arXiv:2211.13435v1 [cs.CV] 24 Nov 2022

A Benchmark of Long-tailed Instance Segmentation with Noisy Labels (Short Version)

Guanlin Li
S-Lab, NTU
guanlin001@e.ntu.edu.sg

Guowen Xu
NTU
guowen.xu@ntu.edu.sg

Tianwei Zhang
NTU
tianwei.zhang@ntu.edu.sg

November 28, 2022

Abstract

In this paper, we consider the instance segmentation task on a long-tailed dataset, which contains label noise, i.e., some of the annotations are incorrect. There are two main reasons making this case realistic. First, datasets collected from real world usually obey a long-tailed distribution. Second, for instance segmentation datasets, as there are many instances in one image and some of them are tiny, it is easier to introduce noise into the annotations. Specifically, we propose a new dataset, which is a large vocabulary long-tailed dataset containing label noise for instance segmentation. Furthermore, we evaluate previous proposed instance segmentation algorithms on this dataset. The results indicate that the noise in the training dataset will hamper the model in learning rare categories and decrease the overall performance, and inspire us to explore more effective approaches to address this practical challenge. The code and dataset are available in https://github.com/GuanlinLee/Noisy-LVIS.

1 Introduction

Instance segmentation models [HGDG17, WZK+20] provide each pixel in an image a label to indicate the specific semantic information, which can be seen as a finer inference of object detection models [RHGS15, LAE+16] and semantic segmentation models [BKC17, CZP+18]. Most of previous works focus on exploring more effective algorithms on balanced and clean datasets, such as MS COCO [LMB+14]. However, instances included in images usually obey a long-tailed distribution [LGG+17, CWG+19, CJL+19], which means most of the instances belong to a small part of categories, which are called "head" classes, and few of instances belong to other categories, which are called "body" and "tail" classes. If users do not modify the training algorithms to fit an unbalanced dataset, the model will give higher confidence to the instances from head classes [LGG+17], which harms the generalizability to the instances from body and tail classes, causing accuracy decrease. On the other hand, the annotations for each item in the dataset are usually given by people who are paid by the big companies, like Google[1] and Meta[2]. In reality, these annotators are not always experts [WRW+09], as the companies wish to decrease the cost. So, there are usually some wrong labels in the dataset, called label noise. For example, in Clothing1M [XXY+15] the percentage of label noise is about 38.5% [SKP+22], and in WebVision [LWL+17] the label noise ratio is about 20.0% [SKP+22]. If users do not adopt any approach to mitigate the impact of the label noise, the accuracy of the model will significantly decrease on the clean test set, because the incorrect labels can provide wrong supervised information during the training process. So, there exists a gap which can cause performance drop when users train a model on a personal dataset, which is unbalanced or contains label noise. To address the challenges in training a model on long-tailed datasets or noisy datasets, many studies have been published, recently.

Similar to exploring how to train classifiers on unbalanced datasets, solving long-tailed challenges in the instance segmentation task raises the attention of the deep learning community. Different from the classification task on long-tailed datasets, backgrounds in the instance segmentation datasets always

[1] https://venturebeat.com/ai/probeat-google-still-needs-you-to-label-photos-for-its-ml/
[2] https://www.reuters.com/article/us-facebook-ai-focus-idUSKCN1SC01T

occupy an overwhelming position, and in foreground, there are head classes and tail classes, which means blindly adopting approaches in classification tasks cannot address the problem in instance segmentation. So, there are a series of new methods [HJT+20, HRCH21, WZZ+21] proposed in solving long-tailed instance segmentation task, including new data augmentation methods, new model architectures and new loss functions. These solutions can work well on long-tailed datasets, without considering the label noise. However, as aforementioned challenges in deep learning, noise labels are more popular in instance segmentation datasets, because there are more instances in the datasets and annotating them exhaustively and precisely is more difficult. Specifically, the noise in instance segmentation datasets source from two aspects, i.e., foreground instances missed annotations (false negative background) and foreground instances mis-annotated into other classes (noisy label). In this paper, we focus on the latter one, which is a more harmful type of noise in the dataset. There are numerous studies focusing on addressing the noisy label problem in classification tasks, such as new model structures [SF15] and new loss functions [GKS17]. And Yang at el. [YML+20] propose a two-stage training fashion to overcome the label noise in the instance segmentation task. On the balanced dataset containing label noise, adopting such methods can successfully overcome the accuracy drop. However, they do not consider a more practical setting, i.e., a long-tailed noisy dataset.

Recently, to fill the gap and make a connection between long-tailed recognition and noisy label learning, several works [WSTL21, KRC21] combine long-tailed learning methods and technologies for solving label noise. The results indicate that this direct solution is helpful to train a model on an unbalanced dataset containing label noise. However, in the instance segmentation task, background objects and foreground objects are required to be handled separately, which is different from the image recognition task, causing that previous methods for long-tailed recognition and noisy label learning cannot be applied to this task directly. On the other hand, we find a lack of relevant datasets, which are both long-tailed and noisy, in the instance segmentation task. It brings the difficulties for other researchers to explore approaches to address these practical challenges. So, in this paper, we attempt to introduce a benchmark for the instance segmentation task on long-tailed and noisy datasets.

Formally, there are two main contributions in this paper, i.e., introducing a new long-tailed noisy dataset and systematically evaluating previously proposed methods on this dataset. Specifically, we build a noisy long-tailed datasets from the LVIS v1 [GDG19], which is a large vocabulary long-tailed instance segmentation dataset containing 1,203 different categories for the foreground objects. We will introduce the details of how we add noisy labels into the LVIS v1 in Section 3. For our second contribution, we evaluate EQL [TWL+20], DropLoss [HRCH21] and Seesaw Loss [WZZ+21] on this new dataset under different noise settings and samplers in Section 4. More evaluation results will be added in the future. From the results under different noise settings and samplers, we find that previous proposed long-tailed instance segmentation losses cannot keep robust under label noise and data sampling methods have a significant impact on the final results, due to distribution shift caused by the label noise. We hope this benchmark and the new proposed dataset can shrink the gap between ideal experimental settings and practical datasets and bring guides for researchers in further exploration.

2 Related Works

2.1 LVIS v1

LVIS v1 is a dataset for Large Vocabulary Instance Segmentation, derived from the Microsoft Common Objects in Context (MS COCO) [LMB+14] with much more fine-grained annotations, which means that the images in LVIS v1 are the same as them in the MS COCO, but the categories of each instance in the images are more precise. Specifically, there are 1,203 different classes in LVIS v1, and in MS COCO, there are only 80 classes. As a large vocabulary dataset, the evaluation process must be accurate. To achieve it, the LVIS v1 adopts positive sets and negative sets for each category[3]. As for the categories in LVIS v1, each of them is a WordNet [Mil95] *synset*, i.e., a word sense specified by a set of synonyms and a definition. Ideally, we can treat the LVIS v1 as an exhaustive and clean long-tailed dataset. And, we will explore the previous long-tailed segmentation methods trained on a noisy dataset, i.e., adding symmetry and asymmetry noise labels into the training set of the LVIS v1.

[3]This can cause that some instances are not annotated and are treated as background. In this paper, we do not consider the false negative samples' influence in the training procedure.

2.2 Long-tailed Instance Segmentation

Compared with the long-tailed object classification task [RYS+20, MJR+21], the long-tailed instance segmentation task is much closer to the long-tailed object detection task [TWL+20]. The reason is that the object detection task can be seen as a sub-task in the instance segmentation task, as some instance segmentation pipeline first generates bounding boxes for each object and then generates masks inside the bounding boxes. Because instance segmentation models need to recognize backgrounds and foregrounds, and generate masks for each foreground object, background-foreground classification plays a key role. However, backgrounds occupy an overwhelming number in the dataset, which means that simply adopting solutions proposed in long-tailed object classification tasks cannot address the background-foreground classification challenges. In long-tailed instance segmentation tasks, there are several new proposed approaches, such as gradient calibration [TWL+20, WZZ+21], two-stage training [WLK+20], data augmentation [GCS+21], and new model structures [ZHL21]. In this paper, we mainly consider evaluating loss functions designed for long-tailed instance segmentation tasks.

2.3 Noisy Label Learning

Various approaches have been proposed to address the label noise from different perspectives, including new model architectures [SF15], regularization terms [JF18], robust loss functions [GKS17], loss adjustment [RLA+15] and sample selection [HYY+18]. [SF15] proposed a new model architecture containing a noise adaptation layer to model the label transition pattern with a noise transition matrix. Furthermore, a generative classification layer is found to be more robust against the label noise [LYL+19]. On the other hand, both adversarial training [GSS15] and label smoothing [LBMK20], which could be seen as regularization, were found to be efficient at improving the model's tolerance to label noise. For the robust loss functions, [GKS17] proved that the mean absolute error (MAE) loss was robust to the label noise, but it harmed the model's generalizability. So, the generalized cross entropy (GCE) [ZS18] keeping the advantages from cross-entropy loss and MAE loss is proposed. [PRM+17] proposed a Backward correction method and a Forward correction method to address the label noise challenge by adjusting the losses. Furthermore, Bootstrapping [RLA+15] was the first label refurbishment method to automatically correct the data labels during the training process. For the sample selection, [HYY+18] adopted two models to adaptively choose samples to train its peer model based on the value of the training loss, which was called the Co-teaching method. Combining with semi-supervised learning, DivideMix [LSH20] trained models on clean dataset and noisy dataset.

Although there are numerous researches focusing on solving the label noise problem, most of them are designed for balanced datasets and the image classification task. In this paper, we first systematically evaluate previous methods of the instance segmentation task on a noisy long-tailed dataset.

3 LVIS v1 with Noise

In this section, we mainly analyze how to add asymmetry (class-related) noise into the training set of the LVIS v1. Because there are 1203 categories, analyzing the semantics of each category manually and gathering classes with similar semantics manually can be impracticable and require a lot of workloads. We create an automated toolkit, to analyze the semantics based on WordNet. We will introduce the analyzing process in the follows.

3.1 Parse synset Categories

To parse the synset categories in the LVIS v1, we adopt `nltk` [BKL09], which is a natural language toolkit. With it, we build a tree to represent the relations between all categories in the LVIS v1, which can be found in Figure 2. Specifically, the numbers in the child nodes represent how many synset categories belong to this node, i.e., their hypernyms are this node[4]. We manually pick 25 hypernyms as super classes and one "others" as a super class to contain the remaining categories, which do not belong to any of the 25 super classes. Then, we divide categories based on their hypernyms into each super class, and there is no overlap between different super classes. The details of the super classes and the number of categories in each super class can be found in Table 1.

[4]We do not plot the categories as the child nodes, because it needs a tree with depth 18. But the toolkit we provide can plot a tree with arbitrary depths.

Name of Super Class	instrumentality				food				living_thing				commodity			
Number of Categories	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**
	150	200	174	524	52	58	29	139	34	60	19	113	23	40	50	113

Name of Super Class	natural_object				decoration				structure				communication			
Number of Categories	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**
	9	15	14	38	5	10	9	24	8	7	9	24	7	9	7	23

Name of Super Class	creation				material				fabric				sheet			
Number of Categories	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**
	6	4	7	17	5	5	5	15	1	6	7	14	1	3	9	13

Name of Super Class	way				strip				plaything				padding			
Number of Categories	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**
	4	3	5	12	0	5	6	11	4	1	4	9	1	2	5	8

Name of Super Class	fixture				surface				measure				causal_agent			
Number of Categories	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**
	1	0	6	7	4	2	0	6	3	2	1	6	1	2	2	5

Name of Super Class	block				float				line				opening			
Number of Categories	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**
	2	1	1	4	1	1	2	4	0	2	1	3	0	1	2	3

Name of Super Class	arrangement				others				**Total**			
Number of Categories	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**	**r**	**c**	**f**	**total**
	0	2	1	3	15	20	30	65	337	461	405	1203

Table 1: Super classes and the number of categories in each super class. We list the number of **r**are, **c**ommon, and **f**requent categories under each super class.

3.2 Noise Label Generation

To add asymmetry (class-related) noise into the training set of the LVIS v1, we randomly pick a new category under the same super class to replace the original label. But, to make the semantics consistency between the original category and the new one, we only consider categories that do not belong to the "others". In Algorithm 1, we describe the process of generating label noise. The Clean Annotation List A is a list containing all instances' categories, which has the following format: $A = [3, 5, 100, 102, \ldots, 90, 1000]$. $A[i]$ represent the i-th element in the list A. The Super Class Dictionary S is a dictionary, whose key is the name of the super class name listed in Table 1, and the value is a list of the categories which belong to the super class, e.g., $S = \{\text{'instrumentality}' : [1, 3, 10, \ldots, 900, 1200], \text{'food}' : [2, 7, \ldots, 1000, 1001], \ldots, \text{'others}' : [90, 500, \ldots, 1203]\}$. We randomly sample a value from a uniform distribution $\mathcal{U}(0, 1)$ between 0 and 1 to fit the Noise Ratio p.

Algorithm 1 Noise Label Generation

Input: Clean Annotation List A, Noise Ratio p, Super Class Dictionary S
for $i = 0 \rightarrow (\text{len}(A) - 1)$ **do**
  $r \leftarrow \mathcal{U}(0, 1)$
  if $r \leq p$ **then**
    $s \leftarrow$ super class of category $A[i]$
    if s is not "others" **then**
      $n \leftarrow$ randomly pick a category from $S[s]$
      $A[i] \leftarrow n$
    end if
  end if
end for
return A

4 Experiments

4.1 Dataset

In this paper, we adopt LVIS v1 to evaluate various methods' performance under different noise settings, which contains 100k images for training and 19.8k images for validation. In LVIS v1, there are 1,203 categories, grouped by their image frequency: rare (1-10 images in the dataset), common (11-100 images in the dataset) and frequent (>100 images in the dataset). For the label noise, we consider two types of the noise, i.e., symmetric (class-agnostic) noise and asymmetric (class-related) noise. When adding symmetric noise into the training set, we randomly pick a label from all 1,203 possible categories to replace the original label. When adding asymmetric noise, we follow the steps in Section 3.2. The results of the validation set are reported from five aspects: average segmentation performance AP, average box performance AP^b and average segmentation performance for rare AP^r, common AP^c and frequent categories AP^f.

4.2 Methods

We evaluate three mainstream long-tailed instance segmentation approaches, including EQL [TWL+20], DropLoss [HRCH21] and Seesaw Loss [WZZ+21]. Specifically, the implementations of EQL and DropLoss are followed the Gumbel Optimized Loss [ADNL22], i.e., we replace the original activation function Sigmoid with Gumbel.

4.3 Implementation Details

We adopt mmdetection [CWP+19] and Pytorch [PGM+19] on 4 V100 GPU cards to implement all experiments. On each GPU card, there are 4 images in one batch. For all methods, we use Mask R-CNN [HGDG17], in which the ResNet-50 [HZRS16] with FPN [LDG+17] is the backbone model, and the 2× schedule and the Normalized Mask [WZZ+21] are adopted to improve the model's performance. The model's details are shown in Table 2. For the training process, we adopt Stochastic Gradient

Method	Model	Backbone	Normalized Mask	Normalized Classifier	Activation
EQL + Gumbel	Mask R-CNN	ResNet-50	✓	✗	Gumbel
DropLoss + Gumbel	Mask R-CNN	ResNet-50	✓	✗	Gumbel
Seesaw Loss	Mask R-CNN	ResNet-50	✓	✓	Softmax

Table 2: Settings for different methods.

Noise	Sampler	AP^b	AP^r	AP^c	AP^f	AP
None	Random	25.40	14.90	23.30	32.40	25.50
	RFS	27.10	18.80	26.60	30.90	26.90
AN (0.2)	Random	21.60	5.60	19.60	28.70	20.70
	RFS	21.40	5.00	19.40	29.00	20.70
AN (0.4)	Random	18.20	4.30	13.90	26.90	17.30
	RFS	18.60	4.00	14.70	27.20	17.80
SN (0.2)	Random	22.80	7.50	21.10	28.90	21.80
	RFS	22.60	6.60	21.40	29.10	21.90
SN (0.4)	Random	19.10	2.80	16.00	27.30	18.20
	RFS	19.30	3.30	16.40	27.50	18.50

Table 3: Results of Seesaw Loss with different samplers.

Descent as our optimizer to train models with 24 epochs. The learning rate is 0.02, and we decrease it by 0.1 at the 16th epoch and the 22nd epoch. The data augmentations include random horizontal flipping and multi-scaling, which are consistent with previous works [WZZ+21, ADNL22]. In inference, we follow the protocol in [WZZ+21], i.e., we resize all images into 1333×800 and use the score thresholds of 10^{-3}. For sampler, we consider RFS sampler [GDG19] and Random sampler, respectively.

4.4 Results

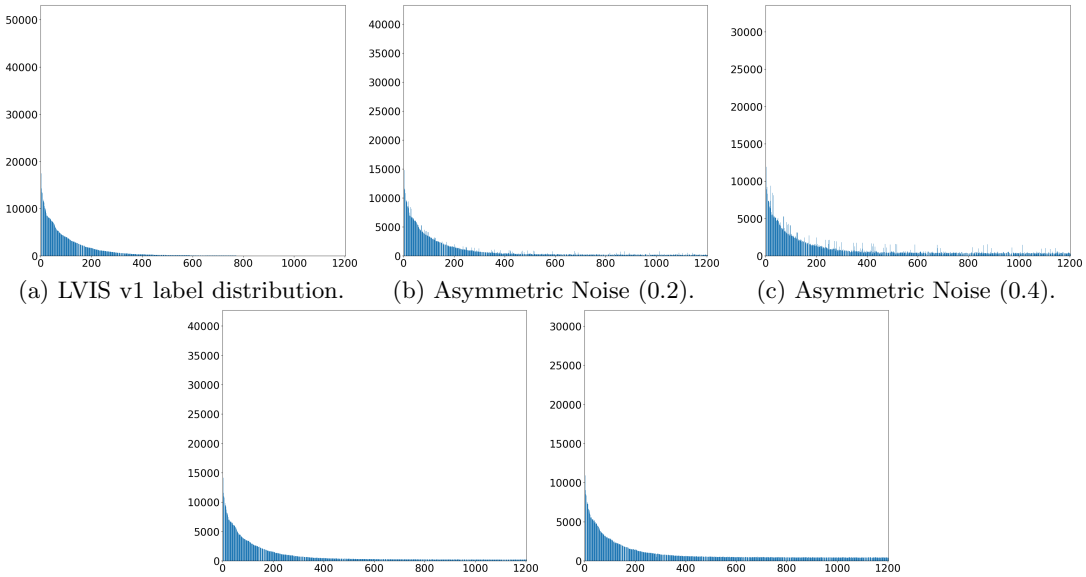
In Figure 1, we compare the label distribution under different noise settings. The x-axis is the category, which is sorted based on the frequency in the original training set of the LVIS v1. We keep the same category order in each image. And the y-axis is the number of instances that are annotated as the corresponding category. Clearly, adding noise will change the label distribution, thus the RFS sampler, which will over-sample low-frequency instances, will be influenced by the shifted label distribution.

We first compare the results of Seesaw Loss with the RFS sampler and the Random sampler in Table 3. For the RFS sampler, we set the over-sample threshold as 0.001. The results indicate that the RFS sampler can significantly increase the performance on the clean long-tailed dataset compared with the Random sampler. When there exist label noise in the dataset, the improvement obtained from the RFS sampler weakens, which can be the distribution shift, causing the sampler samples more instance with incorrect labels.

In Table 4, we present the results under different loss functions, and all of them are equipped with the RFS sampler. The results indicate that methods (DropLoss with Gumbel and EQL with Gumbel) with better performance may not be the best choice for practical when there exists noise in the training labels. On the other hand, different loss functions have different robustness for the asymmetric (class-related) noise and symmetric (class-agnostic) noise. For example, the Seesaw Loss shows more robust under symmetric noise. However, the DropLoss with Gumbel shows more robust under asymmetric noise. This phenomenon indicates that designing a loss for long-tailed instance segmentation dataset with noise is challenging, and it is complex to outperform under all types of noise.

5 Conclusion

In this report, we introduce a tool to automatically analyze the categories in LVIS v1. With it, we add asymmetric noise into the clean training set to create noisy datasets with different noise ratios. Then, we evaluate three mainstream long-tailed instance segmentation loss function on the noisy datasets containing asymmetric noise and symmetric noise, respectively. The results indicate that previous long-tailed loss functions lack the robustness when facing the noisy datasets, and it is important to study how to design new ones to better fit the label noise.



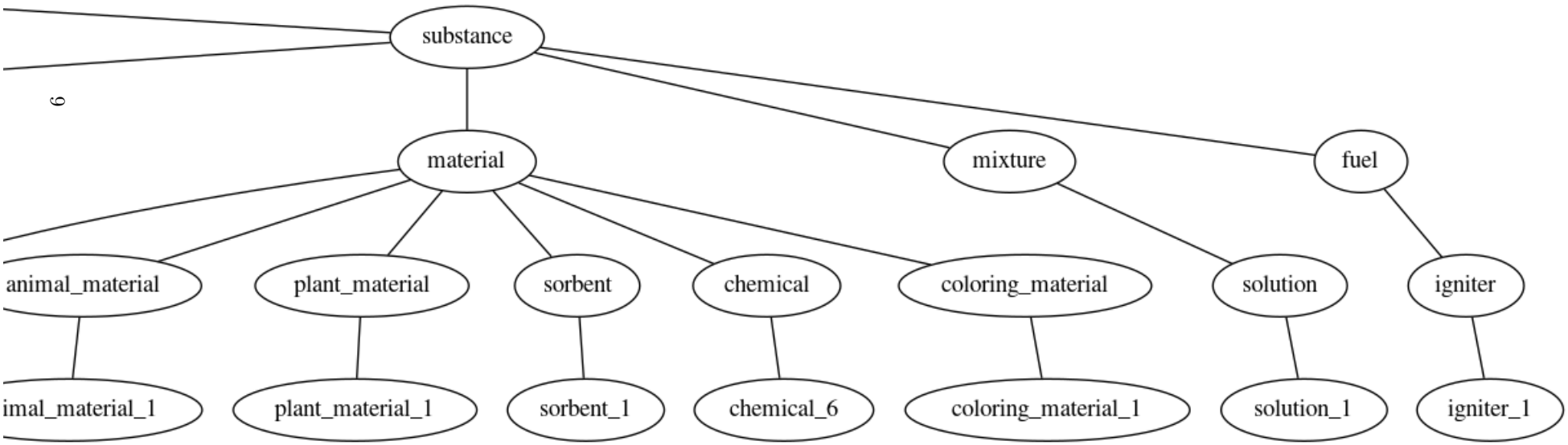
(a) LVIS v1 label distribution. (b) Asymmetric Noise (0.2). (c) Asymmetric Noise (0.4).

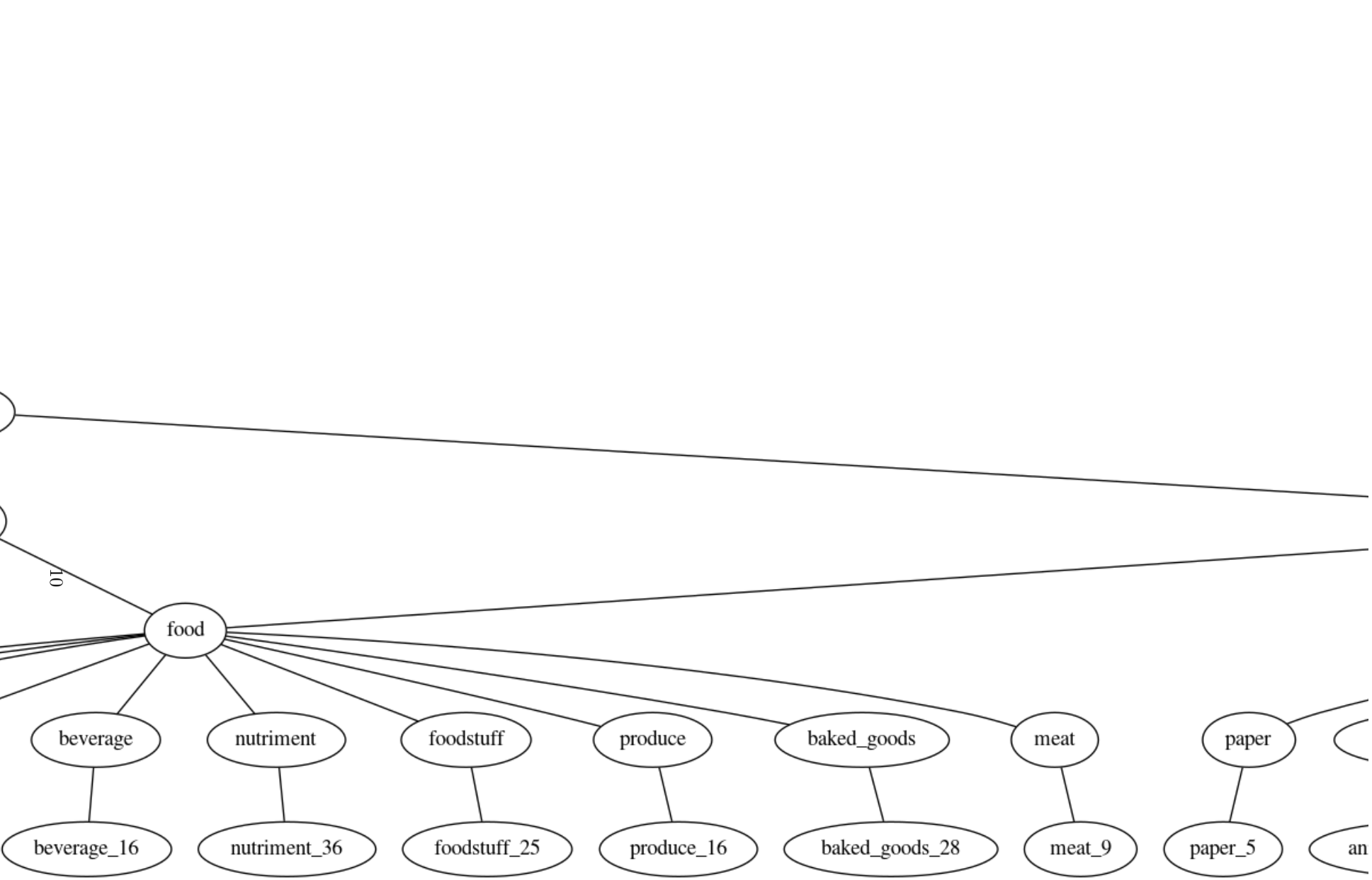
(d) Symmetric Noise (0.2). (e) Symmetric Noise (0.4).

Figure 1: Label distribution under different noise settings.

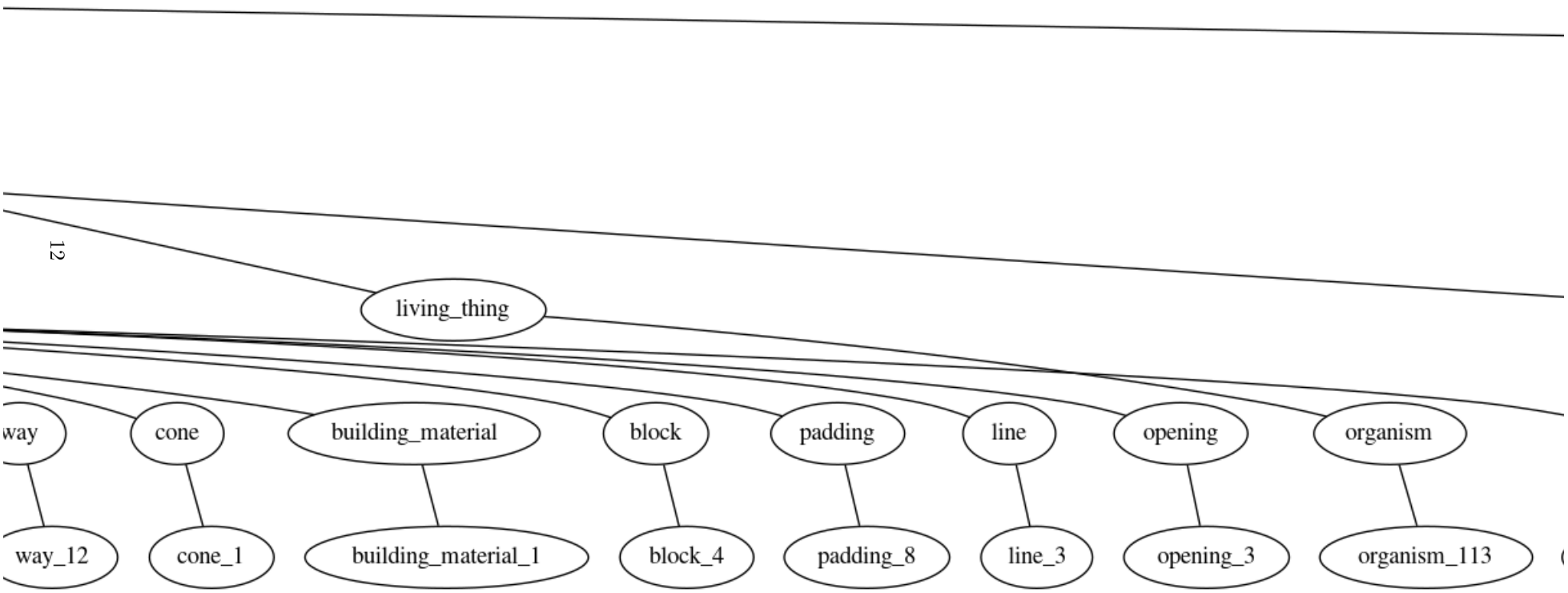
Noise	**Method**	AP^b	AP^r	AP^c	AP^f	AP
None	DropLoss + Gumbel	27.60	19.90	28.00	30.40	27.50
	EQL + Gumbel	26.80	17.30	26.70	31.20	26.90
	Seesaw Loss	27.10	18.80	26.60	30.90	26.90
AN (0.2)	DropLoss + Gumbel	18.90	3.60	16.00	27.70	18.40
	EQL + Gumbel	20.00	4.40	16.30	29.70	19.50
	Seesaw Loss	21.40	5.00	19.40	29.00	20.70
AN (0.4)	DropLoss + Gumbel	14.70	2.30	9.50	25.10	14.30
	EQL + Gumbel	15.90	2.30	9.60	27.60	15.40
	Seesaw Loss	18.60	4.00	14.70	27.20	17.80
SN (0.2)	DropLoss + Gumbel	18.60	3.50	15.30	27.90	18.20
	EQL + Gumbel	19.20	3.30	15.10	29.50	18.70
	Seesaw Loss	22.60	6.60	21.40	29.10	21.90
SN (0.4)	DropLoss + Gumbel	13.80	1.40	7.50	25.00	13.30
	EQL + Gumbel	14.50	1.10	7.40	27.00	14.00
	Seesaw Loss	19.30	3.30	16.40	27.50	18.50

Table 4: Results under different noise settings.


substance
material
mixture
fuel
animal_material
plant_material
sorbent
chemical
coloring_material
solution
igniter
imal_material_1
plant_material_1
sorbent_1
chemical_6
coloring_material_1
solution_1
igniter_1

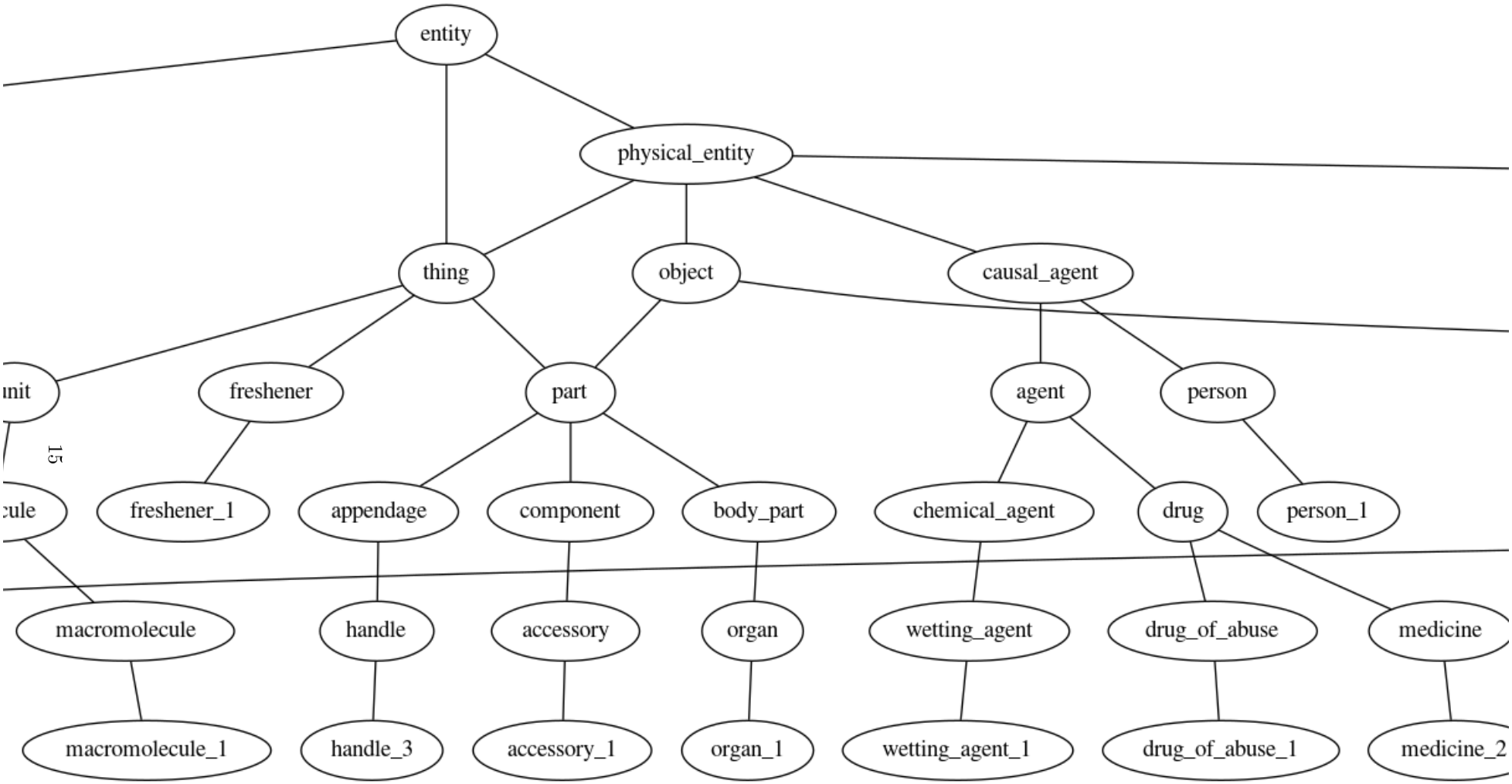

10
food
beverage
nutriment
foodstuff
produce
baked_goods
meat
paper
beverage_16
nutriment_36
foodstuff_25
produce_16
baked_goods_28
meat_9
paper_5

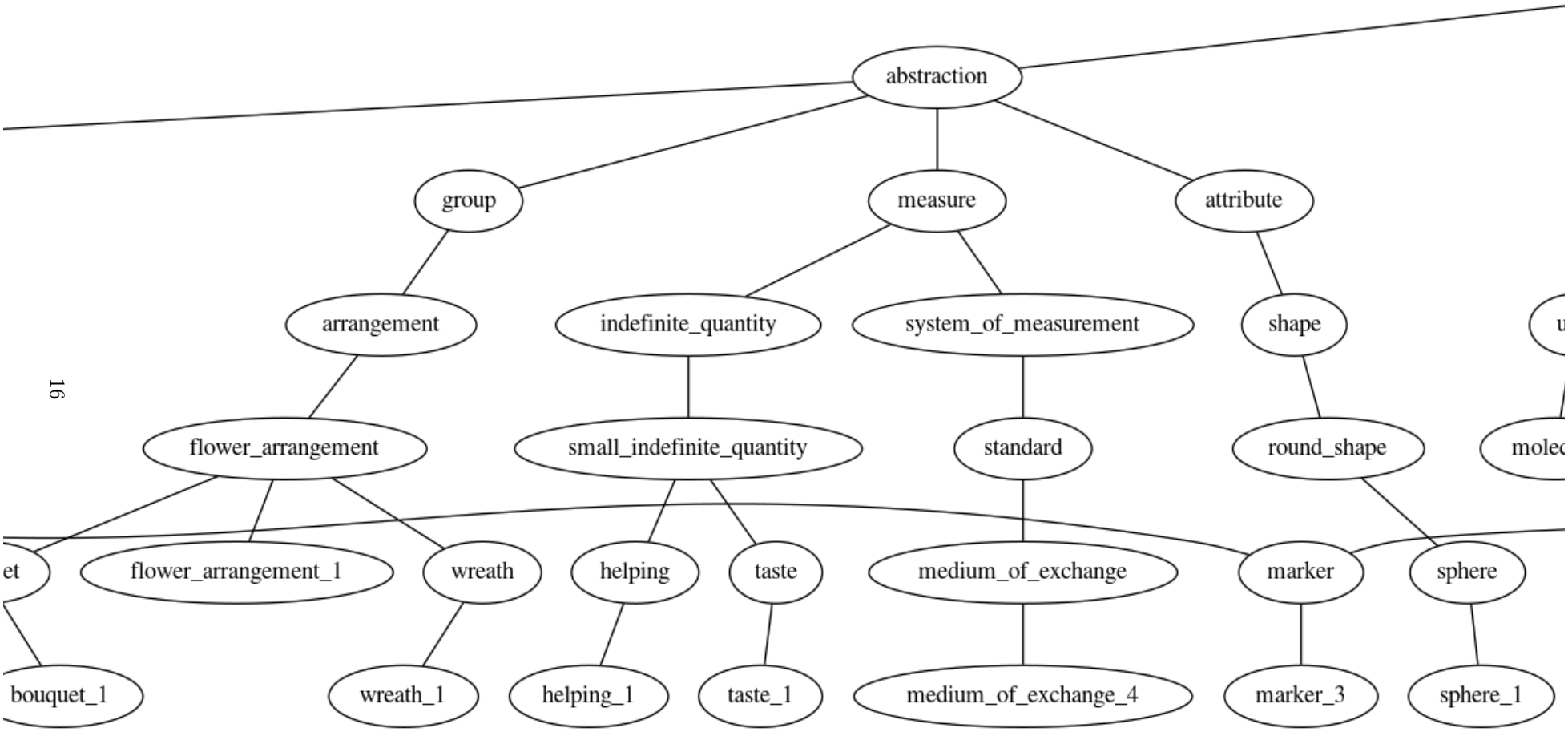
matter
solid
crystal
natural_object
gem
gem_1
chocolate
chocolate_1
seafood
seafood_4
fish
fish_3
breakfast_food
breakfast_food_1
nest
nest_1
plant_part
plant_part_36
covering
covering_40

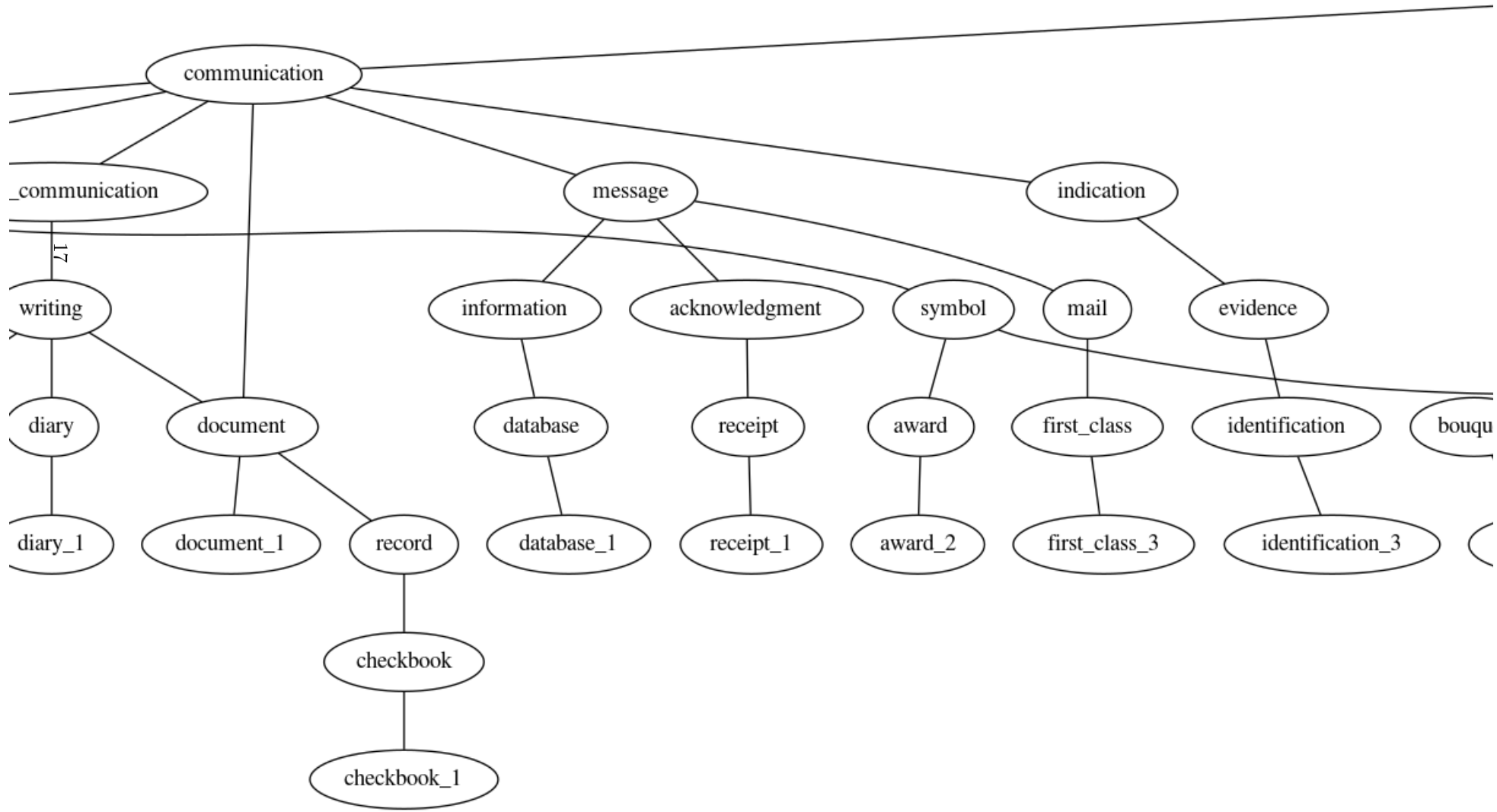

12
living_thing
way
cone
building_material
block
padding
line
opening
organism
way_12
cone_1
building_material_1
block_4
padding_8
line_3
opening_3
organism_113

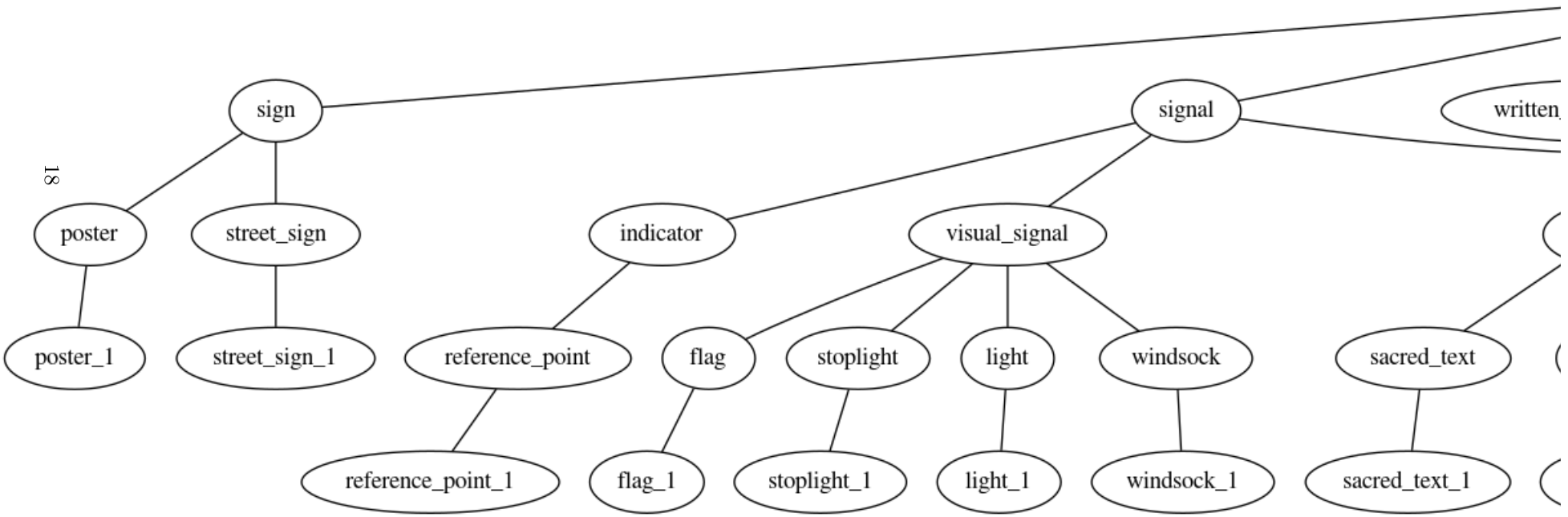
whole
artifact
article
article_9
fixture
fixture_7
creation
creation_17
float
float_4
surface
surface_6
sheet
sheet_13
structure
structure_24
fabric
fabric_14

weight
facility
instrumentality
decoration
commodity
strip
playth
weight_1
facility_1
instrumentality_524
decoration_24
commodity_113
strip_11
plaything_9


entity
physical_entity
thing
object
causal_agent
nit
freshener
part
agent
person
cule
freshener_1
appendage
component
body_part
chemical_agent
drug
person_1
macromolecule
handle
accessory
organ
wetting_agent
drug_of_abuse
medicine
macromolecule_1
handle_3
accessory_1
organ_1
wetting_agent_1
drug_of_abuse_1
medicine_2


abstraction
group
measure
attribute
arrangement
indefinite_quantity
system_of_measurement
shape
flower_arrangement
small_indefinite_quantity
standard
round_shape
et
flower_arrangement_1
wreath
helping
taste
medium_of_exchange
marker
sphere
bouquet_1
wreath_1
helping_1
taste_1
medium_of_exchange_4
marker_3
sphere_1


communication
_communication
message
indication
writing
information
acknowledgment
symbol
mail
evidence
diary
document
database
receipt
award
first_class
identification
diary_1
document_1
record
database_1
receipt_1
award_2
first_class_3
identification_3
checkbook
checkbook_1


sign
signal
written
poster
street_sign
indicator
visual_signal
poster_1
street_sign_1
reference_point
flag
stoplight
light
windsock
sacred_text
reference_point_1
flag_1
stoplight_1
light_1
windsock_1
sacred_text_1

References

[ADNL22] Kostas Alexandridis, Jiankang Deng, Anh Nguyen, and Shan Luo. Long-tailed Instance Segmentation using Gumbel Optimized Loss. In *Proc. of the ECCV*, 2022.

[BKC17] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[BKL09] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[CJL+19] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-Balanced Loss Based on Effective Number of Samples. In *Proc. of the CVPR*, pages 9268–9277, 2019.

[CWG+19] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *Proc. of the NeurIPS*, pages 1565–1576, 2019.

[CWP+19] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *CoRR*, abs/1906.07155, 2019.

[CZP+18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. of the ECCV*, pages 801–818, 2018.

[GCS+21] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proc. of the CVPR*, pages 2918–2928, 2021.

[GDG19] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *Proc. of the CVPR*, pages 5356–5364, 2019.

[GKS17] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proc. of the AAAI*, pages 1919–1925, 2017.

[GSS15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *Proc. of the ICLR*, 2015.

[HGDG17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. of the ICCV*, pages 2961–2969, 2017.

[HJT+20] Xinting Hu, Yi Jiang, Kaihua Tang, Jingyuan Chen, Chunyan Miao, and Hanwang Zhang. Learning to segment the tail. In *Proc. of the CVPR*, pages 14045–14054, 2020.

[HRCH21] Ting-I Hsieh, Esther Robb, Hwann-Tzong Chen, and Jia-Bin Huang. Droploss for long-tail instance segmentation. In *Proc. of the AAAI*, pages 1549–1557, 2021.

[HYY+18] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Proc. of the NeurIPS*, pages 8536–8546, 2018.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proc. of the CVPR*, pages 770–778, 2016.

[JF18] Simon Jenni and Paolo Favaro. Deep bilevel learning. In *Proc. of the ECCV*, volume 11214, pages 632–648, 2018.

[KRC21] Shyamgopal Karthik, Jérôme Revaud, and Boris Chidlovskii. Learning from long-tailed data with noisy labels. *CoRR*, abs/2108.11096, 2021.

[LAE+16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proc. of the ECCV*, pages 21–37, 2016.

[LBMK20] Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *Proc. of the ICML*, volume 119, pages 6448–6458, 2020.

[LDG+17] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *Proc. of the CVPR*, pages 936–944, 2017.

[LGG+17] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *Proc. of the ICCV*, pages 2999–3007, 2017.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proc. of the ECCV*, pages 740–755, 2014.

[LSH20] Junnan Li, Richard Socher, and Steven C. H. Hoi. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *Proc. of the ICLR*, 2020.

[LWL+17] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *CoRR*, abs/1708.02862, 2017.

[LYL+19] Kimin Lee, Sukmin Yun, Kibok Lee, Honglak Lee, Bo Li, and Jinwoo Shin. Robust inference via generative classifiers for handling noisy labels. In *Proc. of the ICML*, volume 97, pages 3763–3772, 2019.

[Mil95] George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[MJR+21] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *Proc. of the ICLR*, 2021.

[PGM+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Proc. of the NeurIPS*, pages 8024–8035. 2019.

[PRM+17] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In *Proc. of the CVPR*, pages 2233–2241, 2017.

[RHGS15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Proc. of the NeurIPS*, 28, 2015.

[RLA+15] Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *Proc. of the ICLR Workshop*, 2015.

[RYS+20] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced Meta-Softmax for Long-Tailed Visual Recognition. In *Proc. of the NeurIPS*, 2020.

[SF15] Sainbayar Sukhbaatar and Rob Fergus. Learning from Noisy Labels with Deep Neural Networks. In *Proc. of the ICLR Workshop*, 2015.

[SKP+22] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. Publisher: IEEE.

[TWL+20] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proc. of the CVPR*, pages 11662–11671, 2020.

[WLK+20] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *Proc. of the ECCV*, pages 728–744, 2020.

[WRW+09] Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier R. Movellan. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Proc. of the NeurIPS*, pages 2035–2043, 2009.

[WSTL21] Tong Wei, Jiang-Xin Shi, Wei-Wei Tu, and Yu-Feng Li. Robust long-tailed learning under label noise. *CoRR*, abs/2108.11569, 2021.

[WZK+20] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Proc. of the NeurIPS*, 33:17721–17732, 2020.

[WZZ+21] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw Loss for Long-Tailed Instance Segmentation. In *Proc. of the CVPR*, pages 9695–9704, 2021.

[XXY+15] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proc. of the CVPR*, pages 2691–2699, 2015.

[YML+20] Longrong Yang, Fanman Meng, Hongliang Li, Qingbo Wu, and Qishang Cheng. Learning with noisy class labels for instance segmentation. In *Proc. of the ECCV*, pages 38–53, 2020.

[ZHL21] Yuhang Zang, Chen Huang, and Chen Change Loy. Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation. In *Proc. of the ICCV*, pages 3457–3466, 2021.

[ZS18] Zhilu Zhang and Mert R. Sabuncu. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In *Proc. of the NeurIPS*, pages 8792–8802, 2018.