

EXIF as Language: Learning Cross-Modal Associations Between Images and Camera Metadata

Chenhao Zheng Ayush Shrivastava Andrew Owens
 University of Michigan

<https://hellomuffin.github.io/exif-as-language>

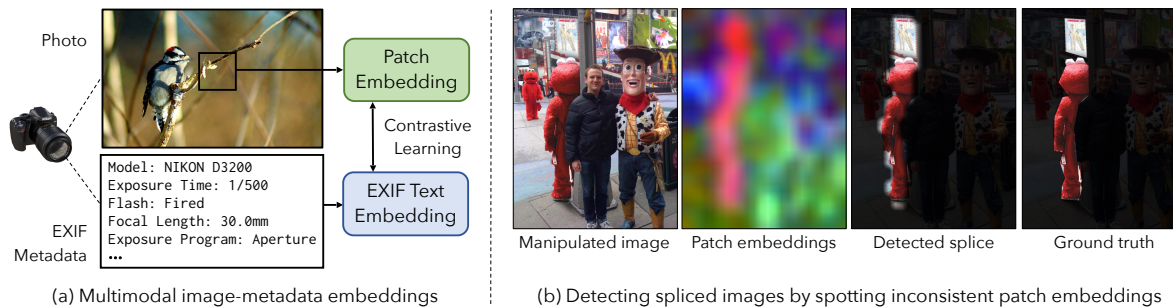


Figure 1. (a) We learn a joint embedding between image patches and the EXIF metadata that cameras automatically insert into image files. Our model treats this metadata as a language-like modality: we convert the EXIF tags to text, concatenate them together, and then processes the result with a transformer. (b) We apply our representation to tasks that require understanding camera properties. For example, we can detect image splicing “zero shot” (and without metadata at test time) by finding inconsistent embeddings within an image. We show a manipulated image that contains content from two source photos. Since these photos were captured with different cameras, the two regions have dissimilar embeddings (visualized by PCA). We localize the splice by clustering the image’s patch embeddings.

Abstract

We learn a visual representation that captures information about the camera that recorded a given photo. To do this, we train a multimodal embedding between image patches and the EXIF metadata that cameras automatically insert into image files. Our model represents this metadata by simply converting it to text and then processing it with a transformer. The features that we learn significantly outperform other self-supervised and supervised features on downstream image forensics and calibration tasks. In particular, we successfully localize spliced image regions “zero shot” by clustering the visual embeddings for all of the patches within an image.

1. Introduction

A major goal of the computer vision community has been to use cross-modal associations to learn concepts that would be hard to glean from images alone [2]. A particular focus has been on learning high level semantics, such as

objects, from other rich sensory signals, like language and sound [58, 62]. By design, the representations learned by these approaches typically discard *imaging* properties, such as the type of camera that shot the photo, its lens, and the exposure settings, which are not useful for their cross-modal prediction tasks [17].

We argue that obtaining a complete understanding of an image requires both capabilities — for our models to perceive not only the semantic content of a scene, but also the properties of the camera that captured it. This type of low level understanding has proven crucial for a variety of tasks, from image forensics [33, 52, 80] to 3D reconstruction [34, 35], yet it has not typically been a focus of representation learning. It is also widely used in image generation, such as when users of text-to-image tools specify camera properties with phrases like “DSLR photo” [59, 63].

We propose to learn low level imaging properties from the abundantly available (but often neglected) *camera metadata* that is added to the image file at the moment of capture. This metadata is typically represented as dozens of Exchangeable Image File Format (EXIF) tags that describe the camera, its settings, and postprocessing operations that

were applied to the image: e.g., Model: “iPhone 4s” or Focal Length: “35.0 mm”. We train a joint embedding through contrastive learning that puts image patches into correspondence with camera metadata (Fig. 1a). Our model processes the metadata with a transformer [75] after converting it to a language-like representation. To do this conversion, we take advantage of the fact that EXIF tags are typically stored in a human-readable (and text-based) format. We convert each tag to text, and then concatenate them together. Our model thus closely resembles contrastive vision-and-language models, such as CLIP [62], but with EXIF-derived text in place of natural language.

We show that our model can successfully estimate camera properties solely from images, and that it provides a useful representation for a variety of image forensics and camera calibration tasks. Our approaches to these tasks do not require camera metadata at test time. Instead, camera properties are estimated implicitly from image content via multimodal embeddings.

We evaluate the learned feature of our model on two classification tasks that benefit from a low-level understanding of images: estimating an image’s radial distortion parameter, and distinguishing real and manipulated images. We find that our features significantly outperform alternative supervised and self-supervised feature sets.

We also show that our embeddings can be used to detect image splicing “zero shot” (i.e., without labeled data), drawing on recent work [8, 33, 54] that detects inconsistencies in camera fingerprints hidden within image patches. Spliced images contain content from multiple real images, each potentially captured with a different camera and imaging pipeline. Thus, the embeddings that our model assigns to their patches, which convey camera properties, will have less consistency than those of real images. We detect manipulations by flagging images whose patch embeddings do not fit into a single, compact cluster. We also localize spliced regions by clustering the embeddings within an image (Fig. 1b).

We show through our experiments that:

- Camera metadata provides supervision for self-supervised representation learning.
- Image patches can be successfully associated with camera metadata via joint embeddings.
- Image-metadata embeddings are a useful representation for forensics and camera understanding tasks.
- Image manipulations can be identified “zero shot” by identifying inconsistencies in patch embeddings.

2. Related Work

Estimating camera properties. Camera metadata has been used for a range of tasks in computer vision, such as for predicting focal length [1, 31, 50, 71], performing white balancing [20, 49] and estimating camera models

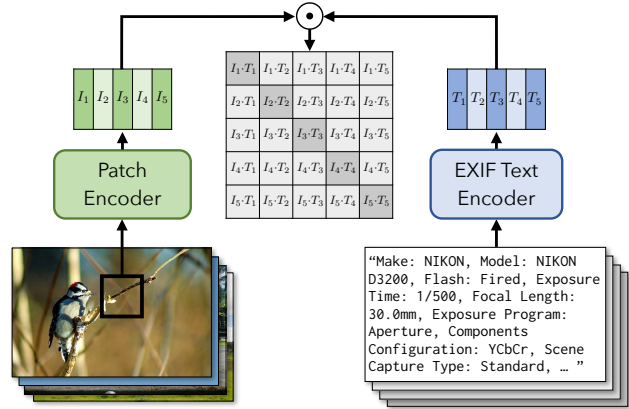


Figure 2. **Cross-modal image and camera metadata model.** We use contrastive learning to associate each image patch with the EXIF metadata that was extracted from its image file. We represent the metadata as text, which is obtained by concatenating the EXIF tags together. We then process it using a transformer.

[7, 39, 74]. It has also been used as extra input for recognition tasks [19, 72]. Instead of estimating camera properties directly (which can be highly error prone [33]), our model predicts an embedding that distinguishes a patch’s camera properties from that of other patches in the dataset.

Image forensics. Early work used physically motivated cues, such as misaligned JPEG blocks [6, 21], color filter array mismatches [3, 4, 23, 78], inconsistencies in noise patterns [37, 47, 48, 61], and compression or boundary artifacts [5, 26, 32, 82]. Other works use supervised learning methods [40, 53, 64, 66, 76, 77, 79–81, 85]. The challenge of collecting large datasets of fake images has led to alternative approaches, such as synthetic examples [53, 79, 81, 85]. Other work uses self-supervised learning, such as methods based on denoising [13], or that detect image manipulations by identifying image content that appears to come from different camera models [7, 9, 54]. Huh *et al.* [33] learned a patch similarity metric in two steps: they determined which EXIF tags are shared between the patches, then use these binary predictions as features for a second classifier that predicts whether two patches come from the same (or different) images. In contrast, we obtain a visual similarity metric that is well-suited to splice localization directly from our multimodal embeddings.

Language supervision in vision. Recent works have obtained visual supervision from language. The formulation includes specific keyword prediction [55], bag-of-word multilabel classification [36], n -gram classification [46] and autoregressive language models [16, 68, 84]. Recently, Radford *et al.* [62] obtained strong performance by training a contrastive model on a large image-and-language dataset. Our technical approach is similar, but uses text from camera metadata in lieu of image captions.

Work in text-to-image synthesis often exploits camera

information through prompting, such as by adding text like “DSLR photo of...” or “Sigma 500mm f/5” to prompts [59]. These methods, however, learn these camera associations through the (relatively rare) descriptions of cameras provided by humans, while ours learns them from an abundant and complementary learning signal, camera metadata.

3. Associating Images with Camera Metadata

We desire a visual representation that captures low level imaging properties, such as the settings of the camera that were used to shoot the photo. We then apply this learned representation to downstream tasks that require an understanding of camera properties.

3.1. Learning Cross-Modal Embeddings

We train a model to predict camera metadata from image content, thereby obtaining a representation that conveys camera properties. Following previous work in multimodal contrastive learning [62], we train a joint embedding between the two modalities, allowing our model to avoid the (error prone) task of directly predicting the attributes. Specifically, we want to jointly learn an image encoder and metadata encoder such that, given N images and N pieces of metadata information, the corresponding image–metadata pairs can be recognized by the model by maximizing embedding similarity. We use full-resolution image patches rather than resized images, so that our model can analyze low-level details that may be lost during down-sampling.

Given a dataset of image patches and their corresponding camera metadata $\{(\mathbf{v}_i, \mathbf{m}_i)\}_{i=1}^N$, we learn visual and EXIF representations $f_\theta(\mathbf{v})$ and $g_\phi(\mathbf{m})$ by jointly training f_θ and g_ϕ using a contrastive loss [57]:

$$\mathcal{L}_i^{V,M} = -\log \frac{\exp(f_\theta(\mathbf{v}_i) \cdot g_\phi(\mathbf{m}_i)/\tau)}{\sum_{j=1}^N \exp(f_\theta(\mathbf{v}_i) \cdot g_\phi(\mathbf{m}_j)/\tau)}, \quad (1)$$

where τ is a small constant. Following prior work [62], we define an analogous loss $\mathcal{L}^{M,V}$ that sums over visual (rather than metadata) examples in the denominator, and minimize a combined loss $\mathcal{L} = \mathcal{L}^{V,M} + \mathcal{L}^{M,V}$.

3.2. Representing the Camera Metadata

This formulation raises a natural question: how should we represent the metadata? The metadata within photos is stored as a set of EXIF tags, each indicating a different image property as shown in Table 1. EXIF tags span a range of formats and data types, and the set of tags that are present in a given photo can be highly inconsistent. Previous works that predict camera properties from images typically extract attributes of interest from the EXIF tags, and cast them to an appropriate data format — *e.g.*, extracting a scalar-valued focal length category. This tag-specific processing limits

EXIF tag	Example values	#values
Camera Make	Canon, NIKON Corporation, Apple	312
Camera Model	NIKON D90, Canon EOS 7	3071
Software	Picasa, Adobe Photoshop, QuickTime1711	
Exposure Time	1/60 sec, 1/125 sec, 1/250 sec	2062
Focal Length	18.0 mm, 50.0 mm, 6.3 mm	931
Aperture Value	F2.8, F4, F5.6, F3.5	137
Scene Capture Type	Landscape, Portrait, Night Scene	5
Exposure Program	Aperture priority, Manual control	9
White Balance Mode	Auto, Manual	3
Thumbnail Compression	JPEG, Uncompressed	3
Digital Zoom Ratio	1, 1.5, 2, 1.2	49
ISO speed Ratings	100, 400, 300	460
Shutter Speed Value	1/60 sec, 1/63 sec, 1/124 sec	1161
Date/Time Digitized	2013:03:28 04:20:46	95932

Table 1. **What information is contained within photo EXIF metadata?** We list several of the most common EXIF tags, along with the common values and number of values they contain in the YFCC100M dataset [73].

the amount of metadata information that can be used as part of learning, and requires special-purpose architectures.

We exploit the fact that EXIF tags are typically stored in a human-readable format and can be straightforwardly converted to text (Fig. 2). This allows us to directly process camera metadata using models from natural language processing — an approach that has successfully been applied to processing various text-like inputs other than language, such as math [45] and code [11]. Specifically, we create a long piece of text from a photo’s metadata by converting each tag’s name and value to strings, and concatenating them together. We separate each tag name and value with a colon and space, and separate different tags with a space. We evaluate a number of design decisions for this model in Sec. 4.4, such as the text format, choice of tags, and network architecture.

3.3. Application: Zero-shot Image Forensics

After learning cross-modal representations from images and camera metadata, we can use them for downstream tasks that require an understanding of camera properties. One way to do this is by using the learned visual network features as a representation for classification tasks, following other work in self-supervised representation learning [12, 29]. We can also use our learned visual embeddings to perform “zero shot” *image splice detection*, by detecting inconsistencies in an input image’s imputed camera properties.

Spliced images are composed of regions from multiple real images. Since they are typically designed to fool humans, forensic models need to rely more on subtle (often non-semantic) cues to detect them. We got inspiration from Huh *et al.* [33], which predicts whether two image patches share the same camera properties. If two patches are predicted to have very different camera properties, then this

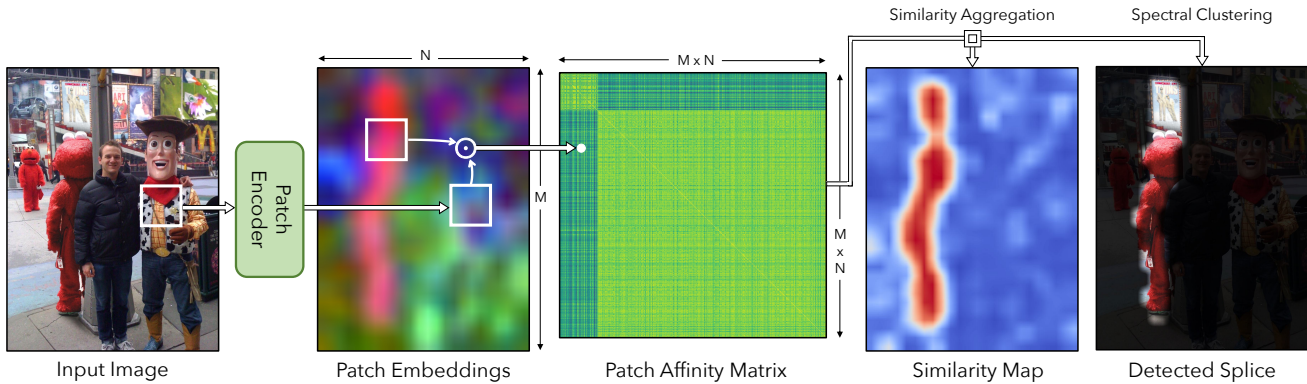


Figure 3. **Zero shot splice localization.** Given a spliced image (left), we compute our cross-modal embeddings for each image patch, which we visualize here using projections onto the top 3 principal components. We then compute the affinity matrix by taking dot product for every pair of patches. We localize the spliced region by clustering these embedding vectors.

provides evidence that they come from different images. In our work, we can naturally obtain this patch similarity by computing the dot product between two patches’ embeddings, since they have been trained to convey camera properties. We note that, unlike Huh *et al.* [33], we do not train a second, special-purpose classifier for this task, nor do we use augmentation to provide synthetic training examples (*e.g.*, by applying different types of compression to the patches).

To determine whether an image is likely to contain a splice, we first compute an affinity matrix $A_{ij} = f_{\theta}(\mathbf{v}_i) \cdot f_{\theta}(\mathbf{v}_j)$ whose entries are the dot product between patches’ normalized embedding vectors. We score an image \mathbf{v} using the sum of the exponentiated dot products between embeddings, $\phi(\mathbf{v}) = \sum_{i,j} \exp(A_{ij}/\tau)$. This score indicates the likelihood that the image is unmodified, since high dot products indicate high similarity in imputed camera properties. To localize the spliced image regions within an image, we aggregate the similarity scores in A_{ij} by clustering the rows using mean shift, following [33]. This results in a *similarity map* indicating the likelihood that each patch was extracted from the largest source photo that was used to create the composite. Alternatively, we can visualize the spliced region by performing spectral clustering via normalized cuts [33, 69], using A_{ij} as an affinity matrix between patches. We visualize this approach in Fig. 3.

4. Results

4.1. Implementation

Architecture. We use ResNet-50 pretrained on ImageNet as our image encoder. We found that the text encoder in models trained on captioning, such as CLIP [62], were not well-suited to our task, since they place low limits on the number of tokens. For the EXIF text encoder, we use DistilBERT [67] pretrained on Wikipedia and the Toronto Book Corpus [86]. We compute the feature representation of the EXIF as the activations for the end-of-sentence token from

the last layer which is layer normalized and then linearly projected into multi-modal embedding space.

Training. To train our model, we use 1.5M full-resolution images and EXIF pairs from a random subset of YFCC100M [73]. We discard images that have less than 10 of the EXIF tags. Because many images only have a small number of EXIF tags available, we only use tags that are present in more than half of these images. This results in 44 EXIF tags (see supplementary for the complete list). In contrast to other work [33], we do not rebalance the images to increase the rate of rare tags. During training, we randomly crop 124×124 patches from high-resolution images. We use the AdamW optimizer [38] with a learning rate of 10^{-4} , weight decay of 10^{-3} , and mixed precision training. We use a cosine annealing learning rate schedule [51]. The batch size is set to 1024, and we train our model for 50 epochs.

Other model variations. To study the importance of metadata supervision on the learned representation, we train a similar model that performs contrastive learning but does not use metadata. The model resembles image-image contrastive learning [12, 29, 33, 87], which has been shown to be highly effective for representation learning, and which may learn low-level camera information [17]. Different from typical contrastive learning approaches, we use strict cropping augmentation so that the views for our model (Eq. 1) come from different crops of the same image, to encourage it to learn low-level image features. We call this model **CropCLR**. Additionally, we evaluate a number of ablations of our model, including models that are trained with individual EXIF tags, that use different formats for the EXIF-derived text, and different network architectures (Table 5).

4.2. Evaluating the learned features

First, we want to measure how well the learned features convey camera properties. Since EXIF file is already embedded with a lot of camera properties such as camera model, focal length, shutter speed, etc., it should be unsur-

Models	Forensics				Radial Distortion	
	CASIA I		CASIA II		Dresden	RAISE
	resize	crop	resize	crop	resize	resize
ImageNet pretrained	0.69	0.64	0.71	0.72	0.23	0.24
MoCo	0.67	0.67	0.68	0.69	0.24	0.28
CLIP	0.71	0.82	0.84	0.81	0.21	0.22
Ours - CropCLR	0.70	0.81	0.86	0.80	0.28	0.32
Ours - Full	0.75	0.85	0.87	0.84	0.31	0.35

Table 2. We do linear probing on top of learned representation to predict two camera related properties that are not presented in EXIF files. The good performance indicates that our model learns general imaging properties. `resize` and `crop` denote the image preprocessing applied.

prising if we can predict those properties from images (we provide such results in Sec. 4.4). Instead, we want to study if the feature learned by the model can be generalized to other imaging properties that are not provided in the EXIF file. Specifically, we fit a linear classifier to our learned features on two prediction tasks: radial distortion estimation and forensic feature evaluation.

We compare the features from our image encoder with several other approaches, including supervised **ImageNet** pretraining [65], a state-of-the-art self-supervised model **MoCo** [29], **CLIP** [62], which obtains strong semantic representations using natural language supervision (rather than EXIF supervision), and finally the **CropCLR** variation of our model. To ensure a fair comparison, the backbone architectures for all approaches are the same (ResNet-50).

Radial distortion estimation. Imperfections in camera lens production often lead to radial distortion artifacts in captured images. These artifacts are often removed as part of multi-view 3D reconstruction [27, 70], using methods that model distortion as a 4th-order polynomial of pixel position. Radial distortion is not typically specified directly by the camera metadata, and is thus often must be estimated through calibration [10], bundle adjustment [70], or from monocular cues [50].

We followed the evaluation setup of Lopez et al. [50], which estimates the quadratic term of the radial distortion model, k_1 , directly from synthetically distorted images. This term can be used to provide an estimate of radial distortion that is sufficient for many tasks [50, 60]. We synthesized the 512×512 images from the Dresden Image Database [25] and RAISE dataset [14] using k_1 parameters uniformly sampled in the range $[-0.4, 0]$. To predict k_1 , we used a regression-by-classification approach, quantizing the values of k_1 into 20 bins (We also show result in regression RMSE metrics in supplementary). We extracted features from different models, and trained a linear classifier on this 20-way classification problem. We provided

them with a 512×512 image as input, and obtained image features from the global average pooling layer after the final convolutional layer (i.e., the penultimate layer of a typical ResNet).

Representation learning for image forensics. We evaluate our model’s ability to distinguish real and manipulated images. This is a task that requires a broader understanding of low level imaging properties, such as spotting unusual image statistics. We use the **CASIA I** [18] and **CASIA II** [43] datasets. The former contains only spliced fakes, while the latter contains a wider variety of manipulations. We again perform linear classification using the features provided by different models. We evaluate two types of preprocessing, resizing and center cropping, to test whether this low level task is sensitive to these details.

In both tasks, we found that our model’s features significantly outperformed those of the other models (Table 2). Our method achieves much better performance than traditional representational learning methods [29, 30, 62], perhaps because these models are encouraged to discard low-level details, while for our training task they are crucially important. Interestingly, the variation of our model that does not use EXIF metadata, CropCLR, outperforms the supervised [30] and self-supervised baselines [29], but significantly lags behind our full method. This is perhaps because it often suffices to use high-level cues (e.g. color histograms and object co-occurrence) to solve CropCLR’s pretext task. This suggests metadata supervision is an important learning signal and can effectively guide our model to learn general imaging information.

4.3. Zero Shot Splice Detection and Localization

We evaluate our model on the task of detecting spliced images without any labeled training data. This is in contrast to Sec. 4.2, which used labeled data. We perform both splice detection (distinguish an image being spliced or not) and splice localization (localize spliced region within an image).

Implementation. For fair evaluation, we closely follow the approach of Huh *et al.* [33]. Given an image, we sample patches in a grid, using a stride such that the number of patches sampled along the longest image dimension is 25. To increase the spatial resolution of each similarity map, we average the predictions of overlapping patches. We consider the smaller of the two detected regions to be the splice.

Evaluation. In splice localization task, we compare our model to a variety of forensics methods. These include traditional methods that use handcrafted features [24, 52, 83], supervised methods [41, 79, 80], and self-supervised approaches [13, 33]. The datasets we use include **Columbia** [56], **DSO** [15], **Realistic Tampering (RT)** [42], **In-the-Wild** [33] and **Hays and Efros** inpainting images [28]. Columbia and DSO are created purely

Style	Method	Columbia [56]		DSO [15]		RT [42]		In-the-Wild [33]		Hays [28]	
		p-mAP	cIoU	p-mAP	cIoU	p-mAP	cIoU	p-mAP	cIoU	p-mAP	cIoU
Handcrafted	CFA [24]	0.76	0.75	0.24	0.46	0.40	0.63	0.27	0.45	0.22	0.45
	DCT [83]	0.43	0.41	0.32	0.51	0.12	0.50	0.41	0.51	0.21	0.47
	NOI [52]	0.56	0.47	0.38	0.50	0.19	0.50	0.42	0.52	0.27	0.47
Supervised	MantraNet [80]	0.78	0.88	0.53	0.78	0.50	0.54	0.50	0.63	0.27	0.56
	MAG [41]	0.69	0.77	0.48	0.56	0.51	0.55	0.47	0.59	0.30	0.61
	OSN [79]	0.68	0.90	0.55	0.85	0.51	0.81	0.66	0.88	0.28	0.57
Unsupervised	Noiseprint [13]	0.71	0.83	0.66	0.90	0.29	0.80	0.50	0.78	0.22	0.53
	EXIF-SC [33]	0.89	0.97	0.47	0.81	0.22	0.75	0.49	0.79	0.26	0.54
	Ours - CropCLR	0.87	0.96	0.48	0.81	0.23	0.74	0.47	0.80	0.26	0.55
	Ours - Full	0.92	0.98	0.56	0.85	0.23	0.74	0.51	0.82	0.30	0.58

Table 3. **Zero shot splice localization.** We evaluate our model on several datasets using permutation-invariant mean average precision (p-mAP) over pixels and class-balanced IOU (cIoU) with optimal threshold selected per image. The result indicates that our model is comparable to state-of-the-art methods, although not specially optimized for this task.

Dataset	Columbia [56]	DSO [15]	RT [42]
CFA [24]	0.83	0.49	0.54
DCT [83]	0.58	0.48	0.52
NOI [52]	0.73	0.51	0.52
EXIF-SC	0.98	0.61	0.55
Ours - CropCLR	0.96	0.62	0.52
Ours - Full	0.99	0.66	0.53

Table 4. **Zero-shot splice detection:** We compare our splice detection accuracy on 3 datasets. We measure the mean average precision (mAP) of detecting whether an image has been spliced.

via image splicing, while Realistic Tampering contains a diverse set of manipulations. In-the-Wild is a splicing image dataset composed of internet images, which may also contain a variety of other manipulations. Hays and Efros [28] perform data-driven image inpainting. The quantitative comparison in terms of permuted-mAP (p-mAP) and class-balanced IoU (cIoU) following [33] are presented in Table 3. We also include splice image detection result in Table 4, where we compare our model to methods that enable splice detection.

Our model ranks first or second place for metrics in most datasets, and obtains performance comparable to top self-supervised methods that are specially designed for this task. In particular, our model significantly outperforms the most related technique, EXIF-SC [33]. We note that both our method and EXIF-SC get relatively low performance on the Realistic Tampering dataset. This may be due to the fact that this dataset contains manipulations such as copy-move that we do not expect to detect (since both regions share the same camera properties). In contrast to methods based on segmentation [13, 79, 80], we do not aim to have spatially precise matches, and output relatively low-resolution localization maps based on large patches. Consequently, our model is not well-suited to detecting very small

splices, which also commonly occur in the Realistic Tampering dataset.

In Fig. 4, we show qualitative results, including both similarity maps and spectral clustering results. In Fig. 6, we compare our model with those of several other techniques. Interestingly, EXIF-SC has false positives in overexposed regions (as pointed out by [33]), since its classifier cannot infer whether these regions are (or are not) part of the rest of the scene. In contrast, our model successfully handles these regions. CropCLR incorrectly flags regions that are semantically different from the background, because this is a strong indication that the patches come from different images. In contrast, we successfully handle these cases, since our model has no such “shortcut” in its learning task.

4.4. Ablation Study

To help understand which aspects of our approach are responsible for its performance, we evaluated a variety of variations of our model, including different training supervision, representations for the camera metadata, and network architectures.

We evaluated each model’s features quality using linear probing on the radial distortion estimation and splice detection task (same as Sec. 4.2). As an additional evaluation, we classify the values of common EXIF tags by applying linear classifiers to our visual representation. We convert the values of each EXIF tag into discrete categories, by quantizing common values and removing examples that do not fit into any category. We average prediction accuracies over 44 EXIF tags to obtain overall accuracy. We provide more details in the supplementary. All models were trained for 30 epochs on 800K images on a subset of YFCC100M dataset. The associated texts are obtained from the image descriptions and EXIF data provided by the dataset.

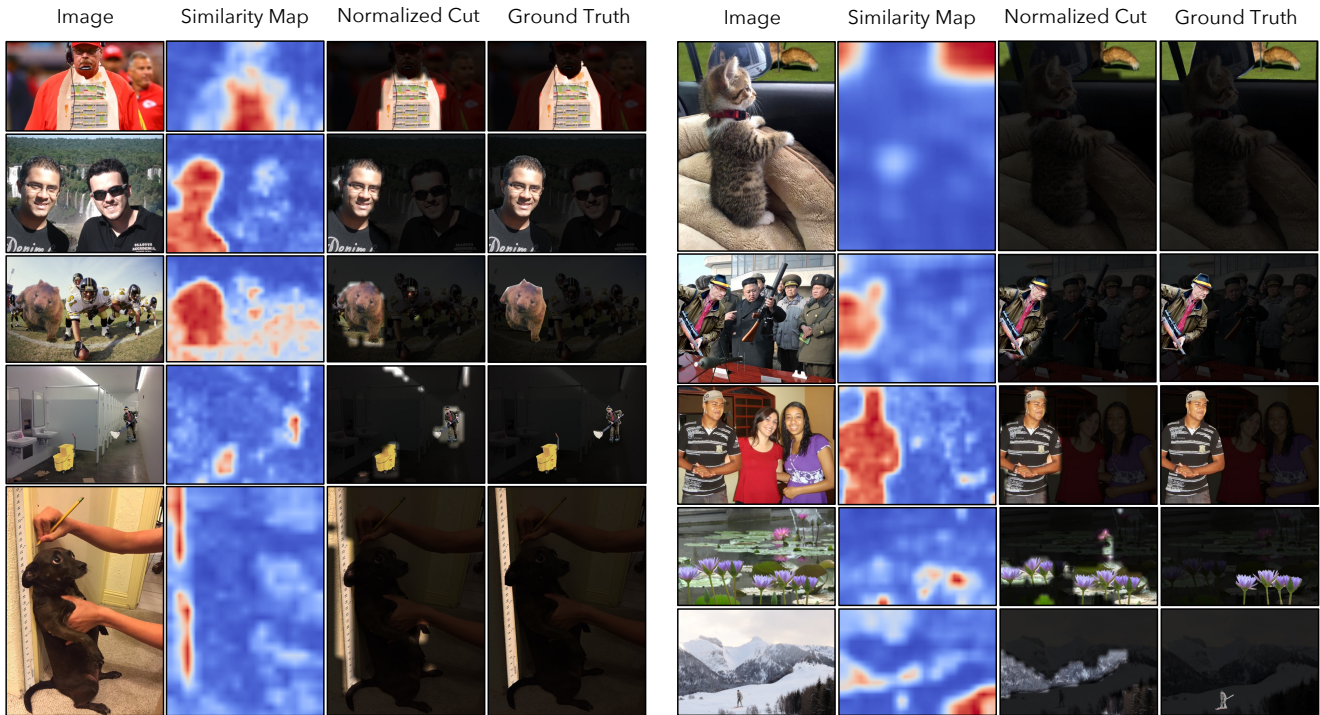


Figure 4

	Method	EXIF	Radial	Forens.
	Majority class baseline	0.12	0.05	0.50
Supervision	All EXIF tags	0.35	0.29	0.85
	CropCLR	0.29	0.22	0.84
	“Camera Model” tag only	0.31	0.27	0.80
	“Color Space” tag only	0.12	0.12	0.61
	YFCC image descriptions	0.15	0.16	0.70
Tag format	Fixed order, w/ tag name	0.35	0.29	0.85
	Fixed order, w/o tag name	0.35	0.28	0.86
	Rand. order, w/ tag name	0.34	0.26	0.77
	Rand. order, w/o tag name	0.33	0.26	0.76
Architecture	DistilBERT, w/ pretrained	0.35	0.29	0.85
	DistilBERT, w/o pretrained	0.36	0.25	0.77
	ALBERT, w/ pretrained	0.33	0.29	0.84
	ALBERT, w/o pretrained	0.34	0.26	0.79

Table 5. **Model ablations.** Downstream accuracy for versions of the model trained with different text supervision, representations of camera metadata, and architectures. We use linear probing to evaluate the average prediction accuracy of EXIF tag values on our YFCC test set, radial distortion estimation on Dresden dataset, and real-or-fake classification on CASIA I dataset. Rows with gray background (replicated for ease of comparison) represent the same model which is our “full” model.

Metadata supervision. We evaluate a variation of our model that trains using the image descriptions provided by YFCC100M in lieu of camera metadata, as well as models supervised by individual EXIF tags (Table 5). For the variations supervised by a single EXIF tag, we chose 14

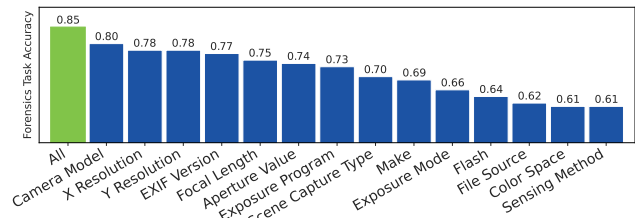


Figure 5. **Per-tag forensics task accuracy.** We train various models supervised by individual EXIF tags, then evaluate the learned representations for splice detection task on CASIA I.

common tags for this experiment, training a separate network for each one. The results of the per-tag evaluation is shown in Fig. 5. These results suggest that having access to the full metadata provides significantly better performance than using individual tags. Moreover, there is a wide variation in the performance of models that use different tag. This may be because the high performing tags, such as Camera Model, convey significantly more information about the full range of camera properties than others, such as Color Space and Sensing Method. These results suggest that a model that simply uses the full range of tags can extract significantly more camera information from the metadata. We also found that the variation trained on image descriptions (rather than EXIF text) performed significantly worse than other models.

Tag format. Since EXIF does not have a natural order of tags, we ask what will happen if we randomize the EXIF tag order during training. Table 5 shows the performance drops

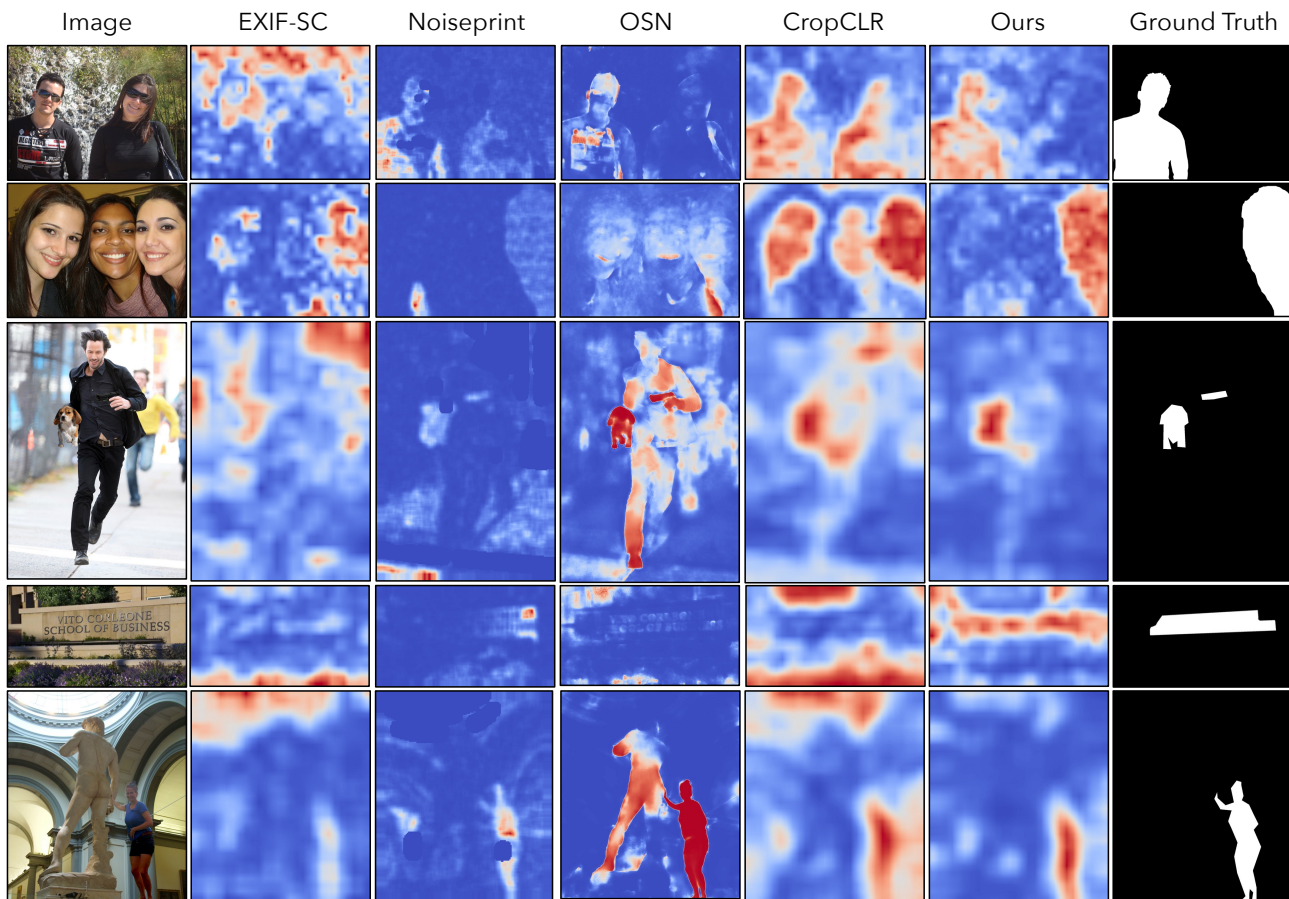


Figure 6. **Qualitative comparison to other methods.** Our method can correctly localize splices in many scenarios where other methods fail. For example, EXIF-SC [33] fails on overexposed image regions; OSN [79] and CropCLR often segment scenes based on semantics.

for all three evaluations in this case. This may be due to the fact that the Transformer model is forced to learn meaningful positional embeddings corresponding to each EXIF tag if their order keeps on changing. We also tried removing the tag names from the camera metadata and just provide the values for those keys, e.g. replacing `Make: Apple` with `Apple`. Interestingly, this model performs on par with the model that has tag names, suggesting that the network can discern information about the tags from the values alone.

Text encoder architecture. To test whether performance of our model tied to a specific transformer architecture, we experimented with two different transformer models, DistilBERT [67] and ALBERT [44]. We see that both architectures obtain similar performance on all three tasks with DistilBERT slightly outperforming ALBERT. We also test how much pretraining the text encoder helps with the performance. From Table 5, we can see pretraining improves performance on the radial distortion and forensics tasks.

5. Discussion

In this paper, we proposed to learn camera properties by training models to find cross-modal correspondences between images and camera metadata. To achieve this, we

created a model that exploits the fact that EXIF metadata can easily be represented and processed as text. Our model achieves strong performance amongst self-supervised methods on a variety of downstream tasks that require understanding camera properties, including zero shot image forensics and radial distortion estimation. We see our work opening several possible directions. First, it opens the possibility of creating multimodal learning systems that use camera metadata as another form of supervision, providing complementary information to high-level modalities like language and sound. Second, it opens applications that require an understanding of low level sensor information, which may benefit from our feature sets.

Limitations and Broader Impacts. We have shown that our learned features are useful for image forensics, which has potential to reduce the spread of disinformation [22]. The model that we will release may not be fully representative of the cameras in the wild, since it was trained only on photos available in the YFCC100M dataset [73].

References

- [1] Ali Azarbayejani and Alex P Pentland. Recursive estimation of motion, structure, and focal length. *IEEE Transactions on*

- Pattern Analysis and Machine Intelligence*, 17(6):562–575, 1995. 2
- [2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 1
- [3] Quentin Bammey, Rafael Grompone von Gioi, and Jean-Michel Morel. An adaptive neural network for unsupervised mosaic consistency analysis in image forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14204, 2020. 2
- [4] Quentin Bammey, Rafael Grompone von Gioi, and Jean-Michel Morel. An adaptive neural network for unsupervised mosaic consistency analysis in image forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14204, 2020. 2
- [5] Jawadul H Bappy, Amit K Roy-Chowdhury, Jason Bunk, Lakshmanan Nataraj, and BS Manjunath. Exploiting spatial structure for localizing manipulated image regions. In *Proceedings of the IEEE international conference on computer vision*, pages 4970–4979, 2017. 2
- [6] Tiziano Bianchi and Alessandro Piva. Image forgery localization via block-grained analysis of jpeg artifacts. *IEEE Transactions on Information Forensics and Security*, 7(3):1003–1017, 2012. 2
- [7] Luca Bondi, Luca Baroffio, David Güera, Paolo Bestagini, Edward J Delp, and Stefano Tubaro. First steps toward camera model identification with convolutional neural networks. *IEEE Signal Processing Letters*, 24(3):259–263, 2016. 2
- [8] Luca Bondi, Silvia Lameri, David Güera, Paolo Bestagini, Edward J Delp, and Stefano Tubaro. Tampering detection and localization through clustering of camera-based cnn features. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017. 2
- [9] Luca Bondi, Silvia Lameri, David Guera, Paolo Bestagini, Edward J Delp, Stefano Tubaro, et al. Tampering detection and localization through clustering of camera-based cnn features. In *CVPR Workshops*, volume 2, 2017. 2
- [10] J-Y Bouguet. Camera calibration toolbox for matlab. http://www.vision.caltech.edu/bouguetj/calib_doc/index.html, 2004. 5
- [11] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. 3
- [12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3, 4
- [13] Davide Cozzolino and Luisa Verdoliva. Noiseprint: a cnn-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2019. 2, 5, 6
- [14] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. Raise: A raw images dataset for digital image forensics. In *Proceedings of the 6th ACM multimedia systems conference*, pages 219–224, 2015. 5
- [15] Tiago José De Carvalho, Christian Riess, Elli Angelopoulou, Helio Pedrini, and Anderson de Rezende Rocha. Exposing digital image forgeries by illumination color classification. *IEEE Transactions on Information Forensics and Security*, 8(7):1182–1194, 2013. 5, 6
- [16] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021. 2
- [17] 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. 1, 4
- [18] Jing Dong, Wei Wang, and Tieniu Tan. CASIA image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*. IEEE, July 2013. 5
- [19] Jiayuan Fan, Hong Cao, and Alex C Kot. Estimating exif parameters based on noise features for image manipulation detection. *IEEE Transactions on Information Forensics and Security*, 8(4):608–618, 2013. 2
- [20] Jiayuan Fan, Tao Chen, and Alex ChiChung Kot. Exif-white balance recognition for image forensic analysis. *Multidimensional Systems and Signal Processing*, 28(3):795–815, 2017. 2
- [21] Hany Farid. Exposing digital forgeries from jpeg ghosts. *IEEE transactions on information forensics and security*, 4(1):154–160, 2009. 2
- [22] Hany Farid. *Photo forensics*. MIT Press, 2016. 8
- [23] Pasquale Ferrara, Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. Image forgery localization via fine-grained analysis of cfa artifacts. *IEEE Transactions on Information Forensics and Security*, 7(5):1566–1577, 2012. 2
- [24] Pasquale Ferrara, Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. Image forgery localization via fine-grained analysis of cfa artifacts. *IEEE Transactions on Information Forensics and Security*, 7(5):1566–1577, 2012. 5, 6
- [25] Thomas Gloe and Rainer Böhme. The’dresden image database’ for benchmarking digital image forensics. In *Proceedings of the 2010 ACM symposium on applied computing*, pages 1584–1590, 2010. 5
- [26] Ashima Gupta, Nisheeth Saxena, and SK Vasistha. Detecting copy move forgery using dct. *International Journal of Scientific and Research Publications*, 3(5):1, 2013. 2
- [27] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 5
- [28] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (ToG)*, 26(3):4-es, 2007. 5, 6
- [29] 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. 3, 4, 5
- [30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceed-*

- ings of the *IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5
- [31] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Jonathan Eisenmann, Matthew Fisher, Emiliano Gambaretto, Sunil Hadap, and Jean-François Lalonde. A perceptual measure for deep single image camera calibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2354–2363, 2018. 2
- [32] Wu-Chih Hu and Wei-Hao Chen. Effective forgery detection using dct+ svd-based watermarking for region of interest in key frames of vision-based surveillance. *International Journal of Computational Science and Engineering*, 8(4):297–305, 2013. 2
- [33] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *ECCV*, 2018. 1, 2, 3, 4, 5, 6, 8
- [34] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011. 1
- [35] Michael R James and Stuart Robson. Straightforward reconstruction of 3d surfaces and topography with a camera: Accuracy and geoscience application. *Journal of Geophysical Research: Earth Surface*, 117(F3), 2012. 1
- [36] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer, 2016. 2
- [37] Thibaut Julliand, Vincent Nozick, and Hugues Talbot. Image noise and digital image forensics. In *International Workshop on Digital Watermarking*, pages 3–17. Springer, 2015. 2
- [38] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4
- [39] Matthias Kirchner and Thomas Gloe. Forensic camera model identification. *Handbook of Digital Forensics of Multimedia Data and Devices*, pages 329–374, 2015. 2
- [40] Vladimir V Kniaz, Vladimir Knyaz, and Fabio Remondino. The point where reality meets fantasy: Mixed adversarial generators for image splice detection. *Advances in Neural Information Processing Systems*, 32, 2019. 2
- [41] Vladimir V Kniaz, Vladimir Knyaz, and Fabio Remondino. The point where reality meets fantasy: Mixed adversarial generators for image splice detection. *Advances in Neural Information Processing Systems*, 32, 2019. 5, 6
- [42] Pawel Korus and Jiwu Huang. Multi-scale analysis strategies in prnu-based tampering localization. *IEEE Transactions on Information Forensics and Security*, 12(4):809–824, 2016. 5, 6
- [43] Akash Kumar and Arnav Bhavsar. Copy-move forgery classification via unsupervised domain adaptation. *arXiv preprint arXiv:1911.07932*, 2019. 5
- [44] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A libbert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019. 8
- [45] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858*, 2022. 3
- [46] Ang Li, Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4183–4192, 2017. 2
- [47] Bo Liu and Chi-Man Pun. Splicing forgery exposure in digital image by detecting noise discrepancies. *International Journal of Computer and Communication Engineering*, 4(1):33, 2015. 2
- [48] Bo Liu, Chi-Man Pun, and Xiao-Chen Yuan. Digital image forgery detection using jpeg features and local noise discrepancies. *The Scientific World Journal*, 2014, 2014. 2
- [49] Yung-Cheng Liu, Wen-Hsin Chan, and Ye-Quang Chen. Automatic white balance for digital still camera. *IEEE Transactions on Consumer Electronics*, 41(3):460–466, 1995. 2
- [50] Manuel Lopez, Roger Mari, Pau Gargallo, Yubin Kuang, Javier Gonzalez-Jimenez, and Gloria Haro. Deep single image camera calibration with radial distortion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11817–11825, 2019. 2, 5
- [51] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 4
- [52] Babak Mahdian and Stanislav Saic. Using noise inconsistencies for blind image forensics. *Image and Vision Computing*, 27(10):1497–1503, 2009. 1, 5, 6
- [53] 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. 2
- [54] Owen Mayer and Matthew C Stamm. Learned forensic source similarity for unknown camera models. In *ICASSP*, 2018. 2
- [55] Yasuhide Mori, Hironobu Takahashi, and Ryuichi Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *First international workshop on multimedia intelligent storage and retrieval management*, pages 1–9. Citeseer, 1999. 2
- [56] Tian-Tsong Ng, Shih-Fu Chang, and Q Sun. A data set of authentic and spliced image blocks. *Columbia University, ADVENT Technical Report*, pages 203–204, 2004. 5, 6
- [57] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3
- [58] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 1
- [59] Guy Parsons. Dall-e 2 prompt book. <http://dallery.gallery/wp-content/uploads/2022/07/The-DALL%C2%B7E-2-prompt-book-v1.02.pdf>, 2022. 1, 3

- [60] Marc Pollefeys, Luc Van Gool, Maarten Vergauwen, Frank Verbiest, Kurt Cornelis, Jan Tops, and Reinhard Koch. Visual modeling with a hand-held camera. *International Journal of Computer Vision*, 59(3):207–232, 2004. 5
- [61] Chi-Man Pun, Bo Liu, and Xiao-Chen Yuan. Multi-scale noise estimation for image splicing forgery detection. *Journal of visual communication and image representation*, 38:195–206, 2016. 2
- [62] 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. 1, 2, 3, 4, 5
- [63] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1
- [64] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019. 2
- [65] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 5
- [66] Ronald Salloum, Yuzhuo Ren, and C-C Jay Kuo. Image splicing localization using a multi-task fully convolutional network (mfcn). *Journal of Visual Communication and Image Representation*, 51:201–209, 2018. 2
- [67] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 4, 8
- [68] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *European Conference on Computer Vision*, pages 153–170. Springer, 2020. 2
- [69] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000. 4
- [70] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846, 2006. 5
- [71] Peter Sturm. On focal length calibration from two views. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II. IEEE, 2001. 2
- [72] Xiaoting Sun, Yezhou Li, Shaozhang Niu, and Yanli Huang. The detecting system of image forgeries with noise features and exif information. *Journal of Systems Science and Complexity*, 28(5):1164–1176, 2015. 2
- [73] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 3, 4, 8
- [74] Amel Tuama, Frédéric Comby, and Marc Chaumont. Camera model identification with the use of deep convolutional neural networks. In *2016 IEEE International workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2016. 2
- [75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2
- [76] Sheng-Yu Wang, Oliver Wang, Andrew Owens, Richard Zhang, and Alexei A Efros. Detecting photoshopped faces by scripting photoshop. In *ICCV*, 2019. 2
- [77] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2
- [78] Xin Wang, Bo Xuan, and Si-long Peng. Digital image forgery detection based on the consistency of defocus blur. In *2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 192–195. IEEE, 2008. 2
- [79] Haiwei Wu, Jiantao Zhou, Jinyu Tian, and Jun Liu. Robust image forgery detection over online social network shared images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13440–13449, 2022. 2, 5, 6, 8
- [80] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9543–9552, 2019. 1, 2, 5, 6
- [81] Bin Xiao, Yang Wei, Xiuli Bi, Weisheng Li, and Jianfeng Ma. Image splicing forgery detection combining coarse to refined convolutional neural network and adaptive clustering. *Information Sciences*, 511:172–191, 2020. 2
- [82] Shuiming Ye, Qibin Sun, and Ee-Chien Chang. Detecting digital image forgeries by measuring inconsistencies of blocking artifact. In *2007 IEEE International Conference on Multimedia and Expo*, pages 12–15. Ieee, 2007. 2
- [83] Shuiming Ye, Qibin Sun, and Ee-Chien Chang. Detecting digital image forgeries by measuring inconsistencies of blocking artifact. In *2007 IEEE International Conference on Multimedia and Expo*, pages 12–15. Ieee, 2007. 5, 6
- [84] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020. 2
- [85] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1053–1061, 2018. 2
- [86] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015. 4

- [87] Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T Freeman. Learning ordinal relationships for mid-level vision. In *Proceedings of the IEEE international conference on computer vision*, pages 388–396, 2015. 4