MOTIVATION.

1. **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.

A. Neural networks pretrained on large image collections have been shown to transfer well to other visual tasks where there is little labelled data, i.e. transferring a model works better than starting with a randomly initialized network every time for a new task, as many visual features can be repurposed. This dataset has as its goal to provide a safer large-scale dataset for such pretraining of visual features. In particular, this dataset does not contain any humans or human parts and does not contain any labels. The first point is important, as the current standard for pretraining, ImageNet and its face-blurred version only provide pseudo-anonymity and furthermore do not provide correct licences to the creators. The second point is relevant as pretraining is moving towards the self-supervised paradigm, where labels are not required. Yet most methods are developed on the highly curated ImageNet dataset, yielding potentially non-generalizeable research.

2. **Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?**

A. The dataset has been constructued by the research group "Visual Geometry Group" at the University of Oxford at the Engineering Science Department.

3. **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.

A. The dataset is created for research purposes at the VGG
research group. Individual researchers have been funded by AWS Machine Learning Research Awards (MLRA), EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines & Systems [EP/L015897/1], the Qualcomm Innovation Fellowship, Innovate UK (project 71653) on behalf of UK Research and Innovation (UKRI) and by the European Research Council (ERC) IDIU-638009.

4. **Any other comments?**

A.

COMPOSITION.

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

A. This dataset only contains photos. In addition we provide tabular meta-data for these images, which contain information such as the creator's username and image capture date.

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

A. The dataset contains 1.4M images, resulting in 181GB as a tar file.

7. **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 (for example, 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 (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).

A. The dataset is a sample of a larger set—all possible digital photographs.
As outlined in Section 3 we start from an existing dataset, YFCC-100M, and stratify the images (removing images with people and personal information, removing images with harmful content, removing images with unsuitable licenses, each user contributes at most 80 images to the dataset). This leaves 1.6M images, out of which we take a random sample of 1.28M images to replicate the size of the ImageNet dataset. While this dataset can thus be extended, this is the set that we have verified to not contain humans, human parts and disturbing content.

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

A. Digital photographs uploaded by users of the flickr platform.

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

A. No. Our dataset deliberately does not contain labels.

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

A. Not from the dataset. Note however that the meta-data that we additionally provide is not complete and might have non-uniform missing values.

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

A. Not applicable: each image stands on its own and we do not provide relationships between these.

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

A. As outlined in the intended usecases, this dataset is meant for pretraining representations. As such, the models derived from training on this dataset need to be evaluated on different datasets, so called

down-stream tasks. Thus the recommended split is to use all samples for training.

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

A. No.

14. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, 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 (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

A. No. The dataset contains links to the publicly hosted mirror of the YFCC dataset on Amazon Web Services.

15. **Does the dataset contain data that might be considered confidential (for example, 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.

A. No.

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

A. No. Besides checking for human presence in the images, the annotators were also given the choice of flagging images for disturbing content, which once flagged was removed.

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

SkipA. Does not relate to people

17. **Does the dataset identify any sub-populations (for example, by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

A.

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

A.

19. **Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race 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.

A.

20. **Any other comments?**

A.

COLLECTION PROCESS.

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

A. The data was collected from the publicly available dataset YFCC-100M which is hosted on the AWS public datasets platform. We have used the meta-data, namely the copyright information to filter only images with the CC-BY licence and have downloaded these using the aws command line interface, allowing for quick and stable downloading. In addition, all files were subsequently scanned for viruses using Sophos SAVScan virus detection utility, v.5.74.0.

22. **What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** How were these mechanisms or procedures validated?

A. Our dataset is a subset of the YFCC-100M dataset. The YFCC-100M dataset itself was created by effectively randomly selecting publicly available images from flickr, resulting in approximately 98M images.

23. **If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?**

A. See the similar question in the Composition section.

24. **Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?**

A. As described, the data was collected automatically by simply downloading images from a publicly hosted S3 bucket. The human verification was done using a professional data annotation company that pays 150% of

the local minimum wage.

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

A. The images underlying the dataset were downloaded between March and June 2021 from the AWS public datasets' S3 bucket, following the download code provided in the repo. However the images contained were originally and taken

anywhere from 2000 to 2015, with the majority being shot between 2010-2014.

26. **Were any ethical review processes conducted (for example, 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.

A. No.

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

SkipA. Does not relate to people

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

A.

28. **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.

A.

29. **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.

A.

30. **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).

A.

31. **Has an analysis of the potential impact of the dataset and its use on data subjects (for example, 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.

A.

32. **Any other comments?**

A.

PREPROCESSING/CLEANING/LABELING.

33. **Was any preprocessing/cleaning/labeling of the data done (for example, 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 remaining questions in this section.

A. After the download of approx. 17M images, the corrupted, or single-color images were removed from the dataset prior to the generation of the dataset(s) used in the paper. The

images were not further preprocessed or edited.

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

A. Yes. The creators of the dataset maintain a copy of the 17M original

images with the CC-BY licence of YFCC100M that sits at the start of our dataset creation pipeline.

35. **Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.

A. We have only used basic Python primitives for this. For the annotations we have used VIA [27, 28]

36. **Any other comments?**

A.

USES.

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

A. In the paper we show and benchmark the intended use of this dataset as a pretraining dataset. For this the dataset is used an unlabelled image collection on which visual features are learned and then transferred to downstream tasks. We show that with this dataset it is possible to learn competitive visual features, without any humans in the

pretraining dataset and with complete license information.

38. **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.

A. We will

be listing these at the repository.

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

A. We believe this dataset might allow researchers and practitioners to further evaluate the differences that pretraining datasets can have on the learned features. Furthermore, since the meta-data is available for the images, it is possible to investigate the effect of image resolution on self-supervised learning methods, a domain largely underresearched thus far, as the current de-facto standard, ImageNet, only comes in one size.

40. **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 dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

A. Given that this dataset is a subset of a dataset that randomly samples images from flickr, the image distribution is biased towards European and American creators. As in the main papers discussion, this can lead to non-generalizeable features, or even biased features as the images taken in other countries might be more likely to further reflect

and propagate stereotypes [84], though in our case these do not refer to sterotypes about humans.

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

A. This dataset is meant for research purposes only. The dataset should also not be used for, e.g. connecting images and usernames, as this might risk de-anonymising the dataset in the long term. The usernames are solely provided for

attribution.

42. **Any other comments?**

A.

DISTRIBUTION.

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

A. No.

44. **How will the dataset be distributed (for example, tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?

A. The dataset
will be provided as a csv along with code hosted on GitHub that allows the user to download the

images in our dataset. In addition, we hope to also host it as a single tarball on our servers.

45. **When will the dataset be distributed?**

A. Starting from July 2021.

46. **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.

A. CC-BY.

47. **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.

A. No.

48. **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.

A. Not that we are are of. Regular UK laws apply.

49. **Any other comments?**

A.

MAINTENANCE.

50. **Who will be supporting/hosting/maintaining the dataset?**

A. The dataset is supported by the authors and by the VGG research group. The main contact person is Yuki M. Asano. We host the dataset on

zenodo: https://zenodo.org/record/5528345.

51. **How can the owner/curator/manager of the dataset be contacted (for example, email address)?**

A. The authors of this dataset can be reached at their e-mail addresses: {yuki,chrisr,vedaldi,az}@robots.ox.ac.uk. In addition, we have added a contact form in which we can be contacted anonymously at

https://forms.gle/tkZugt2DJnFdCE1i6.

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

A. If errors are found and erratum will be added to the website.

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

A. Yes, updates will be communicated via the website. The dataset will be versioned.

54. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the 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.

A. Not applicable.

55. **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 dataset consumers.

A. If even after our verification we find further images that contain humans or problematic content, we will remove those further images from existing splits to preserve the goal of this dataset.

56. **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 dataset consumers? If so, please provide a description.

A. Others are free to reach out to us if their ideas can build on this dataset. All code

will be made available.

57. **Any other comments?**

A.