Multilingual Image Corpus: Annotation Protocol

Svetla Koeva
Institute for Bulgarian Language "Prof. L. Andreychin"
Bulgarian Academy of Sciences
svetla@dcl.bas.bg

Abstract

In this paper, we present work in progress aimed at the development of a new image dataset with annotated objects. The Multilingual Image Corpus consists of an ontology of visual objects (based on WordNet) and a collection of thematically related images annotated with segmentation masks and object classes. We identified 277 dominant classes and 1,037 parent and attribute classes, and grouped them into 10 thematic domains such as sport, medicine, education, food, security, etc. For the selected classes a large-scale web image search is being conducted in order to compile a substantial collection of high-quality copyright free images. The focus of the paper is the annotation protocol which we established to facilitate the annotation process: the Ontology of visual objects and the conventions for image selection and for object segmentation. The dataset is designed both for image classification and object detection and for semantic segmentation. In addition, the object annotations will be supplied with multilingual descriptions by using freely available wordnets.

1 Introduction

We are surrounded by information represented by text, images and video data in multimodal streams. One of the processing tasks for large multimodal data streams is automatic image description (image classification, object segmentation and classification), which is directly connected with the task of image search, as well as with the expansion of the scope for automatic question answering regarding images.

The goal of our project **Multilingual Image Corpus** (MIC 21)[1] is to provide a large image dataset with annotated objects and object descriptions in (at least) 20 European languages. The Multilingual Image Corpus consists of an ontology of visual objects (based on WordNet) and a collection of thematically related images whose objects are annotated with segmentation masks and labels describing the ontology classes. The dataset is designed both for image classification and object detection and for semantic segmentation.

The main contributions of our work are: a) the provision of large collection of high-quality copyright free images; b) the formulation of the Ontology of visual objects based on WordNet noun hierarchies; c) the precise manual correction of automatic object segmentation within the images and the annotation of object classes; and d) the association of objects and images with extended multilingual descriptions based on WordNet inner- and interlingual relations.

We have divided the annotation process into four main stages: a) definition of an ontology of visual objects; b) collection of appropriate images; c) automatic object segmentation; and d) manual correction of object segmentation and manual classification of objects. The annotation protocol includes the Ontology of visual objects and the conventions for image selection and for object segmentation.

The focus of the paper is the annotation protocol which is established to facilitate the manual annotation. We begin with a brief overview of the current state in the field in Section 2. Section 3 is dedicated to the description of the Ontology of visual objects. Dataset collection is described briefly in Section 4. Section 5 provides an outline of the annotation protocol. Finally, conclusions and future directions of our work are presented.

We will show how the presented image dataset benefits from WordNet: providing ontological representation of visual objects based on WordNet noun hierarchies; building interconnectivity of classes by means of the WordNet relations; and ensuring multilinguality by using freely available wordnets.

[1] https://dcl.bas.bg/mic21/

Proceedings of Recent Advances in Natural Language Processing, pages 701–707
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_080

2 Related Work

There is a tradition already established in the image dataset collection and annotation; the available datasets show an increase both in the number of training images and in the number of object classes.

Caltech-256 dataset consists of 30,607 images and covers 256 object categories[2] (classes). The annotation includes bounding boxes, in which the objects are located, and object outlines provided by humans (Griffin et al., 2007). The categories are organised in a taxonomy grouping categories into animate and inanimate and other finer distinctions; for example, the category *electronics* is divided further into *entertainment, computing, home, office* and *others*.

The **CalTech 101 Silhouettes**[3] dataset consists of 4,100 training samples, 2,264 validation samples and 2,307 test samples. The dataset is based on CalTech 101 image annotations. Each image in the CalTech 101 Silhouettes dataset includes a high-quality polygon outline of the primary object in the scene (Marlin et al., 2010).

The **TinyImages** dataset (Torralba et al., 2008) is a large dataset containing 80 million small images (32 x 32 pixels) automatically collected from the Internet using 53,464 nouns from WordNet as queries. The dataset is not available online since it has not been manually evaluated[4].

The **Scene UNderstanding (SUN)**[5] collection contains 899 categories and 130,519 images. SUN annotates scene types and the objects that commonly occur in them. There are 397 categories designed to evaluate numerous state-of-the-art algorithms for scene recognition (Xiao et al., 2010). The **SUN Attribute**[6] dataset consists of 14,340 images from 717 scene categories, and each category is annotated with a taxonomy of 102 attributes (Patterson et al., 2014).

ModaNet[7] is a dataset consisting of annotations of street fashion images. ModaNet provides multiple polygon annotations for each image. Each polygon is associated with a label from 13 meta fashion categories (*bag, belt, footwear, outer, dress*, etc.), where each meta category groups highly related categories to reduce the ambiguity in the annotation process (Zheng et al., 2018).

There are several datasets which have been widely used as a benchmark for object detection, semantic segmentation and classification tasks.

The **PASCAL Visual Object Classes (VOC) 2012**[8] dataset contains 20 object categories including *vehicles, household, animals*, and others: *aeroplane, bicycle, boat*, etc. Each image has pixel-level segmentation annotations, bounding box annotations, and object class annotations (Everingham et al., 2010).

LabelMe is a dynamically developing dataset[9] which contains hundreds of thousands of polygon annotations, thousands of static images and sequence frames with at least one labelled object (Russell et al., 2008). A particular feature of this collection is that it is being developed by users who can add images and categories and can annotated uploaded images. This option however may result in some level of inconsistency based on the decisions of the different users about the annotation protocol. The WordNet noun synonymous sets (synsets) are used to extend the categories, to avoid the inconsistency by means of manual editing and to unify the descriptions provided by different users.

One of the collections that set standards in the increase of datasets sizes is **ImageNet**[10]. The aim is for a the dataset with about 50 million cleanly labelled full resolution images (Deng et al., 2009). Another important feature of this dataset is that it uses WordNet noun hierarchies for image collection and labelling. ImageNet uses 21,841 synsets and contains 14,197,122 annotated images organised by the semantic hierarchy of WordNet (as of August 2014) (Russakovsky et al., 2015).

The taxonomic organisation of nouns in WordNet allows for using more abstract and fine-grained categories when describing objects. WordNet is a semantic network whose nodes host synonyms denoting different concepts and whose arcs, connecting the nodes, encode different types of relations (semantic: genus-kind, part-whole, etc.; extralinguistic: membership in a thematic domain; inter-language: translation equivalents). The idea for organising the lexicon of a given language into a (lexico-)semantic network was first executed in the Princeton WordNet (Miller et al., 1990). Some of the fundamental ideas on which the WordNet

[2] https://www.kaggle.com/jessicali9530/caltech256
[3] https://people.cs.umass.edu/ marlin/data.shtml
[4] https://groups.csail.mit.edu/vision/TinyImages/
[5] https://vision.princeton.edu/projects/2010/SUN/
[6] https://cs.brown.edu/ gmpatter/sunattributes.html
[7] https://github.com/eBay/modanet
[8] http://host.robots.ox.ac.uk/pascal/VOC/
[9] http://labelme.csail.mit.edu/Release3.0/
[10] https://www.image-net.org

is based encompass: a) the use of a semantic network which embraces taxonomies, meronomies and non-hierarchical relations with clearly defined properties which allow for quick and easy automatic processing; b) a different organisation of the lexicon in comparison with the traditional dictionaries where words are ordered alphabetically and the links among semantically related words (such as between sister hyponyms, between a whole and its parts, etc.) are not explicitly presented (Miller, 1986).

The **COCO (Microsoft Common Objects in Context)** dataset (Lin et al., 2014) contains more than 328,000 images with manually annotated object instances (2.5 million)[11]. The dataset has had several releases since 2014 and it addresses object detection, segmentation, keypoint detection and captioning. The different parts of the dataset are annotated with bounding boxes (for object detection) and per instance segmentation masks with 80 object categories; natural language descriptions of the images; keypoints (17 possible keypoints, such as *left eye, nose*); per pixel segmentation masks with 91 stuff categories, such as *grass, wall*; full scene segmentation, with 80 thing categories (such as *person, bicycle, elephant*); dense pose – each labelled person is annotated with a mapping between image pixels and a template 3D model.

The image processing is generally classified as model based (using manually-labelled training data) and search based (using automatically collected training data). The search based approaches might include: effective learning mechanisms for matching a given query (Li and Fei-Fei, 2010); methods for automatic removing of noisy images (Hua and Li, 2015); frameworks combining discovering of multiple textual queries, filtering of noisy textual queries and noisy images (Anvari and Athitsos, 2019; Yao et al., 2020).

In the largest collection of datasets available on the internet 1,455 image datasets are listed[12] (as of August 2021) provided with descriptions and links to the sources and related papers. Among them 134 datasets are designed for semantic segmentation; 104 – for image classification and 102 – for object detection. Ten datasets provide polygon annotations.

To summarise, the tendency in image annotation is from small training datasets to large-scale collections which require crowdsourcing in order to engage a large amount of human effort. Although the number and the diversity of image datasets is constantly expanding still there is a huge demand for more datasets in terms of variety of domains and object classes covered.

3 Ontology of Visual Objects

In current practice, WordNet is usually used in generating text queries for creation of search based image collections. A Visual Concept Ontology is proposed which organises visual concepts (objects or abstract notions that are typically depicted in photos) (Botorek et al., 2014). For the construction of Visual Concept Ontology over 400 "significant" noun synsets (that have at least 300 hyponyms) are extracted from WordNet, then synsets with very "general" meaning such as *entity* or *thing* were removed. This results in 14 top-level ontology classes, which are divided further into 90 more specific classes. Semantically similar synsets are merged into a common class and additional links are established between semantically related synsets such as *roof* and *house*.

We identified 10 thematic domains: Sport, Medicine, Arts, Education, Food, Transport, Clothing, Security, Indoors, Nature. The proposed Ontology includes concepts which are particular for these domains.

Following the strategy for category selection of the ImageNet we applied the rule for no overlapping between the classes: "for any synsets i and j, i is not an ancestor of j" (Deng et al., 2009). Mutually exclusive classes are also defined for other well-known datasets, for example for the COCO thing and stuff classes (Caesar et al., 2018). As it was pointed out, the mutual inclusion might lead to some inconsistencies. An example was given with the PASCAL Context (Mottaghi et al., 2014) classes *bridge* and *footbridge*, which are in a parent-child relation (Caesar et al., 2018). The parent term can replace the child term in some context, but not vice versa, thus: if two images are annotated as *bridge* and *footbridge* respectively, it will not be known whether the parent concept can refer also to the child concept or not.

The Ontology of visual objects has the following components:

Classes which can be represented by visual objects and correspond to the respective WordNet concepts. Among the classes we made a differ-

[11] https://cocodataset.org

[12] https://paperswithcode.com/datasets

entiation between dominant classes and attribute (contextual) classes.

Each thematic domain is represented by several **dominant classes**, which show the main "players" within this domain differentiated by their type or their function. For example, the dominant classes for the domain Sport are: *skier, cricket player, hockey player, volleyball player, swimmer, oarsman*, etc., altogether 31 dominant classes. For the definition of the dominant classes, we use the WordNet sister hyponyms at a certain level (the lowest level allowing classification without specific knowledge for the domain). So far, the selected dominant classes for all thematic domains in focus are 277.

For each dominant class a parent class is selected from the WordNet noun hierarchies and this procedure is repeated consecutively up to the final class which represents a visual object. For example, classes like *basketball player, acrobat, football player*, etc. are hyponyms of *athlete* 'a person trained to compete in sports'. *Athlete* in its turn is a hyponym of *contestant* 'a person who participates in competitions' which is a hyponym of *person*. However, the hypernym of *person* is *organism*, an abstract notion, which is not included in the ontology. As a result of this approach, thousands of annotations will be assigned to objects representing small number of classes, while the annotations with more general classes will be inherited automatically.

Attributes in the ontology are classes related with the dominant ones. The type of the dominant class and the type of attribute class determine the type of the relation between them which expresses the specificity of property attribution: has instrument, wears, uses, has part, etc. For example, the attribute classes for *cricketer* are *cricket bat, cricket ball, cricket helmet, wicket* and *referee*, while for *climber – climbing helmet, chalk bag, claiming backpack*, and so on.

Relations between dominant and attribute classes are not hierarchical. For the definition of attribute classes, we use WordNet relations such as meronymy and morpho-semantic relations between nouns. In many cases, such relations are not overtly established in WordNet and they were additionally inserted in the Ontology.

Finally, we made some evaluation tests for all selected classes with other sources providing lists with concrete objects, such as concreteness ratings (Brysbaert et al., 2018) and acquisition ratings of words (in our case of nouns) (Kuperman et al., 2012). So far, we have identified 1,037 classes grouped in ten thematic domains: Sport, Medicine, Arts, Education, Food, Transport, Clothing, Security, Indoors, Nature.

The **relations** used in the Ontology are relations between classes. Part of the relations and their properties are inherited form WordNet. Additional relations are included in the ontology in case they are not explicit in the WordNet structure. Each class in the Ontology is represented by a unique label, which in most cases is one of the synonyms in the corresponding WordNet synset (in case of ambiguity, a descriptive label is constructed).

Benefits of using an ontology for image labelling can be outlined as follows:

- Selection of mutually exclusive classes.
- Build-in interconnectivity of classes by means of formal relations.
- Easy extension of the proposed ontology with more concepts corresponding with visual objects.

What it more, since wordnets for many languages are linked to Princeton WordNet (Bond et al., 2016), we will provide multilingual descriptions of the images. Freely available wordnets[13] with various lexical coverage for 17 official EU languages (Bulgarian, Croatian, Danish, Dutch, English, Finnish, French, Greek, Italian, Lithuanian, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish, Swedish) and for Albanian, Icelandic, Hebrew and Serbian will be used for a multilingual representation of the selected classes.

4 Image Collection

There are many repositories that can be used for searching and downloading images. Some of the images are assigned with multiple labels or short descriptions, which is used to facilitate the automatic collection of appropriate images. For the selected classes a focused web image search is being conducted to compile a database with images — candidates for annotation. So far, more than 450,000 images were collected from different image providers, which are selected on the basis of the following criteria: the repositories should offer an API and images should be licensed with one of

[13] http://compling.hss.ntu.edu.sg/omw/

the following standards: Universal Public Domain Dedication (CC0 1.0); Attribution 4.0 International (CC BY 4.0) and Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)[14]. Thus, we are avoiding copyrighted material, which might limit the use of our dataset only for academic purposes.

After the collection of images, we perform additional manual selection to ensure the quality of the dataset. The following criteria for selection are observed:

– The image has to contain a clearly presented object described by a given dominant class.

– The object should not (preferably) have occluded parts. If there are occluded parts of the object, they should not be essential for its recognition.

– The target object should be in its usual environment and in a position or use that is normal for its activity or purpose (for example, images in which a skier drinks a beer are not selected).

– The target object should be represented with its inherent attributes (for example, images of a man with wings are not selected).

– The target object should be represented in different positions, photographed from diverse viewpoints and angles and the object background should vary to a sufficient degree (for example, images of a chess player which slightly differ from one another are not selected).

– The instances of the target object should not represent one and the same person, animal or artefact.

– (Preferably) images with up to 10 objects are selected (the objects can belong to different classes or can be instances of one and the same class). If there are images with only one object, then it should be the dominant one.

– Images with small objects, unfocused objects in the background or images with a low quality (low resolution; blurriness caused by an out-of-focus lens, low illumination level, etc.) are not selected.

– Images which represent collages of photos, drawings or are post-processed are not selected.

The final selection of images is triple checked independently by different experts: after the automatic collection, after the automatic generation of segmentation masks and after the manual annotation: correction of the segmentation masks, new polygon outlines and selection of appropriate classes.

[14]https://search.creativecommons.org

5 Annotation Conventions

Our aim is to provide a dataset that will support image classification, instance segmentation and object detection formats. Several open source tools for image annotation (Makes Sense[15], COCO Annotator[16], VGG Image Annotator[17], LabelMe[18], LabelImg[19], etc.) have been evaluated in order to choose the most appropriate one for our purposes. Each annotation tool is usually designed for a specific application and for a specific annotation process. For example, we experimented with the web-based image annotation tool LabelMe to create polygon annotations; with the desktop annotation tool LabelImg to create bounding boxes, etc. To avoid converting annotations for frameworks such as YOLACT[20], DETECTRON[21], etc., which provide segmentation masks and require COCO formatted annotations, we decided to work with the COCO annotator[22]. The COCO Annotator can be containerised, allows for simultaneous work on a project, and offers useful functions that facilitate image annotation: tracking object instances, labelling objects with disconnected visible parts, etc.

It is a known fact that semi-automatic annotation approaches can significantly speed up the annotation process by automatic generation of annotation proposals to support the annotators. The main idea is to reduce the human interaction with the annotation tool and to save time, while maintaining the quality of the annotations. We experimented with Mask R-CNN and YOLACT, which provides instance segmentation on datasets like COCO. Mask R-CNN (He et al., 2017) is an implementation based on Python 3, Keras and TensorFlow. The model generates bounding boxes and segmentation masks for each instance of an object in the image. YOLACT (Bolya et al., 2020) is a framework, which breaks up instance segmentation into two parallel tasks: a) generating a dictionary of non-local prototype masks over the entire image, and b) predicting a set of linear combination coefficients

[15]https://www.makesense.ai
[16]https://github.com/jsbroks/coco-annotator
[17]https://www.robots.ox.ac.uk/ vgg/software/via/
[18]http://labelme.csail.mit.edu/Release3.0/
[19]https://github.com/tzutalin/labelImg
[20]https://github.com/dbolya/yolact
[21]https://github.com/facebookresearch/Detectron
[22]https://github.com/jsbroks/coco-annotator

per instance. Since the number of prototype masks is independent of the number of classes (e.g., there can be more classes than prototypes), YOLACT learns a distributed representation in which each instance is segmented with a combination of prototypes that are shared across classes.

The task for the annotators is to outline polygons for individual objects in the image (either by approving or correcting the automatic segmentation or by creating new polygons) and to classify the objects against the classes from the predefined Ontology.

The annotation follows the following conventions (only the more significant ones are listed here):

– The predicted polygons are accepted or corrected (if necessary) so that they outline the object as well as possible. Every instance of the target object is provided with a polygon.

– All objects from the selected dominant class and attribute classes related with it are annotated with polygons (for example, the *tennis player* and the related objects *racket* and *tennis ball*; *chess player* and the related objects *chessman, chess board*, and *clock*. Other objects can be also annotated if they belong to the predefined Ontology of visual objects.

– Every polygon is required to be as close to the object outline as possible. There is not much information how the overlapping objects should be annotated. The bounding boxes that embrace the estimated extent of the object are not annotated due to the ambiguity and disagreement between the annotators (Lin et al., 2014). One possible solution is to annotate only the visible parts of the objects. We accepted the following conventions: If the objects are included in each other, both objects are annotated; If two objects overlap and the boundaries of the partially occluded object are clear, then the second one is annotated with an estimation for the occluded part (for example, a car behind a road sign); In case the occluded parts can not be determined unambiguously, they are not annotated.

– An object is not annotated if it cannot be recognised for various reasons or less than 10–20 percent of the object is visible.

– If the object can be additionally associated with a different class this is recorded within the metadata (for example, if the *climber* is not a *man* but a *boy, woman* or a *girl*).

The quality control is provided by a second annotator who validates the implementation of the conventions and discusses the quality with the annotation group weekly. If necessary, some of the images are re-annotated.

6 Conclusion and Future Work

The **Multilingual Image Corpus** will provide pixel-level annotations for the selected dominant classes and their parent and attribute classes in ten thematic domains, thus offering more data to train models specialised in object detection, segmentation and classification in these domains. The selected classes for annotation are organised in an Ontology of visual objects that provides options to organise annotated images in different datasets regarding the envisaged tasks.

The Multilingual Image Corpus will be released in autumn of 2021 and will provide: a) a large number of copyright-free images, b) a large number of object classes organised in an ontology, c) a large number of pixel-level annotations; and d) extended image descriptions in (at least) 20 languages based on WordNet. An important result with great significance for the development of different applications for image processing will be the open distribution of the collection.

We are currently planning some experiments with a set of state-of-the-art algorithms on each of the tasks of object detection and segmentation, in order to establish a common baseline for future work.

7 Acknowledgments

The **Multilingual Image Corpus** (MIC 21) project was supported by the European Language Grid project through its open call for pilot projects. The European Language Grid project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 825627 (ELG).

References

Zahra Anvari and Vassilis Athitsos. 2019. A pipeline for automated face dataset creation from unlabeled images. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, page 227–235.

Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. 2020. YOLACT++: Better real-time instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Francis Bond, Piek Vossen, John McCrae, and Christiane Fellbaum. 2016. CILI: the collaborative interlingual index. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 50–57, Bucharest, Romania. Global Wordnet Association.

Jan Botorek, Petra Budíková, and Pavel Zezula. 2014. Visual concept ontology for image annotations. *CoRR*, abs/1412.6082.

Marc Brysbaert, Amy B. Warriner, and Victor Kuperman. 2018. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46:904–911.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2018. COCO-stuff: Thing and stuff classes in context. In *Conference on Computer Vision and Pattern Recognition*, pages 1209–1218.

Jia Deng, Wei Dong, Socher Richard, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. page 248–255.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2010. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338.

Greg Griffin, Alex Holub, and Pietro Perona. 2007. Caltech-256 object category dataset. In *Technical Report 7694*, page 1–20, California Institute of Technology.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. *CoRR*, abs/1703.06870.

Xian-Sheng Hua and Jin Li. 2015. Prajna: Towards recognizing whatever you want from images without image labeling. In *AAAI International Conference on Artificial Intelligence*, page 137–144.

Victor Kuperman, H. Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44:978–990.

Li-Jia Li and Li Fei-Fei. 2010. Optimol: automatic online picture collection via incremental model learning. *International Journal of Computer Vision*, 88(2):147–168.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, Zürich.

Benjamin M. Marlin, Kevin Swersky, Bo Chen, and Nando de Freitas. 2010. Inductive principles for restricted boltzmann machine learning. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, page 509–516.

George Miller. 1986. Dictionaries in the mind. *Language and Cognitive Processes*, 1:171–185.

George Miller, R. Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 2014. The role of context for object detection and semantic segmentation in the wild. In *Conference on Computer Vision and Pattern Recognition*, pages 891–898.

Genevieve Patterson, Chen Xu, Hang Su, and James Hays. 2014. The SUN attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1-2):59–81.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 116:157–173.

Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. 2008. LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173.

Antonio Torralba, Rob Fergus, and William T. Freeman. 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition. *Transactions on Pattern Analysis and Machine Intelligence*, 30:1958–1970.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,*, pages 3485–3492.

Yazhou Yao, Jian Zhang, Fumin Shen, Li Liu, Fan Zhu, Dongxiang Zhang, and Heng Tao Shen. 2020. Towards automatic construction of diverse, high-quality image datasets. *IEEE Transactions on Knowledge and Data Engineering*, 32(6):1199–1211.

Shuai Zheng, Fan Yang, M. Hadi Kiapour, and Robinson Piramuthu. 2018. ModaNet: A large-scale street fashion dataset with polygon annotations. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1670–1678.