Predicting the Politics of an Image Using Webly Supervised Data

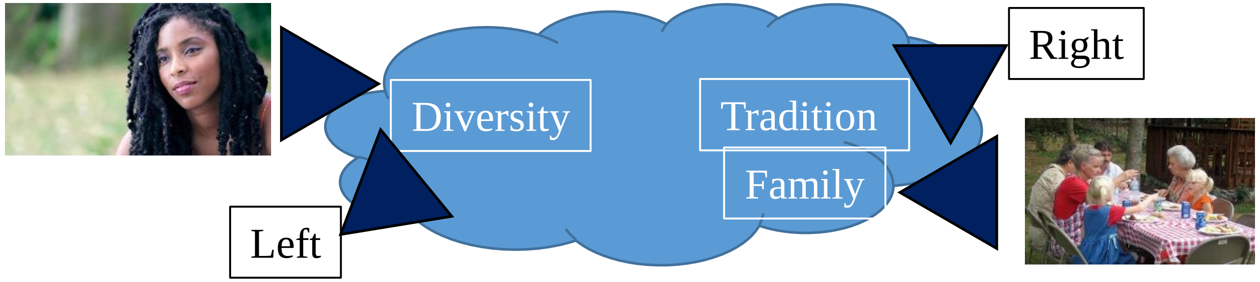
Christopher Thomas and Adriana Kovashka

Published in NeurIPS 2019

OUTLINE

- **Problem introduction**
- Related research
- Dataset
- Our method
- Quantitative results
- Qualitative results

PREDICTING VISUAL POLITICAL BIAS



- We study predicting the **political leaning of an image**
- Certain political sides are associated with certain demographic groups, concepts, people, etc.
  - We want to see whether we can learn this automatically from the data
- Multimodal setting: images + paired ***lengthy*** text articles they appeared with
  - We are interested primarily in ***visual*** bias, not textual

EXAMPLE IMAGES

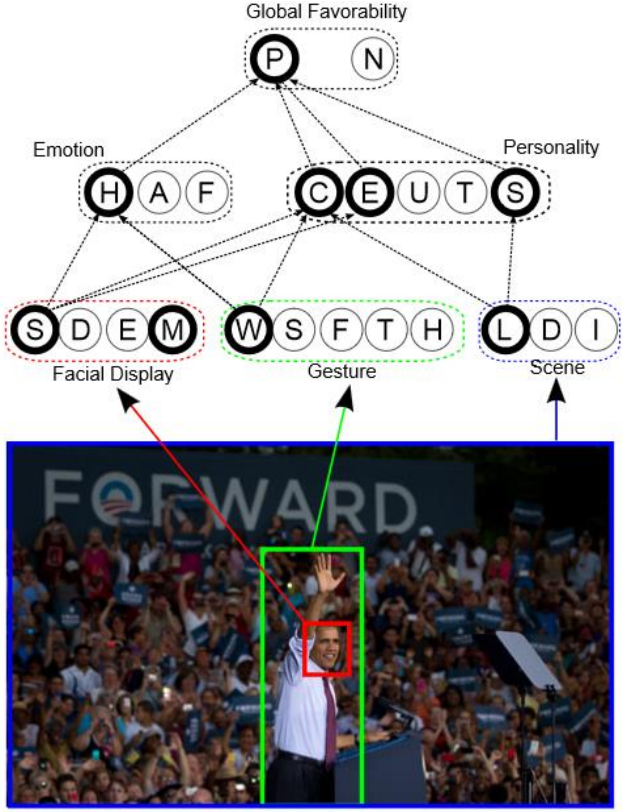
?

OUTLINE

- Problem introduction
- **Related research**
- Dataset
- Our method
- Quantitative results
- Qualitative results

RELATED RESEARCH – VISUAL PERSUASION

- Visual Persuasion: Inferring Communicative Intents of Images
- Uses facial attributes of known politicians to predict whether the image portrays them in a positive or negative light
- We compare against Joo et al. as a baseline
- In contrast, we don't use human chosen attributes / features; instead we leverage the implicit semantics in the auxiliary text domain to guide training



Modeling Persuasive Intents
Joo et al., 2014

Joo, Jungseock, et al. "Visual persuasion: Inferring communicative intents of images." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

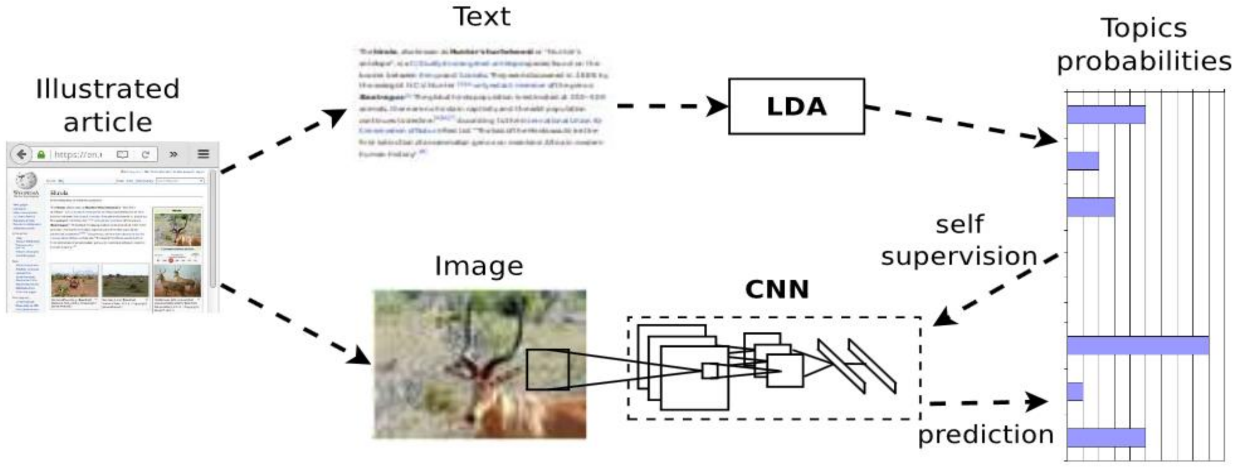
RELATED RESEARCH – POLITICAL FACES

- Same Candidates, Different Faces: Uncovering Media Bias in Visual Portrayals of Presidential Candidates with Computer Vision
- Looked at 13,026 images from 15 news websites about Clinton / Trump during 2016 election
- Looked at visual attribute differences (e.g., facial expressions, face size, skin condition) between the two candidates
- Used crowdsourced workers to rate a subset of 1,200 images and demonstrated that some visual features also effectively shape viewers' perceptions of media slant and impressions of the candidates
  - **We obtain similar results, but we *generate* faces**
- A big difference between this and our work is we consider images beyond known politicians (we also model these differences generatively)

Peng, Yilang. "Same Candidates, Different Faces: Uncovering Media Bias in Visual Portrayals of Presidential Candidates with Computer Vision." *Journal of Communication* 68.5 (2018): 920-941.

Related Work – Priviledged Information

- Self-supervised learning of visual features through embedding images into text topic spaces
- Uses semantic representation in paired text domain to guide training



- Trains CNN to *predict* latent topics from text, then uses the features from the image model to perform classification
- Our dataset / problem is more challenging because of the **many-to-many** relationship with images to topics (image of White House can be paired with text about immigrants, Trump, Obama, military policy, etc.)
    - Thus, directly predicting text embeddings from image doesn't work as well

Gomez, Lluis, et al. "Self-supervised learning of visual features through embedding images into text topic spaces." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

OUTLINE

- Problem introduction
- Related research
- **Dataset**
- Our method
- Quantitative results
- Qualitative results

DATASET COLLECTION

- Used an online resource of biased news sources (from left / right) and politicially contentious issues
    - **20 issues:** Abortion, Black Lives Matter, LGBT, Welfare, etc.

- ***Automatically*** spidered these sites to find pages with images on them and associated text containing the query phrases

- Extracted **images** and **raw text articles** from the sources
    - Used Dragnet text extraction tool which automatically parses HTML for main article text
    - Process is ***noisy***

- Around 1.8M images / articles total

- Dataset is ***highly diverse*** and also ***noisy***

DATA CLEANUP

- Many news sources report on the same visual content – thus many articles feature the same image

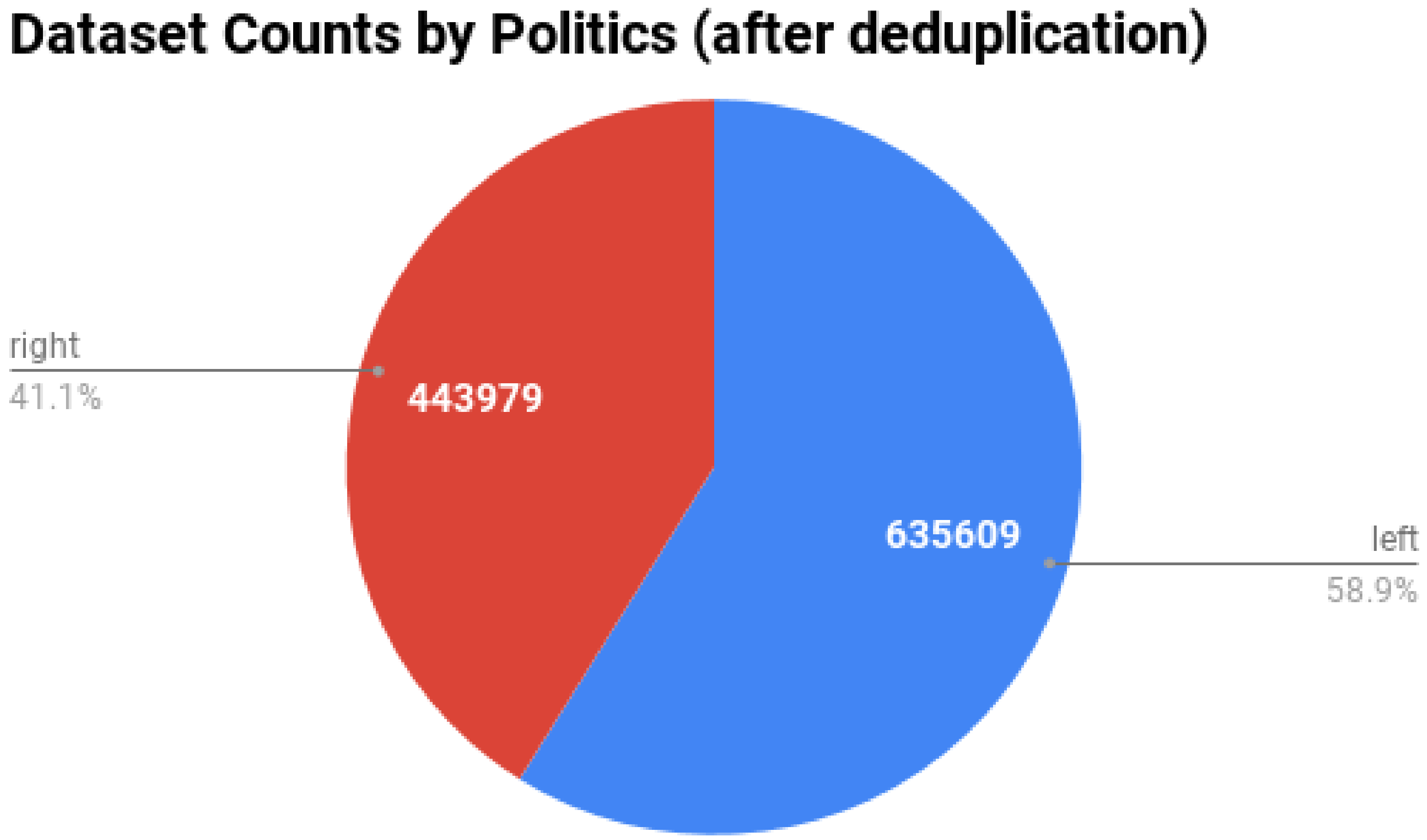
- We extract CNN features for every image in the dataset then we perform approximate KNN search using an off-the-shelf method

- This enables us to find near and exact matches of images

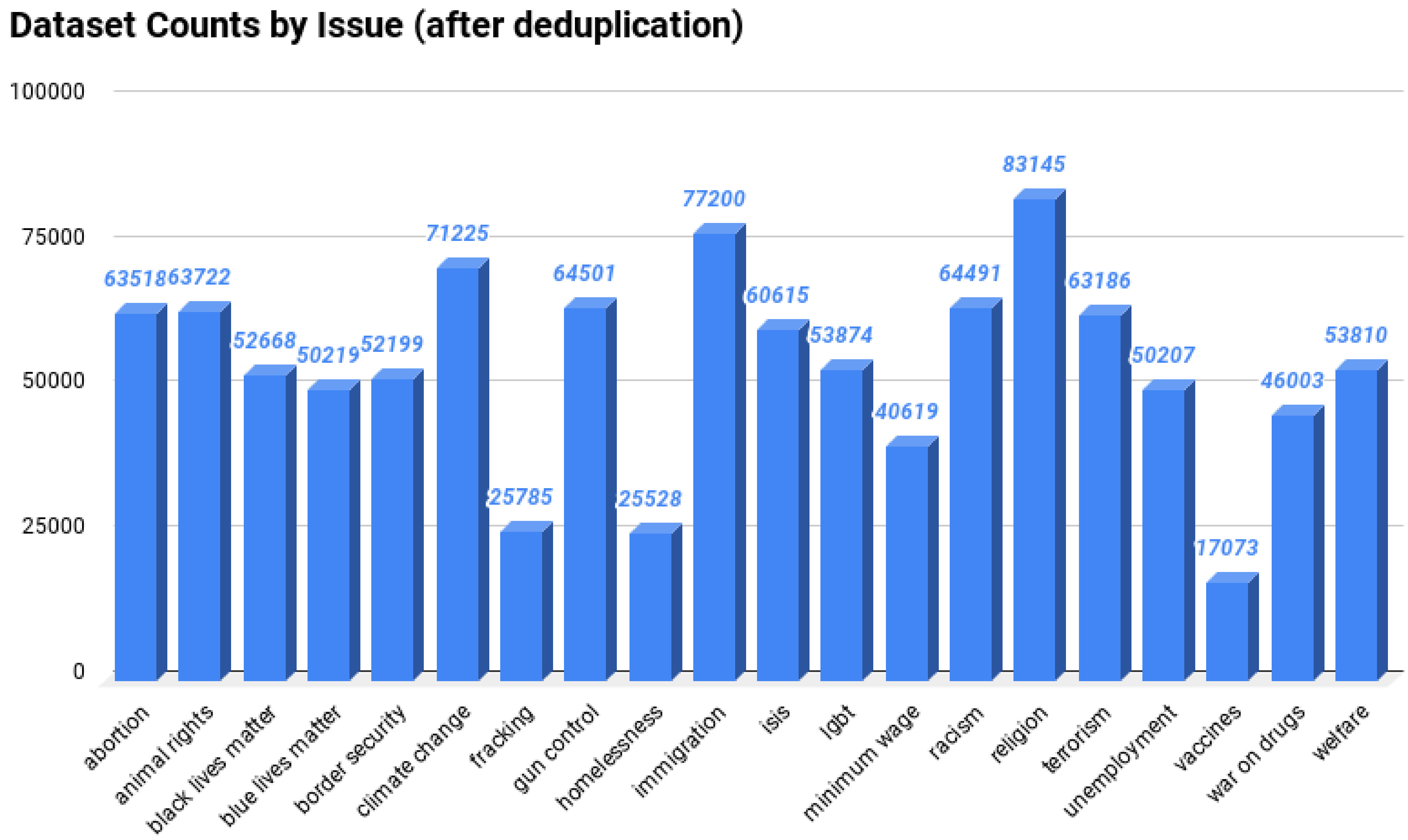
- To form our final dataset, find the side which is most common in the duplicate set and keep one of the instances
  - E.g. 5 times from left, 8 times from right, keep one of the instances from the right and discard all the other instances and their articles

- After cleanup >1M ***unique*** images and paired articles

DATASET DETAILS – BREAKDOWN BY POLITICS



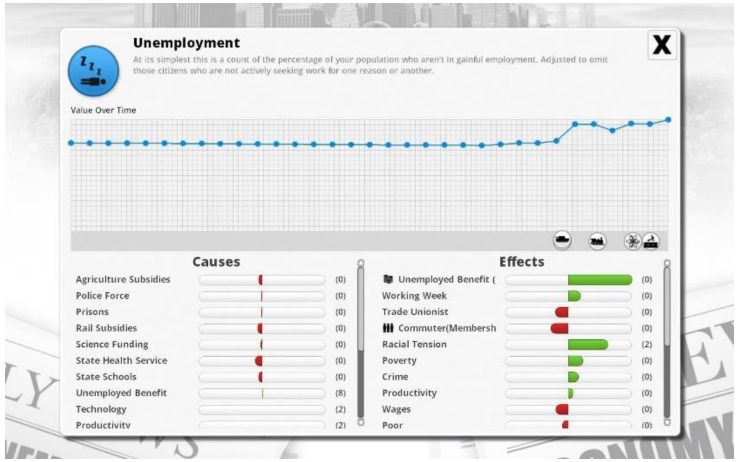
DATASET DETAILS – BREAKDOWN BY ISSUE



DATASET CHALLENGES

- Noise in dataset comes from **automatic harvesting**
  - We assume that any images harvested from a left/right site are of that political label, but they actually may be unbiased or have the reverse bias

- Challenges include:
  - Images may be unrelated to query (i.e. unrelated content on page, ads, etc.)
  - Text may fail to parse correctly or contain headers or other noise
  - Lots of noisy images – text, crops of web pages, clipart illustrations, etc.
  - Images that just aren’t politically biased



CROWDSOURCING

- We ran a large-scale crowdsourcing study on Mturk asking workers to guess the political leaning of images
- We showed 3,237 images to at least three workers each
- 993 images were labeled clearly L/R by at least a majority
- We also asked what **image features** workers used to guess
  - E.g. closeup of face, portrays a public figure, a group or class of people is portrayed in a political way, contained symbols (e.g. swastika), etc.
- We also showed workers the article and asked questions about the *pair*
  - What article text is best aligned with the image
  - Topic of the image and article
  - Finally we asked workers to explain their predictions for a small number
- We manually went through the responses and mined concepts used by humans
  - **Recognized people** and used their knowledge + image's portrayal
  - Used **stereotypical concepts** to guess (e.g. African American = Left)
- Queried Google Images for these concepts and trained an image classifier to detect Mturk stereotypical concepts (used as Human Concepts baseline)

Republican president Guess: R	A heroic memeified photo of Obama comes form liberals Guess: L	Liberal stance: Anti-discrimination for Hispanics Guess: L	This picture is showing trump supporters at a rally. Guess: R
positive picture of Trump Guess: R	A positive picture of Obama Guess: L	Supporting a liberal policy Guess: L	Gun rights supporter are generally right leaning. Guess: R
trump smiling Guess: R	PIC OF OBAMA, LIBERAL PRESIDENT Guess: L	Pro immigration Guess: L	Second Amendment shirt would lean right. Guess: R

THE LEFT LOVES TO PROTEST. Guess: L	Looks like a man cross dressing so that would only be supported by a left winger Guess: L	many black women are more liberal than conservative Guess: L	the image involves voters and the Republicans are very concerned about the threat of voter fraud Guess: R
they like protesting a lot Guess: L	Weirdness embraced Guess: L	Most african american women lean left Guess: L	i chose right because it looks like a voting booth Guess: R
Looks like a leftist political rally Guess: L	Looks like a gay person Guess: L	**Guessed incorrectly**	**Guessed incorrectly**

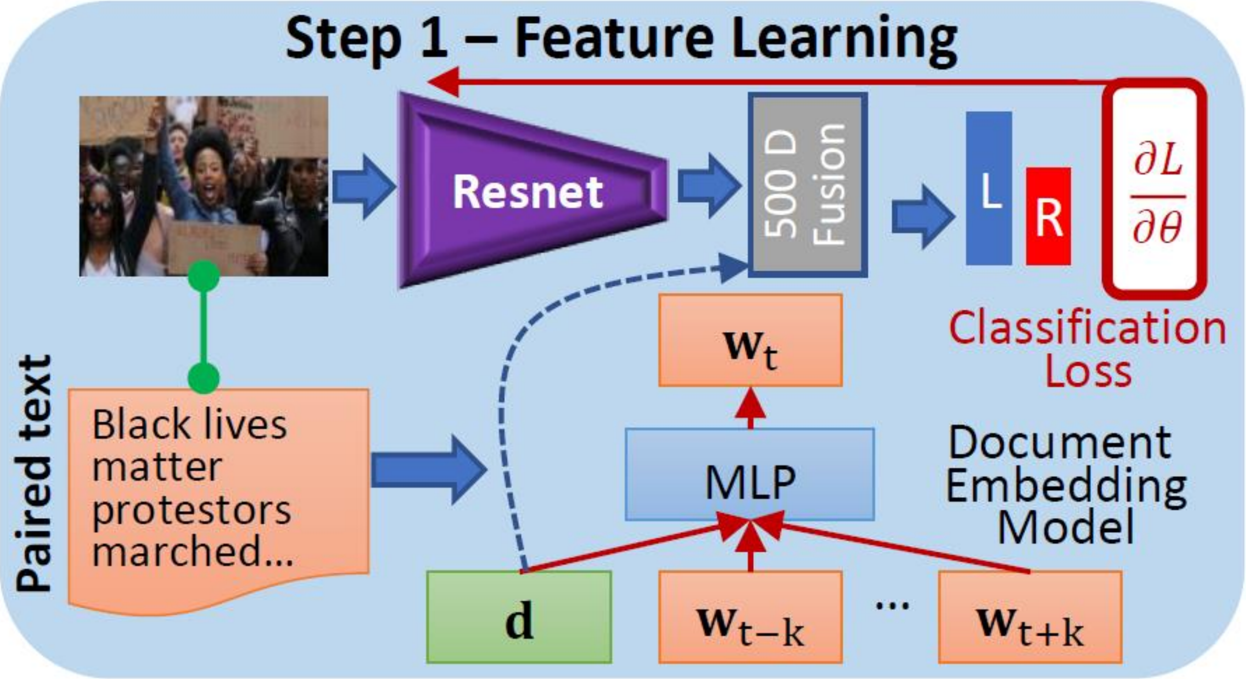
CROWDSOURCING CONSENSUS VS NO CONSENSUS

Examples of images where all workers agree, the majority agree, and for which there was no consensus on the left / right leaning

OUTLINE

- Problem introduction
- Related research
- Dataset
- **Our method**
- Quantitative results
- Qualitative results

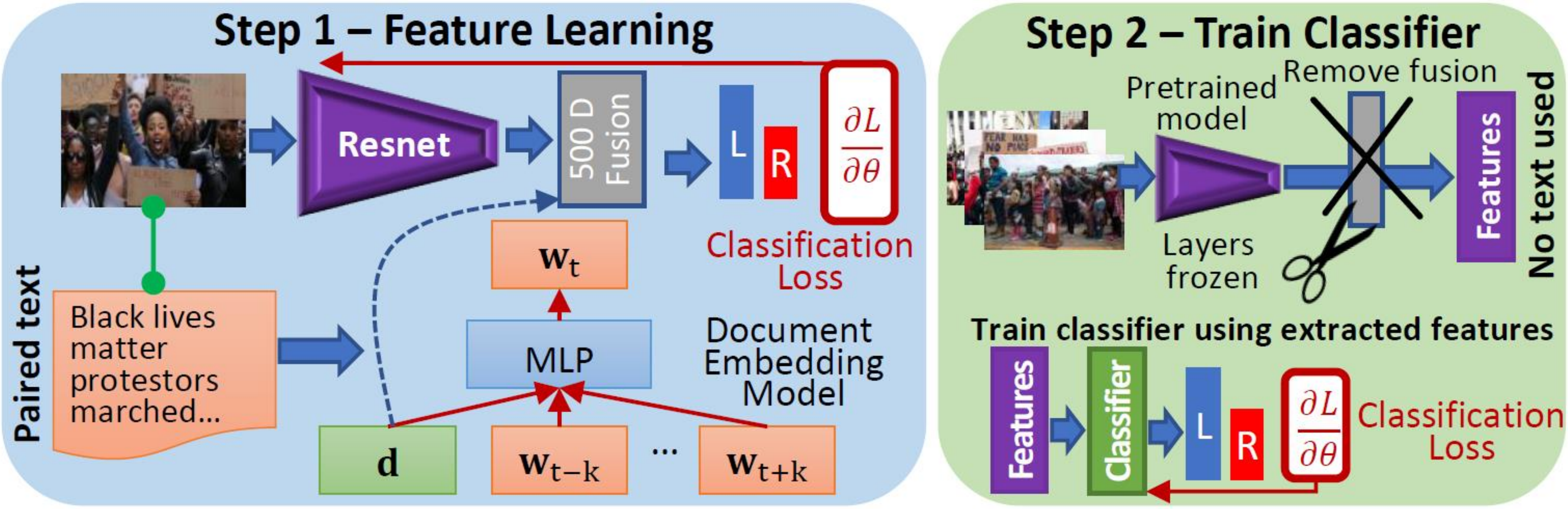
MODEL ARCHITECTURE



- Document embeddings from paired article text act as a source of **privileged information** to help guide training
- Article text is **not** used at test time

- We propose a two-stage approach
- In the first stage, we learn a **document embedding** model from the paired articles
- We then train a Resnet which takes in an image and the document embedding and predicts whether the image-text pair is left/right

MODEL ARCHITECTURE



- In stage two, we **remove the model's dependency on text**
- We remove the multi-modal fusion layer and train a classifier using the features from the CNN trained in stage 1, while freezing the CNN layers
- Our model thus uses **no text at test time**

OUTLINE

- Problem introduction
- Related research
- Dataset
- Our method
- **Quantitative results**
- Qualitative results

Experimental Results – Weakly Supervised

Method	Resnet	Joo	Human Concepts	OCR	Ours	Ours (GT)
Accuracy	0.678	0.670	0.675	0.686	**0.712**	0.803

- Accuracy of predicting Left / Right labels on weakly supervised test set
  - Weakly supervised labels are left / right label of the media source the image came from
- Baselines:
  - **Resnet** – An off-the-shelf 50 layer residual network
  - **Joo et al.** – Uses features presented by Joo et al. for predicting visual persuasion + resnet
  - **Human Concepts** – Features of model trained to predict concepts that MTurkers used
  - **OCR** – Resnet + Optical Character Recognition (uses trained word embeddings of detected words)
- ***Ours (GT)*** *uses text at test time and is thus not purely a visual prediction*
- **Using text domain to guide training of purely visual model improves performance**

EXPERIMENTAL RESULTS – HUMAN LABELS

Feature/Method	ResNet	Joo	Human Concepts	OCR	Ours	Ours (GT)
Closeup	0.567	0.544	0.622	0.578	**0.656**	0.578
Known Person	0.567	0.550	**0.570**	0.560	0.521	0.575
Multiple People	0.722	0.671	0.688	0.730	**0.768**	0.705
No People	0.556	**0.605**	0.494	0.580	0.593	0.667
Symbols	0.558	0.596	0.548	0.577	**0.606**	0.587
Non-Photographic	0.577	0.569	0.584	0.577	**0.585**	0.654
Logos	0.545	0.584	0.597	**0.662**	0.623	0.584
Text in Image	0.629	0.625	0.596	**0.637**	0.607	0.659
Average	0.590	0.593	0.587	0.613	**0.620**	0.626

- We also eval. on human labeled data
  - Images that at least a majority of annotators agreed upon

EXPERIMENTAL RESULTS – HUMAN LABELS

Feature/Method	RESNET	JOO	HUMAN CONCEPTS	OCR	OURS	OURS (GT)
Closeup	0.567	0.544	0.622	0.578	**0.656**	0.578
Known Person	0.567	0.550	**0.570**	0.560	0.521	0.575
Multiple People	0.722	0.671	0.688	0.730	**0.768**	0.705
No People	0.556	**0.605**	0.494	0.580	0.593	0.667
Symbols	0.558	0.596	0.548	0.577	**0.606**	0.587
Non-Photographic	0.577	0.569	0.584	0.577	**0.585**	0.654
Logos	0.545	0.584	0.597	**0.662**	0.623	0.584
Text in Image	0.629	0.625	0.596	**0.637**	0.607	0.659
Average	0.590	0.593	0.587	0.613	**0.620**	0.626

- **Results are sensible**
- **Human Concepts** – Works best on celebrities, politicians, etc.

EXPERIMENTAL RESULTS – HUMAN LABELS

Feature/Method	RESNET	JOO	HUMAN CONCEPTS	OCR	OURS	OURS (GT)
Closeup	0.567	0.544	0.622	0.578	**0.656**	0.578
Known Person	0.567	0.550	**0.570**	0.560	0.521	0.575
Multiple People	0.722	0.671	0.688	0.730	**0.768**	0.705
No People	0.556	**0.605**	0.494	0.580	0.593	0.667
Symbols	0.558	0.596	0.548	0.577	**0.606**	0.587
Non-Photographic	0.577	0.569	0.584	0.577	**0.585**	0.654
Logos	0.545	0.584	0.597	**0.662**	0.623	0.584
Text in Image	0.629	0.625	0.596	**0.637**	0.607	0.659
Average	0.590	0.593	0.587	0.613	**0.620**	0.626

- **Results are sensible**
- **OCR** – Works best on images containing text in the image

EXPERIMENTAL RESULTS – HUMAN LABELS

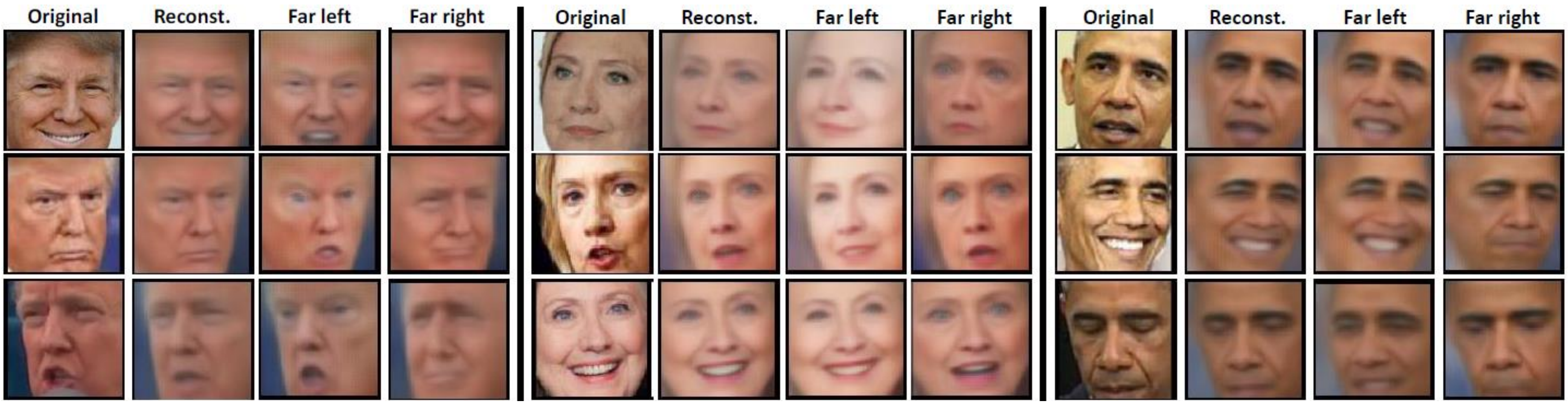
Feature/Method	ResNet	Joo	Human Concepts	OCR	Ours	Ours (GT)
Closeup	0.567	0.544	0.622	0.578	**0.656**	0.578
Known Person	0.567	0.550	**0.570**	0.560	0.521	0.575
Multiple People	0.722	0.671	0.688	0.730	**0.768**	0.705
No People	0.556	**0.605**	0.494	0.580	0.593	0.667
Symbols	0.558	0.596	0.548	0.577	**0.606**	0.587
Non-Photographic	0.577	0.569	0.584	0.577	**0.585**	0.654
Logos	0.545	0.584	0.597	**0.662**	0.623	0.584
Text in Image	0.629	0.625	0.596	**0.637**	0.607	0.659
Average	0.590	0.593	0.587	0.613	**0.620**	0.626

- **Results are sensible**
- **Ours** – Works best on more categories than others and **works best overall**

OUTLINE

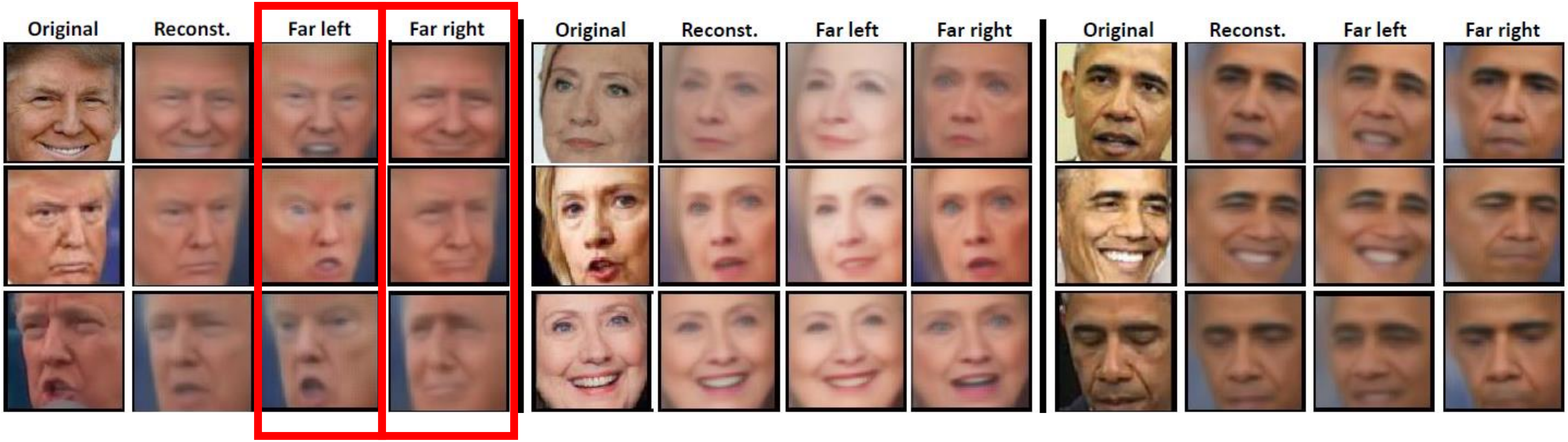
- Problem introduction
- Related research
- Dataset
- Our method
- Quantitative results
- **Qualitative results**

QUALITATIVE RESULTS



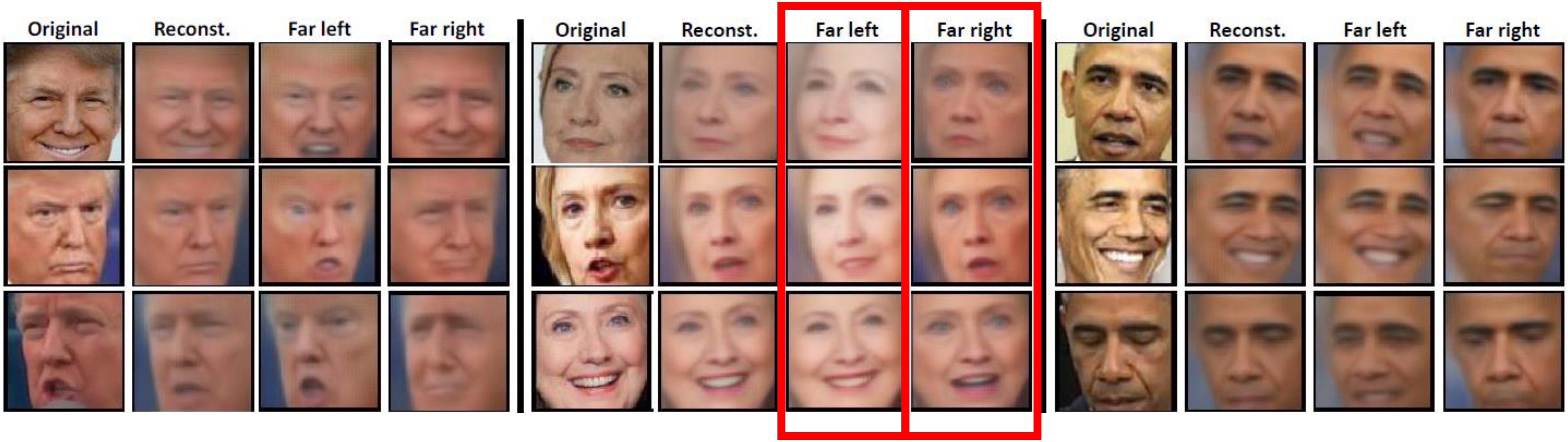
- Trained generative autoencoder on known politicians faces, conditioned on facial semantic attributes / expressions, as well as latent face embedding from autoencoder
- Modify images to be more Left / Right leaning (move embedding towards avg. L/R embedding)
- Trump – Happier on right, angrier/meaner Left
- Hillary – Younger, brighter skin on left, yelling, older on right

QUALITATIVE RESULTS



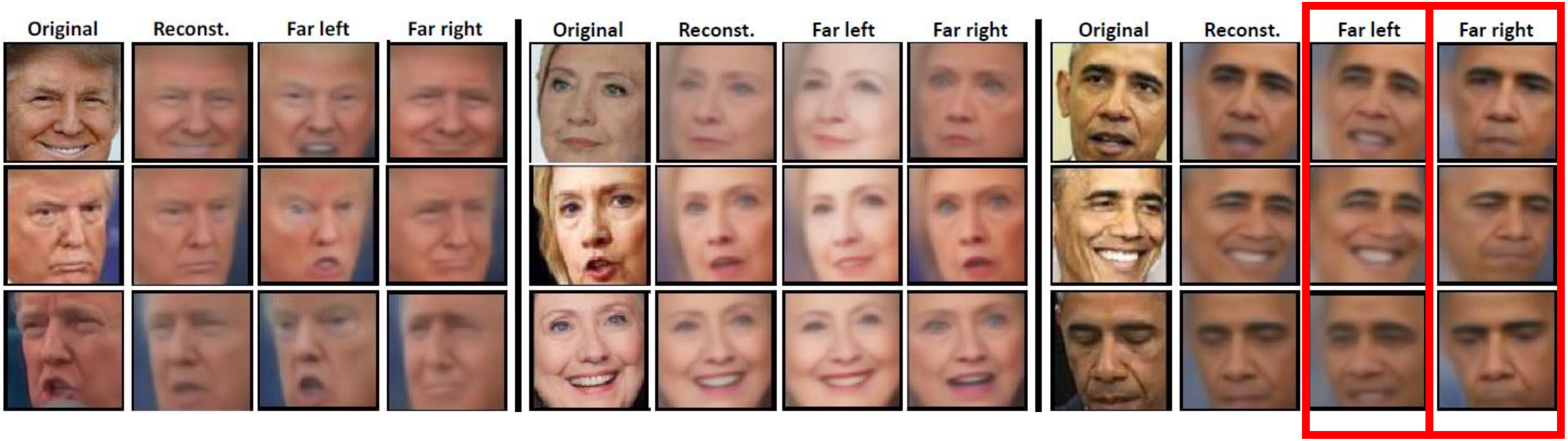
- Trained generative autoencoder on known politicians faces, conditioned on facial semantic attributes / expressions, as well as latent face embedding from autoencoder
- Modify images to be more Left / Right leaning (move embedding towards avg. L/R embedding)
- Trump – Happier on right, angrier/meaner Left
- Hillary – Younger, brighter skin on left, yelling, older on right

QUALITATIVE RESULTS



- Trained generative autoencoder on known politicians faces, conditioned on facial semantic attributes / expressions, as well as latent face embedding from autoencoder
- Modify images to be more Left / Right leaning (move embedding towards avg. L/R embedding)
- Trump – Happier on right, angrier/meaner Left
- Hillary – Younger, brighter skin on left, yelling, older on right

QUALITATIVE RESULTS



- Trained generative autoencoder on known politicians faces, conditioned on facial semantic attributes / expressions, as well as latent face embedding from autoencoder
- Modify images to be more Left / Right leaning (move embedding towards avg. L/R embedding)
- Trump – Happier on right, angrier/meaner Left
- Hillary – Younger, brighter skin on left, yelling, older on right

CLOSEST IMAGES ACROSS L/R BY TOPICS

- We show closest pair of images across the left/right divide
- Note how similar the images in each pair are on the surface, illustrating the challenge of visual bias prediction

WHAT'S IN THE LATENT TEXT SPACE [DOC2VEC]

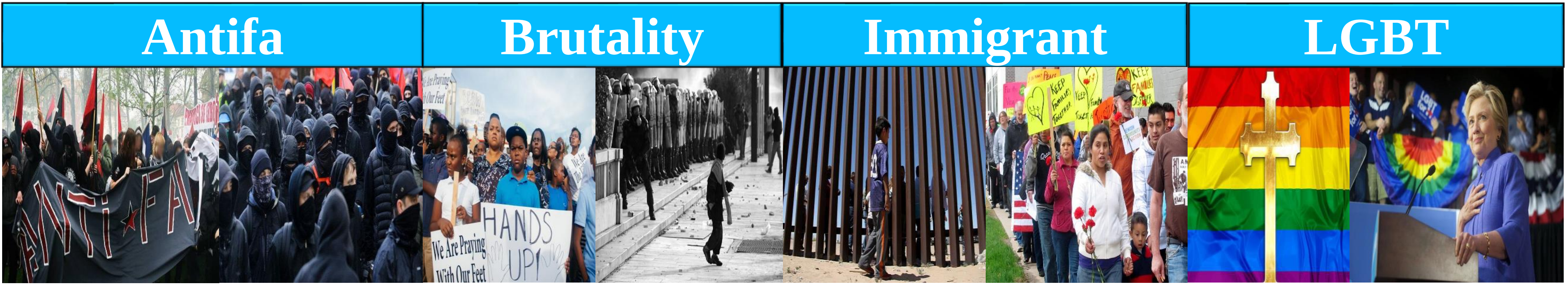
Query:

charlottesville
charleston: 0.7303
parkland: 0.7189
antifa: 0.7135
kkk: 0.7117
ferguson: 0.7038
dallas: 0.6998
confederate: 0.6995
richmond: 0.6956
shooting: 0.6879
horrific: 0.6844
portland: 0.6828
riots: 0.6826
cleveland: 0.6817
heyer: 0.6806
protest: 0.6782
rally: 0.6779

Results

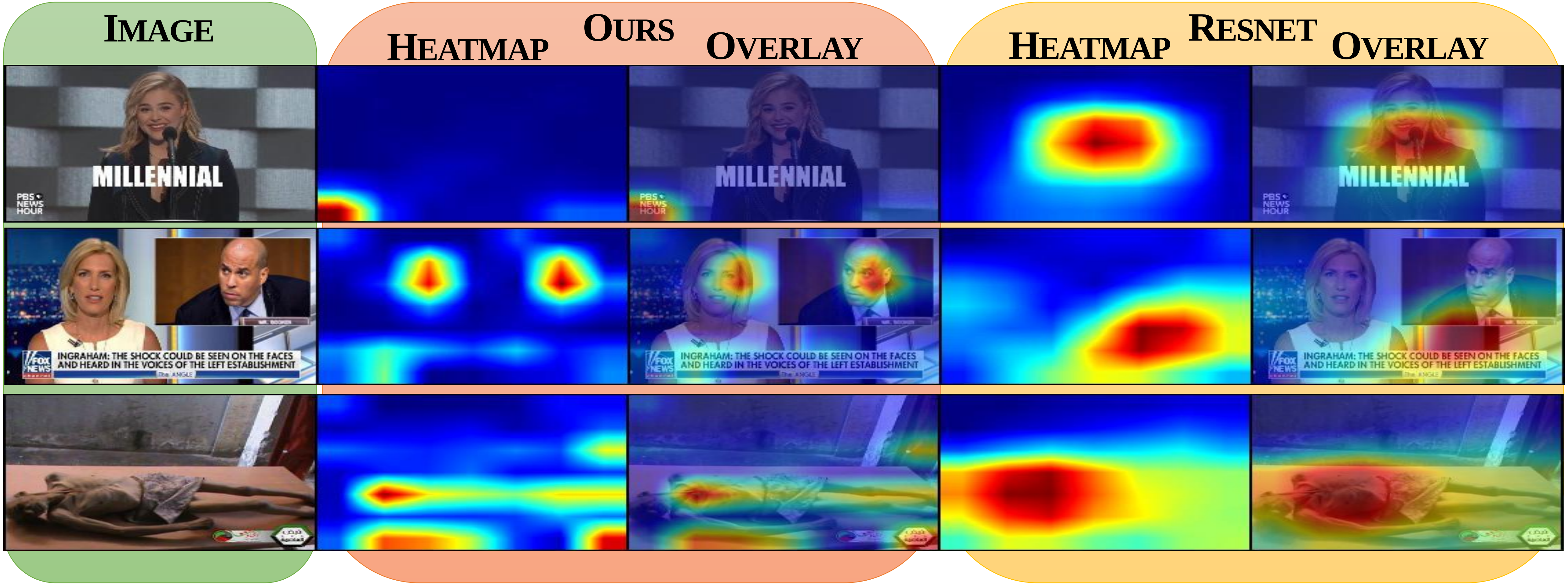
parkland
newtown: 0.7640
hogg: 0.7635
stoneman: 0.7501
nra: 0.7455
charlottesville: 0.7189
shooting: 0.7161
walkout: 0.7135
walkouts: 0.7029
charleston: 0.7002
tragedy: 0.6991
orlando: 0.6986
emma4change: 0.6931
msd: 0.6844
sandyhook: 0.6841
shootings: 0.6795
gun: 0.6752

PREDICTING WORDS FROM IMAGES



- Train a model to **predict individual words from images** given the image and the document embedding
- The model learns **visual cues for each word**, demonstrating the utility of exploiting text, even for purely visual classification
- Black clad protestors → "antifa", Protestors, police → "Brutality", Border wall / Hispanics → "Immigrant", Pride flags → "LGBT"

VISUAL EXPLANATIONS

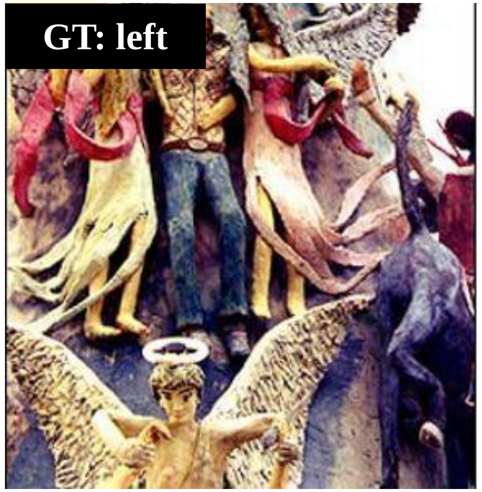


- Our model primarily pays attention to **faces and logos**. The model ignores the face of the person in the first row, but pays attention to the face of the commentator in the second row.
- The model incorrectly predicts the image in the third row; likely because of the logo confuses the model because it likely did not appear in train set and is uncommon

HUMAN VS. MACHINE ABILITY

HUMAN GUESSED, MACHINE FAILED

HUMAN FAILED, MACHINE GUESSED



BOTH FAILED

We show images that humans and/or our model were able/unable to classify. We note the top left image has a subtle country vibe, while the other two images require familiarity with a non-Western church and Emma Thompson to understand, which our classifier misses. On the bottom left, we see our classifier predicts protests, celebrities, and art as left-leaning. Finally, we show a challenging image that fooled both humans and machine.

CONCLUSION

- We collected and release a large dataset of biased images and paired article text
- We performed a large-scale human study and collected annotations on our dataset and studied human intuitions surrounding visual political bias
- We presented an approach for predicting the bias of images
  - Uses auxiliary text domain as a source of **privileged information** to guide training
- We showed both quantitative and qualitative experiments demonstrating our method works
- Use cases of our method include automatically inferring bias of media sources or detecting political ads
- Future work may include improved models of image-text alignment, methods for learning joint image-text embedings under noise, and generating biased images