Exploring Naming Inventories for Architectural Elements for Use in Multi-modal Machine Learning Applications*

Ronja Utescher[1,3], Aaron Pattee[2], Ferdinand Maiwald[1], Jonas Bruschke[4], Stephan Hoppe[2], Sander Münster[1], Florian Niebling[4] and Sina Zarrieß[3,*,†]

[1]Friedrich-Schiller-University Jena, Fürstengraben 1, 07743 Jena, Germany
[2]Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, 80539 Munich, Germany
[3]Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany
[4]Julius-Maximilians-University of Würzburg, Sanderring 2, 97070 Würzburg, Germany

Abstract
Computer vision models are increasingly relevant and useful to Digital History. Next to the increasingly complex neural models, data and data selection are an integral part of this process. In this paper, we examine and extend the data collection practices from a major recent paper in the domain of architectural element classification. We collected an image-text data set for a selection of 56 Baroque landmarks to be analysed in like manner. This different architectural domain yielded insights into the transferability of the original model and data collection procedures. Notably, the architectural domain also has an impact on the availability of classes of architectural elements as well as the performance of the models classifying them.

Keywords
Architecture, Art History, Computer Vision, Machine Learning

1. Introduction

The study of architectural art history has greatly benefited from innovative, computer-aided approaches in recent years. From high-resolution two-dimensional (2D) photos of building edifices, to three-dimensional (3D) models of entire structures, these emerging techniques are laying the foundation for new methodologies in researching architecture [1, 2]. Provided the three-dimensional nature of buildings, research projects have appropriately focused on techniques that produce digital replicas of their forms, such as Structure-from-Motion (SfM) Photogrammetry and Terrestrial Laser Scanning (TLS) [3]. However, one critical aspect in the study of architecture has largely been dormant since the emergence of these technologies, namely, the computer-aided description of architectural elements. 2D images and 3D models were the logical first steps in devising new methodologies for identifying architectural elements, providing precise calculations of their dimensions and structures, buttressed by the unique

COMHUM 2022: Workshop on Computational Methods in the Humanities, June 09–10, 2022, Lausanne, Switzerland
*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.
*Corresponding author.

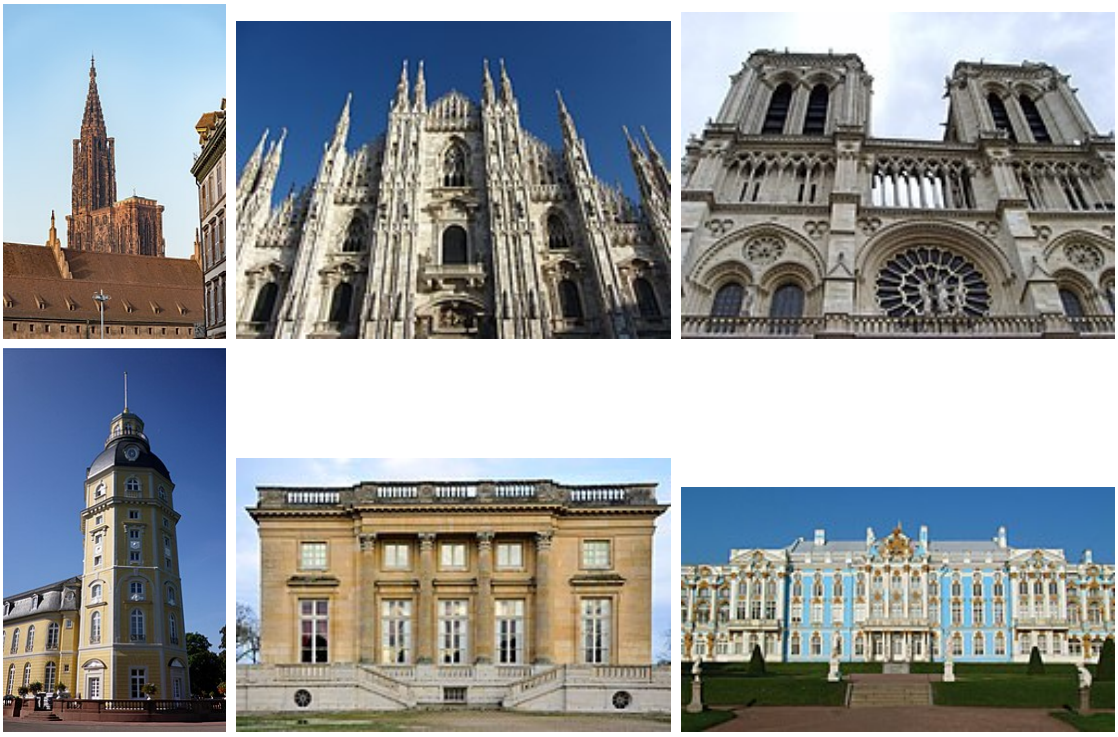

Figure 1: Examples from the *facade* and *tower* classes in Wikiscenes (top) and Wikiscenes-Baroque (bottom)

capability to virtually study a building. What normally follows is a traditional text description of the discoveries achieved by the employment of these techniques, relegating these digital applications as mere means to a more enlightened end. The lamentable result is a digital purgatory of 3D models awaiting their fate in repositories or online databases.

This paper presents one aspect of a larger project seeking to utilise 2D images and 3D models as essential components of a search engine for architectural elements. These digital objects can serve as reference points for future research in architectural art history and archaeology. What is required is a systematic identification of the elements themselves using text descriptors, in order that the digital representations of the elements can be efficiently explored. In many ways, this avenue of research is an evolution of [4], which had expert and non-expert annotators choose phrases from longer textual descriptions with which to index paintings in a collection. For this purpose, we implement the methodology of recent research [5] as the foundation of the link between text and digital representation. The work by [5] represents an important step in integrating 3D models, collections of photographs, and written descriptions of a historic building using state-of-the-art machine learning methods. Given a multi-modal collection of cathedrals, the model learns to detect and classify 10 classes of architectural objects, for example *portals* and *columns*. These classes of objects are however limited by which terms are frequently associated with images in the original source.

In this paper, we take a closer look at the first two steps in their workflow [5]; (1) curating

images and their descriptions from large, open-source collections and (2) selecting the vocabulary of architectural elements for the machine-learning model to classify. We argue that these are important design decisions that have ramifications for the output of the model down the line. Both [5] and the authors of this paper source their images from Wikimedia Commons. While Commons is a free and abundant source of images, the indexing and naming it provides for individual images is comparatively limited and unsystematic.

Our case study, similar to [5], is abstract as it is neither limited to a specific building, nor a design implemented by a specific architect. Rather, it is a collection of baroque monumental buildings largely built between the late 17th and mid-18th centuries, ranging geographically from Portugal to Russia, by a large network of architects. In effect, we construct a new collection which parallels the one introduced in [5]. The majority of the collection consists of 2D images, though this could be supplemented in future research by existing high-resolution TLS models of historic buildings in the German city of Dresden, such as the iconic Zwinger.

For the purposes of this paper, we define a domain according to architectural style. [5] do not explicitly frame the issue this way, but we will briefly talk about the issue here since it is important from a historical and architectural perspective. Wikiscenes and WikiscenesBaroque differ both in architectural style (gothic/baroque) and building function (house of worship/residence of high-ranking dignitaries). [5] define it more so by function, although the Cathedrals lean towards the Gothic style. The paper is structured as follows: In section 2, we discuss the method of the paper, including the Wikiscenes data curation policy and classification model as well as our modifications to said policy. In section 3, we examine the resulting new dataset, *WikiscenesBaroque*. Section 4 details the setup and results of the experiments we conducted in order to assess the influence of this different data on the classification model. Finally, Section 5 summarises the different lessons learnt from the case study.

2. Background

Our goal is to build a machine learning-based system that labels architectural elements in images of buildings, of particular architectural styles assumed in the study of architectural art

Description: The northwestern facade of the Leopold Wing of the Hofburg Imperial Palace in Vienna.

Figure 2: Image with its Wikimedia Commons category hierarchy (left) and text description (bottom). The gold label *facade* is sourced from the description.

history. In Section 2.1, we take a brief look at the uses and requirements of machine learning methods in art history. Sections 2.2 and 2.3 describe [5]'s and our approach respectively.

2.1. Image Classification for Architectural Categories

The documentation of historical built works often goes hand in hand with large collections of image materials, especially for landmarks which are popular objects of study. Machine Learning (ML) methods have the potential to greatly benefit research of these historical landmarks, since they allow the automatic processing of large amounts of data. These ML applications can take the form of labelling data according to predefined categories, or in interactive settings like Image Retrieval. This paper focuses on modelling and data collection in classification settings.

Image and Object Classification models have undergone significant development in the last 10 years. Although there have been a number of new models for Image Classification and other Vision tasks, most models in image processing use a convolutional neural network (CNN) as a backbone (cf [6, 7, 8]. Image classification models generally work with a set classification vocabulary, requiring labelled training data. Datasets in digital history are smaller and more difficult to obtain. There are standard vocabularies for categorising architectural elements, but these are not connected to the datasets. Domain specificity in classification models runs the gamut from models trained on generic vocabularies [9, 10, 11], to domain-specifically trained or fine-tuned models, to models trained on specific landmarks. We use an approach that does not rely on NLP, but on language in terms of the vocabulary which we use to talk about real-world objects and their visual representations.

The method for creating these reconstructions is unsupervised except for the previously mentioned choice in input images for each model. COLMAP also creates a match between the 3D space of the model and the 2D images. [5] exploited these matches in their image classification and segmentation, using a loss function which rewards the model for consistent classification of points in the landmark across different images.

2.2. Landmarks and Image Categories in Wikiscenes

[5] use a combination of automatic data collection and manual refinement which utilises noisy, but freely available data. The authors mine this data for semantic concepts which act as classes for classification and segmentation models based on [12]. This approach bypasses the need for costly manual annotation and exploits the inherent structure of the data.

[5] draw upon Wikimedia Commons as the source of image-and-text data. Any interested party can contribute images, add them to the appropriate Wikimedia Commons (WM) category. This serves as an alternative to other annotation paradigms, such as expert annotation or crowd-sourced annotation. There are ways in which it is noisy; if users utilise the caption and description fields at all, the content and length tends to vary wildly. These user-based annotations are more sophisticated than what could reasonably be collected by laypeople in a crowdworking setting. In this paper, we aim to go into detail about the curation process of a subset of Wikimedia Commons for a number of Baroque Architectural Objects. Wikimedia Commons provides us with a large number of user-uploaded images with open-source licensing. Users are given the opportunity to annotate their images with captions, descriptions, and a

selection of other metadata such as the geolocation and camera specifications.

2.3. AE classes in WikiScenes

The classes in the original dataset are decided upon bottom-up; if there is an architectural element that has a significant number of instances available for training, it is selected as one of the classes for the model. In other words, [5] compute the frequency of all terms from categories, descriptions and captions and manually select a number of architectural element classes from the most frequent terms. Figure 1 showcases examples of two of these classes, *facade* and *tower*.

[5] use image descriptions as well as Wikimedia Commons categories. On Wikimedia Commons, users can add descriptions or captions to their images. However, only a small portion of images have captions/descriptions. This makes *Category* pages the most comprehensive source for text describing images. The images of said landmark are organised into subcategories, sub-subcategories, and so on. Figure 2 shows an example of a Wikimedia Commons Image and its category tree. In this example, the *facade* label can be sourced from the image description. As described in [5], the Wikimedia commons categories themselves are a significant source for concept terms; if a term is used in a category name, all direct member images of the category can be assigned the term as an architectural element (AE) class.

3. Dataset

The original WikiScenes dataset [5] contains data for 99 Gothic cathedrals. We build an analogous dataset of 56 Baroque landmarks, mostly palaces. We investigate how [5] approach transfers to modelling architectural elements in a different domain.

Table 1
Numbers of instances in the train and test sets for WikiscenesBaroque

	garden	facade	hall	statue	court
total	4291 (0.23)	3822 (0.2)	3344 (0.18)	2208 (0.12)	2064 (0.11)
train	3844 (0.23)	3548 (0.21)	2980 (0.18)	1969 (0.12)	1871 (0.11)
test	447 (0.24)	274 (0.15)	364 (0.2)	239 (0.13)	193 (0.1)

	court	stair	gallery	tower	column	fountain
total	2064 (0.11)	929 (0.05)	712 (0.04)	657 (0.04)	375 (0.02)	318 (0.02)
train	1871 (0.11)	829 (0.05)	645 (0.04)	580 (0.03)	323 (0.02)	271 (0.02)
test	193 (0.1)	100 (0.05)	67 (0.04)	77 (0.04)	52 (0.03)	48 (0.03)

3.1. Curating Landmarks and Image Categories

[5] build two point cloud models per landmark; one for the outside and one for the inside of the landmark. The *inside* model is computed from images from the "Interior of [landmark]" category and its leaf categories, white the *outside* model uses the "Exterior of [landmark]" and "Views of [landmark]" categories. These three subcategories are present throughout the original dataset's landmarks, however we did not find this to be universal in our collection of Baroque

Table 2
Numbers of class instances in the original train set and in our cross-domain test set; for big sets, 1500 instances were randomly sampled for our test set

	facade	window	chapel	organ	nave
Wu et al. (train set)	1352 (.14)	874 (.09)	1050 (.11)	628 (.06)	1063 (.11)
cross-test	1500 (.42)	22 (.01)	191 (.05)	119 (.03)	0 (.0)
	tower	choir	portal	altar	statue
Wu et al. (train set)	932 (.10)	1029 (.11)	957 (.10)	956 (.10)	949 (.10)
cross-test	629 (.17)	0 (.0)	173 (.05)	0 (.0)	909 (.26)

landmarks. Out of the 56 landmarks, 22 only had one subcategory that matched the original inside/outside selection process. Overall, there were 38 "interior" categories, and 23 "exterior" (14 + 9 "view").

We selected 56 palaces based upon their architectural similarities to the Zwinger in Dresden, in which all exhibit building phases in the late 17th and early 18th centuries. Additionally, the architects and designers of the selection of buildings all had connections as part of a larger network in which ideas and designs were shared, as many of the major construction efforts occurred between 1700 and 1730 A.D. It was during this time that the Great Northern War raged in which the Baltic and Black Seas, as well as the entire east of Europe served as the battleground for the various armies. Many palatial architects were active military officers involved in constructing bastions and fortresses for and against artillery pieces. As a result, there was a major exchange of ideas and designs during this period which even translated into the construction plans of palaces.

Each landmark has its own hierarchy of categories in Wikimedia Commons. We use the set of all immediate subcategories - the landmark's category being the root - as basis for coming up with a list of terms to blacklist. For each of the 432 categories in this manual selection, we select all its subcategories as well (unless they are in the blacklist). We use the blacklist sparingly, i.e. including as many images as sensible in this step of the annotation. The blacklist is *stamp, in art, painting, collections, plans, aerial, panoramic, plans, signs, maps, things, history, events*, leading to the exclusion of 53 of 432 subcategories. This list is aimed at excluding objects which are not part of the landmark, but located in it (collections, paintings, signs) or associated with it (in art, stamp, things). Also, we exclude non-photographic images (plans, maps) and images with an atypical perspective (aerial, panoramic). Finally, the last two keywords in the blacklist exclude historical images and images depicting certain events This allows us to utilise a larger set of images compared to only collecting the images uploaded in the "Exterior", "Interior" and "Views of" categories This initial selection yields 85,629 images, while using [5]'s method would yield 28,794 images. From these images, we select instances of the AE classes for our cross-domain test set as well as our data for the Baroque-specific model (cf. Section 4).

3.2. Selecting AE Classes

We followed [5] and selected 10 AE classes from the most frequent terms present in the WM Categories and Descriptions. This led to a different set of classes than the original dataset, as shown in Table 3. Out of the 10 AE classes, only 3 (*facade, statue, tower*) are present in both models. Out of the available images, we were able to link 17115 with AE classes for training the model described in the *new domain* setting (Section 4.2).

In order to evaluate [5]'s original model, we also construct an alternate test set using instances of AE classes from the original Wikiscenes dataset. As listed in Table 2, the availability of instances in our Baroque domain is mixed. The classes *nave* and *choir* are entirely absent, while the data contains comparatively many *statues* and *facades*.

4. Experiments

In this section, we introduce the image classification model without 3D Loss from [5] and use it to perform experiments in two novel classification setups.

4.1. Model

The model described in [5] performs two general steps. 1) For each landmark, generate a 3D model from photographs and 2) assemble an architectural vocabulary for identifying elements of the landmark. The [5] model is based on [13]. The implementation uses a resnet50 backbone [14] with an ImageNet [15] pretraining. Their model's backbone extracts both low-level and high-level image features, which are unified in a pooling layer followed by a Global Cue Injection (GCI) module. In [5], the 3D loss provides a small performance boost (3.3% across all images). In this paper, we generally forego the use of 3D Loss as proposed by [5]. We argue that using the baseline classification model is adequate to judge the overall feasibility, since 3D does not fundamentally change the results, instead giving a small boost in accuracy overall.

Table 3
Image-level classification precision w/o 3D Loss on WikiscenesBaroque (WSB)

	garden	facade	hall	statue	court	stair	gallery	tower	column	fountain
precision	60.3	87.8	77.9	45.8	57.6	47.3	74.0	30.2	58.4	45.6

4.2. Classification Setups

We perform experiments in two classification setups. The *baroque domain* setting is analogous to [5]; the *cross-domain* setting tests the performance of the original model in images from a different architectural style.

Baroque Domain In the Baroque Domain task, we train a new image classification model without 3D loss, using our collection of Baroque Wikimedia Commons images. We use a batch size of 16 and 10 epochs for training the model. Like [5], we use a 9:1 split on landmark level, so that the images seen at test time belong to unseen landmarks.

Table 4
Image-level classification precision w/o 3D Loss (baseline) on known Wikiscenes landmarks (WS-K), unknown Wikiscenes landmarks (WS-U) and our cross-domain test set

Test Set	Model	mAP	mAP*	facade	window	chapel	organ
WS-K	baseline	70.8	77.7	87.2	89.2	60.2	89.7
WS-U	baseline	48.3	64.0	71.0	92.2	10.7	57.3
cross-test	baseline	44.2	60.0	88.9	50.8	20.7	30.5

Test Set	Model	nave	tower	choir	portal	altar	statue
WS-K	baseline	85.8	64.1	61.5	68.0	50.0	52.0
WS-U	baseline	71.0	53.4	43.6	31.1	25.8	27.1
cross-test	baseline	-	55.4	-	29.5	-	33.8

Cross-Domain In this experiment, we evaluate the performance of the original model by [5] on images of the landmarks in our data. We refer to this as the *cross-domain* setup because a model trained on Gothic cathedrals is used to classify images of Baroque palaces. We construct a *cross-test* set with instances from our 56 Baroque landmarks. In Table 2, we provide statistics on the distribution of classes in our data. Some classes had no instances or very few in the Baroque dataset, likely due to the domain itself. For the very common class *facade*, we randomly sampled 1500 instances for the test set. Analogously to [5], we evaluate by calculating the precision per class, as well as the mean average precision (mAP) over all instances and the mAP over all landmarks (mAP*).

4.3. Results

Baroque Domain Figure 3 shows some examples of correct and incorrect judgements of the model. Note that while (a) has the gold label *fountain* and there is a fountain in the right foreground of the image, most of the space in the image is taken up by the facade of the building, which is the predicted class of the model.

In Table 3, we report the precision of the model trained on WikiscenesBaroque (WSB) for each AE class. Similarly to Wikiscenes-baseline model (cf. Table 4), this model achieves a high precision for the class *facade*, but also for *hall* and *gallery*. Even though the WSB model is tested on unseen landmarks, its precision on *statues* is higher than the Wikiscenes baseline model (45.8% vs. 33.8%). The model performs worst for *tower* at 30.2%, and under 50% for *statue* (45.8%), *fountain 45.6%* and *stair* (47.3%).

Cross-domain In Table 4, we report the mean average precision (mAP), for each AE class as well as across all images. In [5], the baseline model achieves a higher overall mAP on known landmarks (WS-K) than on images of unknown landmarks (WS-U) with 70.8% vs. 48.3%. The mAP for the cross-test set is lower than for the WS-U test set, however the difference is much smaller (48.3% vs. 44.2%). In the cross-test set, the model performs worst for the *chapel* AE at 20.7%, which is substantially lower than the 60.2% mAP in the WS-K set, but higher than the 10.7% in the WS-U set. On the other hand, the baseline model performs consistently well on the *facade* class. For the *tower* and *portal* classes, the model performance is very similar in the cross-test and WS-U sets.

(a) Gold label: *fountain*, predicted: *facade*

(b) Gold label: *portal*, predicted: *tower*

(c) Gold label: *portal*, predicted: *portal*

(d) Gold label: *statue*, predicted: *statue*

Figure 3: Classification examples

5. Discussion

The automatic annotation of basic architectural elements is beneficial to the digital documentation of Historical Heritage Sites. Architectural classification models like [5] are adept at utilising large amounts of already existing, noisily annotated data, taking concepts from the text annotations and locating them in a 3D model of the landmarks. Approaches like these also require the attention of the researcher handling the model. In this paper, we have examined the various stages of data curation and their effect on the model's performance.

For more specific vocabularies, the text provided in Wikimedia Commons is not a suitable source for labelled data as there are comparatively few instances which are directly annotated. Training a model with more classes and more diverse training data is possible in principle. However, domain-specific models have the advantage of needing less computing power. Additionally, domain experts can advise during the data curation process and steer the model's AE classes towards what is of interest to them.

The results of our experiments suggest that classification and detection of architectural elements works best within domain, with a moderate gap to the cross-domain performance. Images of known landmarks which are seen during training appear to be overall easier to classify, comparing the WS-K to the WS-U and cross-test sets as listed in Table 4. Architectural elements and their visual styles can be highly domain-specific, or even very individualised to particular buildings. For well-documented landmarks, it may well be feasible to have classification models solely for a particular landmark. However, less-resourced landmarks can still benefit from more general models.

Besides under-resourced landmarks, we want to point out the issue of more fine-grained AE classes as a topic for future research. The AE classes which can be mined from resources like Wikimedia Commons are useful for basic classification of images and segmentation of 3D models. There are a great number of distinctions of AE classes within the study of architectural art history, many of which are catalogued in resources like the Art & Architecture Thesaurus [16]. Building Machine Learning models which can identify more elaborate concepts would open up more detailed analysis and documentation of the landmarks and would be an interesting

direction for future research.

6. Conclusion

In summary, this paper as well as [5] suggest that the practice of using online image collections, in particular Wikimedia Commons, yields data that can be used to train models which in turn classify architectural elements. We curate our own dataset, focusing on Baroque landmarks that were constructed in the late 17th and early 18th centuries. The experiments in Section 4 suggest that in our case study, model precision decreases for images of unknown landmarks, especially if they come from a stylistically different domain.

Acknowledgements

This work was supported by a grant from the Federal Ministry of Education and Research (BMBF, grant No. 01UG2120).

References

[1] J.-E. Lutteroth, S. Hoppe, Schloss Friedrichstein 2.0: von digitalen 3D-Modellen und dem Spinnen eines semantischen Graphen, in: Computing art reader: Einführung in die digitale Kunstgeschichte, 2018, pp. 184–198.

[2] P. Sapirstein, Accurate measurement with photogrammetry at large sites, Journal of Archaeological Science 66 (2016) 137–145.

[3] N. Lercari, Terrestrial laser scanning in the age of sensing, Digital methods and remote sensing in archaeology (2016) 3–33.

[4] R. J. Passonneau, T. Lippincott, T. Yano, J. L. Klavans, Relation between agreement measures on human labeling and machine learning performance: results from an art history domain, in: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), 2008.

[5] X. Wu, H. Averbuch-Elor, J. Sun, N. Snavely, Towers of babel: Combining images, language, and 3d geometry for learning multimodal vision, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 428–437.

[6] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014. URL: https://arxiv.org/abs/1409.1556. doi:10.48550/ARXIV.1409.1556.

[7] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. URL: https://arxiv.org/abs/1512.03385. doi:10.48550/ARXIV.1512.03385.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. URL: https://arxiv.org/abs/2103.00020. doi:10.48550/ARXIV.2103.00020.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

[10] H. Caesar, J. Uijlings, V. Ferrari, COCO-stuff: Thing and stuff classes in context, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[11] J. Redmon, A. Farhadi, Yolo9000: Better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[12] N. Araslanov, S. Roth, Single-stage semantic segmentation from image labels, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4253–4262.
[13] N. Araslanov, S. Roth, Single-stage semantic segmentation from image labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4253–4262.
[14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. URL: https://arxiv.org/abs/1512.03385. doi:10.48550/ARXIV.1512.03385.
[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR09, 2009.
[16] P. Harpring, Development of the Getty vocabularies: AAT, TGN, ULAN, and CONA, Art Documentation: Journal of the Art Libraries Society of North America 29 (2010) 67–72.