs231n: Convolutional neural networks for visual recognition, spring 2020. 1

[32] Weihai Li, Yuan Yuan, and Nenghai Yu. Passive detection of doctored jpeg image via block artifact grid extraction. *Signal Processing*, 89(9):1821–1829, 2009. 3

[33] Xufeng Lin and Chang-Tsun Li. Image provenance inference through content-based device fingerprint analysis. In *Information Security: Foundations, Technologies and Applications*, pages 279–310. IET, 2018. 2, 14

[34] Zhiqiu Lin, Jin Sun, Abe Davis, and Noah Snavely. Visual chirality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12295–12303, 2020. 2

[35] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *Advances in Neural Information Processing Systems*, pages 9605–9616, 2018. 3

[36] Laura Lopez-Fuentes, Gabriel Oliver, and Sebastia Massanet. Revisiting image vignetting correction by constrained minimization of log-intensity entropy. In *International Work-Conference on Artificial Neural Networks*, pages 450–463. Springer, 2015. 1, 6

[37] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 8

[38] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. A full-image full-resolution end-to-end-trainable cnn framework for image forgery detection. *IEEE Access*, 8:133488–133502, 2020. 11

[39] Xianzhe Meng, Shaozhang Niu, Ru Yan, and Yezhou Li. Detecting photographic cropping based on vanishing points. *Chinese Journal of Electronics*, 22(2):369–372, 2013. 2, 3

[40] T Nathan Mundhenk, Daniel Ho, and Barry Y Chen. Improvements to context based self-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9339–9348, 2018. 2

[41] Hieu Cuong Nguyen and Stefan Katzenbeisser. Detecting resized double jpeg compressed images–using support vector machine. In *IFIP International Conference on Communications and Multimedia Security*, pages 113–122. Springer, 2013. 3

[42] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016. 2

[43] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5898–5906, 2017. 2, 12

[44] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001. 1

[45] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 1

[46] Lyndsey C Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Scholkopf, and William T Freeman. Seeing the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2035–2042, 2014. 2

[47] A Samii, R Měch, and Zhe Lin. Data-driven automatic cropping using semantic composition search. In *Computer graphics forum*, volume 34, pages 141–151. Wiley Online Library, 2015. 1

[48] 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. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[49] Ramprasaath R Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. *arXiv preprint arXiv:2012.04630*, 2020. 1, 7, 8

[50] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[51] Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T Freeman. Boundless: Generative adversarial networks for image extension. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10521–10530, 2019. 1

[52] James Tompkin. Brown cs231n: Csci 1430: Introduction to computer vision, spring 2020. 1

[53] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 1

[54] Pauline Trouvé, Frédéric Champagnat, Guy Le Besnerais, Jacques Sabater, Thierry Avignon, and Jérôme Idier. Passive depth estimation using chromatic aberration and a depth from defocus approach. *Applied optics*, 52(29):7152–7164, 2013. 2

[55] Basile Van Hoorick. Image outpainting and harmonization using generative adversarial networks. *arXiv preprint arXiv:1912.10960*, 2019. 1

[56] Todd Vorenkamp. Understanding crop factor. *B&H Explora*, 2016. 11

[57] Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1399–1408, 2019. 1

[58] Reg G Willson and Steven A Shafer. What is the center of the image? *JOSA A*, 11(11):2946–2955, 1994. 5

[59] Ido Yerushalmy and Hagit Hel-Or. Digital image forgery detection based on lens and sensor aberration. *International journal of computer vision*, 92(1):71–91, 2011. 2

[60] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5949–5957, 2019. 1