WikiChurches Datasheet

This datasheet accompanies the WikiChurches dataset and follows the template proposed by [1].

I. Dataset Availability

A. How can the dataset be accessed?

The WikiChurches dataset is publicly available at `https://doi.org/10.5281/zenodo.5166986`.

B. How can the dataset be cited?

Björn Barz and Joachim Denzler.
WikiChurches: A Fine-Grained Dataset of Architectural Styles with Real-World Challenges.
arXiv preprint arXiv:2108.06959, 2021.

II. Motivation For Dataset Creation

A. Why was the dataset created?

While the task presented by WikiChurches is architectural style classification, it is neither primarily intended nor particularly suitable as a source of training data for this task. Instead, it was created to serve as a challenging and interesting computer vision benchmark due to the various real-world challenges it poses: fine-grained distinctions between classes based on subtle visual features, a comparatively small sample size, a highly imbalanced class distribution, a high variance of viewpoints, and a hierarchical organization of labels, where only some images are labeled at the most precise level.

B. Has the dataset been used already?

Beyond the simple baseline classification experiment reported in paper accompanying the dataset, it has not been used as of August 2021.

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

WikiChurches could be useful as a benchmark and playground for various research areas, including fine-grained visual recognition [2], [3], data-efficient deep learning [4], dealing with imbalanced training sets, and hierarchical classification with imprecise labels [5].

D. Who funded the creation dataset?

The creator of the dataset was supported by the German Research Foundation as part of the priority programme "Volunteered Geographic Information: Interpretation, Visualisation and Social Computing" (SPP 1894, contract number DE 735/11-1).

III. Dataset Composition

A. What are the instances? Are there multiple types of instances?

Churches buildings, images, architectural styles, distinctive elements of architectural style. One church can have one or more images and styles. An image can have an arbitrary number of bounding boxes enclosing distinctive elements (including zero).

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

Churches	9,346
Images	9,485
Styles	117
Bounding Boxes	631

C. What data does each instance consist of?

Instance	Data
Church	Wikidata ID Name Architectural Styles Year of Construction Country GPS Location
Image	Church ID Original URL URL of Commons Page Commons Categories Uploading User License & Author Information Metadata (capture time, location, ...) SHA1 Hash of Original Image
Style	Wikidata ID Name Parent Style
Bounding Box	Corresponding Image Name of Characteristic Element Coordinates of Top-Left Corner Size (Width & Height)

A description of the files containing this information and their format can be found in the `README.md` file provided with the dataset.

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

Churches and hence their images are associated with one or more architectural styles. Bounding boxes are associated with images and labeled with the name of the architectural element they contain.

E. 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.

Geographical location and year of construction are missing for some churches, since they were not provided in Wikidata.

Not all images have bounding box annotations.

F. Are relationships between individual instances made explicit? If so, please describe how these relationships are made explicit.

Images are linked to churches by filename conventions. Each image filename starts with the Wikidata ID of the church it belongs to, followed by `_wd` and an index number.

Links between bounding boxes and images are made explicit by grouping bounding boxes by image in the file `building_parts.json`.

Styles are explicit attributes of churches. Links from a style to its parent style are explicitly listed in the file `parent_child_rel.txt`.

G. 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).

The dataset is only a small and biased sample of church buildings and architectural styles.

First, it is intentionally limited to churches in Europe and thus does not cover other cultures. The geographical distribution of churches within Europe is highly imbalanced as well. Half of all churches in the dataset are from Germany or France (38% from Germany, 11% from France). Two thirds of the dataset concentrate on as few as four countries: Germany, France, the UK, and Spain. This distribution is not representative of the actual distribution of church buildings across Europe but most likely correlated with the size and level of activity of the local Wikipedia communities and their propensity to enter information in Wikidata.

Similarly, the distribution of styles in the dataset is not representative of reality either. That *Gothic* and *Romanesque* architecture account for the largest portion of the dataset does not imply that they are the most common styles of churches in Europe but simply popular among photographers.

H. 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.

Due to the long-tail class distribution of the full WikiChurches dataset and the varying precision of style labels regarding their level in the taxonomy, we propose several subsets of WikiChurches that alleviate these issues. The subsets have different characteristics and could be useful for different research questions.

WikiChurches-14 comprises all 14 classes from the 1st level of the hierarchy that contain at least 20 images. Churches with more precise labels were re-labeled to their superclass (e.g, *English Gothic* to *Gothic*). Regarding the number of images, this subset still covers 94% of WikiChurches. The class imbalance is still challenging but not impossible to handle (as opposed to the case with only one image per class).

WikiChurches-6 is a subset of WikiChurches-14 restricted to the 6 largest classes comprising more than 200 images. These few classes account for 89% of all images in WikiChurches.

WikiChurches-4 is a subset of WikiChurches-6 limited to the four classes for which we provide bounding box annotations of characteristic visual features: *Romanesque*, *Gothic*, *Renaissance*, and *Baroque*.

WikiChurches-H is intended for studying hierarchical classification. It spans the 19 sub-classes of the four styles from WikiChurches-4 that contain more than 5 images. Churches with 3rd-level labels were re-labeled to their 2nd-level superclass. Additional images from WikiChurches-4 with less precise labels could be incorporated to learn better representations for the 1st-level classes or even all categories.

For WikiChurches-6 and WikiChurches-4, we provide canonical training/validation/test splits. The test split includes about 25% of the churches from each class but at least 100 and at most 500. The validation split is balanced and includes 50 churches from each class. All remaining images are used for the training split. We do not provide canonical splits for the other two subsets since we cannot anticipate the specific needs of the research questions for whose study these subsets could be used.

For each subset and split, a file listing image filenames and the corresponding style label is available in the directory `labels`.

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

Based on a random sub-sample of 200 images, we estimate that 93.5% of the images in the dataset have at least partially

correct labels. About 86.1% are completely correctly labeled, 4.0% are incorrectly labeled and 2.5% of images are not informative enough for making a classification decision.

J. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is self-contained. All images obtained from Wikimedia Commons are provided as part of the dataset, converted to JPEG and with their resolution reduced so that the smaller side is at most 1280 pixels large. Links to the original images are provided but they might become unavailable or be replaced with newer versions. For checking the latter, SHA1 hashes are provided for the original images at the time of dataset creation.

IV. Collection Process

A. 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?

Church metadata was collected from Wikidata [6] by quering its SPARQL [7] API for all churches in Europe with at least one image and at least one architectural style.

The corresponding images were downloaded from Wikimedia Commons.

Bounding box annotations for distinctive architectural elements were created by a domain expert from art history.

B. 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.

Architectural style labels were provided on Wikidata by volunteers from the Wikipedia community.

We assessed the amount of noise in these labels by asking an expert architect to verify the labels of 200 random images from the dataset. About 93.5% of the images were found to be at least partially correctly labeled.

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

All data about European churches available at the time of July 12th, 2017, was obtained.

D. 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 expert from art history who created the bounding box annotations was a contractor and compensated according to the German guidelines on the working conditions for graduate and students assistants.

E. 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.

The metadata about the churches and the images were obtained on July 12th, 2017. The bounding box annotation was completed in December 2017.

V. Data Preprocessing

A. Was any preprocessing/cleaning/labeling of the data done? If so, please provide a description.

All images obtained from Wikimedia Commons were converted to JPEG and resized so that their smaller side has a maximum size of 1280 pixels.

We identified and removed images showing the interior of a church instead of the exterior with the assistance of a pre-trained indoor-outdoor classifier [8], ranking all images by decreasing score for the prediction “indoor”. The images in this ranked list were verified as indoor images manually until the results of the classifier became reliable enough, so that we did not expect a significant amount of further indoor images in the rest of the ranked list.

In a second step, we browsed through all remaining images and removed those that were not photographs of the exterior of a church building. This includes close-ups of individual objects (wall sculptures, pictures, organs etc.), scans of historical drawings, city scenes where the church is just one of many buildings, heavily truncated and occluded buildings, and ruins of former churches.

These two cleaning steps resulted in the removal of over a thousand images. The metadata associated with churches for which not a single image remained was removed from the dataset as well.

B. 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.

The unresized images are not part of the dataset but can still be obtained from Wikimedia Commons. The original links to the files are provided with the dataset.

VI. Dataset Distribution

A. How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

The dataset is hosted on Zenodo, a public open-access hosting service for research data provided by CERN and funded by the European Commission. A DOI has been registered for the dataset.

Each file copy has two replicas located on different disk servers.

B. When will the dataset be released/first distributed? What license (if any) is it distributed under?

The dataset is licensed under the Creative Commons BY-SA 4.0 license and first released in August 2021.

C. Are there any copyrights on the data?

The copyright for the images in the dataset still belongs to the individual authors. Most images require attribution. License and author information is provided with the dataset.

D. Are there any fees or access/export restrictions?

No.

VII. Dataset Maintenance

A. Who is supporting/hosting/maintaining the dataset?

The dataset is hosted on Zenodo, which is provided by CERN which has existed since 1954 and currently has an experimental programme defined for the next 20+ years. CERN is a memory institution for High Energy Physics and renowned for its pioneering work in Open Access. Organisationally Zenodo is embedded in the IT Department, Collaboration Devices and Applications Group, Digital Repositories Section (IT-CDA-DR).

Zenodo is furthermore funded by the European Commission.

B. Will the dataset be updated? If so, how often and by whom?

No new images or church metadata will be added to the dataset. We cannot exclude updates of style metadata or canonical splits, but they are not planned on a regular basis.

C. How will updates be communicated? (e.g., mailing list, GitHub)

A potential update will be registered as a new version on Zenodo.

D. If the dataset becomes obsolete how will this be communicated?

It will not.

E. Is there a repository to link to any/all papers/systems that use this dataset?

No.

F. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

Since WikiChurches is distributed under a Creative Commons license, anyone is welcome to extend and augment it. Any such modification must be licensed under a compatible permissive license as well.

VIII. Legal and Ethical Considerations

A. 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.

No.

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

No.

C. 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.

D. Does the dataset relate to people?

No.

References

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[2] Jong-Chyi Su and Subhransu Maji. The semi-supervised iNaturalist challenge at the FGVC8 workshop. *arXiv preprint arXiv:2106.01364*, 2021.

[3] Riccardo de Lutio, Damon Little, Barbara Ambrose, and Serge Belongie. The Herbarium 2021 Half-Earth challenge dataset. *arXiv preprint arXiv:2105.13808*, 2021.

[4] Robert-Jan Bruintjes, Attila Lengyel, Marcos Baptista Rios, Osman Semih Kayhan, and Jan van Gemert. VIPriors 1: Visual inductive priors for data-efficient deep learning challenges. *arXiv preprint arXiv:2103.03768*, 2021.

[5] Clemens-Alexander Brust, Björn Barz, and Joachim Denzler. Making every label count: Handling semantic imprecision by integrating domain knowledge. In *International Conference on Pattern Recognition (ICPR) 2020*, 2021.

[6] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, September 2014.

[7] Eric Prud'hommeaux and Andy Seaborne. SPARQL query language for RDF. W3C recommendation. `https://www.w3.org/TR/rdf-sparql-query/`, 2008.

[8] 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 (TPAMI)*, 2017.