Automatic universal taxonomies for multi-domain semantic segmentation

Petra Bevandić
petra.bevandic@fer.hr

Siniša Šegvić
sinisa.segvic@fer.hr

Faculty of Electrical Engineering
and Computing,
University of Zagreb,
Zagreb, Croatia

Abstract

Training semantic segmentation models on multiple datasets has sparked a lot of recent interest in the computer vision community. This interest has been motivated by expensive annotations and a desire to achieve proficiency across multiple visual domains. However, established datasets have mutually incompatible labels which disrupt principled inference in the wild. We address this issue by automatic construction of universal taxonomies through iterative dataset integration. Our method detects subset-superset relationships between dataset-specific labels, and supports learning of sub-class logits by treating super-classes as partial labels. We present experiments on collections of standard datasets and demonstrate competitive generalization performance with respect to previous work.

1 Introduction

Semantic segmentation is an important computer vision task with exciting applications in intelligent transportation [14], medical diagnostics [25], remote surveillance [4], and autonomous robots [13]. Current state of the art is based on strongly supervised learning which induces a strong dependence on dense semantic ground truth. Unfortunately, producing dense annotations requires a lot of time and money [6, 43]. There are several datasets of intermediate size [21, 24, 37, 41], but none that is sufficient for delivering robust performance in the wild [37]. Thus, training across several datasets and domains appears as an attractive research direction.

A simple baseline involves per-dataset heads over shared features [9, 15]. Per-dataset predictions can be recombined into a common taxonomy [42], however this is not easily adapted to multi-class problems and overlapping taxonomies [20, 23, 39]. Another baseline concatenates per-dataset taxonomies [9, 22] and feeds them to common softmax. However, this may entail capacity loss due to competition between related logits. A recent approach reconciles a set of taxonomies by pragmatic label adaptation [18] that however has to drop some classes in order to reduce the relabeling effort. Recent work leverages hand-crafted universal taxonomies that allow superclass labels to promote subclass recognition and vice versa [2, 20, 23]. However, this requires human judgment which is expensive and error prone.
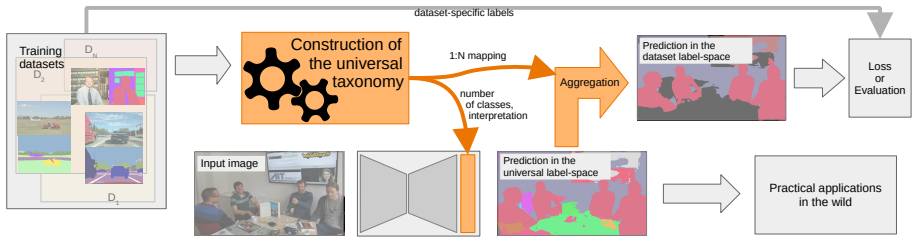
Figure 1: We consider automatic construction of a universal taxonomy from multiple datasets with incompatible labels (orange box). Our method recovers a set of disjoint universal classes, and connects each dataset-specific class to one or more universal classes. These 1:N mappings enable training and evaluation on original labels (top-right). The universal model can be exploited for interpretable inference in the wild (bottom-right).

This paper makes a step further by considering automatic extraction of universal taxonomies over incompatible datasets as sketched in Figure 1. Our method hypothesizes cross-dataset relations through co-occurence analysis. We disambiguate these hypotheses against each other according to mIoU performance on all training datasets. We perform experiments on collections of large semantic segmentation datasets such as Vistas, Ade20k, COCO and WildDash 2. The recovered automatic taxonomies perform comparably to their manual counterparts [2] while outperforming all other baselines [13] by a considerable margin.

2   Related work

We consider semantic segmentation for natural scene understanding (sec. 2.1) by studying cross-domain models which transcend particular training taxonomies (sec. 2.2). We focus on automatic construction of universal taxonomies (sec. 2.3).

2.1   Semantic segmentation

Semantic segmentation classifies each input pixel into one of C known classes [8, 29]. It is one of the most computationally intensive computer vision tasks due to high output resolution. The training footprint often constrains the model capacity [27]. Huge computational complexity leads to very long training times. Consequently, efficient models [12, 25, 26] and knowledge transfer [11] are a good fit for large cross-domain experiments. Besides faster training, they also improve accessibility and decrease environmental impact [28].

2.2   Cross-domain training

Early cross-domain training approaches do not incorporate relations between individual taxonomies. Instead, they either use separate dataset-specific prediction heads on top of shared features [16], or train on a concatenation of particular taxonomies [22]. Naive concatenation has been improved by encouraging cross-talk between logits [9].

Training dense open-set recognition models on positive and negative data may improve detection of unknown [4, 5] or novel classes [32]. This can be viewed as asymmetrical cross-

domain training. The positive domain corresponds to the primary recognition task (e.g. road driving) while the negative domain typically corresponds to anomalies [1, 5, 51].

Some cross-domain approaches propose hierarchical universal taxonomies with distinct nodes for categories and classes [20, 23]. However, this requires complex learning procedures while not offering advantages over flat universal taxonomies.

Incompatible datasets can be unified under a custom common taxonomy by manual relabeling and removal of subclasses [18, 38]. However, these modifications are tedious and destructive. The more datasets one converts, the harder it gets to extend the common taxonomy with new subclasses. This issue can be elegantly solved by constructing a universal taxonomy where each dataset-specific class can be expressed as a union of universal classes [2]. In this case, universal logits can be trained with respect to dataset-specific labels (cf. Figure 1) since dataset posteriors correspond to sums of universal posteriors [7]. The result of such construction allows principled cross-dataset training without any modification of the original datasets. We extend this approach by considering automatic construction of such universal taxonomy from datasets with differing granularities.

2.3 Automatic construction of universal taxonomies

Manual resolution of dataset discrepancies is error prone, especially when the ambition is to train on multiple large-scale datasets with hundreds of symbolic labels. This issue can be elegantly curcumvented by expressing semantic labels with text embeddings instead of categorical distributions [19, 35]. However, a recent study reveals that label semantics often vary across datasets. Their experiments suggest that visual cues outperform label semantics as a tool for recovering cross-dataset relations [33].

Recent work constructs an automatic taxonomy for object detection [42]. Their approach starts by training a model with shared features and separate prediction heads [9, 15]. Subsequently, they freeze the trained features and optimize dataset-specific mappings through linear programming. The resulting cross-dataset mappings outperform their text-embedding counterparts. However, this approach does not handle subset/superset relationships and therefore does not produce a true taxonomy when dataset-specific classes happen to overlap. This situation hampers multi-class performance due to competition between related logits [2].

Cross-dataset relations have also been recovered according to class names [18]. In this setup, superclass logits can be trained with subclass labels [17]. However, this setup can not accommodate the standard multi-class loss, fails if there is a name mismatch [33, 42], and cannot train subclass logits with superclass labels.

Different than all previous work, our method constructs the only flat universal taxonomy which retains all labels in presence of subclass/superclass relations.

3 Method

We consider automatic recovery of a flat universal taxonomy for a given collection of datasets in order to allow cross-domain training of dense prediction models. We propose to automatically discover hierarchical relations between classes of the two datasets and use this information to construct the universal taxonomy as illustrated in Figure 2. We extend pairwise taxonomies for arbitrary tuples of datasets through tournament-style iteration.
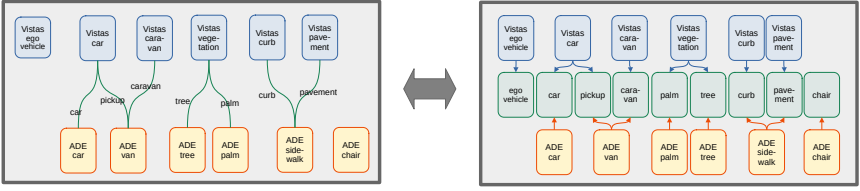
Figure 2: We collect information about cross-dataset relations and expres it as a bipartite graph that connects classes of the two datasets (left). We analyze the bipartite graph in order to recover the universal taxonomy (right). Each edge of the bipartite graph identifies the visual concept shared by the related classes.

Occasionally, our cross-domain models will predict classes which are disjoint from the native taxonomy of the input image. We denote such occurrences as foreign or extra-domain predictions. Conversely, predictions which fall within the native taxonomy are denoted as intra-domain predictions.

3.1   Universal taxonomy for two datasets

Let us consider two dataset-specific taxonomies as $T_a = \{c_i^a\}$ and $T_b = \{c_j^b\}$. We apply a model trained for T_b to training data with T_a groundtruth and the other way round. We collect co-occurrence statistics between ground-truth classes and foreign predictions and store them into two co-occurrence matrices $|T_a| \times |T_b|$ and $|T_b| \times |T_a|$. For convenience, we shall denote the most common foreign prediction for a ground-truth class c as mcfp(c). We shall hypothesize relations between datasets by considering a bipartite graph induced by the mcfp function. The graph has $|T_a| + |T_b|$ vertices which represent classes, and $|T_a| + |T_b|$ edges pointing from a ground-truth class to its most frequent foreign prediction. Hence, each vertex has exactly one outgoing edge. This choice increases the statistical power of our hypotheses and reduces the number of hypotheses and hyper-parameters.

We illustrate our approach on the following two taxonomies: ADE20K = {'ade-road',...} and Vistas = {'vistas-road', 'vistas-zebra',...}. The class 'ade-road' is a superset of 'vistas-road' and 'vistas-zebra'. In Vistas images we shall typically have mcfp('vistas-road') = mcfp('vistas-zebra') = 'ade-road'. On the other hand, in ADE20k images we will have: mcfp('ade-road') = 'vistas-road'. We observe that the mcfp statistic suffices to hypothesize that 'ade-road' is a superset of 'vistas-road' and 'vistas-zebra'.

We analyze the bipartite graph as follows. Cycles of length 2 ($c_i^a \rightarrow c_j^b \rightarrow c_i^a$) indicate overlap. Asymmetric relationships ($c_i^a \rightarrow c_j^b$) suggest a subset hypothesis $c_j^b \supset c_i^a$. Inconsistent triplets $c_i^a \rightarrow c_j^b \rightarrow c_k^a$ where $c_k^a \nrightarrow c_j^b$ suggest a subset and a superset hypothesis $c_j^b \supset c_i^a \land c_k^a \supset c_j^b$. This would mean that $c_i^a \cap c_k^a \neq \emptyset$, which is impossible since input datasets have proper taxonomies. We consider $c_i^a \subset c_j^b$ and $c_j^b \subset c_k^a$ as competing hypotheses which we disambiguate in 3.2. Figure 3 illustrates this procedure on the ADE20K-Vistas example.

We recover the final universal taxonomy from the disambiguated bipartite graph (cf. Figure 4, right). The graph associates each universal class with all incident dataset-specific classes. We thus base the names of universal classes on associated dataset-specific classes: a one-way edge inherits the name of its source vertex, while a two-way edge inherits the names of both adjacent vertices. If we have ade-car→ vistas-car and vistas-car → ade-car, we would
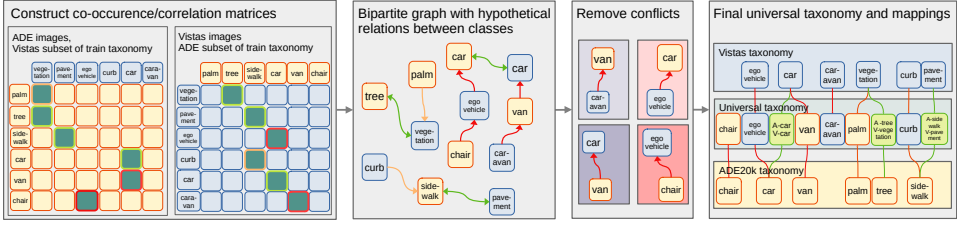
Figure 3: Our method collects two co-occurrence matrices (left) between the ground truth (rows) and foreign predictions (columns). For each row, we locate the strongest column and form a bipartite graph (center-left). We note three 2-cycles (green), two asymmetric relationships (orange), and two inconsistent triplets (red). We form pairs of conflicting hypotheses (shades of the same color, center right) and resolve them by comparing performance on the training dataset (cf. Figure 4). The collected evidence allows us to recover the universal taxonomy (right).

have a universal class named 'ade-car/vistas-car'. Such naming convention provides a degree of interpretability to resulting universal models.

3.2 Conflict resolution with improved naive concatenation

Naive concatenation is a cross-domain pseudo-taxonomy which we obtain by aggregating dataset-specific taxonomies. We use the term pseudo-taxonomy since its members may overlap (e.g. ade-road and vistas-road). Such overlaps require discrimination of semantically related concpets and consequently diminish the effective model capacity.

Naive concatenation performance can be improved by performing post-inference mapping towards the evaluation taxonomy. We define unnormalized classification score of a particular evaluation class $S(c_i^a)$ as the sum of its posterior $P(c_i^a)$ with posteriors of intersecting foreign classes $P(c_j^b)$ [2]:

$$S(c_i^a) = P(c_i^a) + \sum_{c_i^a \cap c_j^b \neq \emptyset} P(c_j^b) \tag{1}$$

The expression $c_i^a \cap c_j^b \neq \emptyset$ is true when c_i^a and c_j^b are in any kind of relation. The model prediction corresponds to $\arg\max_i S(c_i^a)$.

We illustrate the recovery of different dataset-specific scores over ADE20K and Vistas as follows: $S(\text{ade-road}) = P(\text{ade-road}) + P(\text{vistas-road}) + P(\text{vistas-zebra})$, $S(\text{vistas-road}) = P(\text{vistas-road}) + P(\text{ade-road})$ and $S(\text{vistas-zebra}) = P(\text{vistas-zebra}) + P(\text{ade-road})$.

We can use post-inference mapping to compare competing hypotheses arising from inconsistent triplets (§3.1). We create a post-inference mapping for each competing hypothesis and evaluate performance according to (1). We choose the hypothesis with the highest train mIoU performance averaged over all involved datasets. The resolution involves $2N_D N_C$ automatic evaluations where N_C denotes the number of conflicting pairs and N_D indicates the number of training datasets. This procedure is illustrated in Figure 4.

Figure 4: Resolution of contradicting hypotheses according to the train performance of the naive concatenation model with post-inference mapping. The winning hypotheses are selected according to the unnormalized classification score (1) on all train datasets. We resolve one pair of hypotheses at a time. This requires $2N_C N_D$ evaluations where N_C is the number of conflicting pairs and N_D indicates the number of training datasets.

3.3 Universal taxonomy for more than two datasets

Recovering a pairwise universal taxonomy allows us to unify the two involved datasets. The resulting meta-dataset contains images from the two datasets and partial labels in form of unions of universal classes. We can proceed by unifying this meta-dataset with subsequent datasets. However, lack of proper ground-truth precludes recovery of the proposed co-occurrence matrix. We therefore approximate the co-occurrence matrix with a co-incidence matrix between intra-domain and foreign predictions.

To recover a universal taxonomy for more than two datasets, we proceed iteratively. We start by forming pairwise universal taxonomies. We then train naive concatenation models over pairs of meta-datasets and use them to unify the involved meta-datasets. We formulate mapping functions for original datasets as compositions of intermediate mapping functions.

The proposed procedure can be applied to any number of datasets in a straight-forward manner. We have successfully applied this procedure in order to recover the universal taxonomy for the MSeg dataset collection.

4 Experiments

We train semantic segmentation models in multi-domain setups. We promote efficient experimentation [28] by leveraging pyramidal SwiftNet [26] with three shared ResNet-18 [10] backbones and ImageNet pre-training (SNp-rn18). We train on automatic universal taxonomies with partial labels [2, 7, 39]. We train naive concatenation models with the standard NLL loss and the multi-head model with a sum of head-specific NLL losses. Both losses prioritize pixels at semantic boundaries [40]. We perform early stopping with respect to average mIoU validation performance. We attenuate the learning rate between $5 \cdot 10^{-4}$ and $6 \cdot 10^{-6}$ through cosine annealing. We evaluate by mapping foreign predictions to the void class [2, 6].

We train on random crops of 512×512 (§4.2 and §4.3) or 768×768 pixels (§4.1) with horizontal flipping and random scaling between 0.5× and 2×. We favour crops with rare classes and form batches with even representation of all datasets. Our universal models were trained on one Tesla V100 32GB. We train naive concatenation models on two GPUs in order to ensure the same batch size across considered dataset collections. We construct universal

taxonomies by analyzing only the training subsets.

4.1 Unifying dataset pairs

We present experiments on pairs of datasets with incompatible taxonomies. We compare our automatic universal taxonomy with two baselines as well as with a manually constructed universal taxonomy [2]. The two baselines are the naive taxonomy and a model with per-dataset prediction heads and a separate dataset detection head. All models are trained for 100 epochs.

Table 1 presents results of unifying Vistas [24] (road-driving, 65 classes) and WilDash 2 (WD2, road-driving, 25 classes) [37]. We split WD2 into minitrain and minival as in [2]. Vistas has a finer granularity than WD2 with the exception of car types. Our automatic universal taxonomy performs comparably to the manual universal taxonomy while outperforming the multi-head baseline as well as naive concatenation.Interestingly, our automatic taxonomy outperforms the manual taxonomy on some rare Vistas classes, which is likely due to their association with more frequent WD2 classes (e.g. wd-person and vistas-ground-animal, and wd-truck and vistas-trailer).

Taxonomy	#	evals	WD2	Vistas
two heads + dataset recognition	65 + 33 + 2	N/A	54.0	42.2
naive concat	98	N/A	54.8	42.8
manual univ.	67	N/A	56.2	44.4
auto univ. (ours)	67	4	54.6	45.9

Table 1: Evaluation of joint training on WD2 and Vistas. Columns show the number of logits (#), number of tested hypotheses (evals) and mIoU performance on both datasets.

Table 2 pairs Vistas with ADE20K [41] (photos, 150 classes). This experiment also validates two approaches for collecting evidence about visual similarity of dataset-specific classes. We compare separately trained per-dataset models [33] with the naive concatenation baseline. We also validate the two conflict resolution approaches based on co-occurrence and co-incidence matrices. Note also that conflict resolution is not feasible with separate per-dataset models.

We observe that the universal taxonomy produced by separate models has almost as many logits as the naive concatenation baseline. Our automatic universal taxonomy for ADE20k-Vistas has less training logits than its manual counterpart. Interestingly, it hypothesizes less relations than the manual approach (182 < 186) even before the contradicting hypotheses are resolved. This happens because our automatic method connects some classes that are kept separate in the manual taxonomy (e.g. connecting flags with banners or rail tracks with conveyor belts). Coincidence matrices perform similarly to co-occurrence matrices, although their universal taxonomies differ.

4.2 Merging multiple datasets

Table 3 evaluates our universal taxonomy over three datasets. We start from the universal taxonomy ADE20K-Vistas and extend it through unification with COCO (photos, 133 classes) [21]. Due to the huge size of the COCO dataset, we decrease the number of training epochs

Taxonomy	#	evals	ADE	Vistas
naive concat	215	N/A	36.8	41.1
manual univ.	186	N/A	37.4	42.7
auto univ. (separate models, co-occurrence)	213	N/A	37.4	41.7
auto univ. (concat, co-occurrence)	178	24	37.4	42.8
auto univ. (concat, coincidence)	176	26	36.9	42.5

Table 2: Evaluation of joint training on ADE20K-Vistas. Columns show number of logits (#), number of tested hypotheses (evals) and mIoU performance on both datasets. Automatic construction of universal taxonomy with separate per-dataset models underperforms with respect to the taxonomies built with naive concatenation. Collecting evidence through co-occurrence and coincidence performs comparably.

to 20. Before cropping, each image is resized so that its smaller side is 1080 pixels. We train our models on full training datasets, but only use the first 10000 images from each dataset for automatic construction of the universal taxonomy. The table shows that automatic universal taxonomy outperforms naive concatenation while substantially reducing the number of training classes.

Taxonomy	#	evals	ADE	Vistas	COCO
naive concatenation	348	N/A	30.7	32.7	36.5
manual univ.	243	N/A	31.3	39.0	34.6
auto univ.	233	44	30.8	37.4	37.7

Table 3: Joint training on ADE20K, Vistas and COCO. Columns show the number of logits (#), number tested hypotheses (evals) and mIoU performance.

Figure 5 presents a qualitative comparison between our automatic taxonomy and naive concatenation. Our automatic taxonomy succeeds to actualize many good class connections, such as mapping ade-food to {ade-food/coco-donut, coco-pizza, coco-sandwich, coco-hot-dog, coco-carrot, coco-food-other}. Furthermore, it finds some coherent connections we did not initially consider in our manual taxonomy such as mapping 'ade-person' to {'vistas-bicyclist', 'vistas-person/ade-person/coco-person', 'coco-baseball glove', 'coco-tie'}.

4.3 Large-scale experiment on the MSeg dataset collection

The MSeg dataset collection [18] encompasses ADE20K [41], BDD (19 classes) [36], Cityscapes (28 classes) [6], COCO [21], IDD (31 classes) [34], SUN RGBD (37 classes) [30] and Vistas [24]. The authors of the MSeg collection adapt all seven datasets towards a custom universal taxonomy of 194 classes. However, their taxonomy entails an omission of 61 classes in order to contain the relabeling effort. Note also that adding a new class to the MSeg taxonomy would require manual relabelling of all seven datasets.

We start the recovery by unifying dataset pairs: BDD-Cityscapes, IDD-Vistas, and ADE-COCO. We proceed by unifying BDD-City with IDD-Vistas, and ADE-COCO with SUN RGBD. Finally, we construct the universal taxonomy over all 7 datasets. If COCO is among
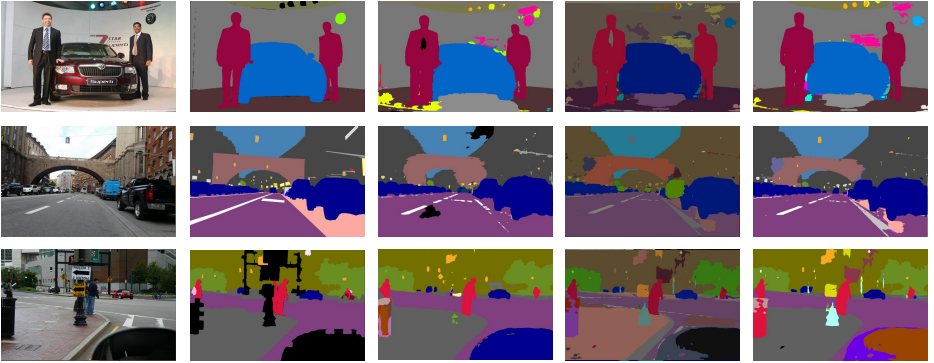
Figure 5: Qualitative comparison of cross-domain models on ADE20K (top), Vistas (middle) and COCO (bottom). We show the input image (column 1), ground truth labels (column 2), predictions of the naive concatenation model (column 3), and predictions of our model in universal (column 4) and dataset-specific labels (column 5). Naive concatenation introduces competition between logits that represent the same visual category. This triggers void predictions (black) on sky and road in the Vistas image. Our universal model finds universal classes that are not present in the corresponding dataset-specific taxonomy and connects them with correct dataset-specific classes: road-marking → road and curb → sidewalk in COCO, tie → person in ADE20K and van → car in Vistas.

the training datasets, we train for 20 instead of 100 epochs.

Table 4 compares our automatic universal taxonomy to the manual universal taxonomy and the MSeg taxonomies. Our automatic taxonomy performs comparably to the manual universal taxonomy [2] while outperforming MSeg taxonomy and naive concatenation. Our automatic taxonomy contains less classes than the manual universal taxonomy. This happens due to a few incorrect associations between rare classes such as equating city-caravan, ade-washer and coco-toaster. Furthermore, our approach brings some debatable but arguably correct decisions due to visual similarity. For example, vistas-pole is associated with bdd-pole/city-pole/vistas-pole/ade-pole, coco-baseball bat, coco-skis, sun-night-stand, and ade-column-pillar.

Interestingly, our approach finds some potentially valid connections that we did not initially consider in our manual taxonomy. For instance, it associates vistas-water to ade-swimming pool (there is often water in swimming pools), city-person to coco-handbag (people carry handbags) and bdd-fence to ade-cradle (cradles often have a safety fence).

Taxonomy	#	evals	ADE	BDD	City	COCO	IDD	SUN	Vistas
naive concat.	469	N/A	27.0	55.6	69.0	29.8	51.3	37.4	33.7
manual univ. [2]	294	N/A	31.0	58.5	72.6	35.4	54.4	41.7	39.1
MSeg original	194	N/A	23.3	59.4	72.6	30.3	42.6	40.2	26.1
auto univ.	243	164	30.7	59.6	72.7	35.6	55.2	42.3	35.8

Table 4: Multi-domain performance evaluation (mIoU %) on the MSeg collection [18]. Unlike [18], we perform evaluation on all classes from the original dataset taxonomies as in [2].

5   Conclusion

We have presented a proof-of-concept for automatic construction of interpretable universal taxonomies for collections of multi-domain datasets with incompatible taxonomies. Our method constructs a set of 1:N mappings which associate dataset-specific classes with their universal counterparts. These mappings establish a hierarchy of visual concepts across particular taxonomies and equip our universal models with a degree of interpretability. The resulting universal taxonomies allow training in the universal label space by treating dataset-specific classes as partial labels.

Our construction approach proceeds by iterative pairwise unification. The unification procedure operates by testing hypothesized relationships between dataset specific classes. We create hypotheses by analyzing a bipartite graph between intra-domain (pseudo) labels and extra-domain predictions. We disambiguate hypotheses according to mIoU performance of the naive concatenation model with post-inference mapping on all involved training datasets.

We evaluate our universal taxonomies in experiments on dataset collections with incompatible taxonomies. We consider collections from the same domain as well as cross-domain collections. We use lightweight models to reduce the training time, yet still succeed to infer coherent relations between classes. Our universal models can deliver both universal and dataset-specific predictions without decreasing inference speed. The reduced number of training logits indicates that our models are more memory-efficient than ad-hoc alternatives.

Our automatic universal taxonomies outperform the naive concatenation baseline and perform comparably to manually designed taxonomies. They are also much more flexible than custom universal taxonomies designed for the standard NLL loss [18, 58], since we can exploit the full training potential of a given dataset collection without any relabeling effort. We observe the best relative performance of our models in large-scale experiments.

Future work should examine ways of streamlining the universal taxonomy construction and explore alternatives for hypothesizing relations between dataset specific classes.

6   Acknowledgement

References

[1] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Dense open-set recognition based on training with noisy negative images. *Image and Vision Computing*, 124:104490, 2022. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2022. 104490. URL https://www.sciencedirect.com/science/article/pii/S0262885622001196.

[2] Petra Bevandić, Marin Oršić, Ivan Grubišić, Josip Šarić, and Siniša Šegvić. Multi-domain semantic segmentation with overlapping labels. In *Proceedings of the*

IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 2615–2624, January 2022.

[3] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and César Cadena. Pixel-wise anomaly detection in complex driving scenes. In *Computer Vision and Pattern Recognition, CVPR*, 2021.

[4] Adrian Boguszewski, Dominik Batorski, Natalia Ziemba-Jankowska, Tomasz Dziedzic, and Anna Zambrzycka. Landcover.ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1102–1110, 2021.

[5] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *International Conference on Computer Vision, ICCV*, 2021.

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[7] Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *The Journal of Machine Learning Research*, 12:1501–1536, 2011.

[8] Clément Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8): 1915–1929, 2013.

[9] Damien Fourure, Rémi Emonet, Elisa Fromont, Damien Muselet, Natalia Neverova, Alain Trémeau, and Christian Wolf. Multi-task, Multi-domain Learning: application to semantic segmentation and pose regression. *Neurocomputing*, 2017. URL https://hal.archives-ouvertes.fr/hal-01507132.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Kaiming He, Ross B. Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *ICCV*, pages 4917–4926. IEEE, 2019.

[12] Yuanduo Hong, Huihui Pan, Weichao Sun, and Yisong Jia. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *CoRR*, abs/2101.06085, 2021.

[13] Ping Hu, Federico Perazzi, Fabian Caba Heilbron, Oliver Wang, Zhe L. Lin, Kate Saenko, and Stan Sclaroff. Real-time semantic segmentation with fast attention. *IEEE Robotics and Automation Letters*, 6:263–270, 2021.

[14] Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1–3):1–308, 2020. ISSN 1572-2740. doi: 10.1561/0600000079. URL http://dx.doi.org/10.1561/0600000079.

[15] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and CV Jawahar. Universal semi-supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5259–5270, 2019.

[16] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and CV Jawahar. Universal semi-supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5259–5270, 2019.

[17] Dongwan Kim, Yi-Hsuan Tsai, Yumin Suh, Masoud Faraki, Sparsh Garg, Manmohan Chandraker, and Bohyung Han. Learning semantic segmentation from multiple datasets with label shifts. *CoRR*, abs/2202.14030, 2022.

[18] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. Mseg: A composite dataset for multi-domain semantic segmentation. In *CVPR*, 2020.

[19] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022.

[20] Xiaodan Liang, Hongfei Zhou, and Eric Xing. Dynamic-structured semantic propagation network. In *CVPR*, pages 752–761, 2018.

[21] 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 *ECCV*, pages 740–755, 2014.

[22] Shota Masaki, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Multi-domain semantic-segmentation using multi-head model. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2802–2807, 2021. doi: 10.1109/ITSC48978.2021.9564940.

[23] Panagiotis Meletis and Gijs Dubbelman. Training of convolutional networks on multiple heterogeneous datasets for street scene semantic segmentation. In *Intelligent Vehicles Symposium*, pages 1045–1050, 2018.

[24] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. Mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, pages 5000–5009, 2017.

[25] Dong Nie, Jia Xue, and Xiaofeng Ren. Bidirectional pyramid networks for semantic segmentation. In *ACCV*, pages 654–671, 2020.

[26] Marin Oršić and Siniša Šegvić. Efficient semantic segmentation with pyramidal fusion. *Pattern Recognition*, page 107611, 2021.

[27] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. In-place activated batchnorm for memory-optimized training of dnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5639–5647, 2018.

[28] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *Commun. ACM*, 63(12):54–63, 2020.

[29] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, 2017. doi: 10.1109/TPAMI.2016.2572683.

[30] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, pages 567–576, 2015.

[31] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. *CoRR*, abs/2111.12264, 2021.

[32] Svenja Uhlemeyer, Matthias Rottmann, and Hanno Gottschalk. Towards unsupervised open world semantic segmentation. *CoRR*, abs/2201.01073, 2022.

[33] Jasper Uijlings, Thomas Mensink, and Vittorio Ferrari. The missing link: Finding label relations across datasets. *CoRR*, abs/2206.04453, 2022.

[34] Girish Varma, Anbumani Subramanian, Anoop M. Namboodiri, Manmohan Chandraker, and C. V. Jawahar. IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *WACV*, pages 1743–1751, 2019.

[35] Wei Yin, Yifan Liu, Chunhua Shen, Anton van den Hengel, and Baichuan Sun. The devil is in the labels: Semantic segmentation from sentences. *CoRR*, abs/2202.02002, 2022.

[36] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018.

[37] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. Wilddash - creating hazard-aware benchmarks. In *ECCV*, 2018.

[38] Oliver Zendel, Matthias Schörghuber, Bernhard Rainer, Markus Murschitz, and Csaba Beleznai. Unifying panoptic segmentation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21351–21360, June 2022.

[39] Xiangyun Zhao, Samuel Schulter, Gaurav Sharma, Yi-Hsuan Tsai, Manmohan Chandraker, and Ying Wu. Object detection with a unified label space from multiple datasets. In *ECCV*, pages 178–193, 2020.

[40] Mingmin Zhen, Jinglu Wang, Lei Zhou, Tian Fang, and Long Quan. Learning fully dense neural networks for image semantic segmentation. In *AAAI*, 2019.

[41] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017.

[42] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *CVPR*, 2022.

[43] Aleksandar Zlateski, Ronnachai Jaroensri, Prafull Sharma, and Frédo Durand. On the importance of label quality for semantic segmentation. In *CVPR*, pages 1479–1487, 2018.