Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please see Adversarial Example under Section 2
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The dataset is available in a Google Drive link and the code will be released on a later date.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] NAO sources images from the OpenImages dataset and is cited.
   (b) Did you mention the license of the assets? [Yes] The annotation is licensed under CC-BY 4.0 and the images are licensed under CC-BY 2.0.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No] The labellers are compensated with money but we are unable to disclose the hourly wage as it is considered trade secrets of Scale AI.

A Appendix

Dataset documentation and intended uses See Appendix B for Datasheet for Dataset of NAO.

URL to website/platform where the dataset/benchmark can be viewed and downloaded by the reviewers The images and dataset can be found on `https://drive.google.com/drive/folders/15P8sOWoJku6SSEiHLEts86ORfytGezi8`

License The author bear all responsibility in case of violation of rights and confirmation of the data license.

Hosting, licensing, and maintenance plan The dataset will be hosted on Google Drive during the review process and will be hosted on scale.com after the review process is completed.

Links to access the dataset and its metadata The images, annotations and metadata can be found on `https://drive.google.com/drive/folders/15P8sOWoJku6SSEiHLEts86ORfytGezi8`

The dataset format The images for NAO are in either JPEG or PNG format. The bounding box annotation follows the MSCOCO data annotation and can be read using `https://github.com/cocodataset/cocoapi`.

Long-term preservation The dataset will be hosted on Google Drive during the review process and will be hosted on `scale.com` after the review process is completed.

Structured metadata We will add the metadata (JSON-LD) on the download page of NAO on `scale.com` once the review process is completed.

Persistent dereferenceable identifier We will add DOI that directs to the download page of NAO on `scale.com` once the review process is completed.

B Datasheet for Dataset

B.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

NAO is created to evaluate the robustness of object detection models, including out-of-distribution objects. NAO causes many object detection models to incorrectly classify the objects. We hope that NAO can motivate future research on robustness and handling out-of-distribution objects.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

NAO is created by the machine learning team of Scale AI.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The ground truth annotation of NAO is funded by Scale AI.

B.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance contains the image and the ground truth bounding boxes with their class labels. Some images contain more than one objects.

How many instances are there in total (of each type, if appropriate)?

Category Name	Count
person	4693
cup	2257
car	752
chair	701
bottle	503
dining table	489
book	315
keyboard	301
bowl	251
fork	243
airplane	214
wine glass	212
dog	210
truck	201
knife	189
cake	178
mouse	176
spoon	163
cell phone	147
potted plant	113
kite	112
handbag	77
tv	76
laptop	75
bicycle	72
sandwich	59
sheep	56
boat	56
bird	55
traffic light	54
backpack	46
couch	41
motorcycle	40
orange	33
sports ball	30
broccoli	29
remote	29
umbrella	29
bench	27
bus	25
vase	25
zebra	24
clock	23

Table 3: Number of objects per object class. Only showing categories with more than 20 objects.

There are 7,936 images with 13,604 objects. Please see Table 3 for number of object per category.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

NAO does not contain all possible instances. NAO is not a random sample of real world images as it is curated to cause object detection models to produce high confidence errors.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance contains an image and ground truth bounding box annotations. The image is in PNG or JPEG format and the bounding box is in MSCOCO (JSON) format to facilitate other researchers to evaluate their models on NAO.

Is there a label or target associated with each instance? If so, please provide a description.

Each image is accompanied by the ground truth bounding boxes.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

There is no information missing from individual instances.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

The images might be taken by the same person.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

It is recommended that NAO to be used exclusively as a test set. However, a train, validation and test set split is available for methods that require training. There are 7,196 images in the train set, 373 in the validation set, and 367 in the test set.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Annotation errors are minimized through reviewers and quizzes that are randomly inserted between real annotation tasks. Labellers with quiz score less than a certain threshold were disabled from working on NAO.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

NAO sources images from OpenImages. We provide metadata to link each image in NAO to the original image from OpenImages.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

No. All images are derived from OpenImages.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No. All images are derived from OpenImages.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

While NAO does not heavily focus on people, "Person" is one of the main object categories in a universal dataset (OpenImages).

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No subpopulation of people is identified.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

The people in the frame might be identified by their Flickr account using the metadata from OpenImages.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

The dataset does not contain any sensitive data.

B.3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The images and bounding boxes are directly observable

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

Please see Section 3 in the main paper for a description of the data collection process.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The dataset is sampled from OpenImages and images with egregious errors from the EfficietDet-d7 are included in NAO.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The creation of NAO include the authors and crowdworkers. The author wrote the instructions, designed the annotation tasks and created the quizzes. The crowdworkers identified if the model predictions are indeed incorrect and put bounding boxes around objects. Crowdworkers are compensated with money but we are unable to disclose the amount as it is considered trade secret.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

It is unknown what timeframe the images were collected but ground truth annotations were collected in October 2020 and May 2021.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

We ensured our work conforms to Neurips ethics review guidelines – potential negative societal impacts (Section 2) and general ethical conduct (Section 3).

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.

The dataset is not directly related to people but people might be included in the images.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The images were sourced from OpenImages which in turns come from Flickr.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

It is unknown whether the creators of OpenImages informed the image uploaders.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Unknown.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Not applicable.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Not applicable.

B.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

No extra processing is done to the images. For the bounding box labeling process, please see Section 3 in the main paper.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Not applicable.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

No.

B.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Not applicable.

What (other) tasks could the dataset be used for?

NAO can be used to evaluate the robustness of object detection model against out-of-distribution objects.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

There is minimal risk because the images are already available in public from OpenImages.

Are there tasks for which the dataset should not be used? If so, please provide a description.

No.

B.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset is distributed through Google Drive and will be distributed on `scale.com`

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

NAO is initially distributed through Google Drive and later distributed on `scale.com` once the paper review process is completed. The images are downloadable as a zip file on Google Drive and will be available as a tarball archive on `scale.com` The bounding box annotations are available in MSCOCO format. DOI will be provided once the dataset is hosted on `scale.com`

When will the dataset be distributed?

The dataset is available now on Google Drive and will be distributed on `scale.com` once the paper review process is completed.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The images copyright belongs to the original Flickr users. The bonding box annotations are made available under the CC-BY 4.0 license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

Not applicable.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Not applicable.

B.7 Maintenance

Who is supporting/hosting/maintaining the dataset?

NAO is hosted on Google Drive and will be hosted on `scale.com` once the paper review process is completed. The maintenance will be done by the authors.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

All questions and comments can be sent to Felix Lau: felix.lau@scale.com

Is there an erratum? If so, please provide a link or other access point.

All changes to the dataset will be announced in the CHANGELOG file.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

The dataset might be updated to correct labeling errors and will be communicated using the CHANGELOG file.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Not applicable.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

We plan to only support and host the latest version of NAO.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Yes. Please email all suggestions or contributions to Felix Lau: felix.lau@scale.com