Political Analysis (2024), **0**, 1–15
doi:10.1017/pan.2024.18

ARTICLE

Improving Computer Vision Interpretability: Transparent Two-Level Classification for Complex Scenes

Stefan Scholz[1], Nils B. Weidmann[1], Zachary C. Steinert-Threlkeld[2], Eda Keremoğlu[1] and Bastian Goldlücke[1]

[1] Center for Image Analysis in the Social Sciences, University of Konstanz, Konstanz, Germany; [2] Luskin School of Public Affairs, University of California, Los Angeles, Los Angeles, CA, USA

Corresponding author: Stefan Scholz; Email: stefan.scholz@uni-konstanz.de

(Received 23 November 2023; revised 8 May 2024; accepted 13 May 2024)

Abstract

Treating images as data has become increasingly popular in political science. While existing classifiers for images reach high levels of accuracy, it is difficult to systematically assess the visual features on which they base their classification. This paper presents a two-level classification method that addresses this transparency problem. At the first stage, an image segmenter detects the objects present in the image and a feature vector is created from those objects. In the second stage, this feature vector is used as input for standard machine learning classifiers to discriminate between images. We apply this method to a new dataset of more than 140,000 images to detect which ones display political protest. This analysis demonstrates three advantages to this paper's approach. First, identifying objects in images improves transparency by providing human-understandable labels for the objects shown on an image. Second, knowing these objects enables analysis of which distinguish protest images from non-protest ones. Third, comparing the importance of objects across countries reveals how protest behavior varies. These insights are not available using conventional computer vision classifiers and provide new opportunities for comparative research.

Keywords: image analysis; computer vision; explainable AI; two-level classification; protest analysis

Edited by: Jeff Gill

1. Introduction

Recent progress in the field of artificial intelligence and computer vision has led to an increasing adoption of image analysis in the social sciences. Images have a number of advantages over textual sources. They are language agnostic, so one can train one model instead of one model per language. They can also improve the measurement of concepts that are typically not mentioned in text, such as violent tactics, crowd composition, or the use of symbols (Abrams and Gardner 2023). These advantages have led to innovative work in political science that studies, for example, the emotional impetus of images (Casas and Williams 2019), altered vote tally sheets (Cantú 2019), or media coverage of politicians (Girbau *et al.* 2024).

Currently, almost all computational image analysis is performed using deep neural networks. While these networks are able to achieve an impressive level of accuracy, it is difficult for the researcher to understand *why* they assign a particular image to a given label or category. This problem is especially pressing when the image is *complex*: the classification of an image that contains many different types of objects is more difficult to understand than the classification of simple images showing single persons, maps, or ballots. As these methods continue to grow in importance in a wide range of research, the necessity of interpreting their operation has become a growing area of research (Rudin 2019). This paper

https://doi.org/10.1017/pan.2024.18 Published online by Cambridge University Press

presents an approach that helps remedy the opacity of vision models, such that they can be explored further for social science applications.

The paper introduces a two-level image classification method to improve computer vision interpretability. First, one creates a feature vector from the objects (“segments” in computer vision terminology) an image contains. Next, a non-visual machine learning classifier uses the feature vectors to identify combinations of objects predictive of the researchers’ outcome of interest. We demonstrate this approach on a new dataset of 141,538 protest images from 10 countries. Mass protest is an increasingly frequent phenomenon: whether the issue is COVID-19 lockdowns in China, womens’ rights in Iran, or indigenous rights in Peru, their global number has increased steeply in recent years (Ortiz *et al.* 2022). The rise of social media has led to a concurrent profusion of images documenting protest, and researchers have started to build protest event datasets from them (Steinert-Threlkeld, Chan, and Joo 2022; Zhang and Pan 2019). Protest images are instances of complex scenes mentioned above: they are frequently composed of different objects such as people, flags, signs, or cars.

This example demonstrates three advantages of the approach. First, the two-level classifier identifies the objects an image contains, allowing for immediate understanding of image content unlike existing pixel-based methods. These objects are human-understandable items such as “car” or “fence,” an improvement over approaches that identify pixel activations of areas of high contrast pixels (Torres 2024). Second, permutation tests reveal which objects distinguish protest from non-protest images, which allow for simple validation tests of the classification. Third, the distribution of object permutation importance across protests shows how symbolic usages varies across different national and political contexts. None of these insights are available using current image classifiers and the pixel-based interpretation techniques applied to them. Researchers seeking to apply our method to their own images can do so with our pre-configured online demonstration tool and the associated API for batch processing (see data availability statement).

2. Image Classification: The Limited Interpretability of Conventional Approaches

Prior to the introduction of multilayered (“deep”) neural networks, computer vision used less flexible statistical models such as color distributions or predefined feature maps. The introduction of deep neural networks, the explosion of training dataset size, and use of specialized hardware has brought forth models that assign proper labels, recognize objects, and analyze faces markedly better than previous methods (LeCun, Bengio, and Hinton 2015).

A convolutional neural network (CNN) consists of a series of layers, ranging from feature extraction to the fully connected layers. The last of these layers then feeds into a function that outputs a vector as long as the number of classes. The output of AlexNet, for example, is 1,000 units long because the dataset it trained on contains 1,000 labels such as dog, cloud, and car (Krizhevsky, Sutskever, and Hinton 2012). In contrast, the output of the model in Cantú (2019) is two units long because the labels are “altered ballot” and “not altered ballot.” This final vector, the output layer, is the model’s estimate of the content of an image. See Webb Williams, Casas, and Wilkerson (2020), Joo and Steinert-Threlkeld (2022), and Torres and Cantú (2022) for more on the functioning of CNNs.

The parameters estimated during training are the values of the kernels and the weights connecting neurons in the fully connected layers. Deep neural networks’ hierarchical structure, numerous feature maps, and use of fully connected layers mean they contain tens of millions of parameters. This complexity makes it impossible to determine *why an image receives the classification it does* by looking at parameter values, a marked contrast to regression models’ coefficients or simpler machine learning classifiers’ parameters, for example, trees in a random forest.

Different methods have been developed to aid the interpretation of computer vision models. Since pixels are the atomic elements of images, like characters in text, it is natural to ask which pixels contribute to a model’s classification of an image. An early example of this approach is deconvolutional neural networks (“DeconvNets”). A DeconvNet estimates important pixels from feature maps (Zeiler and Fergus 2014). For a given feature map and layer, the DeconvNet can show which pixels, such as those

https://doi.org/10.1017/pan.2024.18 Published online by Cambridge University Press

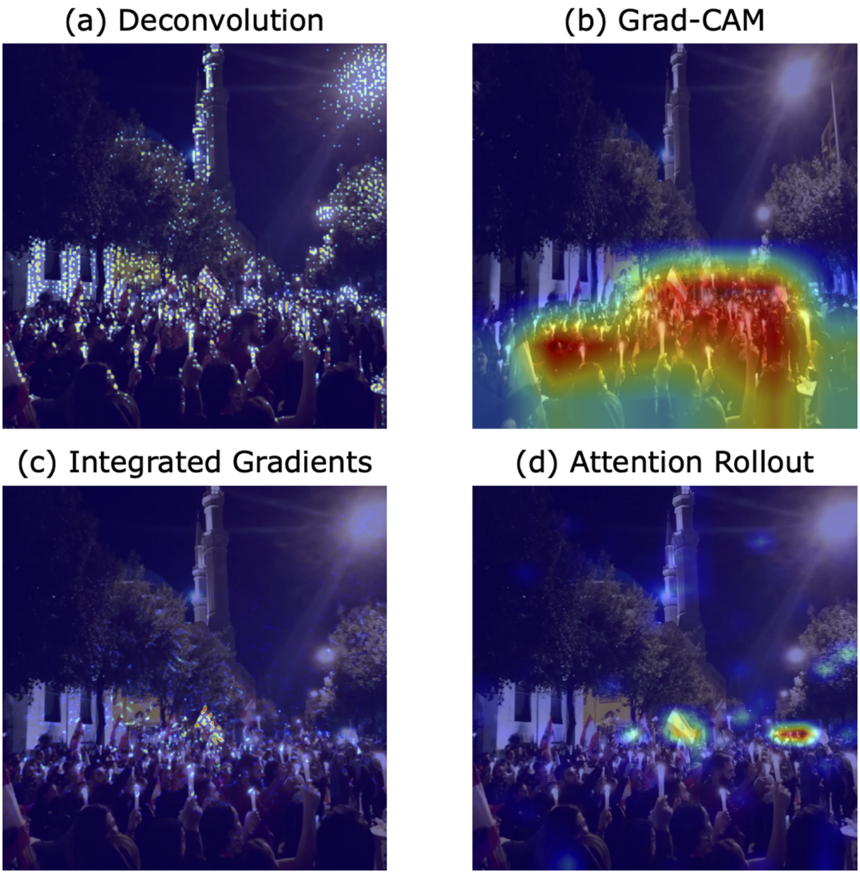


Figure 1. Comparison of visual information extracted from protest image with Deconvolution, Grad-CAM, Integrated Gradients, and Attention Rollout.

associated with a door or face, most activate that feature. The output is the input image at that layer with the non-activated pixels removed. Another approach is gradient class-activation mapping (Grad-CAM; Selvaraju *et al.* 2017). This method sets the class score to 1 for the desired class, for example, "protest," and backpropagates the activation to the pixels in a given image driving that classification; the output is the input image with a pixel heatmap.

A third approach is integrated gradients: the classified image is compared to a baseline image, usually a black square, and each pixel's contribution to the final class score is compared to its prediction when the baseline image is assigned to that class (Sundararajan, Taly, and Yan 2017). The output is the original image with pixels colored by their gradient sum. A fourth approach is attention maps, which are calculated using the attention mechanism of transformer models (Dosovitskiy *et al.* 2021). The attention mechanism of vision transformers (ViTs) allows them to focus on specific areas of images, as opposed to CNNs. The attention weights can be visualized post hoc in attention maps, where the higher the attention weight, the brighter the color of the pixel in the attention map.

Figure 1 shows a protest image with the four pixel-based visual explanations described above.[1] In each case, pixels are colored as a heatmap based on their relative contribution to classifying the image as containing a protest. While the highlighted areas provide plausible cues as to why the image is

[1] As computer vision models, we use the ResNet50 to visualize deconvolution, Grad-CAM and integrated gradients, and ViT models to visualize attention rollout. Both were self-trained on our dataset of protest images.

https://doi.org/10.1017/pan.2024.18 Published online by Cambridge University Press

classified as a protest image, Figure 1 reveals two major shortcomings. First, though important pixels are highlighted, the researcher still must determine the concept or object behind a collection of pixels. For example, in Panel (b) of Figure 1, it is unclear if the protest classification is driven by a group of people, their attire, or the presence of flags. Second, different visual explanation approaches emphasize different parts of the image. Comparing Panels (b) and (d), for example, the latter suggests that classification is driven by different, smaller parts of the image rather than the group of people in the former: a flag, two indecipherable areas on each side of the flag, and even the foliage of a tree.

For comparative research with images, these difficulties pose major obstacles. First, coding criteria should be explicit, such that it is possible to understand why a particular category or label has been assigned. This is a requirement regardless of how the coding was done (human or automatic). In automatic text classification, for example, it is possible to identify words or word combinations that pertain to particular topics, which allows researchers to check the validity of the coding. In image classification, however, there is no natural unit into which images can be decomposed, leading to a second obstacle. Absent an abstract description of the content of an image, it is difficult to compare images across context and cases to see whether and why their content differs.

The method proposed in the next section obviates these problems by identifying specific objects in images (first level) and then using those objects to determine if an image contains the desired concept (second level). Instead of looking at individual pixels, it looks at objects (groups of pixels) in an image. It then uses the differential presence of certain classes of objects to determine whether an image contains the desired concept (for this paper, protest). This approach is different to pixel-alignment methods because it does not need to explain which pixels are aligned with which objects.

3. Two-Level Classification

This paper introduces a two-level process for image classification. The first level creates a vector summarizing the objects contained in each image; the second trains a non-visual classifier on these vectors. This section describes each step in more detail.

3.1. *Creating Feature Vectors from Images*

The first step maps the pixel representation of images to objects, an interpretable lower-dimensional representation. We first describe how to extract the objects from an image, before describing how to turn them into vectors that summarize the objects contained in them.

3.1.1. *Extracting Segments*

In addition to classifying entire images, computer vision models can detect and classify *objects within images*. Object detection refers to estimating bounding boxes around potential objects and classifying these areas into objects of different categories. An extension of object detection is instance segmentation, where rather than simple bounding boxes, pixel masks are provided for the shapes of the detected objects. In recent years, a number of frameworks have been proposed which have increased the accuracy and efficiency of instance segmentation (e.g., Girshick *et al.* 2014; He *et al.* 2017). The most commonly used metric to measure the accuracy is average precision (AP). It rewards correct classifications and precise masks, which means that the higher the AP, the better the framework. These frameworks are trained in a fully supervised fashion on particular datasets and can therefore be readily applied to new images.

The datasets used for training segmentation algorithms define what types of objects these algorithms can later detect. The entire set of these object types (or categories) is called the "vocabulary" of the segmenter. Older approaches such as the *PASCAL Visual Object Classes* (PASCAL VOC; Everingham *et al.* 2010) have small vocabularies (20). More recent ones, such as the *Common Objects in Context* dataset (COCO; Lin *et al.* 2014) has 80 categories, and the *Large Vocabulary Instance Segmentation*

https://doi.org/10.1017/pan.2024.18 Published online by Cambridge University Press

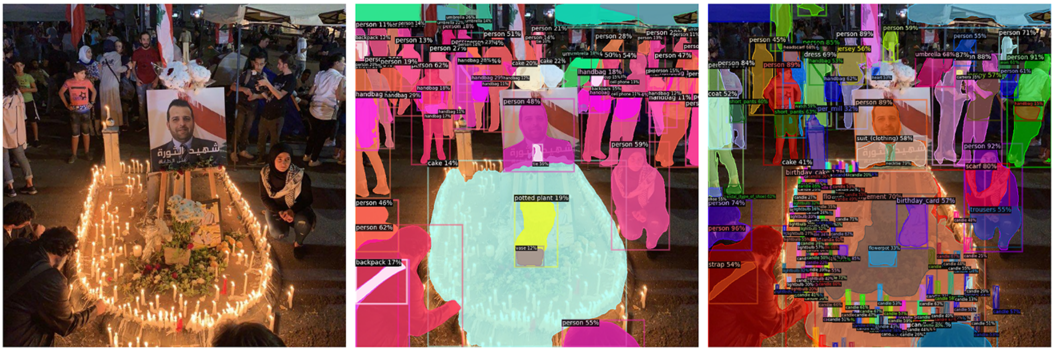


Figure 2. Instance segmentation applied to an image of a candlelight vigil (left) using COCO vocabulary (center) and LVIS vocabulary (right).

dataset (LVIS; Gupta, Dollar, and Girshick 2019) has 1,203; these two datasets are developed on the same set of images. The rest of the paper focuses on the COCO and LVIS datasets.

Figure 2 shows the segments detected in a protest picture of individuals at a vigil with candlelights. The center panel of Figure 2 uses an instance segmentation method trained using the COCO vocabulary, and the right panel shows segments detected with the LVIS vocabulary. Not surprisingly, LVIS detects more segments and more segments of different object types, providing greater detail about the image content than COCO.

3.1.2. Creating Segment Vectors

All instance segmenters generate object positions, categories, and confidence scores for each detected segment of the input image. The position is called the mask; it is a polygon outlining the proposed object. Since the segmentation method cannot always assign a unique predicted object category, it outputs certainty scores (0–1) for each category in the vocabulary on which the detector was trained. Following conventional approaches (e.g., He *et al.* 2017), for each segment, we assign the object category with the highest certainty score to the respective segment. Then, we use this information to create abstract descriptions of the set of objects contained in an image. The collection of objects is used to create a feature vector.

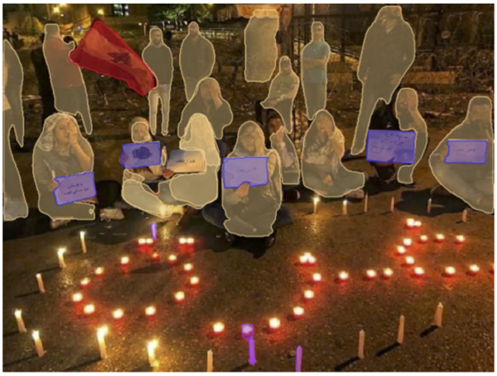
There are a large number of ways to transform the output from the segmentation method into a feature vector. Figure 3 presents the four this paper evaluates. Each entry in the generated vectors corresponds to one type of object from the vocabulary of the segmentation method.

Binary features. A binary feature vector indicates the presence or absence of a certain object category in the image. The top feature vector v_a in Figure 3 shows this construction for five object types.

Count-based features. An extension of the binary feature is to count how many objects of each object category are present. The second vector v_b in Figure 3 illustrates this approach. The model detects 5 segments with a poster, 19 segments with a person, 1 segment with a flag, 4 segments with a candle, and no segments with a gun.

Area-based features. The positional information obtained from the segmenter can also be incorporated into the feature vectors. We do this in two ways. A third type of feature vector uses the maximum area of any object of a given category in the image, assuming that bigger objects are more important. In Figure 3, vector v_c indicates that the largest person on the image occupies 3% of the image area. A fourth feature type uses the sum of the areas identified for each object category and therefore captures what proportion of the entire image objects of each type occupy. The bottom vector v_d in Figure 3, for example, indicates that persons take up 28% of the image.

https://doi.org/10.1017/pan.2024.18 Published online by Cambridge University Press



$$\begin{array}{c|ccccc}
 & \text{Poster} & \text{Person} & \text{Flag} & \text{Candle} & \text{Gun} \\
v_a & [\;1 & 1 & 1 & 1 & 0\;] \\
v_b & [\;5 & 19 & 1 & 4 & 0\;] \\
v_c & [\;0.01 & 0.03 & 0.02 & 0.00 & 0.00\;] \\
v_d & [\;0.03 & 0.28 & 0.02 & 0.00 & 0.00\;]
\end{array}$$

Figure 3. Feature generation from a segmented image (left), with different feature vectors generated from this image (right): binary vector (v_a), count-based vector (v_b), area-max vector (v_c), and area-sum vector (v_d).

3.2. Classification

At the second level, we train a standard machine learning classifier to predict the image labels (protest) from the segment vectors of the images. We rely on four different classifiers: logistic regression, simple decision trees, collections of decision trees, and gradient-boosted decision trees. Logistic regression was chosen since it is widely used by social scientists, and to provide a benchmark against. The tree-based classifiers were selected because they allow the researcher to vary complexity and interpretability.

The simplest decision tree can have a depth of two and classify an observation using one condition. Decision trees can become more complex when the depth of the tree—the number of conditions—is increased. Combining decision trees into an ensemble allows each tree to attempt to correct the misclassifications of its predecessor tree. These gradient-boosted decision trees improve accuracy, though the added complexity inhibits interpretability because any observation now follows multiple trees. In practice, the number of trees—the number of boosting rounds—can reach into the thousands.

4. Application: Coding Protest Images

This section develops and evaluates a two-level classifier using a new dataset of protest images.

4.1. A New Protest Image Dataset

A new dataset of protest images collected from social media is used to test the performance and added value of the two-level classifier. This dataset focuses on large protest episodes in different countries worldwide. To build the dataset, the *Armed Conflict, Location, and Events Dataset* (ACLED; Raleigh *et al.* 2010) was used to identify 46 country-periods with many protests from 2014 to 2021 that are also high-income and populous enough to generate enough social media reporting (Steinert-Threlkeld and Joo 2022). These 46 country-periods were narrowed to 14 based on their Polity IV score and region, with a goal of generating broad coverage of regime types and parts of the world. Twitter is the platform used to obtain images because of its widespread use. This step generates 135 million tweets and 16 million images.

The number of tweets a country produces is a function of a country's population, its gross domestic product, and the duration of a protest. There is therefore a large difference in the number of images per country. To avoid any potential bias, we downsample each country's images to 100,000. For the three countries that have fewer than 100,000 images, all are kept. Supplementary Appendix A.1 presents more details on the selection of countries and images.

A protest is defined as a publicly visible event or action involving at least one person making a political statement or expression. Human coders assigned one of four labels: protest (with high certainty), protest (low certainty), no protest (low certainty), or no protest (high certainty). The use of the different degrees of certainty captures the fact that oftentimes images alone lack the context of a larger protest episode to be interpreted with certainty.[2] Exact and near duplicate images are removed before labeling (see Supplementary Appendix A.2). For more details of the coding procedure, see Supplementary Appendixes A.3 and A.4.

Once labeled, images from Japan, Kazakhstan, Ethiopia, and the Philippines are also excluded because there were fewer than 100 protest images. The remaining 141,538 images are split into 80–20 training and testing sets by country (see Supplementary Appendix A.5 for details). In the training dataset, there are 12,454 protest images and 100,776 non-protest images. In the testing dataset, 3,113 and 25,195, respectively. This relative paucity of protest images matches the distribution found in other work (Steinert-Threlkeld *et al.* 2022).

4.2. *Training the Two-Level Classifier*

Using the new protest image dataset, we construct our two-level classifier using different combinations of vocabularies, feature vectors, and second-level classifiers.

Vocabulary. We use the COCO and LVIS vocabularies introduced earlier. We detect the segments in the COCO vocabulary (Lin *et al.* 2014) with the instance segmentation model by Li *et al.* (2023). This model was validated on the COCO dataset with a mask AP of 0.5230. To detect the segments in the LVIS vocabulary (Gupta *et al.* 2019), we use the instance segmentation model by Zhou *et al.* (2022). This model achieved a mask AP of 0.2497 on the LVIS v1.0 validation set, which is not the highest AP reported in the original paper, but visually has the best results and more reliably recognizes frequent categories. Segments with a confidence score below 0.1 are discarded because visual inspection shows segments below this threshold are not reliable.

Feature Generation. To explore how different kinds of feature vectors affect the classifier performance, the four types of vectors introduced earlier are used: binary features (at least one segment of category on image), count-based features (sum up the number of segments of category on image), the maximum-area features (the area occupied by the largest instance of a category on an image), and the sum-area features (the total area occupied by all instances of a category on the image).

Classification Method. For classification of the feature vectors, four different classification methods are used at the second level: a logistic regression, a simple decision tree, a random forest (Breiman 2001), and an XGBoost gradient-boosted tree (Chen and Guestrin 2016). These classification methods have different hyperparameters that must be chosen. We optimize them using 5-fold cross-validation on the training set; more details are reported in Supplementary Appendix B.

5. Results

This section evaluates how the different design choices of the classifier affect its performance. We then present three sets of results: comparing the two-level classifier to existing conventional computer vision approaches, showing which objects are most important for identifying protest images, and using the by-country importance to understand protests in a comparative context. All results shown here are based on the dataset described above, which combines low and high confidence labels. Supplementary Appendix E shows the results of the subsequent analysis with only high certainty images.

5.1. *Design Choices and Classification Performance*

The full combinatorial space of vocabularies, feature vectors, and second-level classifiers is explored to determine the best two-level classifier, where performance is measured using F1 scores, a commonly

[2] While other work improved classification by combining text and image data (Zhang and Pan 2019), our procedure emulates a scenario where the coding is based on visual material alone.

https://doi.org/10.1017/pan.2024.18 Published online by Cambridge University Press

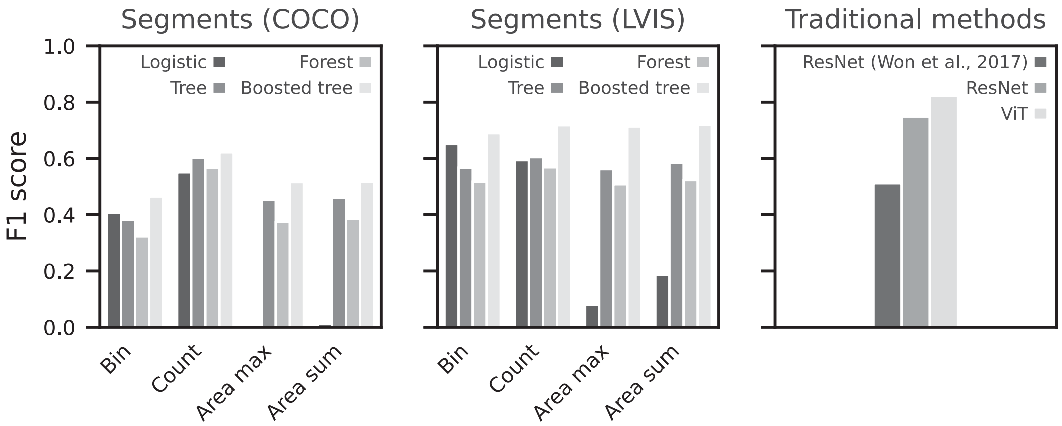


Figure 4. Out-of-sample evaluation of different methods. The figure displays the F1 score achieved on the test set; LVIS area sum with gradient-boosted trees achieves the best F1 score of 0.7203. The logistic regression obtains low F1 scores with the area-based features, making certain bars invisible. Visualization based on Supplementary Table A3.

used metric suitable for imbalanced class distributions. The left two panels of Figure 4 show the results. How do the vocabulary, the types of feature vectors, and the second-level classifier affect predictive performance?

Vocabulary. Comparing models trained on the COCO vocabulary and those trained on the LVIS vocabulary (left and center panels of Figure 4) shows that the latter mostly achieve better results. With the 80 objects categories available from the COCO vocabulary, F1 scores rarely exceed 0.5, while those relying on LVIS achieve F1 scores of up to 0.7203. These results show that the performance for most second-stage models and feature generation methods is improved by incorporating more object categories.

Feature Generation. Feature generation affects the second-stage classifier's performance, though the difference within vocabularies is less than that across them. There are two exceptions. First, count-based features perform particularly well among the models using the COCO vocabulary, and second, logistic models perform poorly for the area feature vectors with the LVIS vocabulary. In all other cases, the creation of the feature vectors does not seem to play a major role. This is not too surprising, since they all encode the presence/absence of different object categories in different ways, which seems to be sufficient for classification.

Classification Method. Logistic regression achieves its best results with the binary- and count-based features, but their results deteriorate significantly with the area-based features. The random forest-based models do not improve the accuracy compared to the logistic regression models for the binary and count feature vectors, but they do for the two area ones. Gradient-boosted trees generally improve the F1 score, in particular for the larger vocabulary. This result matches other work that has found random forests and gradient-boosted trees to outperform logistic regression in classifying rare events (Muchlinski *et al.* 2016; Wang 2019).

We also conduct a more in-depth analysis of our classifier to see if it systematically misses images with particular motives or content. To this end, we cluster the images into 30 topics and report the confusion matrix, precision, recall, and F1 of the classifier for each topic separately. Supplementary Appendix D describes the clustering process in detail and presents the performance of the LVIS area-sum gradient-boosted classifier in these clusters. In the cluster of images showing protests, fires, and smoke, the classification performance is below average, as it is for images with state police. Protests with large crowds, individual banners, or flags, on the other hand, are classified more accurately. However,

https://doi.org/10.1017/pan.2024.18 Published online by Cambridge University Press

we do not see a particularly poor performance for some clusters, which suggests that the classifier does not systematically fail to discover certain kinds of protest images.

5.2. Comparison to Conventional Approaches

To compare the two-level classifier to other computer vision approaches, three other models are used. Relying on earlier work in this area, we use the ResNet50 from Won, Steinert-Threlkeld, and Joo (2017), a CNN trained on a dataset of more than 40,000 online images. The second is the same ResNet50 CNN but trained on this paper's protest image dataset. A larger CNN such as a ResNet101 or ResNet152 could further improve accuracy, but a ResNet50 with the same training times and hardware requirements is used to facilitate comparison. The third model is a ViT, a model that has been shown to outperform CNNs on many computer vision tasks while requiring less computational resources (Dosovitskiy *et al.* 2021). We also use the smaller ViT version ("base"), whose entire weights are also learned during several epochs of training (see Supplementary Appendix B for details on model training).

Figure 4 (right panel) shows the results. The ResNet50 from Won *et al.* (2017) achieves the worst fit of the classifiers, most likely because it was trained on a different, smaller dataset. Training a ResNet50 on our dataset generates an F1 score of 0.7489 on the test data, almost identical to the F1 score of the best segments-based classifier. The ViT model achieves an F1 score of 0.8226 on the test data, and clearly outperforms our two-level classification. Reducing images to identifiable objects may therefore lose information that improves classification. This can happen because some objects that are relevant for protest fail to be included in the generic vocabularies used or because context of objects in the images is lost by extracting the segments only. Nevertheless, while the predictive performance of the segment-based classifier is not as strong as the cutting-edge in computer vision, its performance remains comparable to other established methods. At the same time, it provides different new possibilities, as we demonstrate in the following sections.

5.3. Importance of Features in Protest Images

The two-level classifier's advantage is that, by design, it offers insights into which object categories are related to protest. In a first illustration of this advantage, we examine which types of objects are most likely to appear in protest images. Figure 5 presents the total area occupied by instances from a certain segment category in protest and non-protest images when using the LVIS area-sum gradient-boosted tree classifier. The proportions are calculated for each country separately and then averaged. For clarity, the analysis is limited to the 10 largest segment categories of protest images. Not surprisingly, the largest segments in a protest image are persons: 34% of a protest image is occupied by people versus 22% for non-protest images. Protest images also largely display banners, posters, signboards, jackets, shirts, jerseys, flags, trousers, and cars. It is not always the case that protest images seem to be characterized by larger total areas of these objects types, however. For example, while posters are frequently shown on protest images, non-protest images tend to have larger areas occupied by posters. The reason may be that the presence of a poster alone (as, e.g., in a close-up photo) is not sufficient for an image to be classified as a protest image.

However, the largest categories are not necessarily the most important ones for prediction. To find out how essential an object category is for classifying protest, we calculate the importance of the corresponding features. For this purpose, a feature is randomly permuted to break its relationship to the labels. This random permutation is repeated for each feature, always followed by a reevaluation of the F1 score of the classifier on the training set. Thus, if this random permutation leads to a higher drop in performance for a given feature, this object category is more essential for the model to distinguish a protest image from a non-protest image. Rather than determining the feature importance for all images of all countries at the same time, the feature importance is calculated for each country separately and then averaged. In this way, we evaluate which features seem to be similarly important across all countries irrespective of the number of protest images we have for them.

https://doi.org/10.1017/pan.2024.18 Published online by Cambridge University Press

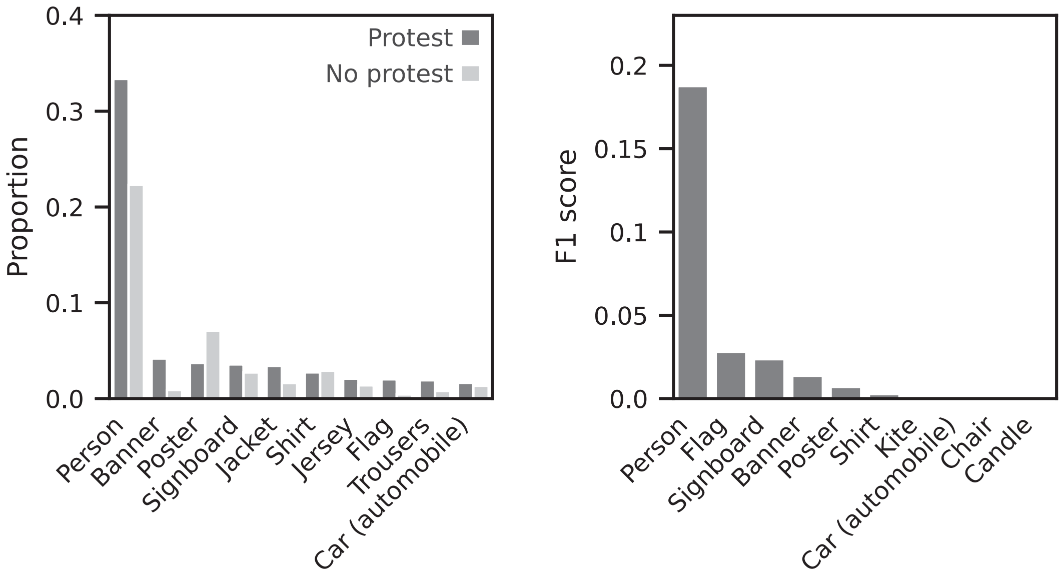


Figure 5. Proportion of segments in protest and non-protest images (left) and importance of area-sum aggregated segments (right).

The right panel of Figure 5 presents the results of this importance test. The model has the largest drop in F1 when the total area occupied by persons is permuted (0.19); followed by flag (0.03), signboard (0.02), and banner (0.01). As expected, almost all of the object types shown in the plot are closely related to protest, confirming the validity of the paper's method. In addition, the largest object categories are not necessarily the most important ones for prediction. For example, most of the clothing items that occupy large areas are not important for prediction, since they provide little additional information beyond the presence of persons, a feature that is already included in the model.

To test whether humans and the model differ in the categories they deem important for protest classification, we conduct an additional validation exercise. For this purpose, 1,000 protest images—100 for each country—are randomly selected, and a human coder was asked to name up to three objects that she considers to be most important in recognizing each image as a protest image. Then, the coder was shown the same image with the LVIS segments highlighted and numbered (without labels). The coder then had to identify which objects she selected in the first step correspond to the segments in the second step. Out of 2,210 objects identified in the first step, the five most important ones were people (776), flag (430), poster (216), signboard (195), and banner (118), closely matching categories that our two-level method identifies. An additional test assesses whether all human-coded objects match the segments the machine identifies by comparing the names given by the coder with the ones assigned by the segmenter. When we match strictly based on identical names, 68% of the human-coded objects are correctly identified by the segmenter. When matching based on a dictionary accounting for membership in the same object categories (i.e., a "child" is a "person"), the successful matching rate increases to 74%. Supplementary Appendix G provides further details on this manual validation. Overall, this exercise shows that both human coders and the segmenter largely use the same objects to classify protest images.

5.4. Country-Specific Importance of Features

Since protesters employ different tactics across different countries and episodes, the predictive importance of certain features may differ across the countries. For this reason, we assess differences across individual countries, comparing the relevance of a particular image feature in one country in relation to its importance in the whole sample. In the following, we compute this country-specific importance of a feature as its deviation from the average importance of the feature. We again use the best-performing

https://doi.org/10.1017/pan.2024.18 Published online by Cambridge University Press

model. We are specifically interested in the cases that have a particularly large deviation (positive or negative) for a feature, some of which we discuss in more detail below.

Protesters frequently hold signs stating demands, but the prevalence of signs at protests varies across countries. Signs demanding free elections were an integral part of the protest surge in Moscow that preceded the highly contested 2019 Moscow City Duma elections (Roth 2019). The images from these protests confirm that the Russian protests indeed stand out as regards the use of posters in comparison to other protest episodes. Panel (a) of Figure 6 shows the relative importance of signs in Russia and other countries and a sample image of a protester displaying a poster during one contentious episode.

Cars present another feature that characterizes protests in a few countries, but is relatively unimportant for most events in our sample. Argentina experienced a surge of protest in 2020 that mainly targeted governmental responses to the COVID-19 pandemic but also expressed discontent with peoples' economic hardship. As protests took place under governmental quarantine measures requiring Argentinians to socially distance, many protesters turned up in cars during some of these events (BBC 2020). Panel (b) of Figure 6 shows a sample image from these protests. In South Africa, on the other hand, cars—while a significant feature (see Panel (b) of Figure 6)—were important for very different reasons. The country experienced a wave of violent unrest during the "July 2021 riots" over former president Zuma's imprisonment. During some of these events, rioters looted and burned trucks and cars (Reuters 2021), making these vehicles a crucial feature for the images in our sample.

As protests often arise as the result of killings or to commemorate deaths, candles are a common feature of protest images. In Lebanon, candles were especially prominent. The "17 October Revolution" in Lebanon describes a wave of nationwide anti-government protests in 2019 that led to the resignation of the government. The protests united citizens and groups across the political and religious spectrum demanding a change of the whole political system. Candles played an important role in this context (see Panel (c) of Figure 6) for two reasons. The shooting of Alaa Abou Fakhr, a peaceful protester, by security forces was followed by public mourning and vigils hold by other protesters who lit candles to commemorate his death (Azhari 2019). In addition, lighting candles was an important symbol used by women's marches that were at the core of the movement (McCulloch and Mikhael 2019). Panel (c) of Figure 6 shows a sample image of an event in Lebanon where candles are central to the scene.

Finally, since tweets contain timestamps, temporal analysis of visual protest features is possible. Figure A4 in Supplementary Appendix F shows how the prevalence of the three most common segments per country changes before, during, and after a country's protest period. This analysis is meant simply as a proof of concept since this type of analysis is not possible with pixel-based methods' approaches to interpretability. Future research using this method may prefer other methods of selecting segments, such as the 10 most common or the 5 most important.

These cases show that protest can look very different across countries. It shows that the importance of features for some episodes of contention can be highly context-dependent, such as cars during COVID-19 lockdowns or the use of candles to commemorate protesters' death. The examples we discuss also show that some features that are rather widespread nevertheless are of particular importance in other cases as the discussion of posters in Russia has highlighted. Overall, our feature-based analysis considers differences between protest happening in different places and different times, and as a consequence also captures untypical instances of contention and helps to understand why they stand out.

6. Conclusion

Images have become increasingly important as a data source for political scientists. Not only has the digital age proliferated their massive spread across the globe, but recent progress in computer vision has also advanced automated analysis of visual material. Existing methods can achieve high accuracy when classifying image content, but their opacity fails to explain *why* classification is made in a certain way. Transparency, however, is important to understand the classification of complex image scenes or assess differences in the same research subject across different cases.

https://doi.org/10.1017/pan.2024.18 Published online by Cambridge University Press

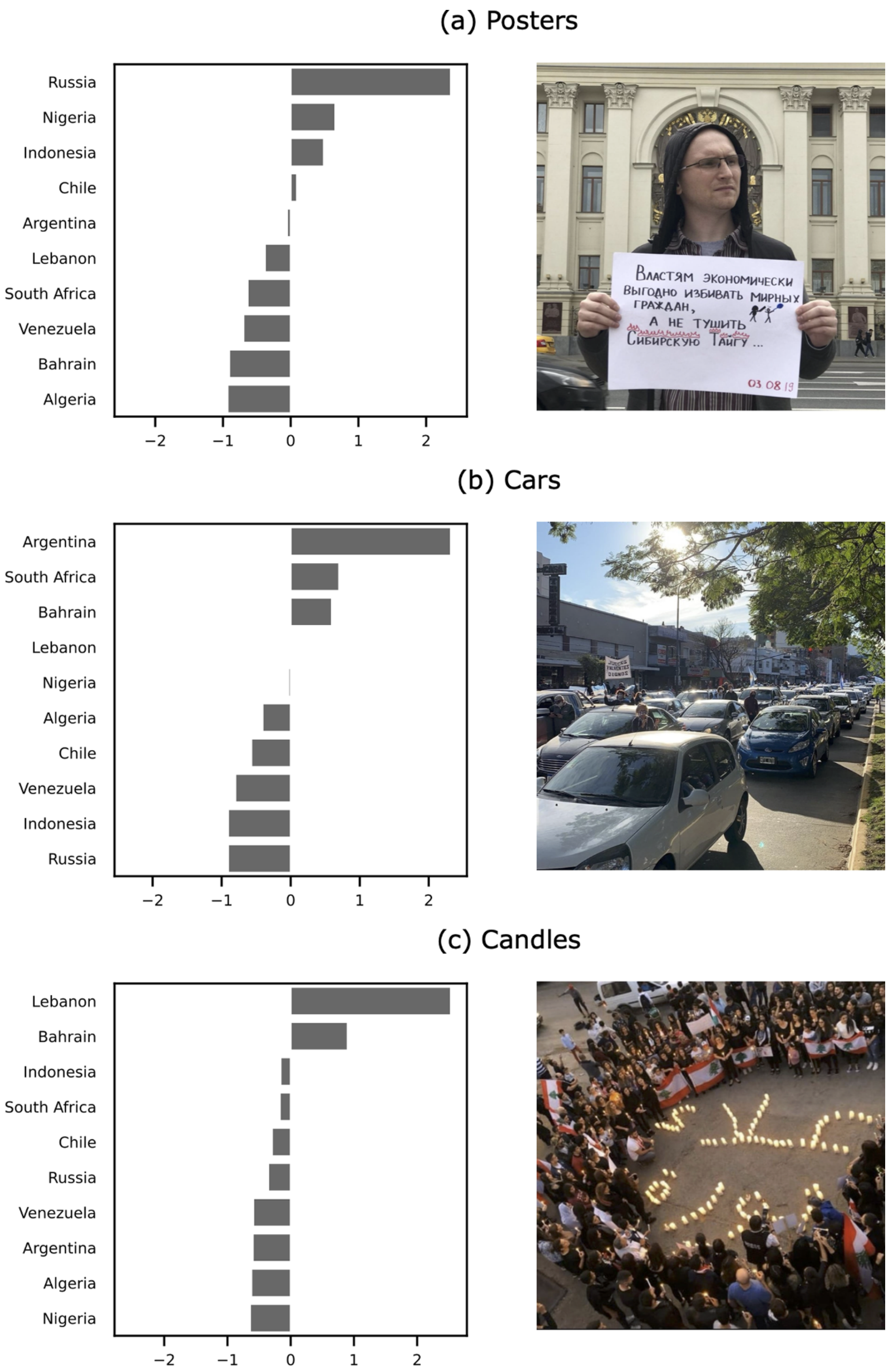


Figure 6. Differences in importance of posters, cars, and candles across different protest episodes. The left column presents the differences in importance of objects. Positive values denote higher importance in relation to the whole sample. The right column shows the examples of protest images: the use of posters in Russia, protest including cars from Argentina, and the use of candles during protest in Lebanon.

https://doi.org/10.1017/pan.2024.18 Published online by Cambridge University Press

This paper presents a new two-level procedure that improves transparency and interpretability of image classification. The first step extracts objects from an image and creates an abstract representation of the image based on these segments. This procedure improves on bag of visual words (BoVW) methods because it identifies human-understandable items in an image; in contrast, BoVW identifies areas of high pixel contrast and represents these areas as vectors based on changes in pixel intensity (Torres 2024). While BoVW is a significant improvement over CNN and transformer classifiers, its reliance on pixel clusters means it is not as interpretable as this paper's method. In the second step, these feature vectors are then used in standard machine learning classifiers to predict the image content. This method and its advantages over previous approaches are demonstrated in an application to protest image analysis. Every day, people take to the streets to protest and images of these protests are shared widely on social media. Existing methods are able to classify accurately this material as protest images, but they make it difficult to understand what particular content in an image depicting a contentious scene leads the algorithm to predict protest. Our method addresses this challenge.

Our two-level approach achieves a slightly lower predictive performance than state-of-the-art image classification methods, but it has three advantages. First, simply knowing which objects are in a protest image is an advance in transparency over conventional computer vision models. Existing approaches emphasize pixels, but pixels do not often map to human-understandable concepts. In addition, the abstract descriptions of the images—that is, the segments they contain—can be shared as part of the replication material, which often does not include the images themselves for privacy and copyright reasons. There is no equivalent for pixel-based vision models. The closest, sharing the image embedding, places an image in multidimensional space but does not reveal anything about what the image contains. Second, the researcher can study which objects and associated features typically lead to an image being classified in a particular way. For protests, people and items such as banners and signs distinguish protest from non-protest images. Third, this method can be used to study protest features specific to particular events. We show how our method can be used to identify objects that are predictive of protest only in particular countries, which enables comparative research on protest tactics. Hence, we demonstrate that the increasing focus on interpretable AI (Guidotti *et al.* 2019) is not only an aim in itself, but can lead to new and improved research that conventional methods cannot be used for.

Several extensions are possible. First, all our feature generation methods use *atomic* objects, that is, some measure of the object independent of other types of objects. Features could also be generated which represent dyadic relationships, for example, how far an object of a certain type is located away from an object of another type. However, the combinatorial explosion of the feature space makes this difficult. Second, researchers can rely on other image segmentation methods, including those provided by commercial actors such as Amazon's Rekognition. We did not do this on purpose, and chose to rely on open-source products that are available to other researchers (at no or low cost), and which allow for the replication of our results. If researchers choose to prioritize speed over transparency and replication, commercial tools can be an option at the first stage. Third, rather than using a generic vocabulary, objects and feature generations could be used that are based on theoretical priors, where one constrains the object (or feature) space because of existing theoretical or qualitative knowledge. In contrast to our inductive approach that requires no prior knowledge, the use of domain-specific segment categories can likely increase performance for particular applications. At the same time, however, constraining will result in a model that is custom-tailored to specific phenomena the researcher is interested in and hence no longer as generally applicable as our method. Fourth, there is ongoing work on open vocabulary segmentation to let the user specify their own vocabulary rather than using a pre-specified one (e.g., Zhou *et al.* 2022). For example, LVIS (or any other vocabulary) does not identify police officers or fire, common components of protest imagery. Fifth, and most importantly, the application of this paper's method is not limited to the study of political protest. In fact, any social science image classification task that relies on complex scenes, such as politicians' campaign imagery (Xi *et al.* 2020), may be performed with this two-level approach.

At a practical level, readers may wonder about the availability of images for research purposes in general. As a result of Elon Musk's purchase of Twitter, the ability to collect large numbers of images

https://doi.org/10.1017/pan.2024.18 Published online by Cambridge University Press

from there on an academic's budget has been severely curtailed. Supplementary Appendix H explains these new restrictions in more detail as well as three alternatives: using images from already downloaded tweets, scraping, or working within the European Union's Digital Services Act. Though research with Twitter is no longer as easy as it used to be, the future of research with images remains as promising as ever.

Acknowledgements. The authors gratefully acknowledge comments from Andreu Casas, Nora Webb-Williams, and Han Zhang, as well as Michelle Torres at the 2023 Annual Meeting of the American Political Science Association.

Funding Statement. This project was funded by a research award from the Dr. K. H. Eberle Foundation.

Data Availability Statement. Replication code, model weights, and data for this article have been published on Dataverse at https://doi.org/10.7910/DVN/TFTEF2 (Scholz *et al.* 2024). To enable users to apply the method to their own images, we provide an interactive demonstration application with reduced functionality (https://huggingface.co/spaces/ciass/protest-segments) and an API (https://huggingface.co/spaces/ciass/protest-segments?view=api) via Hugging Face.

Supplementary Material. For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2024.18.

References

Abrams, B., and P. Gardner. 2023 *Symbolic Objects in Contentious Politics*. Ann Arbor: University of Michigan Press.

Azhari, T. 2019. "One Month on: Hope, Defiance as Lebanon Protests Persist." *Al Jazeera*, November 17.

BBC. 2020. "Covid-19: Protests as Argentina's Cases Pass 900,000." *BBC News*, October 13.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Cantú, F. 2019. "The Fingerprints of Fraud: Evidence from Mexico's 1988 Presidential Election." *American Political Science Review* 113 (3): 710–726.

Casas, A., and N. W. Williams. 2019. "Images That Matter: Online Protests and the Mobilizing Role of Pictures." *Political Research Quarterly* 72 (2): 360–375.

Chen, T., and C. Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Dosovitskiy, A., et al. 2021. "An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale." *International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy.

Everingham, M., L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2010. "The Pascal Visual Object Classes (VOC) Challenge." *International Journal of Computer Vision* 88 (2): 303–338.

Girbau, A., T. Kobayashi, B. Renoust, Y. Matsui, and S. Satoh. 2024. "Face Detection, Tracking, and Classification from Large-Scale News Archives for Analysis of Key Political Figures." *Political Analysis* 32 (2): 221–239.

Girshick, R., J. Donahue, T. Darrell, and J. Malik. 2014. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 580–587.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. 2019. "A Survey of Methods for Explaining Black Box Models." *ACM Computing Surveys* 51 (5): 1–42.

Gupta, A., P. Dollar, and R. Girshick. 2019. "LVIS: A Dataset for Large Vocabulary Instance Segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

He, K., G. Gkioxari, P. Dollar, and R. Girshick. 2017. "Mask R-CNN." *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.

Joo, J., and Z. C. Steinert-Threlkeld. 2022. "Image as Data: Automated Content Analysis for Visual Presentations of Political Actors and Events." *Computational Communication Research* 4 (1): 11–67.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems* 25: 1097–1105.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–444.

Li, F., et al. 2023. "Mask DINO: Towards a Unified Transformer-Based Framework for Object Detection and Segmentation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3041–3050.

Lin, T.-Y., et al. 2014. "Microsoft COCO: Common Objects in Context." In *Computer Vision ECCV 2014*, edited by D. Fleet, et al. Cham: Springer International Publishing.

McCulloch, A., and D. Mikhael. 2019. "As Protests Continue, Lebanon's Sectarian Power-Sharing Stalemate Must End." *The Conversation*, December 20.

Muchlinski, D., D. Siroky, J. He, and M. Kocher. 2016. "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data." *Political Analysis* 24 (1): 87–103.

https://doi.org/10.1017/pan.2024.18 Published online by Cambridge University Press

Ortiz, I., S. Burke, M. Berrada, and H. Saenz Cortés. 2022. *World Protests: A Study of Key Protest Issues in the 21st Century*. Cham: Palgrave Macmillan.

Raleigh, C., A. Linke, H. Hegre, and J. Karlsen. 2010. "Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature." *Journal of Peace Research* 47 (5): 651–660.

Reuters. 2021. "South Africa: 28 Arrests over Protests Sparked by Ex-Presidents Jailing." *The Guardian*, July 10.

Roth, A. 2019. "Thousands March in Moscow Demanding Open City Elections." *The Guardian*, August 10.

Rudin, C. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1 (5): 206–215.

Scholz, S., N. B. Weidmann, Z. C. Steinert-Threlkeld, E. Keremoglu, and B. Goldlücke. 2024. "Replication Data for: Improving Computer Vision Interpretability: Transparent Two-Level Classification for Complex Scenes." https://doi.org/10.7910/DVN/TFTEF2.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2017. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618 626. Venice, IEEE.

Steinert-Threlkeld, Z., and J. Joo. 2022. "MMCHIVED: Multimodal Chile and Venezuela Protest Event Data." In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16, 1332–1341.

Steinert-Threlkeld, Z. C., A. M. Chan, and J. Joo. 2022. "How State and Protester Violence Affect Protest Dynamics." *The Journal of Politics* 84 (2): 798–813.

Sundararajan, M., A. Taly, and Q. Yan. 2017. "Axiomatic Attribution for Deep Networks." In *Proceedings of the 34th International Conference on Machine Learning*, 3319–3328.

Torres, M. 2024. "A Framework for the Unsupervised and Semi-Supervised Analysis of Visual Frames." *Political Analysis* 32 (2), 199–220.

Torres, M., and F. Cantú. 2022. "Learning to See: Convolutional Neural Networks for the Analysis of Social Science Data." *Political Analysis* 30 (1): 113–131.

Wang, Y. 2019. "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data: A Comment." *Political Analysis* 27 (1): 107–110.

Webb Williams, N., A. Casas, and J. D. Wilkerson. 2020. *Images as Data for Social Science Research: An Introduction to Convolutional Neural Nets for Image Classification*, 1st Edn. Cambridge: Cambridge University Press.

Won, D., Z. C. Steinert-Threlkeld, and J. Joo. 2017. "Protest Activity Detection and Perceived Violence Estimation from Social Media Images." In *Proceedings of the 25th ACM International Conference on Multimedia*, 786–794.

Xi, N., D. Ma, M. Liou, Z. C. Steinert-Threlkeld, J. Anastasopoulos, and J. Joo. 2020. "Understanding the Political Ideology of Legislators from Social Media Images." In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14, 726–737.

Zeiler, M. D., and R. Fergus. 2014. "Visualizing and Understanding Convolutional Networks." In *Computer Vision ECCV 2014*, edited by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars. Cham: Springer International Publishing.

Zhang, H., and J. Pan. 2019. "CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media." *Sociological Methodology* 49 (1): 1–57.

Zhou, X., R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra. 2022. "Detecting Twenty-Thousand Classes Using Image-Level Supervision." In *European Conference on Computer Vision*, 350–368.

Cite this article: Scholz, S., Weidmann, N. B., Steinert-Threlkeld, Z. C., Keremoğlu, E. and Goldlücke, B., (2024). Improving Computer Vision Interpretability: Transparent Two-Level Classification for Complex Scenes. *Political Analysis*, 1–15. https://doi.org/10.1017/pan.2024.18