Intentonomy: a Dataset and Study towards Human Intent Understanding

Menglin Jia[1,2] Zuxuan Wu[2,3] Austin Reiter[2] Claire Cardie[1] Serge Belongie[1] Ser-Nam Lim[2]
[1]Cornell University [2]Facebook AI [3]Fudan University

Abstract

An image is worth a thousand words, conveying information that goes beyond the mere visual content therein. In this paper, we study the intent behind social media images with an aim to analyze how visual information can facilitate recognition of human intent. Towards this goal, we introduce an intent dataset, Intentonomy*, comprising 14K images covering a wide range of everyday scenes. These images are manually annotated with 28 intent categories derived from a social psychology taxonomy. We then systematically study whether, and to what extent, commonly used visual information,* i.e.*, object and context, contribute to human motive understanding. Based on our findings, we conduct further study to quantify the effect of attending to object and context classes as well as textual information in the form of hashtags when training an intent classifier. Our results quantitatively and qualitatively shed light on how visual and textual information can produce observable effects when predicting intent* [1]

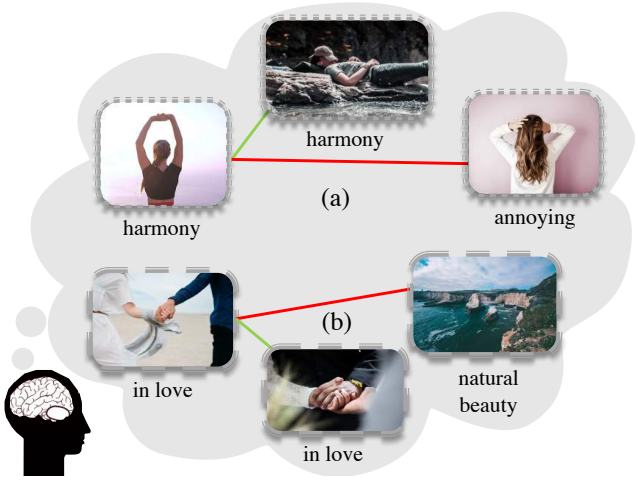


Figure 1. Intent behind images: while (b) shows that the visual motif of holding hands aligns with the common intent of "in love", (a) illustrates that similarity based on visual appearance alone often would lead to an incorrect match with respect to intent.

1. Introduction

Why do we post images on social media platforms like Facebook or Instagram? Are we expressing our feelings to friends and family? Are we seeking to entertain a wide audience? Or is it purely out of habit, or perhaps out of fear of missing out? Images on social media embody more than their explicit visual information, and they tend to be persuasive in commercial ads and even manipulative in the context of political campaigns. Therefore, in the deluge of social media, understanding the intent behind images is critical, especially for tasks like fighting fake news and misinformation [16, 32] on social platforms.

However, understanding human intent behind images from a computer vision point of view is particularly challenging, since it goes beyond standard visual recognition—predicting a set of stuff and thing categories that physically exist in images such as objects [25, 12, 56, 51, 18] and scenes [34, 58, 67]. Additionally, it is a psychological task [41] inherent to human cognition and behavior. It is similar in spirit to visual commonsense reasoning [62, 43] to derive an answer conditioned on the objects and scenes present in images. In certain cases, intent can be inferred rather directly from representative objects and scenes. For example, a couple holding hands or making a heart symbol clearly have the same motive "in love"(Fig. 1(b)). However, the mapping from visual cue to intent is not always one-to-one. Fig. 1(a) shows that two images with completely different contents (a girl facing the ocean *vs*. a person relaxing on a rocky surface, with face covered) can represent the same intent ("harmony"). This goes beyond the usual variability (pose, color, illumination, and other nuisances) traditionally addressed in object recognition pipelines [11, 37]. This brings us to the question: *are objects and their image context sufficient for recognizing the intent behind images*?

In this paper, we introduce a human intent dataset, *Intentonomy*, containing $14K$ images that are manually annotated with 28 intent categories, organized in a hierarchy by psychology experts. To investigate the intangible and subtle connection between visual content and intent, we present a systematic study to evaluate how the perfor-

[1]Intentonomy project page: github.com/kmnp/intentonomy

mance of intent recognition changes as a a function of (a) the amount of object/context information; (b) the properties of object/context, including geometry, resolution and texture. Our study suggests that (1) different intent categories rely on different sets of objects and scenes for recognition; (2) however, for some classes that we observed to have large intra-class variations, visual content provides negligible boost to the performance. Furthermore, our study also reveals that attending to relevant object and scene classes brings beneficial effects for recognizing intent.

In light of this, we further study a multimodal framework for intent recognition. In particular, given an intent category, the framework localizes, in a weakly-supervised manner, salient regions in images that are important for recognizing the class-of-interest. These discovered regions are further reinforced during training using a localization loss to guide the network to focus. In addition, we leverage hashtags as a modality complementary to visual information. We demonstrate through extensive evaluations that properly ingesting visual and textual information helps to boost the performance of intent prediction significantly.

Our work makes the following key contributions: (1) A novel dataset of 14,455 high-quality images, each labeled with one or more human intent. This dataset, which we call *Intentonomy*, offers a total of 28 intent labels supported by a systematic social psychological taxonomy [41] proposed by experts; (2) A systematic study to show how commonly used object and context information, as well as textual information, contribute to intent recognition; (3) We introduce a framework with the help of weakly-supervised localization and an auxiliary hashtag modality that is able to narrow the gap between human and machine understanding of images.

2. Related Work

Prior work on intent recognition has focused on communicative intents in different contexts. Joo *et al*. [19] define 9 dimensions of persuasive intents of a politician implied through a photo (*e.g*., trustworthy). Other works [20, 38, 15, 45] also focus on persuasive intents in political images. Additional related work includes image and video advertisement understanding including topics, sentiment and intent [17, 64, 60], or the motivation behind the actions of people from images [31, 54]. Understanding intent is also a key component in persuasive dialogue systems [35, 9, 61, 57]. In this work, we focus on the behavior of the people who post on social media websites. While a large body of work [23, 1, 24, 2, 42, 40] exists that study the motivations behind the usage of social media, relatively much fewer work exists in the area of computer vision.

The most similar work in terms of understanding human motive in social media is from [22], which introduces a multimodal dataset to understand the document intent in Instagram posts. However, we differentiate our work in terms of goals and methods: (1) we emphasize "visual intent" rather than "textual intent", meaning that we study human motive mainly based on the perceived motives behind images rather than textual data; (2) we systematically analyze how objects and context contribute to the recognition of human motives in the social media domain; (3) our dataset contains more fine-grained categories (28 classes in total) with nine super-categories compared to 8 categories from [22].

Our study on the relationship between intent and content is inspired by [65], which studied the effect of context for object recognition. Other works also proposed context-aware models in various tasks such as object recognition and detection [48, 13, 49, 30, 6, 29, 14, 28, 3, 26, 4], scene classification [59, 5], semantic segmentation [59, 29], scene graph recognition [63], visual question answering [44]. Our work utilizes both object- and scene-level information to distinguish between different intent classes.

3. Intentonomy Dataset

Images Our dataset is built up of free-licensing high-resolution photos from the website Unsplash[2]. We sample images with common keywords that are similar to social media hashtags, including "people","happy", *etc*. The resulting images cover a wide range of everyday life scenes (*e.g*., from parties, vacations, and work).

Intent taxonomy The selection of intent labels is a non-trivial exercise. The labels must form a representative set of motives from social media posts, and it should occur with high enough frequencies in the collection of the dataset. Previous work on motive taxonomies [41] provide a solid foundation for our study. However, not all of the 161 human motives presented in [41] are suitable in the context of social media posts, or can be inferred from single-image inputs. For example, one might need background information about the person inside the image to judge if the intent is "being spontaneous", "to be efficient", "to be on time". Some fine-grained motives in the taxonomy could be merged. For instance, "social group" and "close friends", "making friends" and "having close friends". Wherever possible, we further divide certain motives into sub-motives ("in love" and "in love with animal" for instance), for more granularity. Fig. 2 illustrates our resulting ontology in full with hierarchy information and annotated image examples.

Annotation details Amazon Mechanical Turk (MTurk) was recruited to collect labels of perceived intent by employing a similarity comparison task that we call "unsatisfactory substitutes". We rely on the notion of "mental imagery" [46] – a quasi-perceptual experience that maps example images to a visual representation in one's mind, along with *games with a purpose* [52, 53, 8, 50] as our overall annotation approach. Fig. 3 displays the differences between a standard

[2] Unsplash Full Dataset 1.1.0

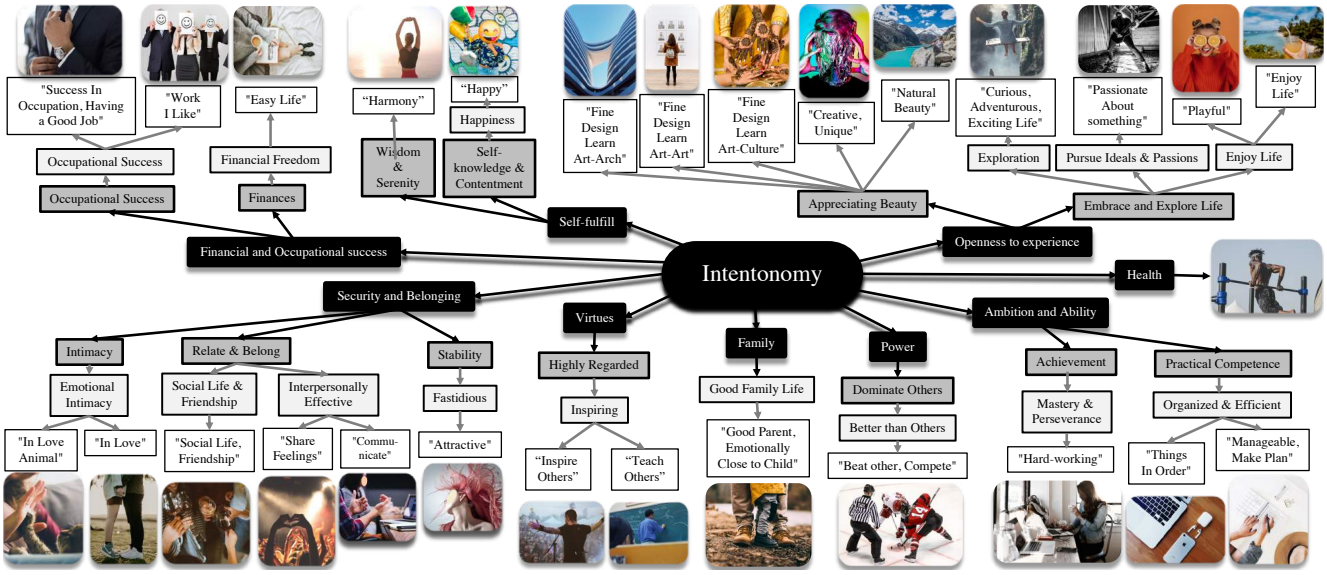


Figure 2. Ontology visualization. We select 28 motive labels from a general human motive taxonomy used in psychology research [41]. There are 9 super-categories in total *(in black box)*, namely "virtues", "self-fulfill", "openness to experience", "security and belonging", "power", "health", "family", "ambition and ability", "financial and occupational success". See the Appendix **??** for dataset statistics.

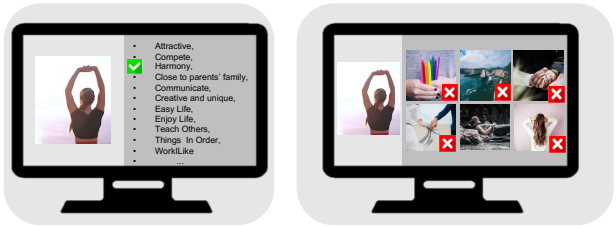


Figure 3. Annotation methods comparison. (a) A standard annotation process: given a image, choose the desired labels from a drop-down list; This approach is time-consuming and highly dependent on the expertise of annotators. (b) Our approach: similarity comparison using "unsatisfactory substitutes" so the annotators can focus on the "swapabilities" of image pairs regarding the intent. The task is to select all the images in the grid that clearly have a different intent than the reference image on the left.

annotation process and ours. Due to space constraints, we leave other details in the Appendix **??**.

Although we have implemented strategies to ensure quality (see the Appendix **??**), we acknowledge that there are inevitable inconsistencies in our training data. Different people have different opinions of perceived intent. Prior work [50] shows that there is at least 4% error rate in popular datasets like CUB-200-2011 [55] and ImageNet [7]. Yet these datasets are still effective for computer vision research. Deep learning is robust to label noise in training set [50, 36]. To this end, we create a highly curated test set by enlisting a single domain-specific taxonomic expert to provide the annotations for both validation (val) and test sets. In our experiments, we regard this expert's opinions as the "gold standard," which allows us to focus on self-consistency in val and test sets, but we acknowledge that challenges remain in terms of resolving matters of disagreement among communities of experts. In the end, Intentonomy dataset has 12,740 training, 498 val, 1217 test images. Each image contains one or multiple intent categories.

4. From Visual Content to Human Intent

Our goal is to investigate systematically how visual content within images contributes to the understanding of human intent. To this end, we disentangle the impact of visual content on intent classification by a series of controlled experiments inspired by the methodology in [65]. More specifically, we study the effect of visual content in terms of object ($\mathcal{O}$) and context ($\mathcal{C}$), and focus on the following fundamental aspects: (1) the amount of content information; (2) three different content properties, including geometry, resolution, and low-level texture. We then analyze the relationships between intent classes and specific things and stuff classes. Fig. 4 and 5 provide an overview of our study under different control settings to analyze how visual information affects intent recognition.

More formally, given an image I, we apply a perturbation either to its objects or context to produce a modified image: $I_x^{(t)} = f(I, t, x)$, $x \in \boldsymbol{X}, t \in \{\mathcal{O}, \mathcal{C}\}$, where $f(\cdot)$ indicates a transformation function as will be introduced below and $\boldsymbol{X}$ is a set of positive integers defining the level of changes. The larger the value of index x, the closer the $I_x^{(t)}$ is to the original images. We now introduce different transformation functions used to see how intent recognition performance changes based on different visual contents.

Amount of content. We control the amount of object

Figure 4. Example images with full content (far left), and image modifications used for the controlled conditions of our study.

Properties	$\lvert\boldsymbol{X}\rvert$	$I_x^{(t)} = f(\cdot)$
`Geometry`	6	$\text{jigsaw}_{(g\times g)}(t)$, $g = 2^{5-x}, x \in [0,5]$
`Resolution`	6	$\text{blur}_{\sigma}(t)$, $\sigma = 2^{5-x}, x \in [0,5]$
`Low-level texture features`	3	$\begin{cases} \text{no t} & x = 0 \\ \mathbb{1}\{\text{texture}(t)\} & x = 1,2 \end{cases}$

Table 1. Content properties investigation. $t \in \{\mathcal{O}, \mathcal{C}\}$.

or contextual information by expanding (or decreasing for context experiments) the bounding boxes (bbox) of detected objects by e pixels:

$$I_x^{(t)} = \begin{cases} \text{bbox}^{(t)} & x = 0 \\ \text{bbox}^{(t)} \pm \text{e} & x \in [1,7] \quad e = 2^x \\ \text{full image} & x = 8 \end{cases}$$

where $\text{bbox}^{(t\in\mathcal{O})}$ denotes the image area within the bounding box, and $\text{bbox}^{(t\in\mathcal{C})}$ is the area outside the bounding box (see two images in Fig. 4(d,e) for an example). A total of 9 variations for both objects and context are included. The larger x indicates that the larger the amount of objects or context are presented.

Content properties. We also study how visual properties impact intent recognition. We analyze the effect of the following properties of $\mathcal{O}$ and $\mathcal{C}$, including:

1. `geometry`: regions of objects or context are broken down to $g \times g$ tiles and randomly re-arranged (we call this operation jigsaw), while the other content component remains intact;
2. `resolution`: convolving the selected content component with a Gaussian function (zero-mean and various values of standard deviation σ);
3. `low-level texture features`: visual textures are constructed using image statistics [33] for the selected content component.

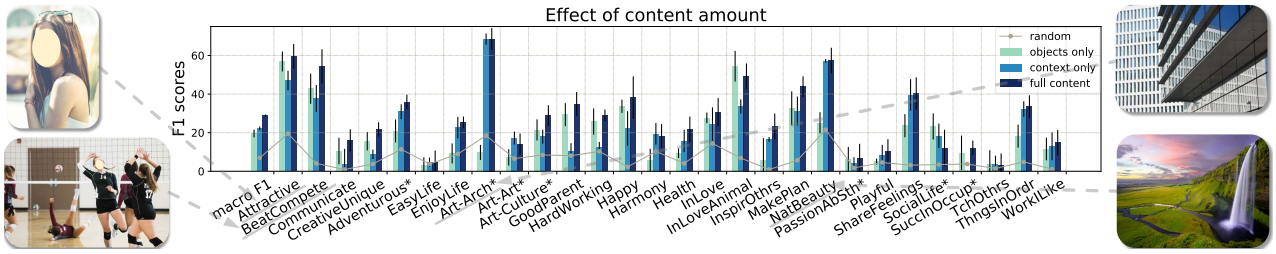
For all three properties, we only modify the selected regions and paste other intact content components to their original locations. Table 1 describes our method in detail.

Analysis and discussion. Given each transformation f, we finetune a pretrained CNN model and obtain the macro F1 score on the modified validation set. Each model is run multiple times to reduce variance. Fig. 5 shows results of the 4 experiments focusing on content size and three properties. In general, we observe a positive correlation between the amount of content and the macro F1 score. We can see in Fig. 5(a) that recognition F1 score decreases when context/object information is removed, for a majority of motive labels (*e.g.* "BeatCompete" and "SocialLife*"), confirming that context and objects clues are both important.

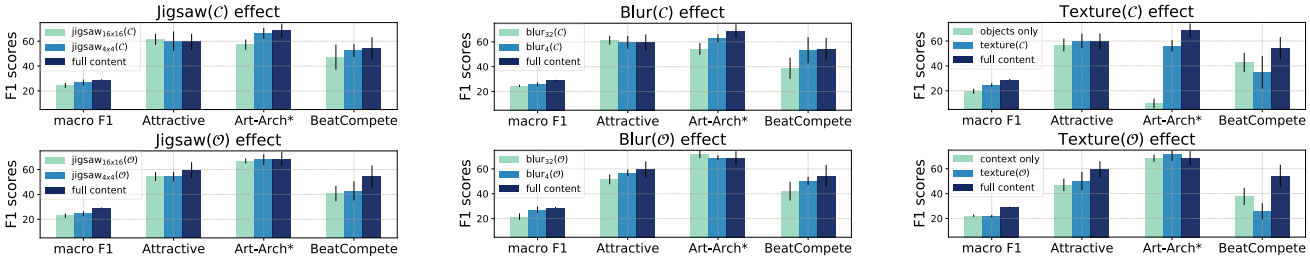
Interestingly, there are some exceptions to this trend where either objects or context, on its own, yield comparable results to the original images. For categories like "attractive" or "in love", object information alone offers comparable F1-scores to full images. In other cases, contextual information achieves decent performance for motives like "appreciate architecture", "natural beauty". Such motives are usually associated with representative gestures that provide strong supervisory signals (*e.g.*, see Fig. 1). These signals usually come from single content module, which we further demonstrate in the next subsection.

In addition, Figs. 5(b)-5(d) demonstrate how content properties affect intent recognition. We see that geometry, blurred effect, and texture features of the content component decrease the intent recognition performance. See macro F1 score and "beatCompete" in Figs. 5(b)-5(d) as an example. Similar to the content size experiment previously, the impact of content properties is different for different classes. The bottom plots of Figs. 5(b)-5(d) show that "Attractive" is sensitive to object manipulation. Motives like "Art-Arch*", on the other hand, have an opposite trend where context contributes more than objects. The recognition results are robust to object manipulation, yet sensitive to context modulation overall. These observations are further illustrative of the varying importance of objects or contextual information for different classes.

Relationship between intent and object/context classes. The above analysis demonstrates different intent categories have different preferences on objects and/or context. We now examine whether there exists relationships between intent categories and *specific* objects/context classes.



(a) Content size.



(b) Content geometry. (c) Content resolution. (d) Content texture.

Figure 5. A study on intent and content. Overall there are three trends among 28 classes, which are presented in Figs. 5(a)-5(d). F1 scores, including average value and standard deviation over 5 runs, and random guess results, for selected classes and selected data variations are displayed. Class names ends with "*" are abbreviated (*e.g.* "Art-Arch*" is short for "appreciate architecture").

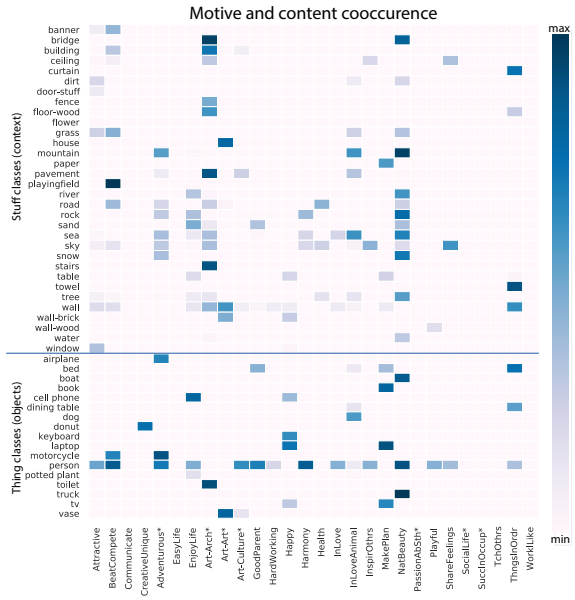


Figure 6. A visualization of Π (Eq. 1), where each entry denotes the correlation between a pair of intent and object/context class.

More specifically, given an image I with a intent label m and a trained intent recognition model, we use class activation mappings [66], to produce a binary mask $\text{CAM}^b(I, m, \tau_{cam})$ (τ_{cam} is a threshold value) to represent the discriminate image regions for class m in the image. We also feed the image to a segmentation model pretrained on the COCO Panoptic dataset [21] to obtain a binary mask $\text{Pano}(I, p, \tau_p)$ (τ_p is a threshold value) for the class p in the COCO dataset. We use the COCO Panoptic dataset [21] because it contains widely used *thing* and *stuff* categories. We then define the correlation between p and m as:

$$\Pi_{p,m} = \frac{\text{CAM}^b(I, m, \tau_{cam}) \cap \text{Pano}(I, p, \tau_p)}{\text{Pano}(I, p, \tau_p)}. \quad (1)$$

Here, objects with high scores tend to be semantically meaningful for the corresponding intent categories. Fig. 6 further validates our findings in the content modulation experiments. While there are intent classes requiring both object and context, certain classes are object-oriented while others are context-oriented. Further, it can be observed that certain intent classes are also more dependent on particular object or context classes. For example, "person" is semantically meaningful for intent like "Attractive" and "inHarmony". It is also consistent that stuff classes like "building", "bridge" can help discriminate classes like "Architecture".

It is worth mentioning that some intent classes (*e.g.* "easyLife", "socialLife") have no or few correlated thing or stuff classes. Indeed, the F1 score for some motive classes are comparable to random guessing (see Fig. 5(a)). We suspect that visual information only is not enough to represent the inherent visual and semantic diversity in those classes.

5. Multimodal Intent Recognition

The study in Sec. 4 demonstrates that different intent classes have different correlation with context and objects, and so using a single "one-size-fits-all" network for intent

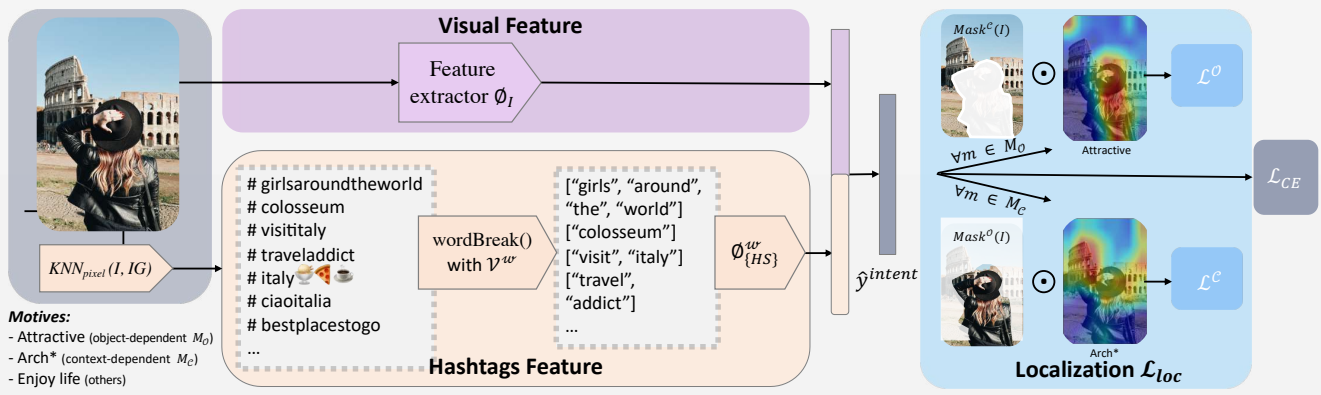


Figure 7. Method overview. Given an image I, we localize important object and context regions for an intent of interest and additionally use hashtags to complement visual information. See texts for more details.

recognition is sub-optimal. To mitigate this issue, we introduce a localization loss that identifies, for each class, regions in images that are important. In addition, as shown, visual information alone is not sufficient for predicting certain classes of intent. To compensate, we also propose to use an auxiliary channel to provide complementary semantic information. The overall framework is presented in Fig. 7.

Object/Context localization. Since different intent classes rely on different visual content (either $\mathcal{O}$ or $\mathcal{C}$), we wish to guide the network to attend to these regions when recognizing a class of interest. In particular, we first split all intent categories into 3 groups based on our study in §4: object-dependent ($M_{\mathcal{O}}$), context-dependent ($M_{\mathcal{C}}$), and others which depends on the entire image. We then use CAM [66] to localize salient regions in a weakly-supervised manner and minimize the overlap area between CAM and the image area that is not a region of interest (Fig. 7).

Formally, given a motive class m and an image sample I, $\text{CAM}(i, m)$ denotes the real-valued version of $\text{CAM}^b(I, m, \tau_{cam})$ (see §4). Let $\text{Mask}^{\mathcal{C}}(I)$ and $\text{Mask}^{\mathcal{O}}(I)$ be the aggregated binary masks in image I that represent all detected thing ($P_{\mathcal{T}}$) and stuff classes ($P_{\mathcal{S}}$), respectively. See examples of $\text{Mask}^{\mathcal{O}}(I)$ and $\text{Mask}^{\mathcal{C}}(I)$ in Fig. 7. The localization loss is then defined as:

$$L^{\mathcal{O}} = \sum_{m \in M_{\mathcal{O}}} \left(\text{CAM}(I, m) \odot \text{Mask}^{\mathcal{C}}(I)\right) \quad (2)$$

$$L^{\mathcal{C}} = \sum_{m \in M_{\mathcal{C}}} \left(\text{CAM}(I, m) \odot \text{Mask}^{\mathcal{O}}(I)\right), \quad (3)$$

where $\odot$ is element-wise multiplication. The final loss $\mathcal{L}_{loc}$ is the summation of all the entries in $L^{\mathcal{O}}$ and $L^{\mathcal{C}}$.

Note that our approach is similar to previous work that addresses contextual bias [39]. Both approaches use CAM as weak annotations to guide training. However, our method does not require a regularization term which grounds CAMs of each category to be closer to the regions from a previously trained model. Therefore, our approach can be trained with a single pass, in an end-to-end fashion.

Hashtags as an auxiliary modality. Visual information is not sufficient for recognizing certain intent categories (see "EasyLife" in Fig. 6). To further improve intent recognition, we resort to language information as a complementary clue for improved performance. Unfortunately, images from Unsplash are not associated with any text information. We instead leverage visual similarities of the Unsplash images to a larger set of images, which do contain associated metadata that loosely describe the semantics within the images. Instagram (IG) is a social media platform that contains billions of publicly available photos, often with user-provided hashtags. This presents an opportunity to weakly relate images with vastly different visual appearances that contains similar semantic information, by means of hashtags.

In particular, we first compute regional maximum activations of convolutions features [10, 47] from the last activation map of a pretrained Resnext-50 (32x4) model (trained on ImageNet-22k [7]) on 7-days of public photos from IG as well as all the images from our intent dataset. Using these embeddings, we then perform a KNN query for each Unsplash image to retrieve the top k matching IG images for each of the images in our intent dataset. Finally, for each matching IG photo, we collect all of the associated hashtags (additional details are in the supplemental material). The collection of all matched hashtags for a given Unsplash image are represented as an unordered set $\boldsymbol{HT}$. See Fig. 7 for examples of fetched hashtags. However, directly using hashtags are challenging because: 1) hashtags can be noisy, much like web-scale data tends to be; 2) a hashtag is usually a concatenation of several words, including multilingual phrases and emojis (*e.g.* `#coffeme`, `#landscapephotography`). There are a large amount of out of vocabulary words if one uses a pre-trained word embedding for the entire hashtag. We thus first break the hashtags down using a known dictionary of words (*i.e.*

#coffeeme → "coffee" "me"). Subsequently, unusual and noisy tokens/hashtags are automatically filtered out.

Formally, given $\boldsymbol{HT}$ for one image sample and a dictionary $\mathcal{V}$, we first segment each hashtag hs into a list of tokens based on the given vocabulary : WordBreak$(hs, \mathcal{V}) = [w]$, $w \in \mathcal{V}$, $hs \in \boldsymbol{HT}$. Separated tokens of one hashtag are mapped to a dense embedding individually, and aggregated into a single representation. Next, all of the resulting hashtag representations are averaged to compute a unified feature for all hashtags associated with a single image. Finally, the hashtag features are concatenated with image features into an integrated representation for classification.

Loss function. To capture the different opinions from crowd annotators, we use cross-entropy loss with soft probability, denoted as $\mathcal{L}_{CE}$, inspired by [27][3]. More formally, our model computes probabilities $\hat{\boldsymbol{y}}^{intent}$ using a softmax activation, and minimizes the cross-entropy between $\hat{\boldsymbol{y}}^{intent}$ and the target distribution $\boldsymbol{y}^{intent}$. $\boldsymbol{y}^{intent}$ is a target vector, where each position m contains the number of crowd workers who labeled the associated image to motive class m, normalized by the total number of crowd workers to indicate a probability distribution.

6. Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of different components of the multimodal framework. More specifically, we report the performance of the following approaches: (1) RANDOM, which is the success rates by random guessing; (2) VISUAL, which finetunes a standard ResNet50 model to classify motives; (3) HASHTAGS (HT), which only uses hashtags to predict intent; (4) VISUAL + HT, which combines visual information and hashtags; (5) VISUAL + $\mathcal{L}_{loc}$, which augments a visual model with the proposed localization loss; (6) VISUAL + $\mathcal{L}_{loc}$ + HT, which denotes our full model. Among them, (2)-(4) are trained using the standard cross-entropy loss only, $\mathcal{L}_{CE}$. When the localization loss $\mathcal{L}_{loc}$ is applied, we sum both $\mathcal{L}_{loc}$ and $\mathcal{L}_{CE}$: $\mathcal{L} = \lambda\mathcal{L}_{loc} + \mathcal{L}_{CE}$. λ is a scalar to determine the contribution of each loss term. Performance are measured using Macro F1, Micro F1, and Samples F1 scores. We repeat each experiment 5 times and report the mean and standard deviation (std).

Table 2 summarizes the results. We can see that the full model achieves a 31.12 macro F1 score, outperforming the VISUAL baseline by +7.76% percent difference, as well as the HT baseline by +57.81% percent difference. Furthermore, compared to the VISUAL only approach, adding the localization loss improves macro F1 score by 5.16%. We also observe that visual and text information are complementary, offering 4.99% and 53.75% gain compared to visual and text only, respectively.

[3]Similar to [27], we also tried sigmoid cross-entropy loss but obtained worse results.

Method	Macro F1	Micro F1	Samples F1
RANDOM	6.94 ± 0.09	7.18 ± 0.10	7.10 ± 0.10
VISUAL	28.88 ± 0.56	37.08 ± 1.07	36.06 ± 1.51
HT	19.72 ± 0.88	29.30 ± 1.62	31.47 ± 1.64
VISUAL + $\mathcal{L}_{loc}$	30.37 ± 0.51 (+1.49)	**38.64 ± 0.95** (+1.56)	37.41 ± 1.51 (+1.35)
VISUAL + HT	30.32 ± 0.62 (+1.44)	37.61 ± 0.85 (+0.53)	**38.98 ± 1.70** (+2.92)
VISUAL + $\mathcal{L}_{loc}$ + HT	**31.12 ± 0.63** (+2.24)	38.49 ± 0.88 (+1.41)	38.77 ± 1.74 (+2.71)

Table 2. Experimental results of different approaches for intent recognition measured in Micro F1, Macro F1, Samples F1 scores. (+ ·) indicate the difference comparing to VISUAL. (+ ·) in green denotes that the difference is larger than the std.

Method	Content		
	$\mathcal{O}$-classes	$\mathcal{C}$-classes	*Others*
RANDOM	7.75 ± 5.47	12.53 ± 5.96	6.05 ± 5.23
VISUAL	34.92 ± 3.63	41.27 ± 3.53	25.34 ± 1.13
VISUAL + $\mathcal{L}_{loc}$	38.82 ± 1.95 (+3.9)	**43.14 ± 3.00** (+1.87)	25.90 ± 1.35 (+0.56)
VISUAL + $\mathcal{L}_{loc}$ + HT	**39.82 ± 1.56** (+4.90)	42.09 ± 2.57 (+0.82)	**26.77 ± 1.13** (+1.43)

Table 3. Results of different approaches in terms of how much object/context information intent categories need. (+ ·) indicate the difference comparing to VISUAL. (+ ·) in green denotes that the difference is larger than the std.

To better understand why $\mathcal{L}_{loc}$ and HT improve visual only model, we break down the intent classes into different subsets based on their content dependency, *i.e.*, object-dependent ($\mathcal{O}$-classes), context-dependent ($\mathcal{C}$-classes), and *Others* which depends on both foreground and background information; (2) difficulty, which measures how much the VISUAL outperforms achieves than the RANDOM results ("easy", "medium" and "hard"). More details are given in the Appx.. Table 3-4 summarize the subset results.

The effectiveness of $\mathcal{L}_{loc}$ We see from Table 3 that when adding the localization loss gains are more significant for $\mathcal{O}$-classes, compared to $\mathcal{C}$-classes and *Others*. The localization loss depends on the area of either object or context regions in the images, and the objects' region, which is used in $\mathcal{L}^{\mathcal{C}}$ in Eq. 3, are typically small[4]. As a result, the $\mathcal{L}_{loc}$ has no significant effect on the final score.

We also conduct a qualitative study to understand why the localization loss helps intent recognition. Results are shown in Fig. 8. We can see that the localization loss helps the model to focus on the correct region of interest for both $\mathcal{O}$- and $\mathcal{C}$-classes, especially when the image is scattered with multiple objects and scenes. Fig. 8(a) confirms that

[4]Note that $\mathcal{L}^{\mathcal{C}}$ minimizes the overlap region between object area and the salient region (CAM).

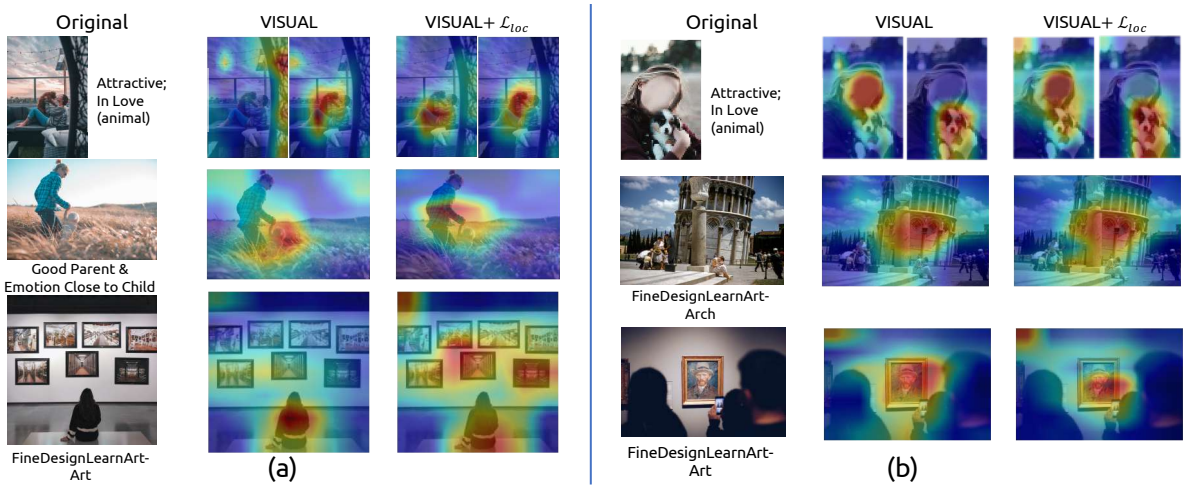


Figure 8. Analysis of the proposed localization loss. (a) VISUAL + $\mathcal{L}_{loc}$ approach learns to isolate appropriate regions of interest, comparing to VISUAL. For example, our method learn to focus on the dog and girl respectively for "In love (animal)" and "Attractive" respectively, which are $\mathcal{O}$-classes. (b) Examples for which both VISUAL + $\mathcal{L}_{loc}$ and VISUAL produce similar visualizations. Both methods focus on the correct region, which are located in center and account for a larger area of the image.

Method	Difficulty		
	Easy	Medium	Hard
RANDOM	19.86 ± 1.28	7.11 ± 3.40	2.81 ± 1.80
VISUAL	61.84 ± 4.90	33.71 ± 2.24	11.73 ± 1.74
HT	63.58 ± 1.79	19.68 ± 1.70	6.63 ± 1.43
VISUAL + HT	**66.67** ± **2.12** (+4.83)	32.93 ± 1.57 (-0.78)	15.52 ± 0.98 (+3.79)
VISUAL + $\mathcal{L}_{loc}$ + HT	66.18 ± 4.56 (+4.34)	**33.86** ± **1.08** (+0.15)	**16.50** ± **1.80** (+4.77)

Table 4. Results of different approaches in terms of how difficult it is for VISUAL to outperforms the RANDOM results. (+ ·) indicate the difference comparing to VISUAL. (+ ·) in green denotes that the difference is larger than the std.

VISUAL + $\mathcal{L}_{loc}$ works well when both object and context information are presented in the image (bottom 2 examples), or the target region of interest is small (top example). We also note in Fig. 8(b) that for images where the region of interest is located in the center, or is relatively large, both VISUAL and our method give good results.

The effectiveness of hashtags. From Table 4, we observe that the model using both images and hashtags outperforms the uni-modal approaches over "easy" and "hard" classes, without hurting the "medium" classes. This suggests there is value in the auxiliary information to help close the semantic gap. Therefore, our results suggest that images and hashtags do in fact complement each other in the motive recognition task. For example, `#love` is directly indicative of the intent label "in love", as is `#workout` of "health" (see Fig. 7 for more hashtag examples). Interestingly, for "easy" classes, HT model outperforms the VISUAL model by 8.2%, however it struggles with the "medium" and "hard" classes (Table 4). This suggests that hashtags provided by users, while noisy, do still contain information about intent to some extent.

It is perhaps counter-intuitive that hashtags do not outperform visual signals entirely. While hashtags seem to capture the essence of human motives (see examples in Fig. 7), careful inspection of the fetched hashtags shows that not all hashtags are useful in practice. Obscurity and ambiguity exist, including typos, slang, inside jokes, and irrelevant information. More effective modeling of hashtags remains an open research problem.

7. Conclusion

In this work, we studied the problem of modeling human motives in social media posts. We introduced a new dataset that taps into mental imagery in a novel annotation game with a purpose to acquire labels from MTurk, and collected a rich image dataset with 28 human motives supported by a social psychological taxonomy. We conducted rigorous studies to explore the connections between content and intent. Our results show that there is still much room for improvement (for context-dependent, and hard classes for example). We therefore hope that the new Intentonomy dataset will facilitate future research to better understand the cognitive aspects of images.

Acknowledgement We thank Luke Chesser and Timothy Carbone from Unsplash for providing the images, Kimberly Wilber and Bor-chun Chen for tips and comments about the annotation process, Kevin Musgrave for the general discussion, and anonymous reviewers for their valuable feedback. This work is supported by a Facebook AI research grant awarded to Cornell University.

References

[1] Tamar Ashuri, Shira Dvir-Gvisman, and Ruth Halperin. Watching me watching you: How observational learning affects self-disclosure on social network sites? *Journal of Computer-Mediated Communication*, 23(1):34–68, 2018. 2

[2] Saeideh Bakhshi, David A Shamma, Lyndon Kennedy, and Eric Gilbert. Why we filter our photos and how it impacts engagement. In *Ninth International AAAI Conference on Web and Social Media*, 2015. 2

[3] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, 2018. 2

[4] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *ICLR*, 2019. 2

[5] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *CVPR*, June 2018. 2

[6] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 33(7):853–862, 2012. 2

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 3, 6

[8] Jia Deng, Jonathan Krause, and Li Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *CVPR*, pages 580–587, 2013. 2

[9] Arie Dijkstra. The psychology of tailoring-ingredients in computer-tailored persuasion. *Social and personality psychology compass*, 2(2):765–784, 2008. 2

[10] Albert Gordo, Jon Almazán, Jérôme Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. *CoRR*, abs/1604.01325, 2016. 6

[11] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 1

[12] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 1

[13] Derek Hoiem, Alexei A Efros, and Martial Hebert. Geometric context from a single image. In *ICCV*, 2005. 2

[14] Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, and Greg Mori. Learning structured inference neural networks with label relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[15] Xinyue Huang and Adriana Kovashka. Inferring visual persuasion via body language, setting, and deep features. In *CVPRW*, pages 73–79, 2016. 2

[16] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *ECCV*, 2018. 1

[17] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *CVPR*, pages 1705–1715, 2017. 2

[18] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *ECCV*, 2020. 1

[19] Jungseock Joo, Weixin Li, Francis F Steen, and Song-Chun Zhu. Visual persuasion: Inferring communicative intents of images. In *CVPR*, pages 216–223, 2014. 2

[20] J. Joo, F. F. Steen, and S. Zhu. Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *ICCV*, pages 3712–3720, 2015. 2

[21] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, pages 9404–9413, 2019. 5

[22] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. Integrating text and image: Determining multimodal document intent in instagram posts. In *EMNLP*, pages 4614–4624, 2019. 2

[23] Chih-Hui Lai. Motivations, usage, and perceived social networks within and beyond social media. *Journal of Computer-Mediated Communication*, 24(3):126–145, 2019. 2

[24] So-Hyun Lee and Hee-Woong Kim. Why people post benevolent and malicious comments online. *Communications of the ACM*, 58(11):74–79, 2015. 2

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[26] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *CVPR*, pages 6985–6994, 2018. 2

[27] D.K. Mahajan, R.B. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 7

[28] Ishan Misra, Abhinav Gupta, and Martial Hebert. From Red Wine to Red Tomato: Composition with Context. In *CVPR*, 2017. 2

[29] R. Mottaghi, X. Chen, X. Liu, N. Cho, S. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 2

[30] Dennis Park, Deva Ramanan, and Charless Fowlkes. Multiresolution models for object detection. In *ECCV*, pages 241–254. Springer, 2010. 2

[31] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Inferring the why in images. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE, 2014. 2

[32] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *EMNLP*, pages 22–32, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. 1

[33] Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70, 2000. 4

[34] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *CVPR*, pages 413–420. IEEE, 2009. 1

[35] Barbara K Rimer and Matthew W Kreuter. Advancing tailored health communication: A persuasion and message effects perspective. *Journal of communication*, 56:S184–S201, 2006. 2

[36] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017. 3

[37] Florian Schroff, Tali Treibitz, David Kriegman, and Serge Belongie. Pose, illumination and expression invariant pairwise face-similarity measure via doppelgänger list comparison. In *ICCV*, Barcelona, 2011. 1

[38] Behjat Siddiquie, Dave Chisholm, and Ajay Divakaran. Exploiting multimodal affect and semantics to identify politically persuasive web videos. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 203–210, 2015. 2

[39] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *CVPR*, 2020. 6

[40] Flávio Souza, Diego de Las Casas, Vinícius Flores, Sun-Bum Youn, Meeyoung Cha, Daniele Quercia, and Virgílio Almeida. Dawn of the selfie era: The whos, wheres, and hows of selfies on instagram. In *Proceedings of the 2015 ACM on conference on online social networks*, pages 221–231, 2015. 2

[41] Jennifer R Talevich, Stephen J Read, David A Walsh, Ravi Iyer, and Gurveen Chopra. Toward a comprehensive taxonomy of human motives. *PloS one*, 12(2):e0172279, 2017. 1, 2, 3

[42] Jennifer R Talevich, Stephen J Read, David A Walsh, Ravi Iyer, and Gurveen Chopra. Toward a comprehensive taxonomy of human motives. *PloS one*, 12(2):e0172279, 2017. 2

[43] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. In *CVPR*, 2016. 1

[44] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question answering. In *CVPR*, pages 1–9, 2017. 2

[45] Christopher Thomas and Adriana Kovashka. Predicting the politics of an image using webly supervised data. In *Advances in Neural Information Processing Systems*, pages 3625–3637, 2019. 2

[46] Nigel J.T. Thomas. Mental imagery. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition, 2019. 2

[47] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. In *Proceedings of the International Conference on Learning Representations*, 2016. 6

[48] Antonio Torralba. Contextual priming for object detection. *International journal of computer vision*, 53(2):169–191, 2003. 2

[49] Antonio Torralba, Kevin P Murphy, and William T Freeman. Using the forest to see the trees: exploiting context for visual object detection and localization. *Communications of the ACM*, 53(3):107–114, 2010. 2

[50] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR*, Boston, MA, 2015. 2, 3

[51] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018. 1

[52] Luis Von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006. 2

[53] Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008. 2

[54] Carl Vondrick, Deniz Oktay, Hamed Pirsiavash, and Antonio Torralba. Predicting motivations of actions by leveraging text. In *CVPR*, pages 2997–3005, 2016. 2

[55] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 3

[56] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1

[57] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *ACL*, pages 5635–5649, 2019. 2

[58] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. IEEE, 2010. 1

[59] Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, pages 702–709. IEEE, 2012. 2

[60] Keren Ye, Narges Honarvar Nazari, James Hahn, Zaeem Hussain, Mingda Zhang, and Adriana Kovashka. Interpreting the rhetoric of visual advertisements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[61] Dian Yu and Zhou Yu. Midas: A dialog act annotation scheme for open domain human machine spoken conversations. *arXiv preprint arXiv:1908.10023*, 2019. 2

[62] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, June 2019. 1

[63] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, June 2018. 2

[64] Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. Equal but not the same: Understanding the implicit relation-

ship between persuasive images and text. *arXiv preprint arXiv:1807.08205*, 2018. 2

[65] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *CVPR*, pages 12985–12994, 2020. 2, 3

[66] B. Zhou, A. Khosla, Lapedriza. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *CVPR*, 2016. 5, 6

[67] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1