Original Article

Sociological Methods & Research
1–54

DOI: 10.1177/00491241221082603
journals.sagepub.com/home/smr
SAGE

Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research

Han Zhang[1] and **Yilang Peng**[2]

Abstract

Automated image analysis has received increasing attention in social scientific research, yet existing scholarship has mostly covered the application of supervised learning to classify images into predefined categories. This study focuses on the task of unsupervised image clustering, which aims to automatically discover categories from unlabelled image data. We first review the steps to perform image clustering and then focus on one key challenge in this task—finding intermediate representations of images. We present several methods of extracting intermediate image representations, including the bag-of-visual-words model, self-supervised learning, and transfer learning (in particular, feature extraction with pretrained models). We compare these methods using various visual datasets, including images related to protests in China from Weibo, images about climate change on Instagram, and profile images of the Russian Internet Research Agency on Twitter. In addition, we propose a systematic way to interpret and validate clustering solutions. Results show that transfer learning significantly

[1] Division of Social Science, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
[2] Department of Financial Planning, Housing, and Consumer Economics, University of Georgia, Athens, GA, USA

Corresponding Author:
Yilang Peng, Department of Financial Planning, Housing, and Consumer Economics, University of Georgia, Athens, GA, 30602, USA.
Email: yilang.peng@uga.edu

outperforms the other methods. The dataset used in the pretrained model critically determines what categories the algorithms can discover.

Keywords

computational social sciecne, machine learning, visual data, image as data, computer vision, unsupervised learning, image clustering

Introduction

Visual information is indispensable for conveying messages. Visual data can take different forms, such as historical images, news photographs, television programs, and social media posts. Scholars have found that images can convey information and trigger emotions, sometimes more powerfully than text (Casas and Williams 2019; Paivio 1990)—as the famous saying suggests, "a picture is worth a thousand words." Visual data have become even more abundant in the current social media age. YouTube and Instagram, two platforms that predominantly circulate visual content, rank first and third on the list of most used social media sites in the United States (Pew Research Center 2019).

Image data are frequently used by sociologists to explore meanings, such as inferring protesters' emotions during protests (Corrigall-Brown and Wilkes 2012; Oleinik 2015), understanding how media frame abortion issues(Rohlinger and Klein 2012), or studying terrorists' visual propaganda strategies (Baele, Boyd, and Coan 2020). In these studies, researchers rely on content analysis techniques by looking at images piece by piece and summarizing common themes by hand. As cultural sociologists have known for a long time, human reading of images is subject to reproducibility issues even for small datasets (DiMaggio, Nag, and Blei 2013). As the amount of image data will only become more plentiful in the future due to the quick rise of visual-information-based social networking platforms, traditional content analysis approaches are increasingly subject to scalability concerns.

Computer vision, a subfield of computer science that aims to train computers to understand digital imagery, has provided social scientists with tools for computational visual analysis (Joo and Steinert-Threlkeld 2018; Peng forthcoming; Torres and Cantu 2022; Williams, Casas, and Wilkerson 2020). An emerging line of social science scholarship uses automated image analysis to answer questions relevant to social scientific research. For example, image

analysis has been used for detecting partisan media bias in news photographs (Peng 2018), predicting poverty from satellite images (Jean et al. 2016), and detecting protests and estimating their size from social media images (Sobolev et al. 2020; Zhang and Pan 2019). The community of computational social science has also provided reviews and tutorials that explain the task of image classification (Joo and Steinert-Threlkeld 2018; Torres and Cantu 2021; Williams, Casas, and Wilkerson 2020).

Still, the introductory work and prior application of computer vision in social scientific research has predominantly focused on *supervised* approaches that train computer algorithms to map input images to predetermined categories. Fewer studies in social scientific research have applied *unsupervised* methods to automatically discover meaningful categories and patterns from images without existing labels. Unsupervised methods have been proven useful for large-scale text data. In particular, topic modeling, an unsupervised technique that automatically finds topics from a collection of textual documents, has become quite popular in recent years (Barbera et al. 2019; Murashka, Liu, and Peng 2021; Roberts et al. 2014). With topic modeling, researchers can efficiently summarize textual data and examine the associations between textual messages and various outcomes. Researchers have used topic models for revealing the structure of scientific knowledge (Song, Eberl, and Eisele 2020a), testing how political messages influence audience thinking (Roberts et al. 2014), and exploring cultural meanings in large-scale newspaper texts (DiMaggio 2015; DiMaggio, Nag, and Blei 2013).

In a similar vein, unsupervised methods of image categorization can also be useful for social scientists to reveal hidden themes and topics in visual media. The computer vision community in computer science has already begun to turn their attention to unsupervised image analysis and has proposed several image clustering methods (Dueck and Frey 2007; Frey and Dueck 2007; Guerin et al. 2017). A few social scientific studies have also applied clustering methods to images, demonstrating the potential of unsupervised methods in discovering meaningful patterns in visual content, such as content categories in Instagram images (Hu et al. 2014; Manikonda and De Choudhury 2017; Peng 2021) and types of gestures in videos of politicians (Kang et al. 2020a). Still, the social scientific community lacks a comprehensive guide on how to perform image clustering on social scientific visual data and the advantages and caveats of different approaches.

To give readers an example of how image clustering can help social scientists investigate visual data, we turn to a dataset called CASM-China (Zhang and Pan 2019) CASM-China contains over 136,330 protests and 302,506

Weibo images associated with the protests in China from 2010 to the middle of 2017.[1] Zhang and Pan (2019) found image data indispensable for identifying offline protests from online social media data because many protesters attempting to mobilize in China avoid posting textual content, which has been shown prone to censorship (King, Pan, and Roberts 2013), but rather post images instead; they nevertheless did not study how protesters use images to mobilize in detail. To explore how protesters use images in Chinese social protests, we randomly selected 60 images from the CASM-China dataset and show them in Figure 1. At first sight, these pictures reveal different aspects of protests in China (with people gathering, holding banners or hand-written petition letters). However, we are not sure whether the sampled images reveal other ways protesters use images for mobilization in the full dataset. We are also not clear about the prevalence of different types of ways that people use images. We need a systematic method to formally identify groups of ways protesters use images during Chinese social protests. In Result, we will show how our proposed unsupervised image clustering algorithm can automatically group images in CASM-China into meaningful categories.

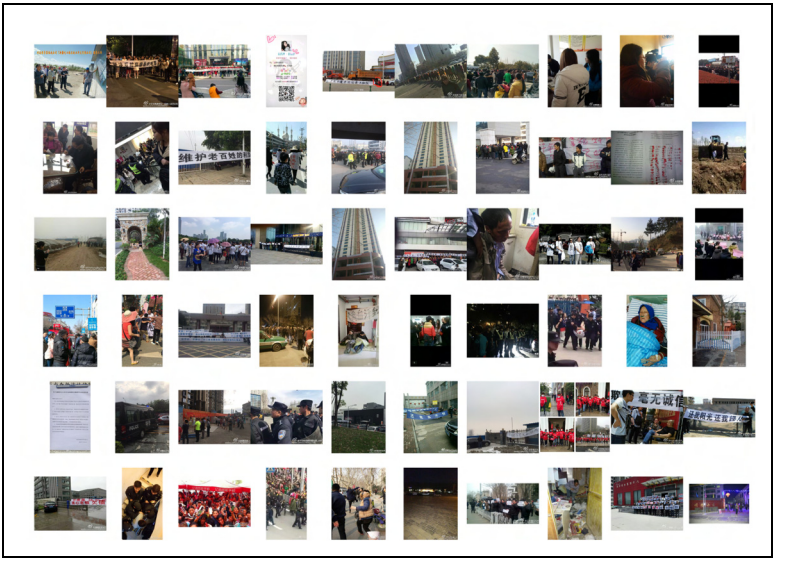

Figure 1. Randomly selected images from the CASM-China protest dataset.

This paper introduces image clustering as an unsupervised approach to image analysis. We propose four steps in image clustering: preprocessing images, transforming images into intermediate representations, clustering, and validating results. Here we use the word "unsupervised" to refer to the main goal of image clustering, which is to automatically discover categories from unlabelled images. Nevertheless, it is possible that all types of machine learning techniques, such as supervised learning, unsupervised learning, and self-supervised learning, can be used in some steps of image clustering, especially extracting intermediate representations of images. In particular, we discuss three specific methods of extracting image representations that efficiently capture the relevant visual information, which has proven to be the key challenge of image clustering. We then test the performance of different image clustering algorithms on a diversity of visual datasets, including images of protests from Weibo, a popular social media platform in China, images published by Instagram accounts about climate change, and profile images of the Russian Internet Research Agency on Twitter. We conclude with the advantages and limitations of different image clustering techniques.

Steps in Performing Image Clustering

We propose four steps in an image clustering task: (1) preprocessing images; (2) extracting intermediate representations; (3) clustering; and (4) interpreting and validating the clustering results (Figure 2). In this section, we briefly overview these four steps to give readers a road map of image clustering. In practice, we found that extracting representations from images is a key challenge in image clustering that critically shapes the categories that emerge from image data. Also, the validation and interpretation of image clusters has its own unique challenges. Therefore, in the following two sections, we discuss the two steps—extracting image representations and interpreting/validating the results in detail.

We first start with how image data are represented. In modern computers, digital images can be stored as matrices with numeric values, also known as *pixel* representations (Parker 2010). For instance, an 8-bit $a \times b$ grayscale image is represented as a matrix X of size $a \times b$, in which each cell (known as a pixel) takes values ranging from 0 (completely black) to 255 (completely white). The $a \times b$ matrix X can also be represented as a *vector* of length $a \cdot b$ by concatenating each row together. In a color image, each pixel is typically represented as a collection of values. For example, RGB (Red, Green, and Blue) color space represents each pixel's color as a combination of red,

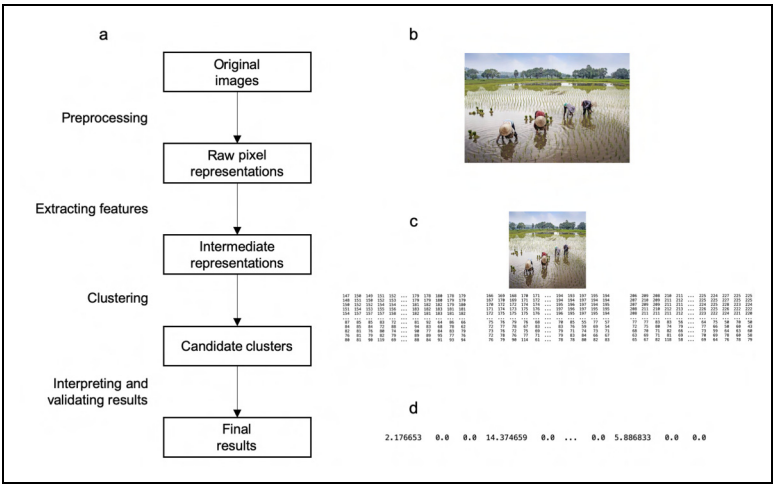

Figure 2. (a) steps in image clustering. (b–d) An example of feature extraction: (b) the original image ($600 \times 400 \times 3$); (c) the resized image ($224 \times 224 \times 3$); (d) the intermediate representation (4096×1) using feature extraction with transfer learning.

green, and blue colors. An RGB color image is thus stored as a three-dimensional matrix of size $a \times b \times 3$.

Step 1: Preprocessing Images: A few preprocessing steps might be taken. First, images from online sources often come in a variety of file formats, such as JPEG, PNG, and GIF. Researchers may transform images into one common format to facilitate further processing. In addition, images often come in a variety of sizes, and researchers need to resize them to the same dimension. Large images might take up a lot of storage and processing power. On the other hand, we also do not want to shrink the image size too small (e.g., 20×20) because that will lose information. Resizing the image resolution to be between 200×200 to 300×300 is the most common practice in the literature. Finally, many image processing methods require the input images to have the same width and height. There are two ways to achieve this: resizing via interpolation or zero-padding (adding columns or rows of all zeros into the pixel matrix of images until they have the same width and height). Hashemi (2019) found that in practice, resizing and zero-padding achieve similar performances. Panel (b) to (c) in Figure 2 visualizes the process to turn an original image into a resized image and its pixel representation.

Step 2: Extracting Intermediate Representations: The pixel representation of images may be too complicated for clustering because it is intrinsically *high-dimensional*. In statistics and machine learning, high-dimensional data means that "the number of features or covariates can even be larger than the number of samples" (Narisetty 2020), whereas low-dimensional data means the opposite—the number of features is smaller than the number of observations (Grimmer, Roberts, and Stewart 2021). For example, the pixel representation of a 800×600 RGB image has a dimension of $800 \cdot 600 \cdot 3 = 1{,}440{,}000$, which easily exceeds the sample size of datasets typically found in social scientific research. The high-dimensional nature of image data yields major challenges for the statistical methods typically used in social science.[2] Moreover, pixel representations of images usually contain *redundant* information. For instance, background objects, light, as well as colors are all irrelevant if the study goal is to tell to tell whether an image contains a cat or a dog; only the contour of the animal is relevant. Therefore. we need to transform images into an intermediate representation that is low-dimensional and could be used for clustering algorithms in the next step. Panel (c) to (d) in Figure 2 visualizes the result of turning the pixel representation of the image into an intermediate representation, a 4096-dimensional vector. We will cover the specific methods of extracting image representations in more detail in Section "Methods for Extracting Intermediate Representations".

Step 3: Clustering: Next, we can apply clustering algorithms to the extracted intermediate representations to group similar images together. There is a massive literature on clustering algorithms; for a review, see Hastie et al. (2009) and Murphy (2012). Clustering algorithms differ regarding whether a unit (here, an image) can belong to a single category only (single membership) or whether it can be assigned to multiple groups with different belonging probabilities (mixed membership). Single-membership algorithms include k-means, affinity propagation, and hierarchical clustering, among others. Mixed-membership clustering algorithms have proven to be especially useful for text and network data (Airoldi et al. 2009; Blei, Ng, and Jordan 2003; Roberts et al. 2014). In this research, we adopted k-means, a popular clustering method that has often been used by social scientists.

Step 4: Interpreting and Validating Results: After the clustering algorithm sorts unlabelled images into clusters, researchers need to further interpret and validate the final solution. Validation is perhaps as important as clustering algorithms itself, as Grimmer and Stewart (2013) advocate: "Validate, Validate, Validate." Validating image clustering results may share some similarities with clustering on other types of data, such as text data, but it also has its unique characteristics. We discuss it in detail in Section "Validation".

Methods for Extracting Intermediate Representations

Having outlined the steps in image clustering, we now move to the key challenge of image clustering: mapping pixel representations of an image to an intermediate representation that has lower dimensions but still preserves important information in images. As noted earlier, while the overall goal of image clustering is unsupervised, it is possible that various machine learning techniques, ranging from supervised, unsupervised, to self-supervised learning, can be used in this step of extracting representations. We introduce one traditional approach (bag-of-visual-words) and two deep learning-based approaches. Figure 3 illustrates the three approaches.

Bag-of-Visual-Words Model

The bag-of-visual-words model is a widely used method to extract image representations (Csurka et al. 2004; Sivic and Zisserman 2003); it has also caught social scientists' attention (Torres and Cantu 2021). The bag-of-visual-words model takes two steps. In the first step, it uses the

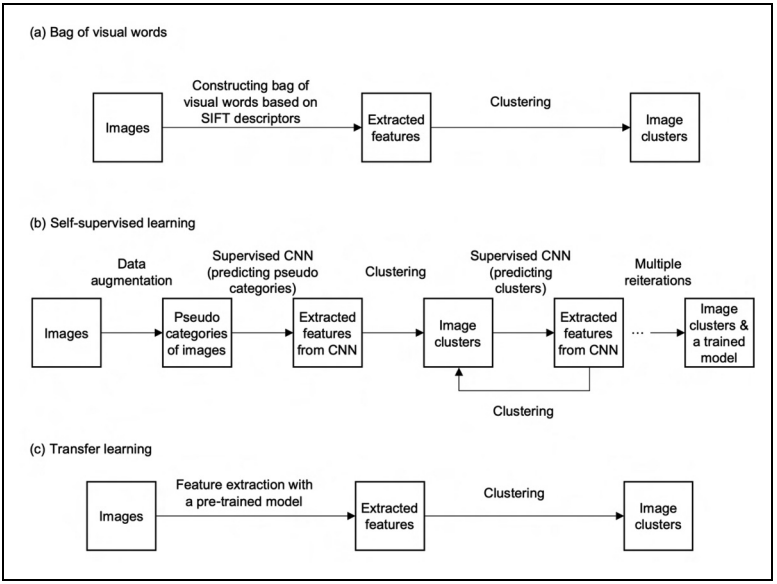

Figure 3. Three approaches of learning image representations and clustering. (a) Bag of visual words, (b) Slef-supervised learning, (c) Transfer learning.

SIFT algorithm to extract scale-invariant basic features (i.e., features do not change invariant to image resizing and rotation) (Lowe 1999, 2004). Basic features of an image include contours of objects (referred as edges in the computer vision literature) and corners (intersection of two edges), among others (Szeliski 2010). In the second step, the bag-of-visual-words model transforms the output of the SIFT algorithm into vectors of the same dimension.[3] It draws inspiration from the bag-ofwords model in natural language processing (Grimmer and Stewart 2013). In text analysis, the bag-of-words model represents each document as a vector of occurrence counts of various words; that is, a distribution over the vocabulary. Analogously, the bag-of-visual-word model represents each image as a vector of occurrence counts of image features extracted from SIFT algorithm; that is, a distribution over the "image" vocabulary.[4]

Deep-Learning Based Methods

The traditional method, such as bag-of-visual-words model, has a predetermined way to combine basic features into more complex patterns (e.g., combinations of edges of different angles). More recent methods of deep learning extract image representations in a way not depending on predetermined rules but based on learning from data. Before we introduce two methods based on deep learning, it might be worthwhile to talk about why we can use deep learning models to perform supervised tasks.

Deep learning uses multilayer neural networks to predict output (categorical or continuous variables) from input (in this case, images). Mathematically, an input matrix X is mapped to output Y through multiple functions (also called layers) $h_i(\cdot)$, $Y = h_d(h_{d-1}(\cdots(h_1(XW_1))W_{d-1})$. Each layer $h_i(\cdot)$ is a nonlinear function applied to a linear weighted sum of its inputs; the weights W_i are learned from the input data; there is a total of d layers.

The pioneers of deep learning argue that the ability to learn simple yet meaningful representations of raw input is the major reason why deep learning outperforms the traditional methods (LeCun, Bengio, and Hinton 2015). The reason lies in that deep learning learns image representations in a hierarchical fashion (LeCun, Bengio, and Hinton 2015). Lower-level neural networks (e.g., $h_1(\cdot)$ and $h_2(\cdot)$ keep basic features such as edges and corners. Upper-level neural networks learn to combine these basic features. The input to the last neural network is $Z = h_{d-1}(\cdots(h_1(XW_1)W_{d-1})$. Z is usually considered the *extracted low-dimensional representation* of the input images under deep learning.

Due to deep learning's ability to extract concise yet meaningful representations, it achieves superior performance on many supervised tasks. Notably, deep learning-based algorithms significantly outperform algorithms that use traditional approaches in the ImageNet Large Scale Visual Recognition Challenge, a task that classifies 1.28 million images into 1,000 object categories (this article refers to this dataset as the ImageNet dataset) (Deng et al. 2009).[5] Three supervised deep learning models are the most widely used in the literature: AlexNet (Krizhevsky, Sutskever, and Hinton 2012), VGGNet (Simonyan and Zisserman 2015), and ResNet (He et al. 2016). Each has achieved good performance, winning first or second place, in the ImageNet Challege in 2012, 2014, and 2015, respectively. These three models differ in their *architecture*, including the form of the h function used and the depth of neural networks (d), among other model details.

Self-Supervised learning. Although standard deep learning models outperform traditional methods in their ability to extract a meaningful representation of raw images, they cannot be directly applied to unsupervised image analysis because researchers do not have labels for input images. *Self-supervised* learning of image features has been recently proposed in the deep learning community to address this limitation (Caron et al. 2018, 2019; Gidaris, Singh, and Komodakis 2018; Yang, Parikh, and Batra 2016)[6] It has also been used to study social scientific problems already (Valensise et al. 2021). The intuition behind self-supervised learning is that we can perform *data augmentation* and create *pseudo*-categories (Gidaris, Singh, and Komodakis 2018). For instance, we can rotate each image in a dataset by 10°, 20°,…, 340°, 350°, effectively expanding the size of the dataset by 35. Then a standard supervised deep learning model can be trained that takes the augmented dataset as input and classifies different rotations of the same image into the same group. After this initial step, the last layer can be extracted as the initial vector of images.

We then apply a predetermined clustering algorithm such as k-means on the initial vectors to group them into K clusters. The new cluster assignment is then used as the new *pseudo*-category, with which we can refit the supervised deep learning model. We iterate between applying a clustering algorithm on the last hidden layer to generate new cluster labels as *pseudo*-category, and using *pseudo*-category to train supervised deep learning models. This iterative process will continue until some convergence threshold is reached. Upon convergence, we obtain both the low-dimensional vector representation of raw images and its cluster assignment for an unlabelled image dataset. The process is called self-supervised because the entire

dataset is used as the training dataset. In contrast, standard supervised learning requires an independent training dataset that has expert-labelled images.

Most studies on self-supervised learning of image representations follow the above iterative process that alternates between training supervised learning algorithms on *pseudo*-category and then clustering the last layer of the learned models. Their differences lie in the choice of deep learning architecture and clustering algorithms. In practice, we choose the method called "DeepCluster," developed by Caron et al. (2018), because it relies on two widely used deep learning architectures (AlexNet and VGG) and the simplest k-means algorithm for clustering.[7]

Transfer learning. Another deep-learning-based method of mapping an image to an intermediate representation is through *transfer learning*. Transfer learning "borrows information" from existing deep learning models trained on external datasets and *repurposes* that model to the datasets at hand (Pan and Yang 2013). The external dataset and the dataset at hand typically are not sampled from the same population. The existing deep learning models are often called *pretrained models*; popular choices of pretrained models include the AlexNet, VGG, and ResNet models trained on the ImageNet dataset.

Transfer learning has been widely used in supervised tasks where researchers have labelled data to predict. Previous research has proposed two types of transfer learning (see Sarkar, Bali, and Ghosh (2018, Chap. 4), TensorFlow (2021) for more details). In one approach called "fine-tuning," researchers keep most layers intact, while allowing the last few layers (i.e., weights) to adapt to the new data. In another approach called "feature extraction with pretrained models," researchers do not need to train their algorithm directly. Instead, researchers feed images into a pre-trained deep learning model, without changing the model's already learned functional forms and weights, and extract features from one of the last few layers. Researchers can use these extracted features in a machine learning model to predict certain labels (Ha et al. 2020).

While these two methods of transfer learning have often been adopted in supervised tasks, we suggest that the second method, "feature extraction with pretrained models," can also be applied to unsupervised image clustering. We perform "feature extraction" by taking the results from one of the final hidden layers as the intermediate representations. Therefore, each image can be transformed into a vector that can be fed into a clustering algorithm. Using the previous notations, imagine someone has fit the deep learning model $Y = h_d(h_{d-1}(\cdots(h_1(XW_1))W_{d-1})$ using X which we refer to as

"pretrained dataset." We have a new dataset X_{new} and we used the estimated $W_1,\cdots W_{d-1}$ to calculate the last hidden layer, and obtain the extracted low-dimensional representation $Z_{new} = h_d{-}_1(\cdots(h_1(X_{new}W_1)W_d{-}_1)$ as the extract feature vector for the new data.

Why can we repurpose pretrained models that are trained on a dataset other than scholar's own to obtain an intermediate representation? As we explained earlier, deep learning models map images into low-dimensional vectors in a hierarchical fashion (LeCun, Bengio, and Hinton 2015). The lower layers in a pretrained model have already "memorized" how to extract basic features from images. Moreover, popular pretrained models, such as those with good performance in the ImageNet Challenge, are trained on gigantic datasets with at least millions of images. The sheer data size allows these models to learn how to extract basic features very well.

Where do social scientists find pretrained models? Computer scientists will often release their trained models alongside their publications. Alternatively, popular software for deep learning research (e.g., TensorFlow (Abadi et al. 2015), Keras (Chollet et al. 2015), and PyTorch (Paszke et al. 2019)) contain off-the-shelf, pretrained models based on popular architectures (AlexNet, VGG, and ResNet) and on the ImageNet dataset. Social scientists have also been using transfer learning models in their research, such as Zhang and Pan (2019) (VGG) and Sobolev et al. (2020) (ResNet).

When applying transfer learning, researchers can choose from a myriad of available pretrained models. We highlight three aspects scholars should consider when choosing their pretrained models.

Supervised vs. Self-supervised Pretrained Models: Researchers first must choose between pretrained models based on supervised or self-supervised tasks. Self-supervised pretrained models learn to distinguish images based on their innate differences (e.g., different rotations) (Caron et al. 2018, 2019; Gidaris, Singh, and Komodakis 2018). On the other hand, supervised models learn a representation of images that are optimal for predicting labels in the training dataset. Even for the same set of images, models trained for predicting one set of categories may not be optimal for predicting another set of categories. For instance, the same images in the ImageNet dataset can be grouped according to whether they contain the same objects (e.g., cat or tree) or whether their scene differs (e.g, outdoor or indoor scenes). Even using the same ImageNet Dataset, a CNN model trained to detect objects is unlikely to produce image representations suited for detecting scene differences, and vice versa (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2015). Therefore, transfer learning using

selfsupervised pretrained models should provide more general representations.[8] However, we note that self-supervised learning is still relatively new, while standard supervised deep learning has a long and mature history. Therefore, most off-the-shelf pretrained models are the results of some supervised tasks.

Training Dataset in Pretrained models: If researchers' dataset differs significantly from the dataset on which pretrained models are trained, pretrained models will likely produce less meaningful presentations (Azizpour et al. 2015). For instance, many popular pretrained models were trained on the ImageNet dataset. If one wants to cluster a dataset full of human faces, a pretrained model (either supervised or self-supervised) trained on ImageNet is unlikely to provide meaningful representations because ImageNet almost does not contain human classes (Zhang and Pan 2019). In this case, it is better to find a pretrained model based on human faces, such as the VGGFace Dataset that contains 2.6 million images and over 2.6 thousand people (Parkhi, Vedaldi, and Zisserman 2015).

Architecture of Pretrained Models: Finally, researchers need to consider what architecture the pretrained model used. A more complex network is usually more powerful in learning image representations. For instance, AlexNet has 8 layers, VGG has 16 or 19 layers, and ResNet has 50, 101, or 152 layers. The more complex model generally performs better on object classification tasks than simple models. In the main results, we used VGG mainly because the self-supervised transfer learning model we used (DeepCluster) is build upon the VGG architecture (Caron et al. 2018). We can compare the self-supervised and supervised transfer learning model trained on the same VGG architecture. We showed the result using ResNet (with supervised transfer learning) in Appendix F.

Validation

We extend the work in unsupervised text analysis, especially topic modeling, to validate image clustering results. Our proposed procedures are easy to use and can help researchers to choose a "optimal" clustering solution among a myriad of clustering solutions.

We focus on *semantic* validity (Grimmer and Stewart 2013; Quinn et al. 2010), namely, whether the images in a cluster constitute a semantically coherent group. We proceed in two ways: visualization and quantitative metrics.

Visualization Through Collated Images

We first recommend researchers visualize clustering solutions in a collated image by 1) sampling M images from each of the K clusters and 2) placing all the images on a $K * M$ canvas (e.g.,). Figure 6 provides an example, with the whole collated image representing one clustering solution and each row representing one cluster. Depending on the research purposes, there are two ways to sample images:

- Representative sampling. One can select M images that are closest to each cluster center (most representative images). This approach can help researchers quickly summarize the main theme of each cluster.
- Random sampling. One can also select a random subset of images from each cluster and see whether they belong to the same category. This approach can help researchers validate the clustering solutions by checking if the images in a cluster are indeed formulating a coherent theme. This approach will be to create a quantitative measure of semantic validity (more shortly).

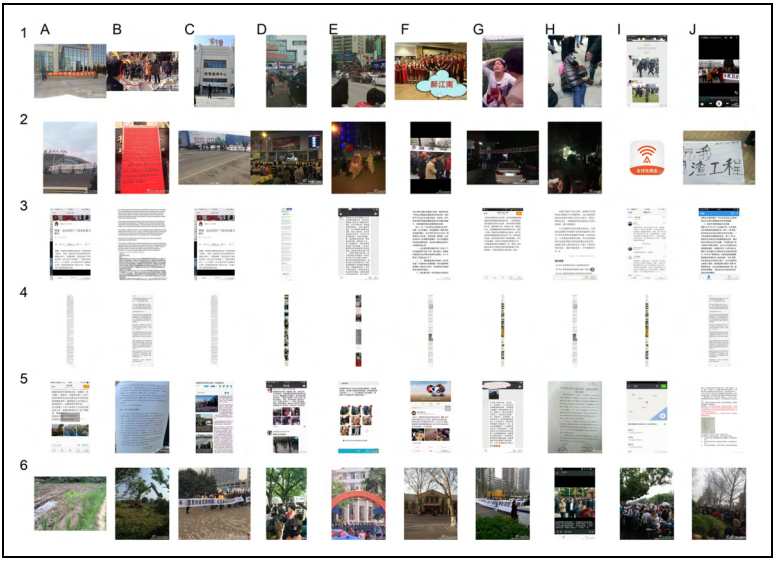

Figure 4. Method 1, Bag-of-Visual-Words model

Inspecting the collated image helps to rule out some obvious unsatisfactory solutions that are not internally coherent (i.e, lacking semantic validity), but as we will show in Section "Performances of Image Clustering Algorithms", it may not always easy for researchers to distinguish two subtly different clustering solutions (e.g., choosing between $K = 6$ or 8). We therefore propose a quantitative measure for semantic validity next.

Quantitative Measure of Semantic Validity: Within-Cluster Consistency

We propose *within-cluster consistency* to quantify how well the clustering solutions identify clusters that are internally consistent. It takes the following steps to calculate within-cluster consistency:

1. Random sampling. First, researchers randomly sample M images from each cluster. We use 20 in practice based on a trade-off between cost and the ability to distinguish different clustering solutions.[9]
2. Visual inspection. By inspecting images from each cluster, researchers can remove clustering solutions that have clusters that are clearly not coherent. This step may be optional, as researchers can still recruit coders to judge the coherency in each cluster (see below). Nevertheless, ruling out some obvious wrong solutions may be a good step to reduce the resources spent on coding. We note that this step should only be used to remove clearly unsatisfactory solutions, but should not be used to establish what clustering solutions are the best.
3. Theme assignment. For each cluster in the candidate clustering solutions, coders browse all sampled images in the cluster and give a label to this cluster, based on what concept the majority of images are portraying. For example, if 12 out of 20 sampled images portray natural scenes, and six out of the 20 sampled images portray indoor meetings (the rest 2 portray separate topics), a *tentative* label "natural scene" is given to the cluster. If no majority theme arise from the cluster (e.g., two topics belong to the same proportion), coders can be given additional images until a majority theme arises. Or they can allow coders to give multiple themes to the same cluster.
4. Theme validation. If there are multiple coders and coders give the same cluster (one set of images) non-identical labels, coders and the researchers will sit down together and decide a final theme for the cluster. For

instance, two coders may give "natural scene" and "outdoor scene" for the same set of images. Researchers and coders need to give a final label to the images. This step is inspired by recent work by Ying, Montgomery, and Stewart (2021), which suggests a similar step for topic modeling result validation. Specifically, following Ying, Montgomery, and Stewart (2021), the final theme should be chosen based on whether it measures the social science concept implied by the label (e.g., whether "natural scene" or "outdoor scene" better fits the study context).

5. Image coding. After a final theme is assigned to each cluster, coders will decide whether each image in the cluster belong to the main theme of the cluster.
6. Within-cluster consistency calculation. For cluster C_j as $\alpha_j = \frac{\sum_{i=\in C_j} I(l_i=\text{mode}(C_j))}{|C_j|}$, where I is an indicator function and $mode(C_j)$ returns the mode of label inC_j (i.e., the most common label in cluster C_j). In plain language, the within-cluster consistency α_j is the proportion of the images in a given cluster that share the most common label in that cluster.

After calculating the within-cluster consistency of each cluster, we proceed with two criteria to rule out unsatisfactory solutions and further choose the best performing ones:

1. First, we need to rule out clustering solutions that have at least one cluster with low withincluster consistency α_j. A small α_j, such as 40%, means that even the most common label only characterizes 40% of the images in that cluster, and the remaining 60% belong to different label groups. In this case, researchers may need to increase K (the number of clusters) to further separate the cluster into two smaller clusters.
2. Next, we calculate average within-cluster consistency across different clusters for each clustering solution. The higher the average consistency, the more images in that category belong to a common theme, and thus the better the clustering solution.

In Section "Performances of Image Clustering Algorithms", we first visually inspect the collated image, calculate within-cluster consistency for different clustering solutions (different K and different methods), and finalize our

clustering solutions. We note that one limitation of our approach is that it is not easy to conduct statistical tests between two clustering solutions.[10]

Performances of Image Clustering Algorithms

We next designed two studies that use real-world datasets to empirically compare the performance of image clustering algorithms. Study 1 compares the performances of the bag-of-visual-words model, self-supervised learning of image representations, and transfer learning. To preview the results, we find that transfer learning significantly outperforms the bag-of-visual-words method and self-supervised learning. Study 2 then focuses on transfer learning and tests how the choices of pretrained datasets impact the clustering results.

Through out this entire section, the results are obtained on a machine with 16 Intel i9-9900K CPU with 3.60 GHz clock rate, and 2 GeForce RTX 2080 SUPER GPU. For programming, we used Python, a popular language in computer vision research.

Study 1: Comparing Different Methods of Image Clustering

Data and methods. Our first dataset comes from CASM-China (Zhang and Pan 2019), as described in the Introduction.

From CASM-China, we selected 2,742 offline protest events in the first half of 2016 that have social media images associated with them.[11] The 2,742 protests contained 14,127 images. We rescaled each image to be of size $224 * 224$ because this is the size of the input used in standard VGG pretrained models.

Method 1 used the bag-of-visual-words model to extract features from images. We expected its performance to be worse than the deep learning-based methods, as described below. Method 2 trained the self-supervised representation learning algorithm (Caron et al. 2018) on the same 14,127 protest Chinese protest images.[12] Method 3 used transfer learning (feature extraction in particular). We used two pretrained models. The first used a standard supervised pretrained model using VGG and ImageNet, and the second used self-supervised pretrained model, "DeepCluster" (Caron et al. 2018).[13]

All three methods used the k-means clustering algorithm. K-means requires a critical parameter, the number of clusters K. Below, we visualize the cluster assignment in collated pictures using $K = 6$ based on our proposed visualization strategy and then present the human validation results with within-cluster consistency, which justifies our choice of K.

Findings from collated images. Figures 4–7 are the collated images. Each row represents a cluster identified by the corresponding method. To highlight the main theme of each category, the images in each category were chosen to be a representative sample by selecting those that are closest to the centroid of the cluster.

Method 1 used bag-of-visual-words model (Figure 4). It produced three clusters of images with textual elements (Clusters 3, 4, and 5). Cluster 6 picks up trees and grasses, which is not an theoretically interesting category for understanding contentious politics. Only Cluster 1 and 2 recognize crowd gathering, but still there are incorrectly classified pictures (e.g., 1.F and 1.G, which display police violence; 2.B, and 2.J, which display hand-written banners; 2.I, which is a Wi-Fi signal irrelevant to protests at all). Considering the fact that we already used images that are closest to the cluster centroids, such that these images should be the most representative, it suggests that Method 1 does not perform very well.

Method 2, which used self-supervised learning (DeepCluster), also did not produce satisfactory clustering solutions (Figure 5). Although it recognized that trees should not be a separate category, it conflated protester's banners

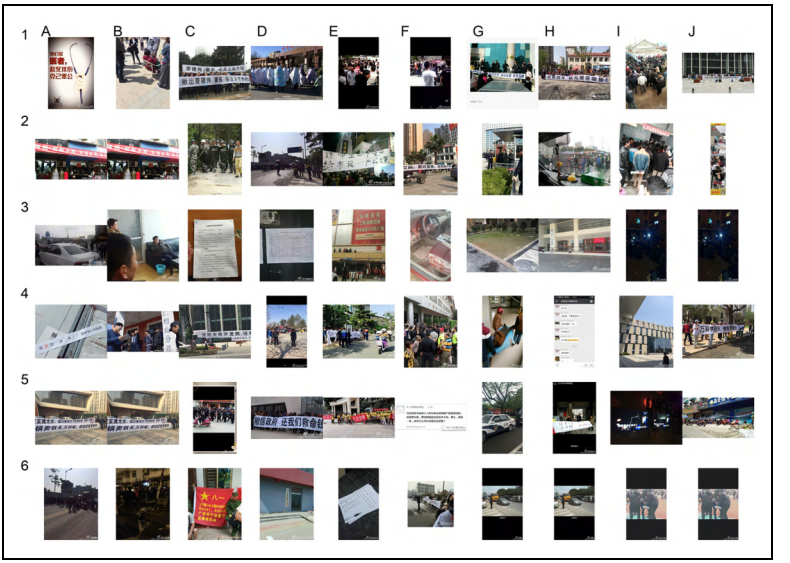

Figure 5. Method 2, self-supervised learning of image representations (Caron et al., 2018).

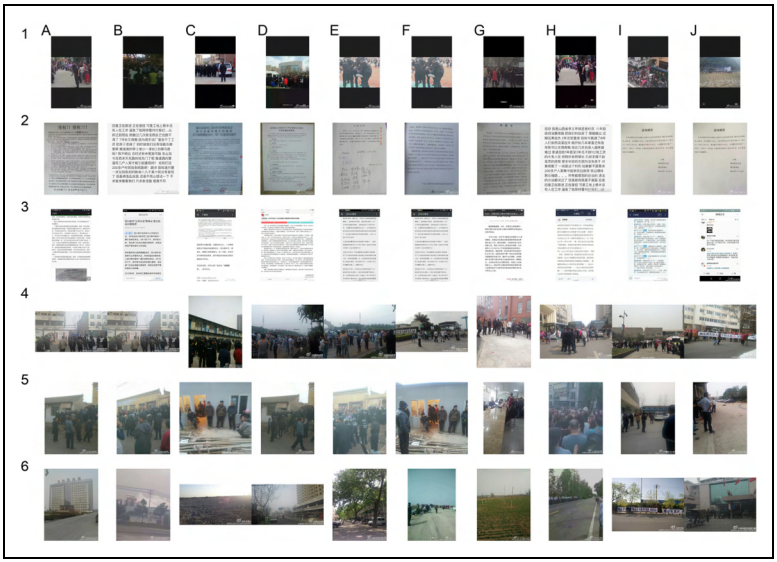

Figure 6. Method 3, transfer learning using supervised pretrained models.

with crowd gathering (Cluster 1), and letters with buildings (Cluster 3). Furthermore, protesters with banners were assigned to different clusters (1, 2, 4, and 5).

Figures 6 and 7 reveal that the transfer learning models (Method 3) both significantly outperformed the two previous methods. Both transfer learning models picked up categories such as gathering of people (Cluster 1 and 5 in Figure 6; Cluster 3 and 5 in Figure 7), hand-written or typed petition letters (Cluster 2 in Figure 6; Cluster 6 in Figure 7), screenshots of text (Cluster 3 in Figure 6; Cluster 4 in Figure 7), and protester with banners.

However, transfer learning based on different pretrained models exhibited differences. Transfer learning using supervised pretrained models treated photos with black backgrounds as a separate category. Transfer learning using self-supervised pretrained models recognized that black borders do not have intrinsic meaning; it did not treat photos with black background as a separate class. Moreover, transfer learning using self-supervised pretrained models recognized that photos about protesters holding banners should be put in the same class (Cluster 4 in Figure 7), while transfer learning using supervised pretrained models did not. Overall, transfer learning based

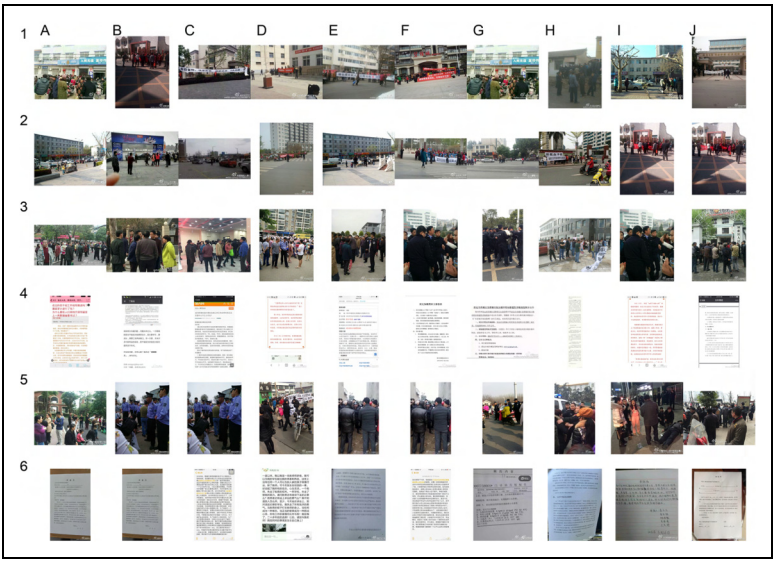

Figure 7. Method 3, transfer learning using self-supervised pretrained models.

on selfsupervised pretrained models produces more meaningful categories. This is expected since selfsupervised learning aims to learn more general features instead of specific features that map images to particular categories in the ImageNet Dataset.

The best model—transfer learning based on self-supervised pretrained models—produced theoretically relevant clusters. Cluster 1 in Figure 7 concerns protesters gathering at the gate of government offices; some protesters held banners. Cluster 2, on the other hand, is relevant to protesters blocking streets. Cluster 3 are also crowd gathering, with a zoomed-in view. Cluster 4 contains screenshots of text, an approach to avoid censorship on Chinese social media, which heavily relies on detecting texts (King, Pan, and Roberts 2013). Cluster 5 is also crowd gathering, but with police presence. Cluster 6 contains petition letters, printed or hand-written. Researchers can potentially apply OCR techniques to extract text from letters and infer protesters' goals.

Quantitative validation. Because visual inspection reveals that the traditional bag-of-visual-words model and self-supervised learning does not yield satisfactory performance, we only instructed research assistants to compare the

results produced by the two transfer learning models. We present a total of six clustering solutions, which vary by the number of clusters (6, 8, and 10 clusters), and two transfer learning models, based on supervised and self-supervised pretrained models. For each cluster produced by a pretrained model, we randomly selected 20 images and had research assistants give a label to each image. We then calculated the within-cluster consistency defined earlier.

Figure 8 shows the human validation results. We proceed in two steps. First, we require the minimum threshold of within-cluster consistency to be 0.5—namely, at least half of the images need to belong to the majority category in that cluster. After this step, only the solutions produced by self-supervised model ($K = 6$ or 8) and supervised model ($K = 8$) that met this minimum threshold were kept and other solutions are discarded. Second, we compared the remaining three clustering solutions with respect to their

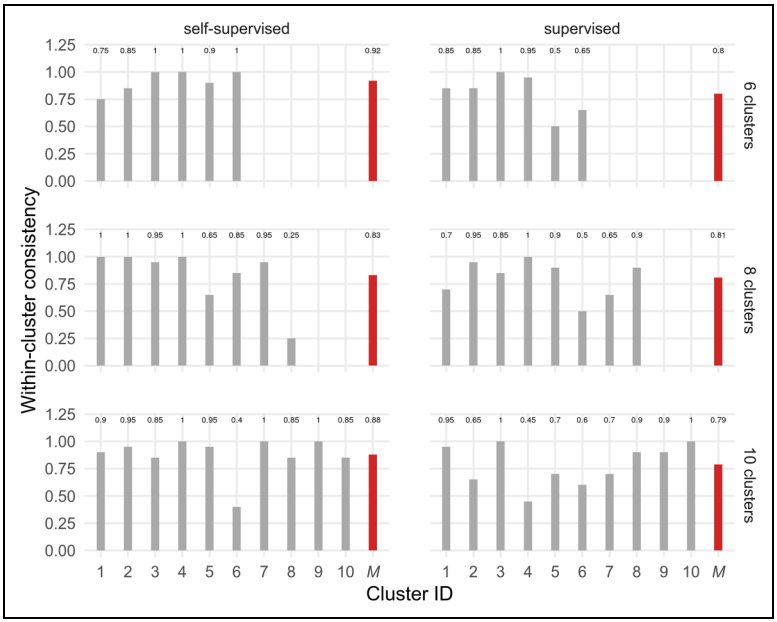

Figure 8. Within-cluster consistency for Chinese protest dataset. Average within-cluster consistency (M) is highlighted in red and the exact values of the average within-cluster consistency is shown on the top of each bar. Six clustering solutions are shown, varying by the number of clusters and the pretrained models.

average within-cluster consistency. Among the three remaining clusters, the self-supervised model with six clusters produced the highest average with-incluster consistency score. Its minimum within-cluster consistency is also significantly larger than the other two remaining solutions (self-supervised model with $K = 8$ and supervised model with $K = 8$). Therefore, we chose the self-supervised model with $K = 6$ as the final clustering solution. This choice is the same as our visual inspection in that self-supervised model slightly outperforms the supervised models. In Appendix C we provided statistical tests to show that our final clustering solution indeed has a statistically higher within-cluster consistency compared with the other five solutions.

Overall, study 1 shows that traditional feature extraction methods (bag-of-visual-words) were less satisfactory. Self-supervised learning using our own dataset, although seemingly attractive and currently popular in computer science, also did not perform well. One possible reason why self-supervised learning did not perform well is that our dataset is still not large enough (n around 10 K), being two scales smaller than the training data used by the two pretrained models (ImageNet dataset, n ¿ 1 M). Moreover, both the bag-of-visual-words and self-supervised learning models took considerable amounts of time to train. In our practice, it took over 24 h to train the selfsupervised learning model (Method 2)[14] and over 3 h to train the bag-of-visual-words model (Method 1). On the contrary, the transfer learning model finished within 5 min.

Moreover, transfer learning is also easier to implement with respect to coding. Python code for implementing Method 3—using the standard supervised pretrained model to extract features—is available in Appendix A. The code demonstrates that we are able to extract vectors within 50 lines of code. Due to the space restriction, we cannot put all codes in the appendix; we will make them public available. Here, we will disclose that Methods 1 and 2 needed significantly more time to implement and could not be finished within 50 lines of code. As traditional approaches and selfsupervised models are slow, harder to implement, and do not have good performance, we do not recommend them to social scientists who plan to apply image clustering methods at this time.

We also compared transfer learning based on supervised and self-supervised pretrained models and found that **self-supervised** pretrained models were slightly better than **supervised** pretrained models in terms of average within-cluster consistency. However, self-supervised learning is still a burgeoning area, and there are fewer off-the-shelf pretrained models. On the other hand, standard software for deep learning, such as Keras, TensorFlow, and PyTorch already have built-in support for using supervised

transfer learning models. Therefore, if scholars cannot find pretrained selfsupervised models, we recommend using pretrained supervised models.

How image clustering is useful in empirical research. While the present research is a methodological piece, it should be useful to demonstrate how image clustering can be used in empirical research (Kang et al. 2020b). We thus offer some examples on how image clustering can inform understanding theoretically interesting questions. First, image clustering might be useful to provide quantitative evidence for existing theories. For instance, one established argument in the qualitative studies of protests in China is that protesters often used disruptive tactics to gain leverage over governments (Cai 2010; Chen 2009, 2012). In particular, blocking government doors and blocking streets are two widely used *disruptive* tactics. Blocking government doors often occurs regardless of whether the targets of the protests are the government itself or other nonstate entities, such as private real-estate developers. Protesters use this tactic to force the government to react (if the government is the target) or intervene and punish the private companies (if the state is not target) (Chen and Cai 2019). The second tactic—blocking streets—is powerful because doing so will attraction attention of the bystander public, who may show sympathy toward protesters and add pressure on the government. In comparison, *violent* tactics in China are often found to be highly ineffective because they bear much higher risk for government repression (Cai 2010); protesters will also consciously limit their actions, trying not to incur physical confrontation with the police (Fu 2017).

Yet, many of the theoretical studies are drawn from qualitative observations but have not gone through scrutiny of quantitative evidence. We plot the number of images in each category in Figure 9 based on our preferred solution (Method 3, transfer learning using self-supervised pretrained models). Confirming the theoretical predictions, protesters gathering at government offices (Cluster 1) and blocking streets (Cluster 2) are the two most prevalent image categories. Crowd gathering with police presence (Cluster 5), which often involves violent confrontation, are considerably less popular than the two categories suggesting tactics. The comparison between images using disruptive tactics versus violent tactics provides quantitative evidence supporting the observations from qualitative studies.

Second, image clustering results may inform new empirical research. As Cluster 4 and 6 have shown, protesters in China often use photos or screenshots of their hand-written/printed petition letters to avoid platform censorship. These texts embedded in the images contain valuable information, especially the claims that were believed to be removed by the censorship

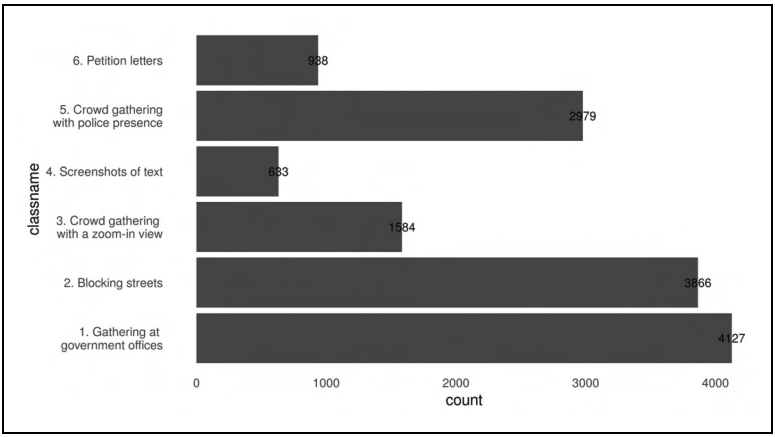

Figure 9. Number of images in each class.

machine, such as criticisms of the government. Traditionally, when researchers study the use of images during protest mobilizations, they mostly examine visual elements such as crowd gathering, fire, and police (Casas and Williams 2019; Oleinik 2015; Steinert-Threlkeld, Chan, and Joo 2021). These text information embedded in information is a neglected aspect (Goebel and Steinhardt 2019). Comparing the text from images to those that they write in text might reflect different strategies the protesters used to mobilize and attract attention from bystanders.

As our discussions have shown, image clustering can be used to measure theoretically interesting concepts (e.g., protest tactics) and to inform new research ideas that have been neglected in previous research (text embedded in images). These two directions will also be our future empirical research plans. These are certainly not the only ways researchers can use image clustering; depending on the dataset and the problem, researchers may discover patterns they have not expected before running image clustering.

Study 2: Comparing Different Pretrained Datasets

Having demonstrated that transfer learning significantly outperforms self-supervised learning from scratch and traditional feature extraction methods such as the bag-of-visual-words model, hereafter we focus on transfer learning. Study 2 examines how the choice of pretrained datasets affects the

clustering results. As discussed in Section "Transfer Learning", the categories of images in the pretrained dataset affect the features memorized in the pretrained model and therefore, what features will be expected to be extracted if we apply the pretrained model on the new data. In transfer learning, it is recommended that the dataset at hand share similarity with the dataset used in the chosen pretrained model.

Selection of datasets. We selected two datasets that can illustrate how the choice of pretrained models impacts the clustering solutions on these two datasets. The first dataset contains images about climate change on Instagram. Previous research has shown that visual messages about climate change often contain a broad range of content themes, including protests, scientific activities, natural landscapes, animals, data visualization, and satellite images (O'Neill and Smith 2014). In February 2020, we curated a list of eleven popular Instagram accounts that extensively publish content related to climate change by using the search keywords "climate change" and "global warming" on Instagram. We then retrieved these accounts' posts using a Python library instaloader.[15] We kept images published before January 31, 2020 for analysis (N = 11,873).

The second dataset is about profile images from disinformation accounts associated with the Russian Internet Research Agency (IRA) on Twitter. In October 2018, Twitter released a dataset that contained accounts it identified as affiliated with IRA. The archive contained these accounts' tweets and associated media. We focused on the profile images (N = 3,709) from these accounts. This dataset is very different from climate change visuals: the majority of the images are human faces, whereas a small subset of the images are icons or logos.

Selection of pretrained models. We selected three pretrained models that used the same architecture (i.e., VGG16) but had been trained on three distinct datasets: ImageNet, Places365, and VGGFace. ImageNet features a diversity of content categories, such as animals (e.g., magpie, jellyfish), natural scenes (e.g., coral reef, lakeside), places (e.g., cinema, restaurant), foods (e.g., pretzels, cheeseburger), vehicles (e.g., aircraft carrier, speedboat), and everyday objects (e.g., joystick, balloon, envelope, volleyball). As visual messages of climate change contain a variety of content, features learned from a diverse dataset like ImageNet should be able to find some categories in climate change visuals. Yet, the ImageNet dataset has almost no categories specifically related to humans, so it may not reveal meaningful categories in the IRA profile images that predominately feature human faces.

In comparison, the Places365 dataset contains about 1.8 million images of 365 scene categories, which include a broad range of indoor (e.g., conference center, classroom, legislative chamber, kitchen) and outdoor settings (e.g., street, skyscraper, forest, ocean) (Zhou et al. 2017). As pictures of climate change communication often feature natural landscapes, urban environments, industry, protests, and conferences, this model should help us discover meaningful categories particularly related to scenes and setting. Still, the Places365 dataset does not contain categories specifically about humans, so we expect that this pretrained model is not very suitable for the IRA profile images.

The VGGFace dataset contains about 2.6 million facial images of over 2.6 thousand people (Parkhi, Vedaldi, and Zisserman 2015). The features extracted should predominantly pertain to human faces. This pretrained model should be less capable of revealing visual categories relevant to the dataset about climate change than the same model trained on ImageNet and Places365. Nevertheless, as the majority of images in the IRA dataset contained human faces, this pretrained model might have better performance for this particular dataset.

To facilitate the comparison among the three pretrained datasets, we retrieved three pretrained models using the same model architecture, VGG16, but trained on different datasets: ImageNet (accessible at https://keras.io/api/applications/), Places365 (https://github.com/ GKalliatakis/Keras-VGG16-places365), and VGGFace (https://github.com/rcmalli/ keras-vggface). We extracted the first dense layer after the convolutional blocks in all three models as our intermediate representations of images.[16] These vectors have 4,096 dimensions. For each model, we performed a principal component analysis and kept the first 200 dimensions for clustering.[17] We then ran k-means clustering with the number of clusters ranging from 5 to 20.

Findings from collated images. We first present the collated images of clustering results to demonstrate how the same deep learning architecture on different pretrained datasets affects the categories machine identified.

Climate Change Visuals: Figures 10–12 show results with $K = 8$ using ImageNet, Places365, and VGGFace, respectively. We randomly selected 20 images from each cluster for inspection.

For climate change visuals, the solution using ImageNet revealed a variety of visual categories (Figure 10). Three clusters (3, 5, and 7) were related to textual messages and graphics, with some variations in design aesthetics. Three clusters (1, 2, and 4) were related to human activities. Cluster 1 captured people, usually several individuals, in outdoor environments, mostly natural scenes or rural areas. These people were also viewed from a distance.

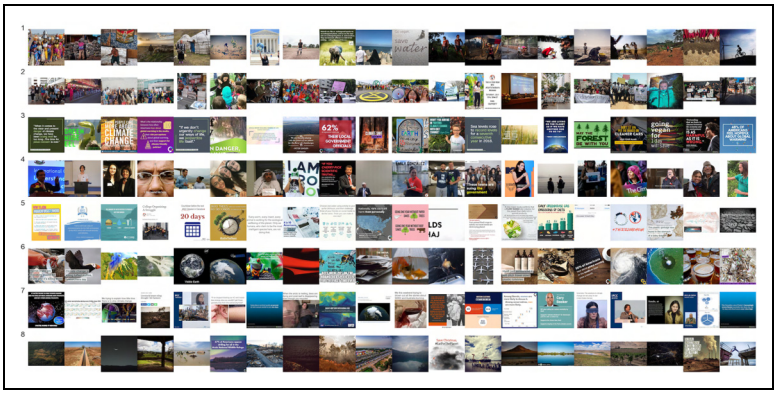

Figure 10. Eight-cluster solution in the climate change dataset based on features extracted from a pretrained model trained on imageNet dataset.

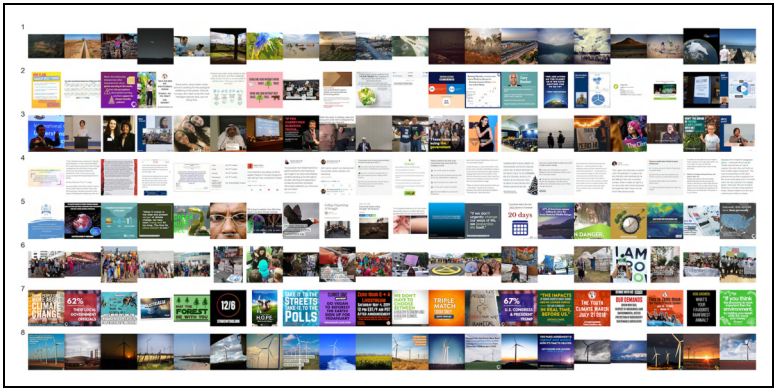

Figure 11. Eight-cluster solution in the climate change dataset based on features extracted from a pretrained model trained on the Places365 dataset.

Cluster 2, in comparison, featured mostly crowds, which were in mostly urban environments such as conferences and streets (protesters). Cluster 4 featured close-ups or medium shots of people, and these images also had a limited number of individuals. Two other clusters also emerged. Cluster 8 featured landscape pictures, usually without people. Cluster 6, however, did not seem to show a clear cohesive theme. It contained a combination of still

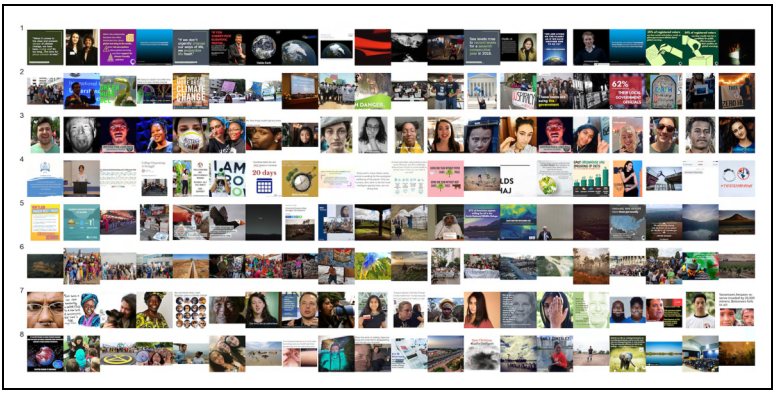

Figure 12. Eight-cluster solution in the climate change dataset based on features extracted from a pretrained model trained on the VGGFace dataset.

objects, animals, and satellite images, potentially because these pictures shared similar visual characteristics. It is also possible that some types of images (e.g., animals, satellite images) only comprised a small percentage of the dataset. If we chose a large number of clusters, this mixed cluster might break down into smaller categories.

The solution using Places365 showed a pattern different from the solution using ImageNet. Four clusters (2, 4, 5, and 7) were related to textual messages and graphics, with variations in design aesthetics. Interestingly, Cluster 4 contained images featuring dense text on a white background, which had not been found in the ImageNet solution. The crucial difference came in how this solution clustered pictures of people. While the ImageNet solution distinguished between pictures of crowds and pictures of several individuals, this Places365 solution mostly grouped pictures of people based on settings. Cluster 3 featured people in indoor settings, such as conference rooms and lecture halls. Cluster 6 mostly featured people in outdoor settings, such as streets. Cluster 1 featured pictures of landscapes. Importantly, pictures of satellite images were also mostly categorized into this cluster. This cluster had both pictures featuring people and pictures without people. This could be because the scene categories in Places365 only concerned the different types of scene, regardless of the presence of people. Finally, Cluster 8 exclusively featured wind farms, a frequently used symbol in climate change messages. This is likely because the Places365 dataset has a scene category—wind farms—and this niche category has distinct visual

characteristics, so this type of image was rediscovered in our dataset regarding climate change. This particular category was absent in ImageNet and VGGFace solutions.

In comparison, the VGGFace solution performed suboptimally. Not surprisingly, this solution discovered two clusters that were entirely related to human faces (Cluster 3 and 7). However, its ability to find other cohesive and meaningful categories was limited. Two clusters (Cluster 1 and 4) predominantly featured images of textual messages and graphics, although these two clusters also mixed in a few other types of images. Cluster 6 featured mostly outdoor scenes, mixing pictures of protests and landscapes. There seemed to be no cohesive themes from the other three clusters (Clusters 2, 5, and 8).

IRA Profile Images: Similarly, we observed that solutions from the three pretrained models also performed differently for the IRA profile images. Figures 13–15 show results with $K = 8$ using ImageNet, Places365, and VGGFace, respectively. We randomly selected 20 images from each cluster for inspection.

For ImageNet, there was a clear division between human (Cluster 1, 2, 3, 7, and 8) and nonhuman clusters (Cluster 4, 5, and 6). Among the clusters featuring humans, the algorithm was able to formulate a cluster featuring solely men (Cluster 1), mostly in suits, a cluster featuring solely women, mostly with long hair (Cluster 7), a cluster featuring people with sunglasses or dark glasses (Cluster 8), a cluster featuring people in environments, usually

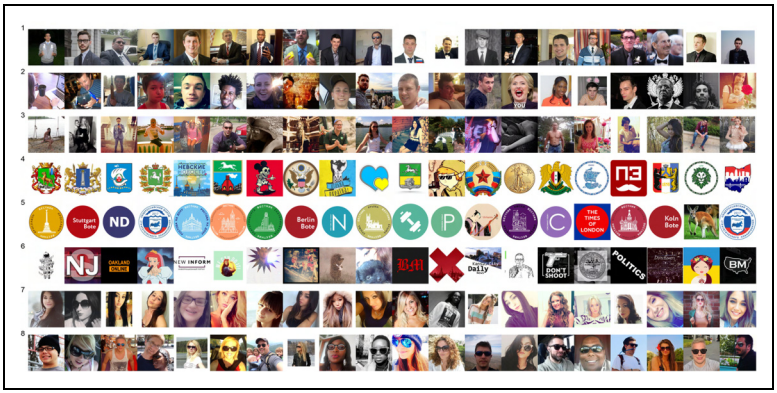

Figure 13. Eight-cluster solution in the IRA dataset based on features extracted from a pretrained model trained on imageNet dataset.

with smaller faces and showing bodies (Cluster 3). There seemed no clear patterns regarding gender, ethnicity, or age differences in Cluster 2. Cluster 4, 5, and 6 mostly featured nonhuman subjects such as icons, logos, or illustrations.

For Places365, the clustering solution was less ideal. First, some clusters even did not have over 20 images, for example, Cluster 4 only had one image, indicating that this pretrained model might result in unbalanced categories. Again, there was a distinction between human (Cluster 1, 3, 4, 5, and 7) and nonhuman clusters (Cluster 2, 6, and 8). Among the clusters featuring humans, the algorithm was able to formulate a cluster featuring mostly women (Cluster 1) and a cluster showing people in outdoor environments (Cluster 3). There seemed no clear patterns regarding other clusters (5 and 7). Cluster 2, 6, and 8 mostly featured nonhuman subjects such as icons, logos, or illustrations.

For VGGFace, the clustering solution was able to find more nuances among categories. Again, there was a distinction between human (Cluster 1, 3–8) and nonhuman clusters (Cluster 2). Among the clusters featuring humans, the algorithm was able to formulate a cluster featuring mostly white men (Cluster 1), two clusters showing mostly white women (Cluster 5 and 6), one cluster featuring mostly African Americans (Cluster 7), and a cluster showing people in environments (Cluster 8). Therefore, compared to the other two pretrained models using ImageNet and Places365, the solution based on VGGFace was more able to pick up visual patterns related to

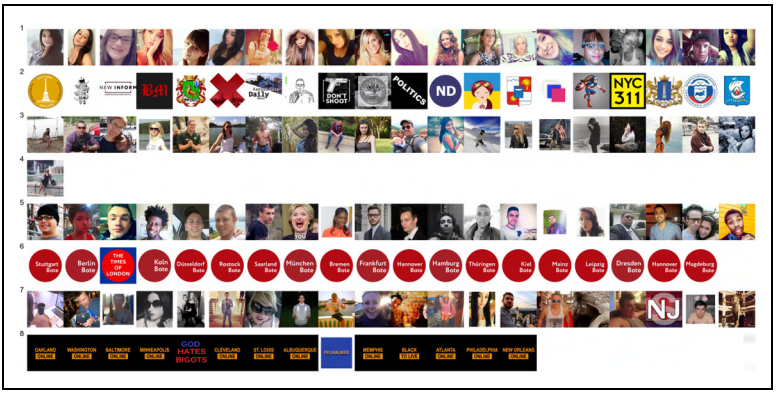

Figure 14. Eight-cluster solution in the IRA dataset based on features extracted from a pretrained model trained on the Places365 dataset.

Figure 15. Eight-cluster solution in the IRA dataset based on features extracted from a pretrained model trained on the VGGFace dataset.

gender and race in human faces. Only one cluster was about nonhuman objects (Cluster 2).

Quantitative validation. Next, we calculated within-cluster consistency for the two datasets, three clustering algorithms (using different pretrained datasets), and K ranging from 6, 8, and 10. Figure 16 shows the human validation results for the climate change dataset. Regardless of K, VGG model trained on Places365 dataset always gives the highest within-cluster consistency, while the same algorithms trained on VGGFace always give the lowest within-cluster consistency.

Figure 17 shows the human validation results for the IRA dataset. Confirming our expectation, VGG model trained on the Places365 dataset yields the lowest within-cluster consistency, regardless of the choice of K. VGG model trained on the ImageNet dataset gives a higher within-cluster consistency than the VGG model on Places365 dataset. VGG model trained on the VGGFace dataset always gives the best within-cluster consistency across different choices of K. It suggests that if one should choose a pretrained model that is trained on a dataset similar to the final dataset used for clustering.[18]

To conclude, Study 2 reveals the choice of pretrained datasets can critically shape the categories found in the clustering solution. Among the three solutions for climate change images, the solution using VGGFace had the worst performance. It was only able to discover categories related to

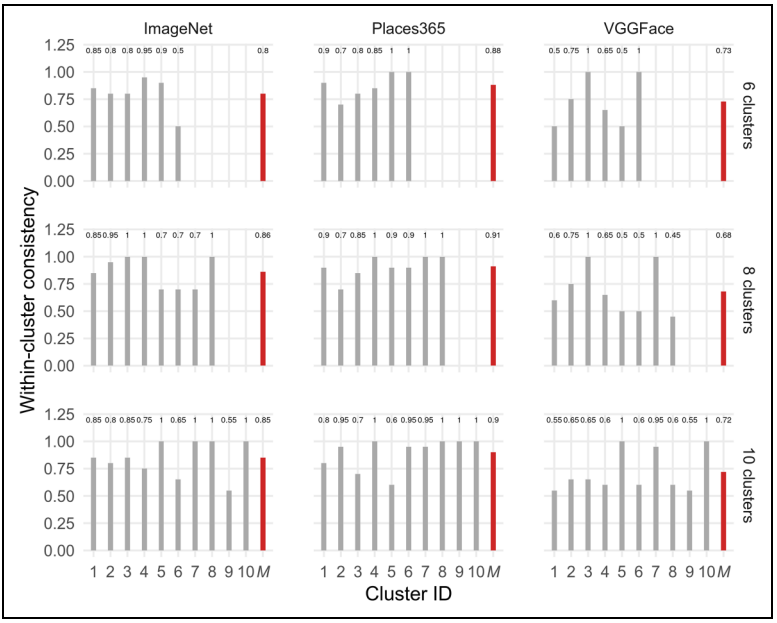

Figure 16. Within-cluster consistency for climate change dataset. Average within-cluster consistency (M) is highlighted in red and the exact values of the average within-cluster consistency is shown on the top of each bar. Nine clustering solutions are shown, varying by the number of clusters and pretrained models.

human faces and missed other visual themes. Both the solution using ImageNet and the solution using Places365 were able to find more meaningful categories, yet they differed on the specific emphasis and categories in the solutions. The solution using Places365 also yields the highest within-cluster consistency based on human coding.

Among the three solutions for IRA profile images, the solution using VGGFace found the most interesting patterns and resulted in the highest within-cluster consistency. For example, this solution was able to formulate categories featuring certain racial or gender groups. Nevertheless, we included this dataset to demonstrate how the choice of pretrained models, and in particular, VGGFace, could influence the categories revealed in image clustering. However, there have already been many well-developed algorithms for facial detection and facial analysis. Therefore, instead of using an unsupervised image clustering approach, it might be more

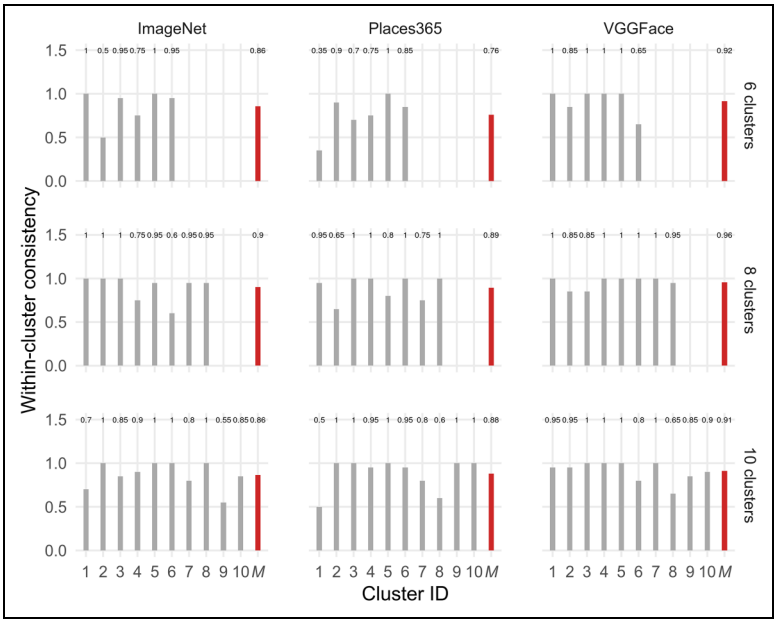

Figure 17. Within-cluster consistency for IRA dataset. Average within-cluster consistency (M) is highlighted in red and the exact values of the average within-cluster consistency is shown on the top of each bar. Nine clustering solutions are shown, varying by the number of clusters and pretrained models.

appropriate to apply a supervised approach to directly analyze facial attributes that might be theoretically meaningful, such as gender, age, and race.

Because our goal in this section is to see how the choice of pretrained dataset matters, we leave the detailed steps of how we choose the best model (which K) in Appendix D. We also employed two coders and calculated the intercoder reliability of these clustering solutions. Appendix E show the intercoder reliability results. For the ones that have high within-cluster consistency (e.g., VGG model trained on VGGFace dataset), the intercoder reliability (Cohen's Kappa) is also high, typically higher than 0.9. Moreover, to add more confidence to our previous observations, for the IRA dataset, VGG models trained on VGGFace dataset generally gave higher intercoder reliability compared with other pretrained dataset choices; VGG models trained on Places365 dataset also have higher intercoder reliability on the climate change dataset.

Other Method Choices

Our studies so far have focused on how different feature extraction methods impact the clustering results, although there might be other methodological choices that impact clustering results, which could be considered by future researchers. First, regarding the clustering algorithm, we have been using a standard algorithm, k-means, throughout the manuscript, but there are other clustering algorithms available. In the Appendix, we presented additional results using agglomerative hierarchical clustering (Figure B1) and Gaussian Mixture model (Figure B2) for the CASM-China dataset. While the additional results look similar to what we have found using the k-means method (Figure 7), we also note that future researchers should also try several clustering algorithms to see if their results are robust under different clustering algorithm choices.

In addition, it is possible to use different model architecture in pretrained models. In this study, we rely on VGG16 consistently for all pretrained models selected, but there are a variety of architecture available. As discussed in Section "Transfer Learning", complex models have better ability to extract meaningful representations and thus lead to better clustering results than simpler models. Appendix F confirms this intuition by comparing clustering results on the CASM-China dataset with pretrained models with three different architecture (i.e., AlexNet, VGG16, and ResNet). We found that other things being equal, transfer learning models trained on AlexNet, an architecture simpler than VGG and ResNet, achieve less satisfactory performance. [19] Future research can more systematically compare pretrained models trained on the same dataset but using different model architectures.

Discussion

In summary, this study introduces the steps in the task of unsupervised image clustering. We argue that one key challenge is to extract concise but meaningful representations of images. We focus on three methods of finding intermediate representations of images: bag-of-visual-words, selfsupervised learning, and transfer learning (in particular, feature extraction via pretrained models). We evaluated these methods on a diversity of datasets. We found that transfer learning significantly outperforms the other two methods, not only with respect to clustering performance but also in terms of the speed of calculation and ease of coding implementation (Study 1). We further discussed and demonstrated some practical considerations in selecting pretrained models for image clustering. We found that the particular pretrained dataset used in transfer learning critically determines the clustering results (Study 2).

This article contributes to the burgeoning study of "image as data" by synthesizing practical steps and methods of unsupervised image clustering for social scientists. Image data are abundant in the current social media age, but so far, the analysis of image data in sociology still heavily relies on researchers looking at images and summarizing common themes through traditional content analysis techniques (Corrigall-Brown and Wilkes 2012; Krippendorff 2004; Oleinik 2015; Rohlinger and Klein 2012). Human reading of images is subject to reproducibility issues even for small datasets, and to scalability concerns for larger image datasets. Much like the advance of text analysis from traditional content analysis to automated, unsupervised methods (especially topic modeling) has fostered a wide range of scholarship capturing meaning in large-scale texts (DiMaggio, Nag, and Blei 2013; Grimmer and Stewart 2013; Wilkerson and Casas 2017), we believe our proposed method also will help future social scientists capture theoretically interesting information in visual data, reduce its complexity, and provide interpretations.

Image clustering does not replace human's reading of images, however. Rather, it provides social scientists a lens to quickly capture meaningful categories from large-scale image datasets and develop these meaningful categories, which can then be used for further theorizing and empirical work. Importantly, we proposed concrete steps (visual inspection through collated images and human coding for within-cluster consistency) to interpret and validate the clustering results. Our article thus strengthens the reproducibility of image analysis by pushing researchers to validate their clustering results instead of offering a post hoc justification for their choice of themes in an image dataset.

It is important to consider the potential biases and ethical implications of computer vision algorithms. Prior research has suggested that machine learning models might incorporate gender, racial, and cultural biases (Zou and Schiebinger 2018). For example, research has shown that facial recognition models often show higher error rates for gender or racial minorities (Grother, Ngan, and Hanaoka 2019; Zou and Schiebinger 2018). Furthermore, many computer vision models have been trained on datasets that predominately contain visual data from Western contexts. Images in the ImageNet dataset used in this study, a dataset popular among computer vision researchers, come from a limited number of cultures, with nearly half of images from the United States (Zou and Schiebinger 2018). In our results, the pretrained models using ImageNet showed satisfactory results for datasets both from Weibo, a Chinese platform and Instagram. Nevertheless, biases in pretrained models may pose a challenge if researchers

are interested in more culture-specific content themes that might not be well reflected in the features learned in some pretrained models.

Although we recommend the use of transfer learning for its effectiveness, speed, and relative simplicity to implement, it may be difficult to choose a pretrained model. In addition, when choosing pretrained models, we recommend researchers use pretrained models that share similarities with the dataset at hand. For example, researchers can evaluate whether the dataset at hand shares similar subjects, sources, and styles with the dataset in the pretrained model. Researchers can also examine the labels in the pretrained dataset and see if these labels potentially map some theoretical concepts. For example, for a dataset of general photographs, some datasets emphasize the presence of specific objects while others code the same photographs regarding settings and scenes. We also note that there has been some progress in quantifying the similarity between pretrained dataset and the target dataset (Cui et al. 2018; Dwivedi and Roig 2019). A future direction is to borrow these measures to inform the selection of pretrained models. Nevertheless, researchers are constrained by the availability of pretrained models. We note that ImageNet, containing over 1000 categories (the full ImageNet contains over 21,841 categories), offers a reasonably good starting point. There are also many pretrained models that utilize this dataset and have been incorporated in popular deep learning packages such as Keras. If researchers cannot find pretrained models that are trained on datasets that are similar to their own, then they have to make a compromise of using a less satisfactory pretrained model.

We close by noting limitations and future directions for research. First, when using pretrained models for supervised tasks, in addition to feature extraction, scholars can also adapt pretrained models to their smaller dataset by keeping some learned parameters and weights constant but allowing others to change with their dataset, a process called "fine-tuning" (Zhang and Pan 2019). However, there has been no standard procedures for fine-tuning pretrained models for unsupervised tasks, which is an opportunity for future methodological research. In addition, this research uses VGG consistently in the paper, but different deep learning architectures could result in various levels of performance. For instance, ResNet is a popular and more complex model than VGG, and there have been studies that used ResNet for self-supervised tasks in the past year (Caron et al. 2020), which could be included in a future comparison. Future research can compare different architecture's performance in more detail. Similarly, there are a wide range of clustering algorithms available in addition to k-means. Future researchers can also more systematically compare the performance of different clustering

algorithms. For instance, k-means are known to have subpar performance when there is a high class imbalance (Liang et al. 2012). Scholars can try various hierarchical clustering algorithms.

We also note that if scholars further use the machine-identified clusters as variables in regression analysis, these choices may bring potential classification errors in machine algorithms, or the intercoder disagreement, in estimating the next-step regression models. There has been some progress in the literature that addresses how to properly account for classification errors in machine learning algorithms (Fong and Tyler 2021; Zhang 2021), or intercoder inconsistencies (Grimmer, King, Superti 2015; Song et al. 2020b), into next-step regression models. So far there has been no literature saying how to adjust for such biases when both ML classification errors and intercoder disagreement exist. And most of the literature are developed with supervised machine learning in mind. Extend these studies into unsupervised clustering is a future research direction.

We included the code to implement transfer learning using VGGNet and ImageNet dataset in Appendix A. Other codes and data to replicate the analyses in the paper will made publicly available. There are already many good tutorials on transfer learning techniques both inside and outside social science (Sarkar, Bali, and Ghosh 2018; Williams, Casas, and Wilkerson 2020). It is most common to use Python language to perform transfer learning with pretrained models, but researchers have started to make it possible to use R to perform transfer learning, too (Ramasubramanian and Singh 2019). We understand that image analysis is relatively new to social scientists, and so we plan to develop software packages (potentially in R) and operation manuals that make it easier for social scientists to use transfer learning models.

Appendix A: Python Code to Perform Feature Extraction from Pretrained Models

Below is the code to extract intermediate representations of images using the pretrained model using a standard VGGNet trained on ImageNet dataset. The extracted vectors will be saved to disk, and scholars can freely choose any implementation of clustering algorithms (e.g., *kmeans* function in R or *Sklearn* package in Python).

```
1  import numpy as np
2  import joblib
3  from keras.preprocessing import image
4  from keras.applications import vgg16
```

```
from keras.models import Model
from keras.applications.vgg16 import VGG16,
preprocess_input

# read in Keras's pretrained VGG model with ImageNet dataset
pretrained_nn = VGG16(weights= "imagenet",
include_top=True)
print (pretrained_nn.summary())

# load the pretrained model to a deep learning model
# fc1 is the first fully connected layer, or the layer after convolutional
layers
feature_model=Model(input=pretrained_nn.input,
output=pretrained_nn.get_layer('fc1').output)

# resize images to 224 * 224
img_size = (224, 224)

# one row per file
imgnamefile = open('img filename.txt', 'r').readlines()

for imgpath in imgnamefile:
      # Load the image from disk
img = image.load_img(imgpath, target_size=img_size)
      # Convert the image to a numpy array
      # image_arrry is of dimension 224 * 224 * 3
      image_array = image.img_to_array(img)

      # normally vgg16 takes a list of images as input;
      # here we have one image
      # so transform the image matrix to 1 * 224 * 224 * 3
image_expand = np.expand_dims(image_array, 0)

      # normalize image data to 0-to-1 range
x_matrix = vgg16.preprocess_input(image_expand)

    # obtain extracted vector
    # it's of dimension 4096 * 1
x_low_vector = feature_model.predict(x_matrix)
```

```
42          # save features to imgsavepath for future use
43  imgsavepath = imgpath + "_extracted_feature.dat"
44          joblib.dump(x_low_vector ,imgsavepath)
```

See Figure B1.

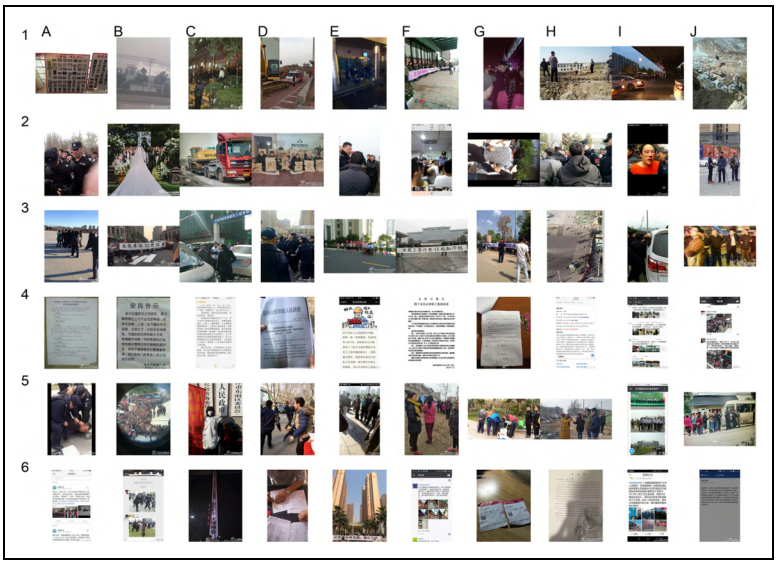

Figure B1. Transfer learning using self-supervised model, using agglomerative hierarchical clustering. Results are similar to Figure 7 where other things are equal but clustering algorithm is k-means.

Appendix B: Comparing Different Clustering Algorithms on Clustering Performance

See Figure B2.

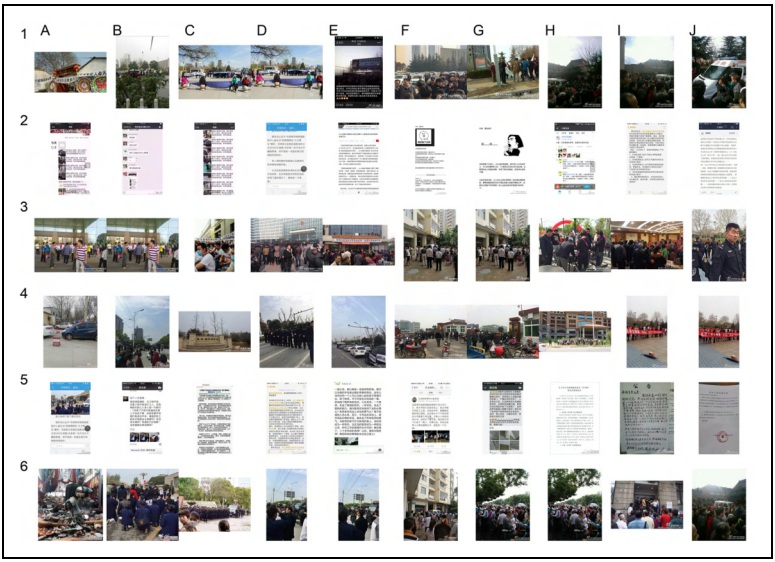

Figure B2. Transfer learning using self-supervised model, using Gaussian mixture model. Results are similar to Figure 7 where other things are equal but clustering algorithm is k-means.

Appendix C: Statistical Tests for Clustering Performances

Table A1 performed two-proportion z-test for comparing model performances. We used twoproportion z-test because within-cluster consistency can be thought as the proportion of images in a cluster that belongs to the main theme of that cluster. Because the proportions are bounded between 0 and 1, it is normally distributed and thus t-test should not be used here. We performed two-sided two-proportion z-test for the mean within-cluster consistency for each pretrained model and the choice of K.

Table A1 shows that in general, the differences between self-supervised and supervised models have a statistically significant difference. In particular, our preferred solution (self-supervised model with $K = 6$) and other clustering solutions have a statistically significant difference. The only exception is that for $K = 10$ and self-supervised model. However, as we discussed in the main text, if we choose $K = 10$, there is one cluster that exhibits poor within-cluster

Table A1. Two Proportion z-Test on Different Dataset, for China Protest Dataset.

	Model		Within-cluster consistency		
Row	1	2	1	2	p-value
1	self-supervised, K = 6	self-supervised, K = 8	0.92	0.83	0.011 *
2	self-supervised, K = 6	self-supervised, K = 10	0.92	0.88	0.219
3	self-supervised, K = 6	supervised, K = 6	0.92	0.8	0.002 **
4	self-supervised, K = 6	supervised, K = 8	0.92	0.81	0.002 **
5	self-supervised, K = 6	supervised, K = 10	0.92	0.79	0 ***
6	self-supervised, K = 8	self-supervised, K = 10	0.83	0.88	0.126
7	self-supervised, K = 8	supervised, K = 6	0.83	0.8	0.51
8	self-supervised, K = 8	supervised, K = 8	0.83	0.81	0.652
9	self-supervised, K = 8	supervised, K = 10	0.83	0.79	0.288
10	self-supervised, K = 10	supervised, K = 6	0.88	0.8	0.025 *
11	self-supervised, K = 10	supervised, K = 8	0.88	0.81	0.033 *
12	self-supervised, K = 10	supervised, K = 10	0.88	0.79	0.004 **
13	supervised, K = 6	supervised, K = 8	0.8	0.81	0.895
14	supervised, K = 6	supervised, K = 10	0.8	0.79	0.884
15	supervised, K = 8	supervised, K = 10	0.81	0.79	0.64

Each row lists the two different models' algorithms, *K*, within-cluster consistency, and whether the differences between two within-cluster consistency measures are statistically significant. "0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1".

performance, preventing us from using that solution. These comparisons confirm our choice of using self-supervised model and $K = 6$ as the final clustering solution.

Appendix D: Choosing the Best Clustering Solution for the Climate Change Dataset

We proceeded by first ruling out clustering solutions whose within-cluster consistency did not reach a minimum threshold (0.5). This exercise discarded VGGFace ($K = 6$ and 8) and ImageNet ($k = 6$). For the remaining six clustering solutions, VGGFace ($K = 10$) did not form a clear cluster that had a consistent theme; most of the clusters barely reached 0.5, and it had the lowest average within-cluster consistency. This observation affirms that our previous argument: when the pretrained dataset is relatively dissimilar from the images

to be analyzed(as a dataset of faces, such as VGGFace, is dissimilar to climate change images), it is unlikely that the extracted features from the pretrained model can be clustered into coherent themes. The pretrained model based on Places365 dataset ($K = 8$) yielded the highest average within-cluster consistency. Its minimum within-cluster consistency is also the largest among all clustering solutions. Therefore, we chose the pretrained model based on the Places365 dataset ($K = 8$) as the final clustering solution.

Appendix E: Intercoder Reliability

We had two independent coders. After the main theme of a cluster is determined, each coder will code the images independently (step 4) as whether the image belongs to the main theme of that cluster or not. For instance, if the main theme is "male with glasses", the two labels given to each image in that cluster will be "male with glasses: yes" or "male with glasses: no". Then we calculated the intercoder reliability across the two measures.

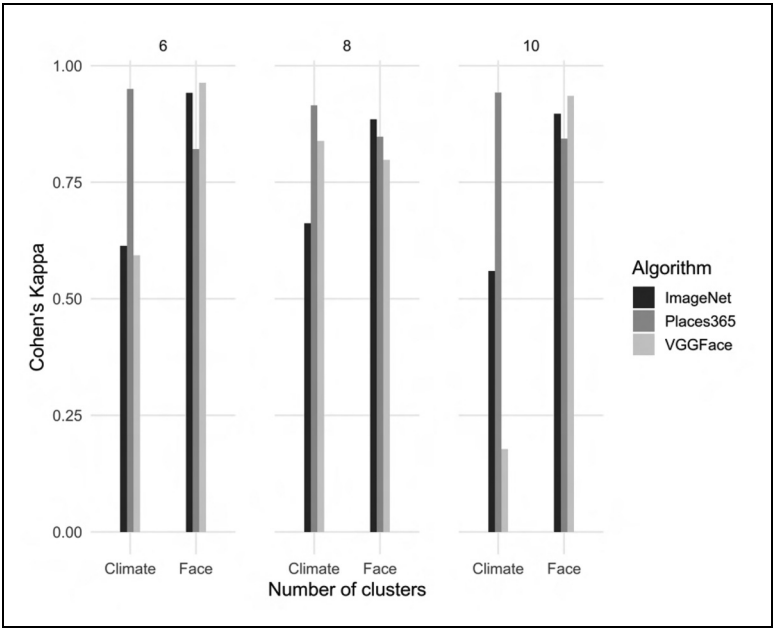

Figure B3. Cohen's Kappa measure of intercoder reliability.

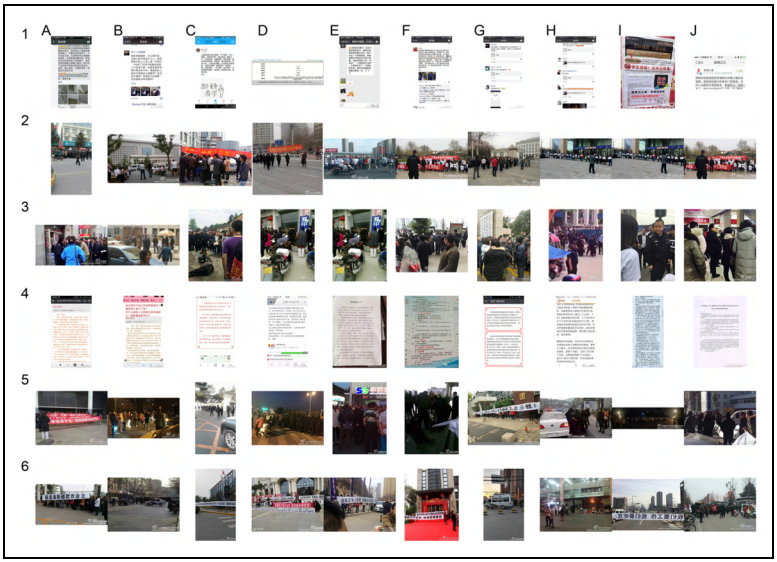

Figure B4. Transfer learning using self-supervised pretrained models (AlexNet as the architecture), on protest dataset.

Figure B3 plots the Cohen's Kappa measure of intercoder reliability. In general, VGG model trained on the VGGFace Dataset yields higher intercoder reliability measures for the IRA Face dataset. On the other hand, VGG model trained on the Places365 dataset yields higher intercoder reliability measures for the Climate Change datsaet. This makes sense, given that a model trained on a dataset which is similar to the ultimate target dataset should yield better performance, as we already discussed in the main text.

Appendix F: Comparing Different Architectures on Clustering Performance

We used the same dataset as in Study 1, CASM-China. We applied a pretrained self-supervised learning model. The pretrained self-supervised learning model used with the architecture AlexNet, a less complicated predecessor of VGG (Krizhevsky, Sutskever, and Hinton 2012). The results are shown in Figure B4. The only difference between Figure B4 and Figure 7 in the main text is that the former relies on AlexNet architecture and the latter relies on VGG architecture. Because VGG is more complex in its architecture, in

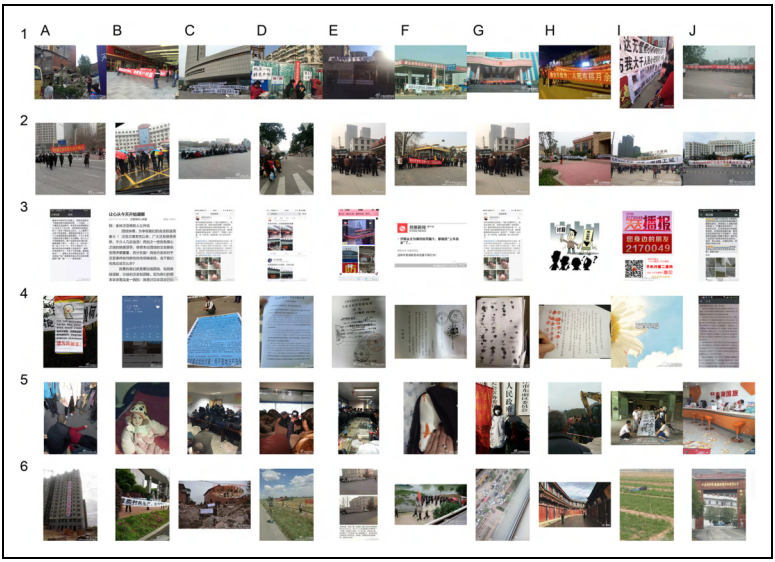

Figure B5. Transfer learning using supervised pretrained models (ResNet as the architecture), on Protest Dataset.

theory, we should expect it to perform better at clustering images. Indeed, Figure B4 shows that the AlexNet architecture, the simpler architecture, confuses some images. For instance, Cluster 4 conflates screenshots of text with pictures of petition letters. Cluster 6 also conflates crowd gatherings in front of buildings and on roads. The simpler pretrained model indeed produces an inferior clustering solution than a more complex pretrained model.

We also tried transfer learning using a supervised pretrained model with a more complex architecture, ResNet (with 152 layers). We did not use self-supervised learning model because the original authors of DeepCluster did not train their model with ResNet (Caron et al. 2018) but used AlexNet and VGG, which are simpler in model architecture than ResNet. The results are shown in Figure B5. Figure B5 and Figure 4 used the same dataset and k-means clustering algorithms but differ in the architecture of the pretrained models. ResNet model differs from VGG model in that it recognizes that black borders are not a meaningful feature. The clusters it finds are meaningful.

However, there are still several problems. For instance, Cluster 6 contains both gathering and photos of buildings/lands, which is not extremely clear. Cluster 5 finds the gathering of people, but it also has some pictures with only one man (F) or a child (B).

Overall, we found that using the same pretrained dataset and model, VGG performs well in finding meaningful representations, but AlexNet's performance is not good enough. ResNet, on the other hand, outperforms VGG model, other things being equal. These findings confirm the intuition that if a pretrained model used a more complex model architecture, its ability to extract meaningful features is better.

Acknowledgments

The authors thank Charles Crabtree, Erik P. Bucy, Brandon Stewart, Michelle Torres, Cristian Vaccari, and participants in the 2020 American Political Science Association Conference, the 2021 Asian PolMeth Conference, and the 2020 International Communication Association Preconference on Visual Politics for their comments on earlier versions of this paper. The authors also thank Qianru Huang, Yongxin Ke, and Wen Wen for their assistance in data analysis. This paper received a Top Paper Award from the Computational Methods Division of International Communication Association.

Authors' Note

The dataset in Study 1 and the climate change dataset in Study 2 can be accessed at Harvard Dataverse (https://doi.org/10.7910/DVN/VSOH5H). We note that the images in Study 1 are about protests in China and may contain identifying information, so we make Study 1's images restricted and available upon request. The IRA dataset in Study 2 is publicly accessible at Twitter (https: //transparency.twitter.com/en/reports/information-operations.html). The scripts for analysis can be found at https://github.com/yilangpeng/image-clustering

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Han Zhang https://orcid.org/0000-0003-2912-8780
Yilang Peng https://orcid.org/0000-0001-7711-9518

Notes

1. Zhang and Pan (2019) designed a two-stage deep learning algorithm to detect offline protests based on both text and images in social media posts from Weibo, the Chinese alternative of Twitter.
2. For instance, classical OLS regression cannot be identified if the number of variables exceeds the number of observations.
3. The SIFT algorithm will map different images into vectors of different lengths, and thus cannot be directly used for standard clustering algorithms such as k-means.
4. Technically, the bag-of-visual-words model first concatenates all features produced by the SIFT algorithm, performs clustering on these features, and finds meaningful clusters of SIFT features. These meaningful features are used as the "vocabulary." It then calculates each SIFT vector's occurrence frequency with respect to the vocabulary. The histogram of each SIFT vector over the vocabulary is then used as the extracted representation of an image. This step is also referred to as "codebook generation" in the literature. Further details can be found in standard computer vision textbooks such as (Szeliski 2010).
5. The ImageNet project is a large image database that has more than 14 million pictures of 21,841 categories (http://image-net.org/about-stats). The ImageNet project runs the ImageNet Large Scale Visual Recognition Challenge, which uses a subset of the full ImageNet project.
6. Some authors also call this unsupervised learning of image features (Gidaris, Singh, and Komodakis 2018). We choose to call this branch of research self-supervised learning to distinguish it from our main purpose of the article—unsupervised image clustering.
7. Yang, Parikh, and Batra (2016) uses a combination of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) as the architecture and hierarchical clustering algorithms.
8. Unless the category used by a supervised pretrained model is very similar to the expected category of images.
9. Other work in topic modeling often used 5 to 10 documents if the purpose is to inspect whether the top documents in a topic indeed exhibit coherent topics (Ying, Montgomery, and Stewart 2021). We use 20 because we need to calculate percentage of images that belongs to the main theme of the cluster, and it will be too small to choose 10. In the results, we show that this choice of 20 indeed let us to distinguish between different clustering solutions (Section "Quantitative Validation"). Alternatively, if one finds that 20 images are not enough, they should certainly increase the number.
10. It is also popular among practitioners to get a large number of clusters and then refine the results by combining several smaller clusters into a larger category (Hu et al. 2014; Manikonda and De Choudhury 2017; Peng 2021; Roberts et al. 2014). To draw an analogy, in topic modeling, scholars can first get a solution

with a large number of topics and then classify these topics into a smaller set of theoretically meaningful categories (Song, Eberl, and Eisele 2020a).

11. We selected these protests because these predicted events have been validated by research assistants manually.
12. Technically, we trained 50 epochs using Caron et al. (2018)'s model with VGG architecture.
13. Both models were trained based on VGG architecture and ImageNet dataset. The first transfer model relies on Keras' native transfer learning model (https://keras.io/api/applications/. The second transfer model was downloaded from the author's website (https://github.com/facebookresearch/deepcluster).
14. Technically, we trained 50 epochs.
15. The package can be accessed from https://instaloader.github.io/. These eleven accounts are @climate.change.communication, @climatechangetruth, @climatereality, @climatesavemovement, @climemechange, @cnnclimate, @everydayclimatechange, @ipcc @nasaclimatechange, @noaaclimate, @thisiszerohour. On average, they had 112477.5 followers (Median = 64904) at the time of data collection.
16. Technically, the fc1 layer in ImageNet, fc1 in Places365, and fc6 in VGGFace.
17. We tried to perform k-means on the original 4,096-dimension vectors. The performance is similar, but the speed is significantly slower if we use 4,096 dimensional vectors.
18. As a side note, when $K = 8$ and the pretrained datasetis VGGFace, the within-cluster consistency is the highest, which help us choose the K presented in the previous section.
19. Sometimes this choice will be constrained by the availability of pretrained models. For instance, when DeepCluster, the self-supervised learning model was first developed, the authors did not use ResNet as their architecture (Caron et al. 2018).

References

Abadi, Martin, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. "TensorFlow: LargeScale Machine Learning on Heterogeneous Systems." Software available from tensorflow.org.

Airoldi, Edoardo M., David M. Blei, Stephen E. Fienberg, and Eric P. Xing. 2009. "Mixed Membership Stochastic Blockmodels." Pp. 33–40 in *Advances in*

Neural Information Processing Systems, Vol. 21, edited by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Red Hook, NY: Curran Associates, Inc.

Azizpour, Hossein, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. 2015. "Factors of Transferability for a Generic Convnet Representation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38:1790–802.

Baele, Stephane J., Katharine A. Boyd, and Travis G. Coan. 2020. "Lethal Images: Analyzing Extremist Visual Propaganda from ISIS and Beyond." *Journal of Global Security Studies* 5:634–57.

Barbera, Pablo, Andreu Casas, Jonathan Nagler, Patrick J. Egan, Richard Bonneau, John T. Jost, and Joshua A. Tucker. 2019. "Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data." *American Political Science Review* 113:883–901.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.

Cai, Yongshun. 2010. *Collective Resistance in China: Why Popular Protests Succeed or Fail*. Redwood City, CA: Stanford University Press.

Caron, Mathilde, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. "Deep Clustering for Unsupervised Learning of Visual Features." Pp. 139–56 in *Computer Vision – ECCV 2018*, edited by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, Lecture Notes in Computer Science. Cham: Springer International Publishing.

Caron, Mathilde, Piotr Bojanowski, Julien Mairal, and Armand Joulin. 2019. "Unsupervised Pre-Training of Image Features on Non-Curated Data." Pp. 2959–68 in Proceedings of the IEEE/CVF International Conference on Computer Vision.

Caron, Mathilde, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments." *Advances in Neural Information Processing Systems* 33:9912–9924.

Casas, Andreu and Nora Webb Williams. 2019. "Images That Matter: Online Protests and the Mobilizing Role of Pictures." *Political Research Quarterly* 72:360–75.

Chen, Xi. 2009. "The Power of "Troublemaking": Protest Tactics and Their Efficacy in China." *Comparative Politics* 41:451–71.

Chen, Xi. 2012. *Social Protest and Contentious Authoritarianism in China*. Cambridge, United Kingdom: Cambridge University Press.

Chen, Chih-Jou Jay and Yongshun Cai. 2019. "Targets Matter: Managing Social Protests in China."

Chollet, François, et al. 2015. "Keras." https://keras.io.

Corrigall-Brown, Catherine and Rima Wilkes. 2012. "Picturing Protest: The Visual Framing of Collective Action by First Nations in Canada." *American Behavioral Scientist* 56:223–43.

Csurka, Gabriella, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cedric Bray. 2004. "Visual Categorization with Bags of Keypoints." Pp. 1–22 in Workshop on Statistical Learning in Computer Vision, ECCV.

Cui, Yin, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. 2018. "Large Scale Finegrained Categorization and Domain-specific Transfer Learning." Pp. 4109–18 in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Deng, Jia, Wei Dong, Richard Socher, Li-jia Li, Kai Li, and Li Fei-fei. 2009. "Imagenet: A Large-Scale Hierarchical Image Database." Pp. 248–255 in 2009 IEEE conference on computer vision and pattern recognition.

DiMaggio, Paul. 2015. "Adapting Computational Text Analysis to Social Science (and Vice Versa)." *Big Data & Society* 2.

DiMaggio, Paul, Manish Nag, and David Blei. 2013. "Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of Us Government Arts Funding." *Poetics* 41:570–606.

Dueck, Delbert and Brendan J. Frey. 2007. "Non-metric Affinity Propagation for Unsupervised Image Categorization." Pp. 1–8 in 2007 IEEE 11th International Conference on Computer Vision. IEEE.

Dwivedi, Kshitij and Gemma Roig. 2019. "Representation Similarity Analysis for Efficient Task Taxonomy & Transfer Learning." Pp. 12387–96 in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Fong, Christian and Matthew Tyler. 2021. "Machine Learning Predictions as Regression Covariates." *Political Analysis* 29:467–484.

Frey, Brendan J. and Delbert Dueck. 2007. "Clustering by Passing Messages Between Data Points." *Science* 315:972–6.

Fu, Diana. 2017. "Disguised Collective Action in China." *Comparative Political Studies* 50:499–527.

Gidaris, Spyros, Praveer Singh, and Nikos Komodakis. 2018. "Unsupervised Representation Learning by Predicting Image Rotations." in International Conference on Learning Representations.

Goebel, Christian and H. Steinhardt. 2019. "Better Coverage, Less Bias: Using Social Media to Measure Protest in Authoritarian Regimes."

Grimmer, Justin, Gary King, and Chiara Superti. 2015. "The Unreliability of Measures of Intercoder Reliability, and What to Do About It." https://polisci.ucsd.edu/_files/hand.pdf.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annual Review of Political Science* 24:395–419.

Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21:267–97.

Grother, Patrick, Mei Ngan, and Kayee Hanaoka. 2019. *Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects*. Gaithersburg, MD: National Institute of Standards and Technology.

Guerin, Joris, Olivier Gibaru, Stéphane Thiery, and Eric Nyiri. 2017. "CNN features are Also Great at Unsupervised Classification." https://arxiv.org/abs/1707.01700.

Ha, Yui, Kunwoo Park, Su Jung Kim, Jungseock Joo, and Meeyoung Cha. 2020. "Automatically Detecting Image–Text Mismatch on Instagram with Deep Learning." *Journal of Advertising* 50:52–62.

Hashemi, Mahdi. 2019. "Enlarging Smaller Images Before Inputting into Convolutional Neural Network: Zero-Padding vs. Interpolation." *Journal of Big Data* 6:98.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman, T. Hastie, J. Friedman, and R. Tibshirani. 2009. *The Elements of Statistical Learning*. Vol. 2. Berlin/ Heidelberg, Germany: Springer.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. "Deep Residual Learning for Image Recognition." Pp. 770–8 in 2016 IEEE Conference on Computer Vision and Pattern Recognition.

Hu, Yuheng, Lydia Manikonda, Subbarao Kambhampati, et al. 2014. "What we Instagram: A First Analysis of Instagram Photo Content and User Types." In Eighth International AAAI Conference on Weblogs and Social Media.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science* 353:790–4.

Joo, Jungseock and Zachary C. Steinert-Threlkeld. 2018. "Image as Data: Automated Visual Content Analysis for Political Science." https://arxiv.org/abs/1810.01544.

Kang, Zhiqi, Christina Indudhara, Kaushik Mahorker, Erik P. Bucy, and Jungseock Joo. 2020a. "Understanding Political Communication Styles in Televised Debates via Body Movements." Pp. 788–93 in European Conference on Computer Vision. Springer.

Kang, Zhiqi, Christina Indudhara, Kaushik Mahorker, Erik P. Bucy, and Jungseock Joo. 2020b. "Understanding Political Communication Styles in Televised Debates via Body Movements." Pp. 788–93 in Computer Vision – ECCV 2020 Workshops, edited by Adrien Bartoli and Andrea Fusiello, Lecture Notes in Computer Science. Cham: Springer International Publishing.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107:326–43.

Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage Publications.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." Pp. 1097–105 in *Advances in Neural Information Processing Systems*, Vol. 25, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Red Hook, NY: Curran Associates, Inc.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521:436–44.

Liang, Jiye, Liang Bai, Chuangyin Dang, and Fuyuan Cao. 2012. "The *K*-Means-Type Algorithms Versus Imbalanced Data Distributions." *IEEE Transactions on Fuzzy Systems* 20:728–45.

Lowe, D. G. 1999. "Object Recognition from Local Scale-Invariant Features." Pp. 1150–7 in Proceedings of the Seventh IEEE International Conference on Computer Vision.

Lowe, David G. 2004. "Distinctive Image Features from Scale-Invariant Keypoints." *International Journal of Computer Vision* 60:91–110.

Manikonda, Lydia and Munmun De Choudhury. 2017. "Modeling and Understanding Visual Attributes of Mental Health Disclosures in Social Media." Pp. 170–81 in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.

Murashka, Volha, Jiaying Liu, and Yilang Peng. 2021. "Fitspiration on Instagram: Identifying Topic Clusters in User Comments to Posts with Objectification Features." *Health Communication*36:1537–1548.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: The MIT Press.

Narisetty, Naveen Naidu. 2020. "Chapter 4 – Bayesian Model Selection for High-Dimensional Data." Pp. 207–48 in *Handbook of Statistics*, edited by Arni S. R. Srinivasa Rao and C. R. Rao, vol. 43 of Principles and Methods for Data Science. Amsterdam, Netherlands: Elsevier.

Oleinik, Anton. 2015. "On Content Analysis of Images of Mass Protests: A Case of Data Triangulation." *Quality & Quantity* 49:2203–20.

O'Neill, Saffron J. and Nicholas Smith. 2014. "Climate Change and Visual Imagery." *Wiley Interdisciplinary Reviews: Climate Change* 5:73–87.

Paivio, Allan. 1990. *Mental Representations: A Dual Coding Approach*. Oxford, United Kingdom: Oxford University Press.

Pan, Weike and Qiang Yang. 2013. "Transfer Learning in Heterogeneous Collaborative Filtering Domains." *Artificial Intelligence* 197:39–55.

Parker, Jim R. 2010. *Algorithms for Image Processing and Computer Vision*. Hoboken, NJ: John Wiley & Sons.

Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. 2015. "Deep Face Recognition." Pp. 41.1–41.12 in Procedings of the British Machine Vision Conference 2015. Swansea: British Machine Vision Association.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." Pp. 8024–35 in *Advances in Neural Information Processing Systems*, Vol. 32, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. D. Alche-Buc, E. Fox, and R. Garnett. Red Hook, NY: Curran Associates, Inc.

Peng, Yilang. 2018. "Same Candidates, Different Faces: Uncovering Media Bias in Visual Portrayals of Presidential Candidates with Computer Vision." *Journal of Communication* 68:920–41.

Peng, Yilang. 2021. "What Makes Politicians' Instagram Posts Popular? Analyzing Social Media Strategies of Candidates and Office Holders with Computer Vision." *The International Journal of Press/Politics* 26:143–166.

Peng, Yilang. Forthcoming. "Athec: A Python Library for Computational Aesthetic Analysis of Visual Media in Social Science Research." *Computational Communication Research*.

Pew Research Center. 2019. "Social media Fact Sheet." https://www.pewresearch.org/internet/fact-sheet/social-media/

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54:209–28.

Ramasubramanian, Karthik and Abhishek Singh. 2019. "Deep Learning Using Keras and TensorFlow." Pp. 667–88 in *Machine Learning Using R: With Time Series and Industry-Based Use Cases in R*, edited by Karthik Ramasubramanian and Abhishek Singh. Berkeley, CA: Apress.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural topic models for open-ended survey responses." American Journal of Political Science 58:1064–1082S.

Rohlinger, Deana A. and Jesse Klein. 2012. "Visual Landscapes and the Abortion Issue." *American Behavioral Scientist* 56:172–88.

Sarkar, Dipanjan, Raghav Bali, and Tamoghna Ghosh. 2018. *Hands-On Transfer Learning with Python: Implement Advanced Deep Learning and Neural Network Models Using TensorFlow and Keras*. Birmingham, United Kingdom: Packt Publishing.

Simonyan, Karen and Andrew Zisserman. 2015. "Very Deep Convolutional Networks for LargeScale Image Recognition." In Proceedings of the Third International Conference on Learning Representations.

Sivic and Zisserman. 2003. "Video Google: A Text Retrieval Approach to Object Matching in Videos." Pp. 1470–7 in Proceedings Ninth IEEE International Conference on Computer Vision.

Sobolev, Anton, M. Keith Chen, Jungseock Joo, and Zachary C. Steinert-Threlkeld. 2020. "News and Geolocated Social Media Accurately Measure Protest Size Variation." *American Political Science Review*114:1343–1351.

Song, Hyunjin, Jakob-Moritz Eberl, and Olga Eisele. 2020a. "Less Fragmented Than We Thought? Toward Clarification of a Subdisciplinary Linkage in Communication Science, 2010–2019." *Journal of Communication* 70:310–34.

Song, Hyunjin, Petro Tolochko, Jakob-Moritz Eberl, Olga Eisele, Esther Greussing, Tobias Heidenreich, Fabienne Lind, Sebastian Galyga, and Hajo G. Boomgaarden. 2020b. "In Validations We Trust? The Impact of Imperfect Human Annotations as a Gold Standard on the Quality of Validation of Automated Content Analysis." *Political Communication* 37:550–72.

Steinert-Threlkeld, Zachary, Alexander Chan, and Jungseock Joo. 2021. "How State and Protester Violence Affect Protest Dynamics." *The Journal of Politics* 84.

Szeliski, Richard. 2010. *Computer Vision: Algorithms and Applications*. Berlin/ Heidelberg, Germany: Springer Science & Business Media.

TensorFlow. 2021. "Transfer Learning and Fine-Tuning | TensorFlow Core." https:// www.tensorflow.org/tutorials/images/transfer_learning

Torres, Michelle and Francisco Cantu. 2022. "Learning to See: Convolutional Neural Networkś for the Analysis of Social Science Data." *Political Analysis*30:113–131.

Valensise, Carlo Michele, Alessandra Serra, Alessandro Galeazzi, Gabriele Etta, Matteo Cinelli, and Walter Quattrociocchi. 2021. "Entropy and Complexity Unveil the Landscape of Memes Evolution." *Scientific Reports*11:20022.

Wilkerson, John and Andreu Casas. 2017. "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges." *Annual Review of Political Science* 20:529–44.

Williams, Nora Webb, Andreu Casas, and John D. Wilkerson. 2020. *Images as Data for Social Science Research: An Introduction to Convolutional Neural Nets for Image Classification*. Cambridge, United Kingdom: Cambridge University Press.

Yang, Jianwei, Devi Parikh, and Dhruv Batra. 2016. "Joint Unsupervised Learning of Deep Representations and Image Clusters." Pp. 5147–56 in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Ying, Luwei, Jacob M. Montgomery, and Brandon M. Stewart. 2021. "Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures.." *Political Analysis* 5:2020.

Zhang, Han. 2021. "How Using Machine Learning Classification as a Variable in Regression Leads to Attenuation Bias and What to Do About It." https://osf.io/preprints/socarxiv/453jk/

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

Zhou, Bolei, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. "Places: A 10 Million Image Database for Scene Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40:1452–64.

Zou, James and Londa Schiebinger. 2018. "AI can be Sexist and Racist—It's Time to Make it Fair." *Nature* 559:324–326.

Author Biographies

Han Zhang is an Assistant Professor in the Division of Social Science at The Hong Kong University of Science and Technology. He obtained his PhD in sociology from Princeton University, and his BS in Computer Science and BA from Peking University. His research interests include computational social science, social movements, and social networks. His past research has been published in *Sociological Methodology* and *Chinese Sociological Review*. His research won the Mayer N. Zald Distinguished Contribution to Scholarship Student Paper Award from the Section on Collective Behavior and Social Movements of American Sociological Association, and Top Paper Award from the Computational Methods Division of International Communication Association.

Yilang Peng (PhD, Annenberg School for Communication, University of Pennsylvania) is an Assistant Professor in the Department of Financial Planning, Housing and Consumer Economics at the University of Georgia. His research areas include computational social science, visual communication, computer vision, and science communication. His research has appeared in venues such as the *Journal of Communication, New Media & Society, Communication Research*, and the *Proceedings of ACM Conference on Human Factors in Computing Systems*. His recent research is funded by the National Science Foundation and examines how visual attributes influence credibility perceptions of misinformation.