International Journal of Computer Vision (2021) 129:2978–3003
https://doi.org/10.1007/s11263-021-01506-3

Predicting Visual Political Bias Using Webly Supervised Data and an Auxiliary Task

Christopher Thomas[1] · **Adriana Kovashka**[1]

Received: 27 January 2020 / Accepted: 24 June 2021 / Published online: 27 August 2021

Abstract
The news media shape public opinion, and often, the visual bias they contain is evident for careful human observers. This bias can be inferred from how different media sources portray different subjects or topics. In this paper, we model visual political bias in contemporary media sources at scale, using webly supervised data. We collect a dataset of over one million unique images and associated news articles from left- and right-leaning news sources, and develop a method to predict the image's political leaning. This problem is particularly challenging because of the enormous intra-class visual and semantic diversity of our data. We propose two stages of training to tackle this problem. In the first stage, the model is forced to learn relevant visual concepts that, when joined with document embeddings computed from articles paired with the images, enable the model to predict bias. In the second stage, we remove the requirement of the text domain and train a visual classifier from the features of the former model. We show this two-stage approach that relies on an auxiliary task leveraging text, facilitates learning and outperforms several strong baselines. We present extensive quantitative and qualitative results analyzing our dataset. Our results reveal disparities in how different sides of the political spectrum portray individuals, groups, and topics.

Keywords Weak supervision · Noisy data · Unsupervised discovery · Curriculum learning · Privileged information · Image-text alignment · Visual rhetoric

1 Introduction

One of the goals of the media is to inform, but in practice, the media also shapes opinions (Happer and Philo 2013; Philo 2008; Angermeyer and Schulze 2001; Gilens 1996; Schill 2012; Muñoz and Towner 2017). The same issue can be presented from multiple perspectives, both in terms of the text written, and the visual content chosen to illustrate the article. For example, when speaking of immigration, left-leaning sources might showcase the struggles of well-meaning immigrants, while right-leaning sources might portray the misdeeds of law-breaking immigrants. The topics portrayed are also a strong cue for the left or right bias of the source media—for example, tradition is primarily seen as a value on the right, and diversity on the left (Edsall 2012).

In this paper, we present a method for recognizing the political bias of an image, which we define as whether the image came from a left- or right-leaning media source. This requires understanding: (1) what visual concepts to look for in images, and (2) how these visual concepts are portrayed across the spectrum. Note that this is a very challenging task because many of the concepts that we aim to learn show serious visual variability within the left and right. For example, the concept of "immigration" can be illustrated with a photo of a border wall, children crying behind bars while detained, immigration agents, protests and demonstrations about the issue, politicians giving speeches, etc. Human viewers account for such within-class variance by generalizing what they see into broader semantic concepts or themes using prior knowledge, deduction, and reasoning.

Communicated by Judy Hoffman.

This material is based upon work supported by the National Science Foundation under Grant Numbers 1566270 and 1718262. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Funding was also provided by a Nvidia hardware grant.

✉ Christopher Thomas
chris@cs.pitt.edu

Adriana Kovashka
kovashka@cs.pitt.edu

[1] Department of Computer Science, University of Pittsburgh, Pittsburgh, PA, USA

Fig. 1 **Top:** Can you guess whether each image appears in a far-left or far-right media source? *Use your bias:* What are the left and right stereotypically associated with? See the end of this caption for answers. **Bottom:** In Stage 1, our method relies on text paired with images to guide the model towards learning relevant *visual* semantics. In Stage 2, we freeze the feature extraction part (shown with transparent fill) and learn a purely visual classifier using features extracted from our Stage 1 model. At test time, our method makes purely visual classifications, without requiring any text for inference. **Answers:** Top: Left, Right, Left; Bottom: Right, Left, Left

On the other hand, modern convolutional architectures learn by discovering recurring textures or edges representing objects in the images through backpropagation. However, the same objects might appear and be discussed *across* the political spectrum, meaning that the simple presence or absence of objects is not a good indicator of the politics of an image. Thus, model training may fall into poor local minima due to the lack of a recurring discriminative signal. Further, it is not merely the presence or absence of objects that matters, but rather *how* they are portrayed, often in subtle ways.

In order to capture the visual concepts necessary to predict the politics of an image, we propose a method which uses an auxiliary channel at training time, namely the article text that the image is paired with. Our method contains two stages, as shown in Fig. 1. In the first one, we learn a document embedding on the articles, then train a model to predict the bias of the image, given the image and the paired document embedding. To be successful on this task, the model learns to recognize visual cues which complement the text embedding and suggest the politics of the image-text pair. The text serves as privileged information to guide learning. At test time, we want to recognize bias from images alone, without any article text. Thus, in the second training stage, we use the first stage model as a feature extractor and train a linear bias classifier on top.

Since recognizing the right semantic and visual concepts amidst intra-class variance requires large amounts of data, we train our approach on webly supervised data: the only labels are in the form of the political leaning of the source that the image came from. However, for testing purposes, we collect human annotations of bias (political leaning) and test on images where annotators agreed on the label. We experimentally show that our method outperforms numerous baselines on both a large held-out webly supervised test set, and the set of human-annotated images.

We present many qualitative results, studying different types of bias inherent within our dataset, including both visual and text bias. Our results show different political groups present different subjects (incl. politicians, political groups, individuals, etc.) in significantly disparate ways. We also present generative results in which we explicitly model, and then generate, faces exhibiting the disparities our method captures.

Ethical ramifications We believe that recognizing the political bias of a photograph is an important step towards building computer vision systems that are aware of matters of social importance. Such awareness is necessary if we hope to use computer vision systems to automatically tag or describe images in a socially sensible manner (e.g. for the visually impaired) or to summarize large collections of potentially biased visual content. Social media companies or search engines may deploy such techniques to automatically identify the political bent of images or news articles being spread or linked to. This can be used to create a more balanced presentation of news. Progress in understanding political bias has already been made in this space in domains other than images. For example, Facebook automatically determines users' political leanings from site activity and pages liked (Merrill 2016). Other works have studied predicting political affiliation from text (Conover et al. 2011; Wong et al. 2016; Volkova et al. 2014) or even MRI scans (Schreiber et al. 2013). However, *visual* bias understanding has been greatly underexplored. While some work examines *visual persuasion* (Joo et al. 2014; Hussain et al. 2017) or how political figures are portrayed in the media (Peng 2018; Messing et al. 2016), none analyze predicting the political leaning of general images as we do.

The goal of our work is not to enable or further discrimination or reinforce stereotypes about individuals or groups. Rather, our work seeks to use machine learning techniques to *reveal* disparities in visual media which already exist. By raising awareness, we hope individual consumers of media are better able to approach material they are presented with (with a more skeptical eye) and question whether the portrayal of a subject they are seeing is politically skewed. Our work can also be used to combat, rather than reinforce, bias. One of many possibilities is a "balanced" image search engine, where our method is used to predict the political bias for each image returned. Studies (Noble 2018) show that search engine algorithms may perpetuate bias. The bias

score accompanying each image could be directly presented to the user. Another possible option would be to explicitly present users with images from both sides of the political spectrum, allowing the user to get a broader view of the subject. By returning images from across the political spectrum and/or explicitly revealing the inherent bias of images, we can help users be more informed consumers of visual media.

To summarize, our contributions are as follows:

- We collect and make available[1] a very large dataset of biased images with paired text, and a large amount of diverse annotations regarding political bias.
- We propose a novel weakly supervised method for predicting the political leaning of an image by using noisy auxiliary textual data at training time.
- We perform detailed experimental analysis of our method on both webly-supervised and human-annotated data, and demonstrate the factors humans use to predict bias in images.

2 Related Work

Our work relates to a number of subfields of machine learning, including weakly supervised learning, learning with privileged information, and curriculum learning. We briefly describe some relevant work below.

Weakly supervised learning Recently, weakly supervised approaches have been proposed for classic topics such as object detection (Oquab et al. 2015; Cinbis et al. 2016; Zhou et al. 2016; Wei et al. 2018; Ye et al. 2019), action localization (Wang et al. 2017a; Richard et al. 2017), etc. Researchers have also developed techniques for learning from potentially noisy web data, e.g. (Chen and Gupta 2015). Also related is work in unsupervised discovery of patterns and topic modeling. For example, Singh et al. (2012), Zhou et al. (2010) use an iterative clustering-detection pipeline to discover patterns that occur frequently but are discriminative. Li et al. (2018, 2017) and Sicre et al. (2017) leverage deep networks to mine discriminative patterns. Jae Lee et al. (2013) and Doersch et al. (2012) discover patterns informative for the architectural style of a city or the evolving design of cars over the decades. Both of these rely on finding clusters of image patches that are compact in terms of the top-level weak label (e.g. "Paris" or "1950s car"), i.e. clusters that primarily contain samples from a given label, and ignore clusters with near-uniform label distribution.

Our work is in the weakly supervised discovery setting, in the sense that other than noisy left/right labels, our method does not receive information about what makes an image left- or right-leaning. In contrast to these works, our problem exhibits much larger within-class variance (with left and right being the classes of interest). Unlike objects and styles, the differences between left and right live in semantic space as much as they do in visual space, thus these methods do not guarantee success. Nevertheless, we borrow intuitions from these methods and help our methods by focusing them on the higher-level semantics of the problem.

Curriculum learning Also relevant are self-paced and curriculum learning approaches (Jiang et al. 2015; Pentina et al. 2015; Zamir et al. 2017; Zhang et al. 2017; Jiang et al. 2018). These attempt to simplify learning by finding "easy" examples to learn with first. We too employ a type of curriculum learning. We first train a multi-modal classifier to predict bias, using the assumption that the relation between text and bias is more direct. We then leverage this model as a feature extractor by adding an image-only politics classifier on top. Thus, our method focuses the model on relevant visual concepts using text.

We compare against several methods which use a curriculum-based approach in our experiments: Joo et al. (2014) and Gomez et al. (2017) both learn relevant semantic concepts on a separate, auxiliary training task, which aid the classifier in performing inference on the target task. Because prior work (Orr 1997; He et al. 2019) has shown that using a larger-batch size improves classification performance on noisy data by smoothing the gradient, we thus compare against a baseline curriculum-learning approach designed to alleviate the problem of noisy minibatches. We freeze the lower-layers of the model after training and then perform a second stage of training of just the classifier using all features in the train set for optimization, which we show slightly improves performance.

Privileged information Our method exploits a similar intuition as privileged information methods (Vapnik and Izmailov 2015; Sharmanska et al. 2013; Hoffman et al. 2016; Motiian et al. 2016; Elliott and Kádár 2017; Gomez et al. 2017; Borghi et al. 2018; Lambert et al. 2018) that use an extra feature input at training time. These approaches use tied weights (Borghi et al. 2018) or multitask training (Elliott and Kádár 2017), or compute summary statistics (Sharmanska et al. 2013; Lambert et al. 2018), to guide learning. The closest such method to ours is Gomez et al. (2017) which uses an approach trained to predict text embeddings from images. The model's predicted features are then applied on visual-only data for image classification by training a linear SVM. However, directly predicting text embeddings from images is much more challenging on our data because of the many-to-many relationship of images with topics (e.g. image of the White House can be paired with text about Trump's children, border control, LGBT rights, etc.). We compare against Gomez et al. (2017)'s approach in Sect. 5.2.

[1] Our dataset, code, and additional materials are available here: http://www.cs.pitt.edu/~chris/politics.

Connecting images and text Predicting text from images has received sustained attention (Vinyals et al. 2015; Donahue et al. 2015; Johnson et al. 2016; Venugopalan et al. 2017; Pedersoli et al. 2017; Chen et al. 2017; Dai et al. 2017; Anderson et al. 2018; Eisenschtat and Wolf 2017; Ye et al. 2019). Common approaches for connecting image and text include projecting images and text to a common feature space (Faghri et al. 2018; Kiros et al. 2015; Eisenschtat and Wolf 2017; Ye and Kovashka 2018; Thomas and Kovashka 2020; Alayrac et al. 2020). Attention and co-attention, where a method discovers which parts of a sentence refer to which parts of an image in an unsupervised way, have also been shown to help in vision-language tasks (Lu et al. 2016; Xiong et al. 2017). Recently, researchers have leveraged transformer architectures to devise joint image-text embeddings that perform well on a variety of visual reasoning tasks (Lu et al. 2019; Tan and Bansal 2019; Chen et al. 2020).

To learn the political bias of images, we use an auxiliary pre-training task where images and corresponding text cooperate. However, our domain is unique from the above settings in that articles that are paired with our images are orders of magnitude longer. We include a number of results which analyze the image-text connection in this work, including finding the most visually consistent words, predicting individual words for images, and discovering concepts and training visual prediction models using webly supervised data for those concepts.

Visual rhetoric Our work also belongs to a recent trend of developing algorithms to analyze visual media and the strategies that a media creator uses to convey a message. Joo et al. (2014, 2015), Yoon et al. (2020) analyze the skills and characteristics that a politician is implied to have through a photo, e.g. "competent"; we adapt their method as a baseline in our setting. Peng (2018) study differences in facial portrayals between presidential candidates, and Wang et al. (2016, 2017b) examine visual differences between supporters of the left or right. We learn to *generate* faces from the left and right. Further, we examine differences in general images rather than just faces. Hussain et al. (2017) and Ye et al. (2019) predict the persuasive messages of advertisements, but persuasion in political images is more subtle: there is usually no slogan telling the viewer what to do or believe. These works are based on careful and expensive human annotations, while we aim to discover facets of bias in a weakly supervised way. Also related is work showing how to infer personal political beliefs from images of the subject's face (Kosinski 2021), and how politicians are portrayed in racially biased ways (Messing et al. 2016), however these only exploit facial features. Most related to our work is Xi et al. (2020) but this work comes from sources with more clear agenda (social media accounts of well-known politicians), and only uses visual features, while we leverage metadata (text) in a privileged information learning setting at training time.

Bias prediction in language Prior work in NLP has discovered indicators of biased language and political framing (i.e. presenting an event or person in a positive or negative light) (Recasens et al. 2013; Baumer et al. 2015; Card et al. 2015; Liu et al. 2019; Akyürek et al. 2020). For example, Recasens et al. (2013) and Baumer et al. (2015) use carefully designed dictionary, lexical, grammatical and content features to detect biased language, using supervision over short phrases. We leverage Recasens et al. (2013)'s technique to discover biased word usage in our dataset. However, it is not clear what "lexicon" of biased content to use for images. Others in NLP have studied predicting political affiliation from text (Pennacchiotti and Popescu 2011; Cohen and Ruths 2013; Colleoni et al. 2014; Conover et al. 2011; Wong et al. 2016; Volkova et al. 2014), mainly in the context of social media, and conducted misinformation analysis (Baly et al. 2018; Karimi and Tang 2019; Potthast et al. 2018; Jin et al. 2016; Khattar et al. 2019). However, work on *visual* framing is significantly more limited.

Fairness in machine learning We investigate the bias in how events, topics, and people are portrayed in the media. This type of bias is directly related to bias in human perceptions that people of a particular group (demographic, political, etc.) have certain qualities or beliefs. This bias over human qualities is evident in data that can be used to train machine learning algorithms, and has thus been tackled in a few prior works (Burns et al. 2018; Zhao et al. 2017; Ryu et al. 2017; Bechavod and Ligett 2017; Bolukbasi et al. 2016). For example, Burns et al. (2018) ensure that the same classifier is equally likely to fire on images of men and women when the relevant property (e.g. "snowboarding") is present. In contrast, rather than *debiasing* models, we aim to *model* and predict the *type* of political bias.

Other works (Sen et al. 2015; Olteanu et al. 2019; Nguyen et al. 2014; Eickhoff 2018) have analyzed the bias inherent in human annotated data introduced by crowdsourcing annotations. Otterbacher et al. (2018) show that sexist workers are less likely to find image search results biased. Dong et al. (2012) show that different ethinic groups tend to label the same images differently. In contrast to these works, we show *how* media sources already believed to be politically biased then exhibit that bias, both in terms of the visual content they choose to accompany article text and in terms of the text of the article itself. We also explicitly ask our workers to provide their rationale for their predictions and then *leverage* the stereotypical and biased notions used by the workers to model bias in visual media.

Comparison with our prior work This paper is a significant expansion of Thomas and Kovashka (2019). We compare against new baselines on the task of political bias prediction: Zhang et al. (2019), Gomez et al. (2017), a curriculum method and OCR-based method. We also include a new method and show new results for predicting the politics of *faces* rather

than full images. We include a finer-grained quantitative analysis of differences in facial portrayals across left/right than was present in our earlier work. We provide extensive new analysis of the connection between image and article text, e.g. discovering visually consistent words, predicting words from images, discovering words indicative of bias, the impact on humans' prediction of bias before/after seeing text, and predicting whether images and text are correctly paired. We also explain the ethical uses of our work and better position our work in the context of related work.

3 Dataset

Because no dataset exists for this problem, we assembled a large dataset of images and text about contemporary politically-charged topics. We got a list of "biased" sources from `mediabiasfactcheck.com` which places news media on a spectrum from extreme left to extreme right. We used a list of current "hot topics" e.g. immigration, LGBT rights, welfare, terrorism, the environment, etc. from Peck and Boutelier (2018). We crawled the media sources that were labeled left/right or extreme left/right for images using each of these topics as queries. After identifying images associated with each keyword and the pages they were on, we used Peters and Lecocq (2013)'s method to extract articles. The method splits webpages into a sequence of blocks based on the document structure and then predicts for each block whether it is part of the main article set using features such as link and word density. We use the publicly available implementation.[2] We obtained 1,861,336 images total and 1,559,004 articles total. We manually removed some boilerplate text (headers, copyrights, etc.) which leaked into some articles. However, because of the large diversity of HTML formats across the media sources, boilerplate text could not be completely removed in all cases.

3.1 Data Deduplication

Because sources cover the same events, some images are published multiple times. To prevent models from "cheating" by memorization, all experiments are performed on a "deduplicated" subset of our data. We extract features from a Resnet (He et al. 2016) model for all images. Because computing distances between all pairs is intractable, we use (Malkov and Yashunin 2016) for approximate kNN search ($k = 200$). We set a threshold on neighbors' distances to find duplicates and near-duplicates. We determine the threshold empirically by examining hundreds of kNN matches to ensure all near-duplicates are detected. From each set of duplicates, we select one image (and its associated article) to remain in our "deduplicated" dataset while excluding all others. If the same image appeared in both left and right media sources, we keep it on the side where it was more common, e.g. one left source and three right sources would result in preserving one of the image-text pairs from the right sources. We break ties randomly, i.e. if an image appears equally on the left and right, we randomly assign it to either the left or right by choosing one of the image-text pairs with the randomly chosen label. By including such examples, our model is forced to explicitly learn in the presence of subjectivity and noise by, for example, making its predictions less confident on such examples. This may help our classifier achieve a more realistic model of the subjective aspects of bias than it would by only training on data with objective, obvious bias. After removing duplicates, we are left with 1,079,588 unique images and paired text on which the remainder of this paper is based.

3.2 Crowdsourcing Annotations

We treat the problem of predicting bias as a weakly supervised task. For training, we assume all image-text pairs have the political leaning of the source they come from. In Sect. 5.3, we show that this assumption is reasonable by leveraging human labels, though it is certainly not correct for all images/text, e.g. a left-leaning source may publish a right-leaning image to critique it, or a photo in a biased source may contain no bias at all (e.g. an image of a cat). In order to better explore the viability of the weak labels, and understand human conceptions of bias, we ran a large-scale crowdsourcing study on Amazon Mechanical Turk (MTurk). We asked workers to guess the political leaning of images by indicating whether the image favored the left, right, or was unclear. In total, we showed 3,237 images to at least three workers each. We show examples of different levels of agreement in Fig. 2. In total, 993 were labeled with a clear left/right label by at least a majority. The remaining images were labeled as some combination of "Unclear" labels with "Left"/"Right" labels, e.g. "UUL" or "ULR".

We also asked our annotators what image features they used to make their guess. The features workers could choose (and the count of each agreed upon) was: closeup-90 (closeup of specific person's face), known person-409 (portrays public figure in politically-relevant way), multiple people-237 (group or class of people), no people-81 (scenes or objects associated with parties, e.g. windmill/left, gun/right), symbols-104 (e.g. swastika, pride flag), non-photographic-130 (cartoons, charts, etc.), logos-77 (logo of e.g. CNN, FOX, etc.), and text in image-267 (e.g. text on protest signs, captions, etc.). We also asked workers to provide a free-form text explanation of their political bias prediction for a small number of images. We extracted semantic concepts from these explanations and later used them to train one of

[2] https://github.com/dragnet-org/dragnet.

Fig. 2 We asked workers to predict the political leaning of images. We show examples here where all annotators agree, the majority agree, and where there was no consensus

our baseline methods (Sect. 5.1). Humans often mentioned using the positive/negative portrayal of public figures and the gender, race and ethnicity of photo subjects. We provide a demonstration of differences in portrayal across left/right in Sect. 5.4. Absent these cues, workers used stereotypical notions of what issues the left/right discuss or their values. For example, for images of protests or college women, annotators might guess "left".

We next showed workers the image's article and asked a series of questions about the image-text pair, such as the political leaning of the *pair* (as opposed to image only), the topic (e.g. terrorism, LGBT) the pair is related to, and which part of the article text is best aligned with the image. We computed agreement scores and found that 2.45 out 3 annotators agreed on the bias label of an image on average (including the "unclear" label), while 1.71 out of 3 agreed on topic, on average.

To ensure quality, we used validation images with obvious bias to disqualify careless workers. We restricted our task to US workers who passed a qualification test verifying familiarity with recent news and persons in the news, who had $\geq$ 98% approval rate, and who had completed $\geq$1,000 HITs. In total, we collected 14,327 sets of annotations (each containing image bias label, image-text pair bias label, topic, etc.) at a cost of $4,771. We include a number of experimental results on this human annotated set of images in Sect. 5.3.

Note that we did not take into account the annotators' personal political bias. The correlation between personal bias and bias labels could be explored in future work. However, in this project, we wanted to minimize making bias data collection uncomfortable due to asking personal questions and potentially violating privacy.

3.3 Relation of Weak and Human Annotations

In order to ensure that our weakly-supervised labels are actually capturing a meaningful signal which approximates human understandings of political bias, we perform the following test of weak-to-human label correlation. We evaluated the impact of text on humans' bias predictions. To do so, we compared how humans *changed* their predictions (made originally using the image only) after they saw the text paired with the image.

We found that when workers picked a left/right label, the label was strongly correlated with the weakly supervised label. Moreover, after seeing the text, humans became even more correct with respect to the noisy labels, switching many "unclear" predictions to the "correct" label (i.e. the noisy label). Specifically, in Table 1, we show the number of images labeled left/right before/after showing the worker the text paired with the image. Rows represent the image-only label of humans, and the columns represent the label after seeing both the image and the paired text. Any off-diagonal number represents a change in labeling between seeing the image only and seeing the image and text. We highlight the weak label in blue for left, and red for right. The row/column of blue in the left of the table, and the row/column of red in the right of the table, shows human annotations that agreed with the weak label, after seeing the image or both image and text. Red shading on the left, and blue shading on the right, shows the sum of "incorrect" votes (where human and weak labels disagree) in the setting of humans seeing either just the image, or both image and text. As a whole, we see that alignment between human and weak labels is strong, especially when humans see both images and text. In this setting, weak and human labels agree on 179 images on the left (with only 32 labels being of opposite value), and 92 on the right (with 74 being of opposite value).

When the weakly supervised label is Left, for example, we can see that of the 82 people who initially voted Left, 67 kept their initial vote after seeing the text, with only 15 changing their vote ("incorrectly," i.e. diverging from the weak, source-derived bias). Of the 49 who voted Right initially (diverging from the weak label), only 22 kept their initial ("incorrect") vote, while a significant 17 changed their vote to Left ("correct"), while 10 changed it to Unclear. Finally, for the 310 who initially voted an image was Unclear, 95 changed their vote to Left ("correct") after seeing the text, and 8 changed it to Right. When the weak label is Left, we see that while 82 initially voted left, after seeing the text 179 voted left. When

Table 1 Counts of how many users labeled an image Left/Right/Unclear

		Weakly Supervised = Left				Weakly Supervised = Right			
		Human Label After Seeing Image + Text							
		Left	Right	Unclear	SUM	Left	Right	Unclear	SUM
Human Label On Image Only	Left	**67**	**2**	**13**	**82**	28	**20**	6	54
	Right	**17**	22	10	49	**9**	**25**	**2**	**36**
	Unclear	**95**	8	207	310	37	**47**	121	205
	SUM	**179**	32	230	441	74	**92**	129	295

Rows show the label of the human on the image alone, while columns show the label after seeing the text. We further divide the table into two larger columns, which represent images with a weak (source-derived) label of Left/Right. Our results show the text helps annotators, and that our weakly-supervised labels are meaningful. We shade rows and columns corresponding to the "correct" (aligned with source) label

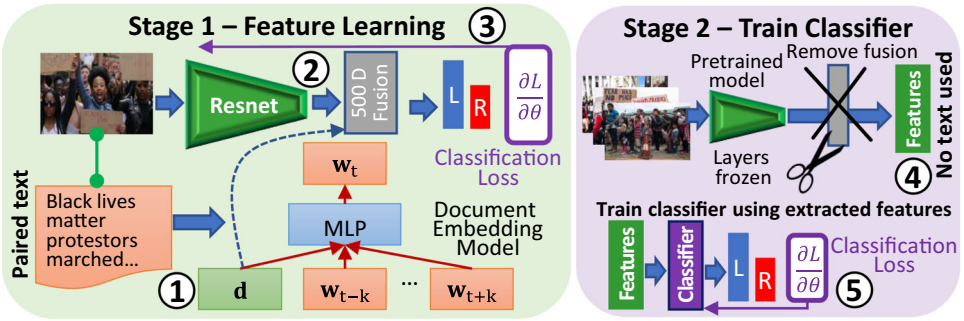

Fig. 3 We propose a two-stage approach. In stage 1, we learn visual features jointly with paired text for bias classification. In particular, we (1) learn a document embedding, then using this embedding model, we extact a representation for the text in our political articles, (2) pass an image from the article through a ResNet, then (3) train a model based on both the text and image inputs. In stage 2, we remove the text dependency by training a classifier on top of our prior model using purely visual features. In particular, we (4) extract visual features from the two-input stage-1 model, and (5) train a simple bias classifier based on these features. We show that this approach significantly outperforms directly training a model to predict bias

the weak label is Right, 36 agreed with the weak label before, and 92 after seeing the text. In other words, we can conclude that after seeing the (disambiguating) text, annotators do in fact align more with the weak label of the image, which indicates that the weakly supervised label captures a meaningful notion of bias.

Overall, this analysis indicates that: 1) our noisy labels are a good approximation of the true bias of the images (and thus can be used for training a method); and 2) the paired text is useful for predicting bias (a result also later borne out by our experiments).

4 Approach

We hypothesize that the complementary text domain provides a useful cue to guide the training of our visual bias classifier. Some aspects of bias could be semantic or textual (e.g. presence/absence of the word "donations"), while other aspects of bias are more visual. The text of the articles includes words that semantically correlate with political bias, e.g. "unite", "medicaid", "donations", "homosexuality", "Putin", "Antifa" and "brutality" strongly correlate with left bias according to our model, while "defend", "retired", "NRA", "minister" and "cooperation" strongly correlate with right bias. However, our method will ultimately be used to classify bias from images alone. We encourage it to capture the more visual aspects of bias, by factoring out these semantic concepts into the auxiliary text domain, and making it unnecessary to capture concepts which the textual channel captures. We thus enable our model to learn complementary visual cues.

We use information from the visual pipeline, and fuse it with the document embedding as an auxiliary source of information. Because we are primarily interested in *visual* political bias, we next remove our model's reliance on text features, but keep all convolutional layers fixed. We train a linear bias classifier on top of the first model, using it as a feature extractor. Thus, at *test time,* our model predicts the bias of an image *without using any text*. We illustrate our method in Fig. 3.

4.1 Method Details

We capture the implicit semantics of an image by leveraging the association between images and text. Let

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{a}_i, \mathbf{y}_i\}_{i=1}^{N} \tag{1}$$

denote our dataset $\mathcal{D}$, where $\mathbf{x}_i$ represents image i, $\mathbf{a}_i$, represents the textual article associated with the i^{th} image, and $\mathbf{y}_i$ represents the political leaning of the image. In the first stage of our method, we seek the following function:

$$f(\mathbf{x}_i, \Omega(\mathbf{a}_i)) = \mathbf{y}_i \tag{2}$$

where $\Omega(.)$ represents transforming the article text into a latent feature space. We train Doc2Vec (Le and Mikolov 2014) offline on our train set of articles to parameterize Ω. Specifically, Ω is trained to maximize the average log probability

$$\frac{1}{T}\sum_{t=1}^{T} \log p(\mathbf{w}_t | \mathbf{d}, \mathbf{w}_{t-k}, \ldots, \mathbf{w}_{t+k}) \tag{3}$$

where T is the number of words in article $\mathbf{a}$ (we omit the index i to simplify notation), p is the probability of the indicated word, $\mathbf{w}_t$ is the learned embedding for word t of article $\mathbf{a}$, $\mathbf{d}$ is the learned document embedding of $\mathbf{a}$ (200D), and k is the window around the word when training the model. We use hierarchical softmax (Morin and Bengio 2005) to compute p. We train Doc2Vec on our corpus of news articles, and observe more intuitive embeddings than from a pretrained model.

We show examples of the learned Doc2Vec space in Table 2. In the top row of the table, we show several query words which we embed using our model. We then compute the distance from each query word to all other learned words in our dataset's vocabulary and rank the words in order of increasing distance. Thus, retrieved words near the top are more closely related to the query word in the learned space than words below. We observe meaningful relationships within the space which are model can potentially exploit. For example, for the topic "Stoneman" (a school shooting), the model learned that "Parkland" (another school shooting), "NRA" (the National Rifle Association which protested gun measures following the shooting), "gunman", and "shooter" all relate to the broader topic of school shootings and violence in general. By providing this semantic supervision to our model, we wish to discover relevant *visual* cues which relate to the broader subject matter, and which are predictive of the politics of the image.

After training, we compute Ω for a given article $\mathbf{a}$ by finding the embedding $\mathbf{d}$ that maximizes Eq. 3. Ω thus projects each article into a space where the resulting vector captures the overall latent context and topic of the article. We provide $\Omega(\mathbf{a})$ to our model's fusion layer for each train image. The fusion layer is a linear layer which receives concatenated image and text features and learns to project them into a multimodal image-text embedding space which is finally used by the classifier.

The formulation of $f_\theta(.)$ described above requires that the *ground-truth* text be available at test time and also does not ensure that our model is learning *visual* bias (i.e. the classifier may be relying primarily on text features and ignoring the visual channel completely). To address this problem, in the second stage of our method, we finetune f_θ to directly predict the politics of an *image only*, without the text, as follows:

$$f'_{,\theta'}(\mathbf{x}_i) = \mathbf{y}_i \tag{4}$$

Specifically, we freeze the trained convolutional parameters of f_θ and add a final linear classifier layer to the network, whose parameters are denoted θ'. Because f_θ's convolutional layers have already been trained jointly with text features, they have already learned to extract visual features which complemented the text domain; we now learn to use those features *alone* for bias prediction, as shown in Fig. 3.

4.2 Additional Method for Faces

We next explore whether the same people were shown in disparate ways across the political spectrum. We thus began by detecting faces in our dataset using the regularly updated DLIB library's (King 2009) CNN-based face detector.[3] The detector consists of a three layer convolutional network which runs over a spatial pyramid using a sliding window approach to predict whether a face is present in each window. We use the pre-trained model publicly released by DLIB. Observationally, we found there is strong visual variability in the faces that left/right-leaning sources choose for popular figures, such as Donald Trump, Barack Obama and Hillary Clinton. We later provide quantitative and qualitative demonstrations in Sect. 5.4.

We also seek to capture the semantics behind these differences in facial portrayals. To do so, we leverage existing datasets containing labeled facial attributes and expressions. We train two residual networks on the datasets of Liu et al. (2015) and Mollahosseini et al. (2017), and use them to predict facial attributes and expressions for every image in our dataset. After detecting faces in our dataset, we *recognize* faces of known political figures because we expect popular political figures to recur throughout the dataset and be indicative of bias. In order to decide which figures to recognize, we leverage the text paired with images. We ran Honnibal and Montani (2017)'s named entity recognizer on our text

[3] http://dlib.net/dnn_mmod_face_detection_ex.cpp.html.

Table 2 Word relationships learned by our trained document embedding

Charlottesville	Clinton	dreamers	fascist	FBI	FOX	Obama	Stoneman	supremacist	terrorism	Trump
parkland	o'reilly	daca	fascism	cia	nbc	trump	parkland	supremacists	extremism	obama
antifa	maher	immigrants	racists	comey	cbs	bush	nra	supremacy	islamophobia	bush
ferguson	obama	undocumented	racist	doj	abc	reagan	gunman	nationalist	extremists	duterte
dallas	bush	aliens	nationalist	irs	breitbart	erdogan	shooter	house	racism	erdogan
rally	huckabee	immigration	extremist	investigation	cable	bashar	morning	privilege	extremist	sterling
nfl	merkel	deportation	supremacist	mueller	fake	clinton	separating	dana	fascism	reagan
islamophobia	trump	illegally	democrat	intelligence	buzzfeed	duterte	shootings	fascist	fbi	corbyn
berkeley	blasio	deferred	supremacy	flynn	cnn	macron	cbs	extremist	bigotry	macron
spencer	davis	shutdown	bigotry	wikileaks	hannity	carter	sheriff	conspiracy	immigration	clinton
shootings	treasury	amnesty	supremacists	dhs	outlet	vice	ripped	racist	shootings	bashar
tweeted	benghazi	bipartisan	islamophobia	epa	msnbc	obamacare	outrage	evangelical	russia	cameron

The top row are query words and the words below are the nearest words in the learned space

articles and narrowed the list of detected "Person" entities to the 96 most frequent politicians (and other celebrities) to form a vocabulary of "known" faces. We downloaded images for each face and used Schroff et al. (2015) to perform face recognition on the detected faces.

Formally, let f_a, f_e, f_r be our facial attribute, expression, and recognition networks, respectively. For each image in our dataset, $\mathbf{x}^i$, we obtain automatically predicted 40 attributes, 8 expression labels, and one-hot identity labels as follows:

$$\left\{\mathbf{x}^i, f_a\left(\mathbf{x}^i\right), f_e\left(\mathbf{x}^i\right), f_r\left(\mathbf{x}^i\right)\right\}_{i=1}^{N}. \tag{5}$$

We later use these predicted facial attribute and expression features for analysis and as input for our baseline networks for predicting the bias of faces.

In addition to "Person" entities detected by our named entity recognizer in the text, we also examine the nationalities, religious, and political groups (NORP) entities detected by our recognizer. As we did with person entities, we use the detected vocabulary of NORPs and download visual data for each of the top 200 NORPs from Google Image Search. We train a separate residual network (He et al. 2016) to perform image classification on this train set. We then use this model to provide predictions of each concept on each image in our dataset: $P(n_j|\mathbf{x}_i)$ denotes the probability that image $\mathbf{x}_i$ exhibits NORP n_j. We then use the probability of each predicted concept, as a feature vector $\mathbf{x}_i = [P(n_1|\mathbf{x}_i), P(n_2|\mathbf{x}_i), \ldots, P(n_{|N|}|\mathbf{x}_i)]$, where N is our vocabulary of NORPs. We later use these predictions for analyzing how different sides of the political spectrum portray different NORPs.

Note that we *do not use demographic information to predict bias*. Instead, we use it to show that some demographic factors (e.g. certain ethnic groups) are portrayed in notably different ways on the left and right in our crawled dataset. We believe this observation about portrayals in the media indicates a problem with the media, and our goal is to point out the problem so it may eventually be addressed.

4.3 Implementation Details

All methods use the Resnet-50 (He et al. 2016) architecture and are initialized with a pretrained Imagenet model. We train all models using Adam (Kingma and Ba 2015), with learning rate of 1.0e-4 and minibatch size of 64 images. We use cross-entropy loss and apply class-weight balancing to correct for slight data imbalance between L/R. We use an image size of 224x224 and random horizontal flipping as data augmentation. We use Xavier initialization (Glorot and Bengio 2010) for non-pretrained layers. We use PyTorch (Paszke et al. 2017) to train all image models. For our text embedding, we use Řehůřek and Sojka (2010), with $\mathbf{d} \in \mathcal{R}^{200\times 1}$ and train using distributed memory (Le and Mikolov 2014) for 20 epochs with window size $k = 20$, ignoring words which appear less than 20 times.

5 Experiments

We present experimental results on a number of tasks. We introduce the baselines we compare against for politics prediction, in Sect. 5.1. In Sect. 5.2, we present our results for predicting left/right bias on full images for a variety of methods, and perform a detailed analysis of factors the model uses for prediction. We test on a large held-out test set from our dataset, whose left/right labels come from the leaning of the news source containing the image. We also perform ablations of our method using weakly-supervised labels to test the soundness of our method and experimental design for politics prediction. Next, in Sect. 5.3 we test our methods using the per-image labels provided by humans. We show results on test images for which a majority of human annotators agreed on the bias. We discuss the relationship between our weakly-supervised and human labels and analyze how humans reason about visual bias.

Because we find humans strongly relied on identifying public figures and how people were portrayed in guessing the politics of an image, we then perform an analysis of faces alone, in Sect. 5.4. We first train models to predict the bias of faces. We show results for both well-known politicians and for faces in general. We then analyze the differences in facial portrayals across the left/right for a variety of facial features. We present results showing that faces are portrayed significantly different for popular figures and ethnic groups on opposite ends of the political spectrum.

In Sect. 5.5, we perform a similar analysis of the text paired with images, to discover how the text itself manifests political bias. We note political figures and some ethnic groups that appear disproportionately on one side vs. the other. Similarly, we leverage existing techniques for discovering biased word usage in language to analyze our dataset.

In Sect. 5.6, we present several results exploring the relationship between image and text. We show the most "visually consistent" words in our dataset (i.e. the words where the paired image content is more consistent across images which the word was paired with). We also show results for directly predicting the words that appeared in the article given an image.

Finally, in Sect. 5.7, we examine the topic annotations (e.g. abortion, gun rights, etc.) within our dataset. We also show visual consistency across topics, with some visually grounded topics (e.g. gun rights) being more consistent in visual space than abstract topics, illustrating the challenging semantic nature of the problem of modeling visual political bias.

5.1 Methods Compared

We show the accuracy of each method on predicting left/right bias. Note that we apply some baselines to either full images or faces. For example, "facial semantics" only applies to faces. Similarly, OCR is not applicable to faces. We compare against the following baselines:

- RESNET (He et al. 2016)—A standard 50-layer classification Resnet, trained for left/right classification.
- CURRIC—Our approach is a two-stage curriculum method, which first learns visual features coupled with text features and then learns to predict bias without the text. We wanted to see whether a standard Resnet trained in the same way but without text inputs, would gain any benefit. We thus first train a Resnet on our task. We then freeze the lower layers and train *only* the classifier on all train features in the second stage. Optimizing over all train features at once vs. minibatches can mitigate noisy gradients from our diverse data (Orr 1997; He et al. 2019).
- JOO (Joo et al. 2014)—Adaptation of Joo et al.'s method for our task. We use Joo et al. (2014)'s dataset to train predictors for 15 attributes and nine "intents" (qualities the photo subject is estimated to have, e.g. trustworthiness, competence). We then use the predictions for these attributes and intents on images from our dataset as additional features to a Resnet to predict a left/right leaning.
- HUMCONC—We use the manually extracted vocabulary of bias-related concepts (e.g. "confederate", "African-American") from the human-provided explanations (Sect. 3.2) and download data for each from Google Image Search. We train a separate Resnet to predict concepts, and use it on each image in our dataset: $p(c_j|\mathbf{x}_i)$ denotes the probability that image $\mathbf{x}_i$ exhibits concept c_j. We use the confidence of each concept, as a feature vector to predict bias.
- OCR—We use (Minghui Liao and Bai 2018) to recognize free-form scene text in images. Because images contain words not found in the default lexicon (e.g. "Manafort"), we create our own lexicon from the 100k most common words in our articles. We use Garbe (2019) for spelling correction. We represent each recognized word as its learned word embedding, denoted $\mathbf{w}'_i$, weighed by the confidence of the recognition $p\left(\mathbf{w}'_i\right)$ as provided by the recognition model. The feature is thus given by $\frac{1}{n}\sum_{i=1}^{n} p\left(\mathbf{w}'_i\right)\mathbf{w}'_i$.
- GOMEZ (Gomez et al. 2017)—Similar to our method, Gomez leverages text to guide the learning process, without requiring text at test time. Gomez first trains a Resnet to predict the text embedding of the article paired with the image, from the image alone. Note that in our case, we do not predict the text embedding, but rather use it as a source of auxiliary information. In the second stage, a classifier is trained to predict the left/right label from the model's features.
- ZHANG (Zhang et al. 2019)—We leverage neighbors' features to assist in inferring the political label. The intuition is that images in our dataset are ambiguous, hence neighbors may make the learning task easier. We compute nearest neighbors for each image in visual space and formulate the inference problem as a graph. We compute attention using the features of neighboring images from the last Resnet layer.
- FACIALSEM—We predicted facial attributes, expressions, and identities for every face in our dataset (see Sect. 4.2). We create a feature vector by appending the predicted identity of the portrayed person to the facial attributes and expressions, resulting in the following vector which is fused with Resnet image features: $\mathbf{x}_i = [f_a(a_1|\mathbf{x}_i), \ldots, f_a(a_m|\mathbf{x}_i), f_e(e_1|\mathbf{x}_i), \ldots, f_e(e_n|\mathbf{x}_i), f_r(p_1|\mathbf{x}_i), \ldots, f_r(p_o|\mathbf{x}_i)]$, where $f_a(a_j)$ and $f_e(e_j)$ denote the confidence of attribute/expression a_j or e_j being present in image $\mathbf{x}_i$. Further, $f_r(p_k|\mathbf{x}_i)$ is a 1 or 0 depending on whether person identity marker p_k is predicted in $\mathbf{x}_i$.

For reference, we also show three methods which use the ground truth text paired with the images *at test time*. We thus consider them upper-bounds to the task of visual-only prediction.

- TEXT uses the document embeddings computed from the text paired with the image, without using the image at test time.
- WORDS is a two-stage topic-based method. We train a Resnet to predict, for the 1000 most visually consistent words (see Sect. 5.6), which words appeared in the first two sentences of the image's paired article. To make the task easier, we also condition the model on the Doc2Vec vector of the article text. We then train a second model using *just the predicted words* to predict the left/right label for the image.
- IM+TEXT uses the text paired with the images (to compute a document embedding), in addition to the image. It is the same as the first stage of our approach (see Fig. 3, left), without the addition of the image classifier layer in step 2.

All methods use the same residual network architecture. For methods relying on additional features, we use the fusion architecture in Fig. 3.

5.2 Evaluating on Weakly-Supervised Labels

In this section, we present our results for predicting the political leaning of visual media. In Table 3, we show the results

of evaluating our methods on 75,148 held-out images with weakly-supervised labels. Our method performs best overall. The top two performing methods rely on semantics discovered in the text domain (OURS and OCR). OCR is unique in that it is able to explicitly use text information at test time, by discovering text within the image and then using word embeddings. OURS improves over OCR by 2.6%. The improvement of OURS over RESNET is 3.4%. This amounts to classifying an additional ∼2,555 images correctly. We also observe that CURRIC performs nearly 1% better than RESNET. One reason for this is because of the high visual diversity of our dataset. A classifier that is being trained while the lower layers of the model continue to change (the model keeps shifting features because it is unable to settle on consistent patters) must continually readjust itself to the changing features. However, by freezing the lower layers of the model, we allow the classifier to optimize for a stationary set of visual features over the entire dataset. The classifier is thus able to obtain a better and more stable classification, resulting in a slight gain in performance, but still worse than our method.

We observe that JOO, which leverages features learned on an external dataset, performs worse than RESNET. This is likely because Joo et al. (2014)'s data mainly features closeups of politicians, while ours contains a much broader image range, thus the predicted features are not useful in our setting. Further, relying on the concepts humans identified (HUMCONC) actually slightly *hurt* performance compared to RESNET. This may be because of a disconnect between humans' preconceived notions about left/right and those required by the dataset.

In addition, we note that GOMEZ performs much worse than our method, even though both try to exploit information in the paired text domain as a source of privileged information. We believe one reason for this is the many-to-many relationship of images with topics (e.g. image of the White House can be paired with text about Trump's children, border control, LGBT rights, etc.). Thus, it is much harder to predict the document embedding paired with the image since the text could be about many different issues, hence the GOMEZ model's features are not discriminative of politics. We observe that ZHANG, which relies on nearest neighbors computed in image space, also performs poorly. One reason for this is because our problem lies highly in semantic, rather than visual space. Thus, *visual* nearest neighbors are not necessarily indicative of the politics of an image.

Quantitative ablations To test the soundness of our method and experimental design, we performed several ablations. We first tested the importance of the second stage of our method (right side of Fig. 3). We used IM+TEXT, the result of the first stage, and instead of performing stage 2, we removed the dependency on text by zeroing out all text embedding weights in the fusion layer. We evaluated on our weakly-supervised test set and obtained 0.677, a result significantly worse than our full method, underscoring the importance of stage 2. We next tested how the performance of our method varied given the length of the article text. We trained our method with the first k sentences and obtained these results: $k = 1 \rightarrow 0.672$, $k = 2 \rightarrow 0.669$, $k = 5 \rightarrow 0.668$, $k = 10 \rightarrow 0.669$. All choices of k performed worse than using the full article (0.712). We finally examined how reliant our method was on training images from a particular media source (i.e. to test if the model was learning non-generalizable, source-specific features). We experimented with leaving out all training data harvested from a few popular sources. The result was (before → after excluding): CNN (0.873 → 0.866), TheBlaze (0.746 → 0.742), DailyCaller (0.703 → 0.667), DemocraticUnderground (0.713 → 0.700), NewsMax (0.685 → 0.628), CommonDreams (0.647 → 0.636), Breitbart (0.607 → 0.566). We observed only a slight decrease for most sources, suggesting our method is not dependent on seeing the source at train time.

5.3 Evaluating on Human Labels

We next tested our methods on test images which at least a majority of MTurkers labeled as having the same bias, i.e. those that humans agreed had a particular left/right label. We described this dataset in Sect. 3.2. Because workers also labeled images with what features they used to make their prediction, we break down each method's performance by feature. We show the result in Table 4. We only include the competitive methods from Table 3 (those with at least 65% accuracy) for brevity.

OURS performs best on average across all categories and performs best (or ties) on four out of eight categories. Categories where OURS is outperformed are reasonable: OCR performs best or second-best when text can be relied on in the image, i.e. "logos" and "text in image". We note that while the overall result for OCR approaches OURS, OURS works better on a broader set of images than OCR and is thus a more general method for predicting *visual* bias. OURS is also outperformed by HUMCONC when humans relied on a known face (politician, celebrity, etc.). This may be because HUMCONC relies on external training data (Sect. 5.1) which feature many known individuals, e.g. "rappers" and "founding fathers". Perhaps counterintuitively, JOO outperforms our method when the prediction depends on scene context ("no people"), but note that some of the attributes that JOO uses do capture the scene/background (e.g. indoor, background, national flag, etc.). Further, unlike OURS, this method uses an external human-labeled dataset to learn features, including the scene attributes. CURRIC improves upon RESNET (whose features it uses in its second stage of training) in nearly every category and performs best or ties in two categories. This result suggests that label noise and high visual diversity

Table 3 Accuracy on weakly-supervised labels with the best visual-only prediction method in bold, and second-best in italics

Method	ResNet	Curric	Joo	HumConc	OCR	Gomez	Zhang	Ours	Text	Words	Im+Text
Accuracy	0.678	*0.687*	0.670	0.675	0.686	0.547	0.566	**0.712**	0.825	0.626	0.803

These results are computed on full images, hence face-specific methods are excluded

Table 4 Accuracy on human consensus labels with the best visual-only prediction method in bold, and second-best in italics

Feature/Method	ResNet	Curric	Joo	HumConc	OCR	Ours	Text	Im+Text	# Ims
Closeup	0.567	0.589	0.544	*0.622*	0.578	**0.656**	0.667	0.578	90
Known Person	*0.567*	0.558	0.550	**0.570**	0.560	0.521	0.558	0.575	409
Multiple People	0.722	*0.738*	0.671	0.688	0.730	**0.768**	0.709	0.705	237
No People	0.556	0.531	**0.605**	0.494	0.580	*0.593*	0.642	0.667	81
Symbols	0.558	0.587	*0.596*	0.548	0.577	**0.606**	0.625	0.587	104
Non-Photographic	0.577	**0.585**	0.569	*0.584*	0.577	**0.585**	0.631	0.654	130
Logos	0.545	*0.636*	0.584	0.597	**0.662**	0.623	0.546	0.584	77
Text in Image	0.629	**0.652**	0.625	0.596	*0.637*	0.607	0.648	0.659	267
Average	0.590	0.610	0.593	0.587	*0.613*	**0.620**	0.628	0.626	

The results are computed on full images grouped into eight categories by our human annotators

within minibatches may prevent the classifier from converging on the best local minima for a given set of features. By fixing the model's layers and optimizing the classifier layer across the entire train set at once, the classifier converges on a better solution. This technique can be applied to any method to potentially improve their performance as well.

In terms of the "upper bound" methods, we note that Im+Text performs significantly worse on human labels vs. weakly-supervised labels. This is likely because some words in text point to a specific bias (e.g. abortion vs. pro-choice), which the model may be over-relying on to predict the bias of the image. In contrast, the relationship between image features and bias is often more ambiguous. Further, because of the noisy data collection, Im+Text may have learned to exploit dataset-specific features (e.g. author names, header text, etc.) for prediction, which does not actually translate into humans' commonsense understanding of political bias. This also explains why Im+Text does not improve upon Text alone on average (but it does for five of eight categories).

We next test whether our assumption that all images harvested from a right- or left-leaning source exhibit that type of bias is reasonable. Several results computed from our ground-truth human study suggest that our web labels are a reasonable approximation of bias. First, we observe that the relative performance of the methods across Tables 3 and 4 is roughly maintained; Ours is best, followed by OCR and Curric essentially tied. The results are also sound, e.g. when humans used text, OCR tends to do better, which indicates the model's concept of bias correlates with humans'. Earlier, in Table 1, we showed that human labels agree with our weak labels more, when text information is presented to disambiguate the image's bias.

Human label consensus's effect on performance In an additional experiment, we explored the difference between the performance of our method on images on which the *majority* of humans agreed vs. those on which humans *unanimously* agreed. We found that our method worked better when humans unanimously labeled the images vs. simple majority (gain of 4.8%). This suggests that as humans become more certain of bias, our model (trained on noisy data) also performs better.

Detecting images with ambiguous bias Our result in Table 1, showed that humans become more correct with respect to our weakly-supervised labels after seeing the text. Next, we learn to predict whether a human will change their political label of the image *after seeing the text*. Thus, we model whether the politics of the image is *predictable without the text* (i.e. unambiguous). We use the F_1 score rather than accuracy due to class imbalance. We find that our model is fairly accurate at predicting whether images are unambiguous (0.731), but less accurate at detecting ambiguous images (0.308), i.e. images that humans change their label on. In other words, given an image and the text paired with the image, our model is able to accurately able to predict when the human label of the image *will not* change after humans see the text. These are likely simpler cases where the image has more apparent political bias. However, the model performs worse at predicting when the human label of the image *will* change, after seeing the text. For example, if the model suspects the image has a right bias, but suspects the text has a left bias, it is difficult for the model to decide whether humans will rely more on the image

Table 5 Words for both the left and the right that had the highest predictive weight (i.e. the word's appearance caused the classifier to be more likely to predict that category)

Left							
bob	television	unite	views	speakers	irs	medicaid	putin
homosexuality	outlets	gary	enforce	donations	doj	opposition	broadcast
speaks	antifa	adhere	westminster	lobby	achievements	networks	pelosi
reactions	labour	venezuela	supporter	meeting	memoir	warrant	outlet
brutality	misleading	hall	sharon	prominent	illegal	angela	referring
raped	absurd	berkeley	spoke	donald	qaeda	karl	rejected
brad	quit	roe	intelligence	candidate	evan	hosted	comedian
Right							
teresa	defend	hopeful	survivor	indicted	immigration	colleagues	retired
theresa	refuse	political	roger	caucus	nba	bipartisan	williams
rand	nra	withdraw	trump	minister	racist	ratings	longtime
sexually	fox	joins	cruz	deputy	unilaterally	sentenced	denial
dana	pleaded	declaring	exposing	victories	planned	ballot	hannity
russians	juan	guests	hashtag	cooperation	establishment	chancellor	network
sarah	recording	blaming	deportation	roy	supporting	don	erdogan

(unambiguous label) or the text (ambiguous label) to make their final decision for the image. These cases are likely more semantically complex (and therefore more difficult to model) and require one to reconsider what the image is intended to portray in the context of the text.

Politically discriminative words Given the strong performance of models relying on text and the fact that workers became more confident after seeing the text, we wanted to discover words in the text which are indicative that the image-text pair leans left or right. We used the classifier from the WORDS model. In Table 5, we show words for both the left and the right that had the highest predictive weight (i.e. the word's appearance caused the classifier to be more likely to predict that category). We note several interesting results: "bob" (likely from Robert Mueller's name) and "unite" (from the Unite The Right protest) are among the strongest predictors of "left", while "teresa" (likely from Teresa May) and "immigration" strongly indicate right. Many of the words used by the model suggest topics frequently mentioned by their respective sides, but which are not mentioned by the other side, possibly because they are politically damaging/advantageous to one side. For example, "irs", "putin", "doj", and "antifa" are predictive of left, while "nra", "trump", "fox" predict right. In sum, this result allows us to see disparities in the issues covered in left vs. right articles, as well as the different words used by the articles which are politically discriminative.

Visual explanations We wanted to see whether we could interpret how our model learned to perform bias classification. We used Grad-CAM++ (Chattopadhay et al. 2018) to compute attention maps on images that humans annotated. We show the result in Fig. 4. We observe that our model pays the most attention to logos and faces of public figures. We see the model only focuses on the "PBS" logo in the first row, but pays attention to both the "Fox News" logo and the face of the well-known commentator in the second row. We believe that because our model was trained with the topic information provided via the text embedding during stage one, the visual component of the model learned to focus on learning visual features that complemented the text (such as logos and faces). Ultimately these features work better than those found by the standard ResNet model.

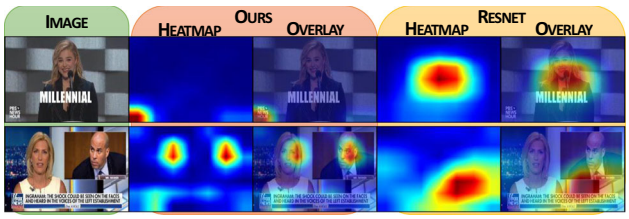

Fig. 4 We show visual explanations using Chattopadhay et al. (2018). We note that our model looks to logos and faces of public figures, while the baseline uses objects (e.g. microphone) and scene type (e.g. city in background)

5.4 Evaluating on Faces

Many workers noted how politicians were portrayed in making their decision (Sect. 3.2). We thus analyze how well our methods could do at predicting the politics of faces alone, in the absence of any context from the image. We thus trained models to predict the political bias of the faces we detected

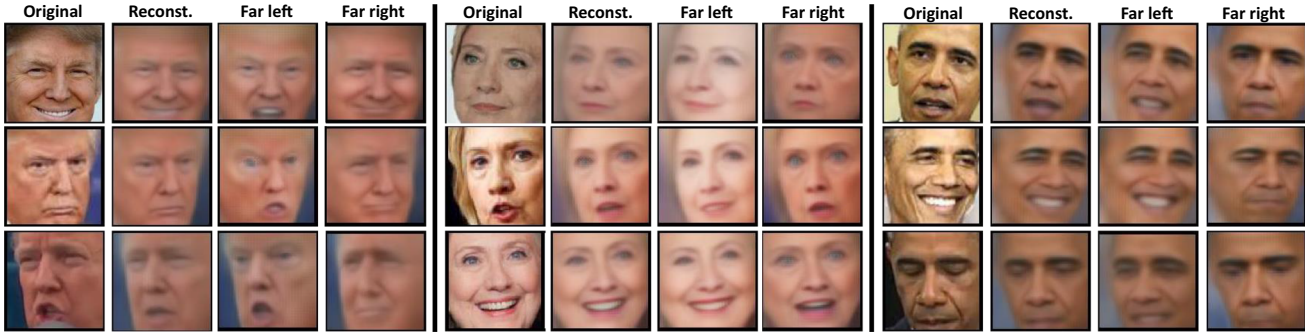

Fig. 5 We modified photos to be more left/right-leaning, using a generative model trained on our noisy data. We show the model's "reconstruction" of each face next to the original sample, followed by the sample transformed to the far left and right

in our images (see Sect. 4.2). We assume all detected faces in an image have the same political bias as the image itself (e.g. a right leaning image with 10 detected faces results in 10 individual samples all with the "right" weakly supervised label).

Predicting the bias of all faces We present our results in Table 6 for a subset of the previously evaluated baselines as well as the face-specific method FACIALSEM which relies on predicted facial attributes, expressions, and identity. Note that the OCR model is inapplicable to cropped faces because those do not contain text. We observe that FACIALSEM substantially improves over other baselines and achieves the strongest performance (0.607). We also observe that OURS performs on par with (slightly better than) RESNET and JOO.

One possible reason for the lack of performance gain of our main method on faces is the lack of context from which the model can learn. In the full image setting, our method has a complete view of the image and the text (in stage 1) and is thus able to learn how the concepts in the image complement the text. However, in the face setting, our model has no visual context and is unable to learn relevant visual features to complement the shared text embedding. We note that even though the model sees no context outside of the cropped face, the FACIALSEM model is able to predict the political leaning of the face with 60.7% accuracy, which suggests that faces are portrayed in a biased manner which the models are capturing.

Predicting the bias of well-known vs. lesser-known faces We next show the accuracy of predicting the bias of a media source, based on of different types of faces, in Table 7. We wanted to test whether well-known public figures are portrayed in a substantially different way compared to lesser-known or unknown figures. We show our model's accuracy at predicting the politics of Obama/Trump faces (most well known), then on a much larger set of 96 politicians we detected in text as described in Sect. 4.2 (less well-known), then faces that were classified as being one of 196 nationalities, religions or groups (NORPs) (unknown person other than a known category or nationality), and finally all faces (most unknown). We observe our model is remarkably

Table 6 For a subset of methods, we show the accuracy of predicting the politics of *just faces* detected in our images and evaluate on weakly-supervised labels. We show the best visual-only prediction method in bold

Method	RESNET	JOO	FACIALSEM	OURS	IM+TEXT
Accuracy	0.588	0.579	**0.607**	0.590	0.723

accurate for known political figures and that performance decreases as the face in question becomes less well known. Our results strongly suggest that public figures, and to a lesser extent, nationalities, religious, and political groups, are portrayed in politically biased ways. This is sensible because content creators may attempt to disparage or elevate political figures and groups of people who are politically opposed or aligned to their position.

Modeling facial differences across politics So far we have seen strong evidence that faces are presented in substantially different ways across the political spectrum, particularly political figures. We next seek to actually *visualize* the differences in how well-known individuals are portrayed within our dataset. To this end, we trained a generative model to modify a given Trump/Clinton/Obama face, and make it appear as if it came from a left/right leaning source. We use a variation of the autoencoder-based model from Thomas and Kovashka (2018), which learns a distribution of facial attributes and latent features on ads, not political images. We train the model using the features from the original method on faces of Trump/Clinton/Obama detected in our dataset. To modify an image, we condition the generator on the image's embedding and modify the distribution of attributes/expressions for the image to match that person's average portrayal on the left/right, following Thomas and Kovashka (2018)'s technique. We show the results in Fig. 5. Observe that Trump and Clinton appear angry on the far-left/right (respectively) end of the spectrum. In contrast, all three appear happy/benevolent in sources supporting their own party. We also observe Clinton appears younger in far-

Table 7 Accuracy of predicting the politics of different types of faces (from most well-known to least well-known)

Face Type	Obama / Trump	One of 96 Politicians	Any NORP	All Faces
Accuracy	0.830	0.820	0.670	0.590

Table 8 Facial attributes and expressions which significantly differed (shown in blue) across the left/right per politician

	Barack Obama	Chuck Schumer	Donald Trump	Hillary Clinton	Mitch Mcconnell	Nancy Pelosi	Paul Ryan
Facial Attributes							
5 o'Clock Shadow	■	■	■				
Attractive	■			■			
Bags Under Eyes	■		■	■			■
Bald	■			■			
Big Lips	■		■			■	
Big Nose	■		■	■	■		■
Chubby	■			■	■		
Double Chin	■			■	■		
Gray Hair	■			■			
Heavy Makeup				■			
High Cheekbones	■		■	■	■		
Mouth Slightly Open	■	■	■				
Mustache	■		■				
Narrow Eyes	■		■	■		■	■
No Beard	■		■				
Pointy Nose			■	■			
Receding Hairline				■	■		
Rosy Cheeks				■		■	
Smiling	■			■			
Young	■		■	■	■	■	
Facial Expressions							
Anger	■		■	■			
Contempt	■				■		
Disgust				■		■	
Fear		■	■	■	■	■	
Happy	■		■	■		■	
Neutral			■		■		■
Sad	■	■	■	■	■		
Surprise			■	■	■		
Arousal	■		■	■		■	
Valence	■		■	■		■	

left sources. In far-right sources, Obama appears confused or embarrassed. These results further underscore that our weakly supervised labels are accurate enough to extract a meaningful signal.

Discovering biased features for public figures We next show in Table 8 *which* features are quantitatively different for *which* politicians, using a subset of all features. We predicted facial attributes and expressions for the most frequent politicians which appeared in our dataset. We then performed a per-feature T-test to discover which attributes and expressions are portrayed differently across the left and the right for each politician. We highlight cells in blue whose feature differences in portrayal are significant ($p \leq 0.05$) across the political spectrum. Empty cells indicate the difference across left/right was not significant.

We observe that Obama, Trump, and Clinton have the most facial differences. We observe a number of significant differences which were also reflected in our generation results (Figure 5). We see Hillary Clinton differs in "attractive", "bags under eyes", "chubby", "double chin", "heavy makeup", and numerous other attributes which suggest she is being portrayed as older and less attractive on one side vs. the other. We observe similar attribute patterns for Obama and Trump, with Obama and Trump likewise being portrayed differently in terms of their age ("young") and attractiveness (e.g. "5 o'clock shadow", "bags under eyes"). For facial expressions, we see Obama, Trump, and Clinton all differ in the "anger" "happy", and "sad" facial attributes, as well as their facial expressions' arousal and valence scores. As was shown from our generation results, negative expressions (e.g. "anger", "disgust", etc.) are used to portray figures from the opposite side of the spectrum, while positive expressions (e.g. "happy") are used for political figures on the same side. Interestingly, we also note significant differences in both the arousal and valence scores for several politicians. Arousal is a measure of the intensity of a given facial expression and measures whether a given face is exciting/agitating vs. calm/soothing, while valence is a measure of the "pleasantness" of the face (Mollahosseini et al. 2017). Thus, our results suggest that not only are the expressions themselves different, but the degree to which those expressions are shown is also different (i.e. through their arousal) as well as the overall pleasantness of the face.

Table 9 Frequently detected NORP's facial semantic attributes and their differences in portrayal across the left/right

	African American	Arab	Asian	Muslim	Mexican	Hispanic	White
Facial Attributes and Expressions							
Median p-val.	**0.000**	**0.003**	**0.018**	0.071	0.099	0.185	0.370

Table 10 Top-15 names across the left/right which were mentioned most on one side, relative to the other side

Left	Right
Richard Spencer	Brett Kavanaugh
Milo Yiannopoulos	Justin Trudeau
Scott Pruitt	Jesus Christ
Michael Flynn	Nancy Pelosi
Alex Jones	George Soros
Karl Marx	Joe Biden
Richard Bertrand Spencer	Rush Limbaugh
Moon Jae In	Barack Obama
Colin Kaepernick	Pope Francis
Steve Bannon	Al Gore
Jared Kushner	Jeremy Corbyn
Betsy Devos	Bill Clinton
Adolf Hitler	Ronald Reagan
Michael Cohen	Chuck Schumer
Doug Jones	Ron Paul

Discovering biased features for NORPs We next expand the analysis that we performed in Table 8, to faces detected to be NORPs by our classifier described in Sect. 4.2. We predict facial attribute and expression values on each face and discover features which significantly differ in their portrayal across left/right. We found many NORPs had features which significantly differed. We thus show a condensed version of the table for a subset of NORPs which were most commonly detected in our dataset. In Table 9, we show the median p-values of the features (we found that the average p-value was too strongly influenced by several features with large p-values). We highlight significant differences ($p \leq 0.05$) in bold. We see that "African American", "Arab" and "Asian" all have median p-values which are significant, indicating at least half of the features significantly differ across the left and the right. We observe that the most significant differences occur with "African American" and the least significant p-value occurs for the "White" category, which implies this category's portrayal is most uniform across left/right. This result shows that groups of people, primarily minority groups, are portrayed in significantly different ways in left vs. right media sources. While this result is expected, it does quantitatively demonstrate a problem.

5.5 Bias in Text

In this section, we extend our analysis of political bias to the text paired with each image, without considering the image. We observe in Table 3 that the TEXT model is highly accurate at predicting bias, suggesting that the text contains a highly discriminative signal. We thus wish to understand precisely how the text is biased, both in terms of disparities in the frequency in which certain subject matter is discussed, as well as the choice of words to discuss those subjects. We first consider what political figures are mentioned disproportionately on each side of the political spectrum. We then consider the use of language by each side known to be biased from prior research.

Public figures with disproportionate mentions in text In Sect. 4.2, we described how we performed named entity recognition on our text dataset and discovered frequently mentioned names which we then used to train a face recognition model. We also wanted to discover what names were lopsided in their frequency of occurrence on each side of the spectrum. We counted the number of occurrences for each name on the left vs. the right. Because of data imbalance between the left and the right, we normalized the number of occurrences of a name on each side by the total number of names mentioned on that side. In Table 10 we show the names with the largest difference between sides. We observe extreme and polarizing figures are mentioned disproportionately, e.g. Richard Spencer, Alex Jones, Adolf Hitler are mentioned much more on the left vs. the right. In contrast Brett Kavanaugh, Rush Limbaugh, Bill Clinton, etc. are mentioned more on the right relative to the left. These results are sensible. For example, biased sources on the left may attempt to smear the right with Richard Spencer (a neo-Nazi), Milo Yiannopoulos (an alt-right figure), and Alex Jones (a conspiracy theorist). Discussing these figures disproportionately on one side suggests that relatively obscure public figures are being overemphasized for potentially politically biased reasons. Similarly, the right more frequently mentions Brett Kavanaugh (a Supreme Court justice accused of sexual misconduct) and George Soros (a large donor to political causes on the left). For Kavanaugh, right sources were likely trying to rally support behind his nomination to the Supreme Court. Right sources have also frequently attacked Soros's funding of leftist political causes with conspiracy theories (Vogel et al. 2018).

Table 11 Most disproportionately used known biased words from Recasens et al. (2013) by the left/right

Left										
report	people	thing	work	king	way	very	white	right	try	revolution
say	movement	fascist	lack	fight	struggle	act	content	world	system	start
need	march	see	happen	comment	rights	know	make	write	national	society
film	racist	different	country	live	support	war	win	take	regime	justice
post	human	social	pat	article	action	violence	call	nationalist	quality	point
Right										
tax	end	child	year	use	law	government	conservative	lie	state	liberal
migration	left	life	illegal	policy	form	increase	abortion	provide	cost	author
pass	school	free	rate	serve	claim	believe	man	business	day	terrorist
new	vote	aim	old	order	prove	heart	individual	formation	church	economic
come	result	marriage	term	former	religious	faith	service	far	case	fact

Imbalanced biased word usage Recasens et al. (2013) studies the problem of detecting bias in text. The authors consider edits to Wikipedia made to remove biased language and develop a lexicon of words which suggest a biased or non-neutral point of view (e.g. McMansion vs. large home, murder vs. kill, pro-life vs. anti-abortion, etc.). We count the number of times words appearing in the biased word lexicon were used by both the left and the right in our dataset. We then show the "biased" words that are used most by one political side relative to the other. The most skewed words across the left/right were "report" (for the left), most likely in connection to the Mueller Report, while "tax" is used most by the right. On the left, we observe words which indicate potentially biased characterizations, e.g. revolution, movement, struggle, fascist, racist, nationalist, etc. On the right, we observe biased language about a different set of issues, e.g. lie, migration, illegal, abortion, terrorist, etc. Collectively, our results presented in Tables 10 and 11 reveal that each side of the political spectrum has a set of "hot-button" issues which they use to either galvanize their audience for their cause or which they use to attack the opposite side.

5.6 Exploring the Relationship Between Image and Text

Our method for predicting political bias leverages the text paired with articles to guide the training of our purely visual model. We thus seek to better understand the relationship between our images and text. We first discover and illustrate words whose visual representation is most consistent throughout our dataset. We then further examine the WORDS method (described above in Sect. 5.1), which directly predicts words from an image and discover which words the model is able to predict best. Finally, we study whether we can directly model the complex relationship between images and text within our dataset by training a model to predict whether a given image-text pair is properly aligned.

Modeling word-level visual consistency We have argued that one of the challenges of modeling political bias in images is that the relationship between images and semantic topics and text is highly complex. For example, an image of the White House could be paired with an article about immigration or one about the US-Afghanistan war. Thus, unlike traditional image captioning tasks where the text directly describes the literal content of the image, the visual grounding of the text in the image is non-literal and consequently more challenging for a model to grasp. However, because we are exploiting our text to guide training of our models, we wished to discover the most "visually consistent" words in our dataset, that is, words whose visual expression in the image is most consistent across different images. Our goal is similar to Hessel et al. (2018), which seeks to model the visual concreteness of topics within multimodal data. To discover the most visually consistent words in our dataset, we first performed tokenization using Spacy (Honnibal and Montani 2017). Then, for each word, we created a list of images in which that word appeared in the first two sentences. Next, again for each word, we performed k-means clustering, with $k = 5$, which we determined worked well empirically. The intuition behind performing k-means in our case is that many words may appear visually inconsistent if one simply takes the average distance between all pairs of images for a given word, because their visual grounding could be multimodal. We compute the visual consistency v for word w as:

$$v_w = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} \| x_i^j - c_j \|^2}{\sum_{j=1}^{k} n_j} \tag{6}$$

where k is the number of clusters, n_j is the number of images in cluster j, x_i^j are image features which have been assigned to cluster j by k-means and c_j is the centroid of cluster

Table 12 Most "visually consistent" words in our data, in decreasing order (left to right, top to bottom). See text for details

fox	cnn	gop	host	republican	republicans	donald	candidate	senate
clinton	hillary	democrats	trump	presidential	secretary	barack	interview	attorney
democratic	president	conservative	obama	immigration	liberal	committee	speech	campaign
election	house	twitter	congress	immigrants	party	leader	vote	bill
executive	racist	meeting	abortion	prime	george	paul	asked	white
conference	debate	minister	press	administration	chief	calling	john	washington

Fig. 6 Most visually consistent words and image examples from the two tightest visual clusters computed for each word

j. Equation 6 essentially measures how tightly the visual features for a given word fit the 5-modal distribtion induced by our clustering. We compute this metric for the 10,000 most common words.

We show the most visually consistent words in Table 12. We observe that news organizations, political groups, and candidates dominate: FOX, CNN, GOP, Republican, Donald, Clinton, Trump, Barack, etc. Several political topics also emerge as visually consistent, e.g. immigration, immigrants, abortion. We next wanted to see what did our visually consistent images for each word actually look like. For six of our most visually consistent words, we sample images from the top-2 tightest visual clusters computed for each word and present them in Fig. 6. We observe that three of the top four (FOX, CNN, host) most consistent words primarily feature images of people on newscasts. We note that the most visually similar images are not necessarily semantically similar, as the news broadcasts are presenting a variety of unrelated topics. The tightest clusters for the word "GOP" feature portrait shots of political figures on the right (from top to bottom: Donald Trump, Mike Pompeo, Brett Kavanaugh, and Ted Cruz). For the last two words ("Republican" and "Donald"), we observe that the model has placed cartoons and illustrations closest together, in addition to clusters of political figures.

Predicting visually-consistent words from images We next wanted to see how well a model could exploit this word-level visual consistency. We previously used the predictions from this model to train a word-based politics predictor in Table 3 (the WORDS model). We show the F_1 score of predicting words from an image on our test set in Table 13. We choose F_1 score because multiple words can be paired with each image. We observe that our model performs better at predicting visually consistent words on average vs. non-consistent words. We observe numerous words which appeared in Table 12 have relatively higher F_1 scores relative to other words, with all the highest scoring words appearing in the table as being visually consistent. For example, we see "president" : 0.434, "trump" : 0.590, "donald" : 0.413, "immigration" : 0.497, and "abortion" : 0.534. However, we observe that the visual consistency of images associated with a single word does not guarantee discriminativity. In other words, just because images associated with a word all share similar visual content, does not imply that all images with that type of visual content are exclusively associated with that particular word. For example, we observe relatively poor performance at predicting the word "CNN" and "FOX", even though these words have visually consistent images. This is likely because the model has trouble differentiating between many different news programs, given their similar visual content: the model

Table 13 Per-word F_1 scores of a model trained to predict whether each word is/is not present in the image's article given the image and text embedding

Word	F_1 score	Word	F_1 score	Word	F_1 score	Word	F_1 score	Word	F_1 score	Word	F_1 score
trump	0.590	abortion	0.534	immigration	0.497	president	0.434	hillary	0.430	gay	0.429
donald	0.413	clinton	0.378	immigrants	0.357	supreme	0.341	obama	0.317	republican	0.309
news	0.302	fox	0.302	party	0.299	republicans	0.295	racist	0.274	presidential	0.273
democratic	0.272	media	0.263	candidate	0.259	bill	0.257	white	0.252	illegal	0.252
election	0.248	conservative	0.246	justice	0.244	democrats	0.240	campaign	0.238	senate	0.232
tuesday	0.225	speech	0.223	deal	0.222	administration	0.218	house	0.217	debate	0.216
paul	0.211	vote	0.204	foreign	0.203	political	0.200	minister	0.199	washington	0.191
thursday	0.189	conference	0.188	voters	0.187	meeting	0.186	twitter	0.183	night	0.181
cnn	0.171	prime	0.169	congress	0.168	barack	0.162	host	0.157	committee	0.157

We consider a dictionary of the top-1000 most visually consistent words and show the performance of the model on the best-performing words below

Fig. 7 We train a model to predict words from images. The model learns relevant visual cues for each word, demonstrating the utility of exploiting text (in this case, as an extra input)

may recognize a newsanchor at a desk, but then become confused as to whether the image is from CNN, FOX, MSNBC, ABC, etc.

High-response images for visually consistent words In Fig. 7, we show examples of images that were among the top-100 strongest predictions for that word. We observe, for example, that the model strongly predicts "antifa" for black-clad protestors and protestors holding banners. The model predicts "brutality" for images with African American protestors and for police scenes. The model predicts the word "immigrant" for images containing a border wall and Hispanic individuals, and "LGBT" for pride flags and rainbow like banners.

5.7 Visual Variability Across Political Topics

Each image in our dataset is also labeled with the political topic (e.g. abortion) that the media source was queried with when the image was scraped. We now explore the topic annotations on our dataset. We first present results on predicting the political topic of an image. We then discover topics which are most visually consistent in their portrayal across images. Finally, we illustrate the difficulty of classifying images as left/right, by showing images which are closest in visual space from each political side within each topic.

Predicting political topics from images We trained a model to predict the weak political topic label for each image in our training set (assuming each image exemplifies the topic of the parent article), given the image and the document embedding of the text. To ensure that the weakly supervised topic labels were actually capturing the real political issue of the images, we evaluated our model on our set of human annotated data. Each image can be labeled as being related to multiple topics, so we compute F_1 score rather than accuracy. We present the results in Table 14. We find that our model is able to predict most topics fairly accurately. For example, we observe that our model is most accurate at predicting images of "abortion" and "gun control". This makes sense because images about these topics share common scenes and objects: images about abortion often feature protest scenes or images of babies, while gun control images often feature firearms.

Table 14 F_1 score of predicting the political topics of an image-text pair on human annotations. Note that the same image-text pair can be labeled with multiple issues

Topic	F_1	Topic	F_1
Abortion	0.688	ISIS	0.555
Animal Rights	0.540	LGBT	0.540
Black Lives Matter	0.426	Minimum Wage	0.504
Blue Lives Matter	0.053	Racism	0.526
Border Security	0.465	Religion	0.547
Climate Change	0.480	Terrorism	0.544
Fracking	0.455	Unemployment	0.511
Gun Control	0.627	Vaccines	0.596
Homelessness	0.527	War On Drugs	0.545
Immigration	0.578	Welfare	0.192
Average		0.534	

Visually consistent topics We next analyze which topics had the most consistent purely visual expression i.e. without considering the text. We computed the 20 nearest neighbors in visual space for several hundred randomly chosen images from our dataset. We then computed the entropy of the topic distribution of the retrieved neighboring images and sorted the results in order of increasing entropy. We show the result in Fig. 8, with the first row showing the query image and the next three rows showing the top-3 closest images to the query in visual space. We see the left two columns all feature feature firearms. The retrieved neighbors in the first two columns are extremely consistent in their topic annotations and are almost all labeled "gun control". The third column also features military/law enforcement holding firearms, but are much more diffuse in terms of the neighbors' topics (e.g. ISIS, foreign policy, terrorism, etc.). The queries and their neighbors to the right are even more diffuse in terms of topics, e.g. the protest images (second to last column) all feature protests (or political rallies), but are about a number of disparate topics from welfare to immigration, even though they are close in visual space. Thus, predicting the political topic of an image is complex in that it requires not only recognizing the objects and

Fig. 8 Images where neighbors in visual space are most consistent in terms of their topic. We show that some topics (e.g. gun control) have a consistent visual expression, while other topics are less visually cohesive

scene type of an image (e.g. protest), but actually reasoning how the objects and individuals relate in more nuanced ways.

6 Conclusion

We assembled a large dataset of biased images and paired articles and presented a weakly supervised approach for inferring the political bias of images. Our method leverages the image's paired text to guide the model's training process towards relevant semantics in a way which ultimately improves bias classification. Our method demonstrates the potential of using an auxiliary semantic space, e.g. for abstract tasks such as video summarization and visual commonsense reasoning. We demonstrate the contribution of our method and dataset both quantitatively and qualitatively, including on a large crowdsourced dataset. We performed a detailed experimental analysis demonstrating how bias in the media is expressed both visually and textually.

Our work has several broader contributions. First, we believe studying and recognizing visual bias in images is an important step in building socially-informed machine learning systems. By recognizing how data is biased, researchers can actively work to combat biased portrayals learned by their models. Further, by automatically recognizing biased depictions, we can actively combat bias within the media. Our work can be used to expose and make users aware of discrimination and stereotypical portrayals of individuals or groups. For example, one possible solution is to automatically flag content for users so that they can become more informed that the perspective they are being presented with is non-neutral. Similarly, our work can be used to quantify not only what media sources (through the images they publish) are biased, but the types of bias that each media sources purvey. Our work has implications for social media companies which may seek to prevent the spread of discriminatory content on their platforms. By revealing bias within content presented to users, we ultimately hope to help both users and publishers become more informed consumers of visual media.

References

Akyürek, A. F., Guo, L., Elanwar, R., Ishwar, P., Betke, M., & Wijaya, D. T. (2020). Multi-label and multilingual news framing analysis. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8614–8624).

Alayrac, J. B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., & Zisserman, A. (2020). Self-supervised multimodal versatile networks. In *Neural Information Processing Systems (NeurIPS)*.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018) Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Angermeyer, M. C., & Schulze, B. (2001). Reinforcing stereotypes: How the focus on forensic cases in news reporting may influence public attitudes towards the mentally ill. *International Journal of Law and Psychiatry*.

Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., & Nakov, P. (2018). Predicting factuality of reporting and bias of news media sources. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 3528–3539).

Baumer, E., Elovic, E., Qin, Y., Polletta, F., & Gay, G. (2015) Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1472–1482).

Bechavod, Y., & Ligett, K. (2017). Penalizing unfairness in binary classification. arXiv preprint arXiv:1707.00044.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to

homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems (NIPS)*

Borghi, G., Pini, S., Grazioli, F., Vezzani, R., & Cucchiara, R. (2018) Face verification from depth using privileged information. In *British Machine Vision Conference (BMVC)*. Springer.

Burns, K., Hendricks, L.A., Darrell, T., & Rohrbach, A. (2018) Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision (ECCV)*.

Card, D., Boydstun, A., Gross, J. H., Resnik, P., & Smith, N. A. (2015). The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing* (Volume 2: Short Papers) (pp. 438–444).

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V. N. (2018) Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Winter conference on applications of computer vision (WACV)* (pp. 839–847). IEEE.

Chen, T. H., Liao, Y. H., Chuang, C. Y., Hsu, W. T., Fu, J., & Sun, M. (2017). Show, adapt and tell: Adversarial training of cross-domain image captioner. In *Proceedings of the IEEE international conference on computer vision (ICCV)*.

Chen, X., & Gupta, A. (2015). Webly supervised learning of convolutional networks. In *Proceedings of the IEEE international conference on computer vision (ICCV)* (pp. 1431–1439).

Chen, Y. C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020). Uniter: Learning universal image-text representations. In *European Conference on Computer Vision (ECCV)*.

Cinbis, R. G., Verbeek, J., & Schmid, C. (2016). Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 39*(1), 189–203.

Cohen, R., & Ruths, D. (2013). Classifying political orientation on twitter: It's not easy! In *Seventh International Association for the Advancement of Artificial Intelligence (AAAI) Conference on Weblogs and Social Media*.

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication, 64*(2), 317–332.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011). Predicting the political alignment of twitter users. In *IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and IEEE Third International Conference on Social Computing (SocialCom)* (pp. 192–199). IEEE.

Dai, B., Fidler, S., Urtasun, R., & Lin, D. (2017) Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE international conference on computer vision (ICCV)*.

Doersch, C., Singh, S., Gupta, A., Sivic, J., & Efros, A. (2012). What makes Paris look like Paris? *ACM Transactions on Graphics, 31*(4), 10.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Dong, Z., Shi, C., Sen, S., Terveen, L., & Riedl, J. (2012). War versus inspirational in forrest gump: Cultural effects in tagging communities. In *Sixth international AAAI conference on weblogs and social media*.

Edsall, T. B. (2012). Studies: Conservatives are from mars, liberals are from venus. https://www.theatlantic.com/politics/archive/2012/02/studies-conservatives-are-from-mars-liberals-are-from-venus/252416/.

Eickhoff, C. (2018). Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM international conference on web search and data mining* (pp. 162–170). ACM.

Eisenschtat, A., & Wolf, L. (2017). Linking image and text with 2-way nets. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Elliott, D., & Kádár, Á. (2017). Imagination improves multimodal translation. In *Proceedings of the eighth international joint conference on natural language processing* (Volume 1: Long Papers) (pp. 130–141).

Faghri, F., Fleet, D. J., Kiros, J. R., & Fidler, S. (2018) Vse++: Improved visual-semantic embeddings. In *British Machine Vision Conference (BMVC)*.

Garbe, W. (2019). Symspell. https://github.com/wolfgarbe/SymSpell.

Gilens, M. (1996). Race and poverty in Americapublic misperceptions and the American news media. *Public Opinion Quarterly, 60*(4), 515–541.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics (AISTATS)* (pp. 249–256).

Gomez, L., Patel, Y., Rusinol, M., Karatzas, D., Jawahar, C. V. (2017). Self-supervised learning of visual features through embedding images into text topic spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Happer, C., & Philo, G. (2013). The role of the media in the construction of public belief and social change. *Journal of Social and Political Psychology, 1*(1), 321–336.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 770–778).

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., & Li, M. (2019). Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp 558–567).

Hessel, J., Lee, L., & Mimno, D. (2018). Quantifying the visual concreteness of words and topics in multimodal datasets. In *North American Association for Computational Linguistics*.

Hoffman, J., Gupta, S., & Darrell, T. (2016). Learning with side information through modality hallucination. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 826–834). IEEE.

Honnibal, M., & Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing (to appear).

Hussain, Z., Zhang, M., Zhang, X., Ye, K., Thomas, C., Agha, Z., Ong, N., & Kovashka, A. (2017). Automatic understanding of image and video advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Jae Lee, Y., Efros, A. A., & Hebert, M. (2013). Style-aware mid-level representation for discovering visual connections in space and time. In *Proceedings of the IEEE international conference on computer vision (ICCV)* (pp. 1857–1864).

Jiang, L., Meng, D., Zhao, Q., Shan, S., & Hauptmann, A. G. (2015). Self-paced curriculum learning. In *Twenty-ninth association for the advancement of artificial intelligence (AAAI) conference on artificial intelligence* (Vol. 2, p. 6).

Jiang, L., Zhou, Z., Leung, T., Li, L. J., & Fei-Fei, L. (2018). Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the international conference on machine learning (ICML)* (pp. 2309–2318).

Jin, Z., Cao, J., Zhang, Y., Zhou, J., & Tian, Q. (2016). Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia*, *19*(3), 598–608.

Johnson, J., Karpathy, A., Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*

Joo, J., Li, W., Steen, F. F., & Zhu, S. C. (2014) Visual persuasion: Inferring communicative intents of images. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Joo, J., Steen, F. F., & Zhu, S. C. (2015). Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *Proceedings of the IEEE international conference on computer vision (ICCV)*.

Karimi, H., & Tang, J. (2019). Learning hierarchical discourse-level structure for fake news detection. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers) (pp. 3432–3442).

Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019) Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference* (pp. 2915–2921).

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research, 10,* 1755–1758.

Kingma, D. P., & Ba, J. (2015) Adam: A method for stochastic optimization. In *Proceedings of the international conference on learning representations (ICLR)*.

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2015). Unifying visual-semantic embeddings with multimodal neural language models. In: *TACL*.

Kosinski, M. (2021). Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports, 11*(1), 1–7.

Lambert, J., Sener, O., & Savarese, S. (2018). Deep learning under privileged information using heteroscedastic dropout. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the international conference on machine learning (ICML)* (pp. 1188–1196).

Li, H., Ellis, J. G., Zhang, L., & Chang, S. F. (2018). Patternnet: Visual pattern mining with deep neural network. In *Proceedings of the 2018 ACM on international conference on multimedia retrieval* (pp. 291–299). ACM.

Li, Y., Liu, L., Shen, C., & Van Den Hengel, A. (2017). Mining mid-level visual patterns with deep cnn activations. *International Journal of Computer Vision (IJCV), 121*(3), 344–364.

Liu, S., Guo, L., Mays, K., Betke, M., & Wijaya, D. T. (2019). Detecting frames in news headlines and its application to analyzing news framing trends surrounding us gun violence. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)* (pp. 504–514).

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision (ICCV)*.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in neural information processing systems (NeurIPS)* (pp. 13–23).

Lu, J., Yang, J., Batra, D., & Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. In *Advances in Neural information processing systems (NIPS)* (pp. 289–297).

Malkov, Y. A., & Yashunin, D. A. (2016) Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

Merrill, J. B. (2016). Liberal, moderate or conservative? see how facebook labels you. The New York Times https://www.nytimes.com/2016/08/24/us/politics/facebook-ads-politics.html.

Messing, S., Jabon, M., & Plaut, E. (2016). Bias in the flesh: Skin complexion and stereotype consistency in political campaigns. *Public Opinion Quarterly, 80*(1), 44–65.

Minghui Liao, B. S., & Bai, X. (2018). TextBoxes++: A single-shot oriented scene text detector. *IEEE Transactions on Image Processing, 27*(8), 3676–3690. https://doi.org/10.1109/TIP.2018.2825107.

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*.

Morin, F., & Bengio, Y. (2005). Hierarchical probabilistic neural network language model. In *Tenth international workshop on artificial intelligence and statistics (AISTATS)* (Vol. 5, pp. 246–252). Citeseer.

Motiian, S., Piccirilli, M., Adjeroh, D. A., & Doretto, G. (2016). Information bottleneck learning using privileged information for visual recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1496–1505). IEEE.

Muñoz, C. L., & Towner, T. L. (2017). The image is the message: Instagram marketing and the 2016 presidential primary season. *Journal of Political Marketing, 16*(3–4), 290–318.

Nguyen D, Trieschnigg D, Doğruöz AS, Gravel R, Theune M, Meder T, De Jong F (2014) Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In: Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers (COLING), pp 1950–1961

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data, 2,* 13.

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2015). Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 685–694).

Orr, G. B. (1997). Removing noise in on-line search using adaptive batch sizes. In *Advances in neural information processing systems* (pp. 232–238).

Otterbacher, J., Checco, A., Demartini, G., & Clough, P. (2018). Investigating user perception of gender bias in image search: The role of sexism. In *The 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 933–936). ACM.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in pytorch. In *Advances in neural information processing systems workshops (NIPS-W)*.

Peck, T., & Boutelier, N. (2018). Big political data. https://www.isidewith.com/polls.

Pedersoli, M., Lucas, T., Schmid, C., & Verbeek, J. (2017). Areas of attention for image captioning. In *Proceedings of the IEEE international conference on computer vision (ICCV)*.

Peng, Y. (2018). Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. *Journal of Communication, 68*(5), 920–941.

Pennacchiotti, M., & Popescu, A. M. (2011). A machine learning approach to twitter user classification. In *Fifth international association for the advancement of artificial intelligence (AAAI) conference on weblogs and social media*.

Pentina, A., Sharmanska, V., & Lampert, C. H. (2015). Curriculum learning of multiple tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 5492–5500).

Peters, M. E., & Lecocq, D. (2013). Content extraction using diverse feature sets. In *Proceedings of the 22nd international conference on world wide web (WWW)* (pp. 89–90). ACM.

Philo, G. (2008). Active audiences and the construction of public knowledge. *Journalism Studies, 9*(4), 535–544.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (Volume 1: Long Papers) (pp. 231–240).

Recasens, M., Danescu-Niculescu-Mizil, C., Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (Volume 1: Long Papers) (Vol. 1, pp. 1650–1659).

Řehůřek, R., Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*, ELRA, Valletta, Malta (pp. 45–50). http://is.muni.cz/publication/884893/en.

Richard, A., Kuehne, H., & Gall, J. (2017). Weakly supervised action learning with RNN based fine-to-coarse modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 754–763).

Ryu, H. J., Mitchell, M., & Adam, H. (2017). Improving smiling detection with race and gender diversity. arXiv preprint arXiv:1712.00193.

Schill, D. (2012). The visual image and the political image: A review of visual communication research in the field of political communication. *Review of Communication, 12*(2), 118–142.

Schreiber, D., Fonzo, G., Simmons, A. N., Dawes, C. T., Flagan, T., Fowler, J. H., & Paulus, M. P. (2013). Red brain, blue brain: Evaluative processes differ in democrats and republicans. *PLOS ONE, 8*(2), 1–6. https://doi.org/10.1371/journal.pone.0052970.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 815–823).

Sen, S., Giesel, M. E., Gold, R., Hillmann, B., Lesicko, M., Naden, S., Russell, J., Wang, Z. K., & Hecht, B. (2015). Turkers, scholars, arafat and peace: Cultural communities and algorithmic gold standards. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing* (pp. 826–838). ACM.

Sharmanska, V., Quadrianto, N., & Lampert, C. H. (2013). Learning to rank using privileged information. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 825–832). IEEE.

Sicre, R., Avrithis, Y. S., Kijak, E., & Jurie, F. (2017). Unsupervised part learning for visual recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3116–3124).

Singh, S., Gupta, A., & Efros, A. A. (2012). Unsupervised discovery of mid-level discriminative patches. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 73–86). Springer.

Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 5103–5114).

Thomas, C., & Kovashka, A. (2018). Persuasive faces: Generating faces in advertisements. In *Proceedings of the British machine vision conference (BMVC)*.

Thomas, C., & Kovashka, A. (2019). Predicting the politics of an image using webly supervised data. In *Advances in neural information processing systems (NeurIPS)* (pp. 3625–3637).

Thomas, C., & Kovashka, A. (2020). Preserving semantic neighborhoods for robust cross-modal retrieval. In *European Conference on Computer Vision (ECCV)* (pp. 317–335). Springer.

Vapnik, V., & Izmailov, R. (2015). Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research (JMLR), 16*(2023–2049), 2.

Venugopalan, S., Anne Hendricks, L., Rohrbach, M., Mooney, R., Darrell, T., & Saenko, K. (2017). Captioning images with diverse objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3156–3164).

Vogel, K., Shane, S., & Kingsley, P. (2018). How vilification of george soros moved from the fringes to the mainstream. https://www.nytimes.com/2018/10/31/us/politics/george-soros-bombs-trump.html. Accessed January 15, 2020.

Volkova, S., Coppersmith, G., & Van Durme, B. (2014). Inferring user political preferences from streaming communications. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (Volume 1: Long Papers) (Vol. 1, pp. 186–196).

Wang, L., Xiong, Y., Lin, D., & Van Gool, L. (2017a). Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 4325–4334).

Wang, Y., Feng, Y., Hong, Z., Berger, R., & Luo, J. (2017b). How polarized have we become? a multimodal classification of trump followers and clinton followers. In *International conference on social informatics*.

Wang, Y., Li, Y., & Luo, J. (2016). Deciphering the 2016 us presidential campaign in the twitter sphere: A comparison of the trumpists and clintonists. In *Tenth international association for the advancement of artificial intelligence (AAAI) conference on web and social media* (pp. 723–726).

Wei, Y., Shen, Z., Cheng, B., Shi, H., Xiong, J., Feng, J., & Huang, T. (2018). Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 434–450).

Wong, F. M. F., Tan, C. W., Sen, S., & Chiang, M. (2016). Quantifying political leaning from tweets, retweets, and retweeters. *IEEE Transactions on Knowledge and Data Engineering, 28*(8), 2158–2172.

Xi, N., Ma, D., Liou, M., Steinert-Threlkeld, Z. C., Anastasopoulos, J., & Joo, J. (2020). Understanding the political ideology of legislators from social media images. *Proceedings of the International AAAI Conference on Web and Social Media, 14,* 726–737.

Xiong, C., Zhong, V., & Socher, R. (2017). Dynamic coattention networks for question answering. In *Proceedings of the international conference on learning representations (ICLR)*.

Ye, K., Honarvar Nazari, N., Hahn, J., Hussain, Z., Zhang, M., & Kovashka, A. (2019). Interpreting the rhetoric of visual advertisements. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. https://doi.org/10.1109/TPAMI.2019.2947440.

Ye, K., & Kovashka, A. (2018). Advise: Symbolism and external knowledge for decoding advertisements. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ye, K., Zhang, M., Kovashka, A., Li, W., Qin, D., Berent, J. (2019). Cap2det: Learning to amplify weak caption supervision for object detection. In *Proceedings of the IEEE international conference on computer vision (ICCV)*.

Yoon, J., Joo, J., Park, E., & Han, J. (2020). Cross-domain classification of facial appearance of leaders. In *International conference on social informatics* (pp. 440–446). Springer.

Zamir, A. R., Wu, T. L., Sun, L., Shen, W. B., Shi, B. E., Malik, J., & Savarese, S. (2017). Feedback networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1808–1817). IEEE.

Zhang, J., Wu, Q., Zhang, J., Shen, C., & Lu, J. (2019) Mind your neighbours: Image annotation with metadata neighbourhood graph co-attention networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2956–2964).

Zhang, Y., David, P., & Gong, B. (2017). Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 2020–2030).

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2921–2929).

Zhou, F., De la Torre, F., & Cohn, J. F. (2010). Unsupervised discovery of facial events. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2574–2581). IEEE.