

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053

FINDING AGREEMENT IN DISAGREEMENT: SIMULTANEOUS LABEL ALIGNMENT AND MULTI-DATASET TRAINING WITH SLAMDUNKS

Anonymous authors

Paper under double-blind review

ABSTRACT

Multi-dataset training is a key strategy for improving the versatility and robustness of deep models, but its effectiveness is often hindered by unaligned and contradictory dataset taxonomies. These inconsistencies introduce training noise and prevent effective knowledge sharing. To address this, we propose SLAMDUNKS, a framework for simultaneous multi-dataset training and label alignment. Its core is a shared feature extractor trained with two competing heads: a gating head that determines which dataset-specific classes should be shared, and a classification head that maps samples to the emerging shared taxonomy. To rigorously evaluate alignment quality, we introduce a synthetic benchmark where ground-truth relations are modeled as bipartite graphs. Our method demonstrates remarkable precision, perfectly recovering the true taxonomy (a Graph Edit Distance of 0) for same-domain datasets. Across more challenging cross-domain pairs, SLAMDUNKS achieves an Average Precision of 0.8, outperforming the state-of-the-art by 0.1 to 0.2 and validating its superior alignment capabilities.

1 INTRODUCTION

Machine learning has driven remarkable progress in domains such as natural language processing DeepSeek-AI (2025), medicine Schmidt et al. (2024), and climate science Eyring et al. (2024). A key driver of this success is labeled data, which connects raw samples to meaningful applications. However, even within the same domain, datasets often reflect different labeling perspectives. For example, an animal can be classified by species, diet, or reproductive strategy. A platypus may therefore be grouped with mammals (e.g., lions), carnivores (e.g., venus flytraps), or egg-laying species (e.g., birds). This diversity produces specialized datasets with overlapping content that could, in principle, be combined for better training Zhou et al. (2022); Bevandić et al. (2024). Yet combining datasets requires mapping samples into a common taxonomy, a process that is rarely straightforward: labeling policies may be inherited Krešo et al. (2018), redesigned Lambert et al. (2020), or left ambiguous, and automation remains difficult.

Vision–language models offer partial solutions by aligning classes through names or descriptions Li et al. (2022), or dynamically relabeling into new taxonomies with foundation models Sun et al. (2025). However, these approaches often require human intervention or rely on manually defined taxonomies, and purely language-based alignment is prone to inconsistencies. For example, in COCO Lin et al. (2014), tie is a separate class, while in ADE20K Zhou et al. (2017) it belongs to person Bevandić & Šegvić (2022). Both datasets also include carpet and floor, but differ in how carpeted floors are annotated Uijlings et al. (2022). Even recent open-vocabulary methods employ ad hoc class renamings Ghiasi et al. (2022); Xu et al. (2023), which may improve performance but are rarely systematic or sufficiently discussed. Such inconsistencies undermine rigorous evaluation by obscuring the semantic relation between training and test data.

Language can nonetheless provide a useful signal for discovering class relations, but models must also observe dataset samples directly. Prior work shows that unified taxonomies can be constructed if class mappings are known Bevandić et al. (2022), or recovered when relationships are specified Bevandić & Šegvić (2022). We consider two classes related if they share at least one underlying

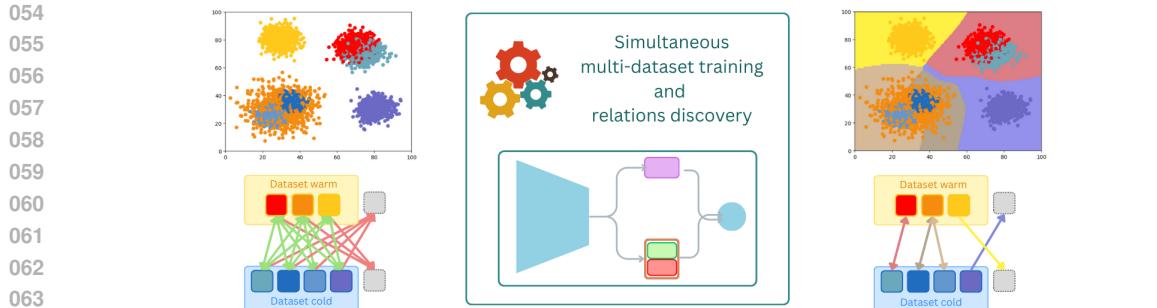


Figure 1: Overview of our proposed approach. Given a dataset pair, we start from the assumption that all the classes may be related to each other or to an additional rest of the world class. Our model has two competing heads on top of a shared feature extractor. Our custom loss results in a model which finds the correct semantic relations and is able to reason into a recovered shared taxonomy.

concept—for instance, mammals and egg-laying species are linked by the platypus, and the shared concept of an egg-laying mammal can itself be treated as a standalone class.

We propose SLAMDUNKS (Synchronising Labels Across Multiple Datasets for UNinterrupted Knowledge Sharing), a model that simultaneously discovers semantic relations and trains a classifier in the unified taxonomy. Starting with two datasets with only dataset-specific labels, we assume each class may relate to any class in the other dataset or remain standalone. Each potential relation defines a candidate in the universal taxonomy, establishing an upper bound on trainable categories. Training proceeds over all candidates, and only valid concepts remain active. The resulting model classifies samples into the unified taxonomy while revealing class relations (Figure 1).

Evaluating such methods is equally important. Standard metrics based on single-dataset performance Lambert et al. (2020); Bevandić & Šegvić (2022) can be misleading, as improvements may simply reflect model capacity rather than genuine alignment Rong et al. (2024). To address this, we propose an evaluation framework with controlled ground-truth relations. By constructing custom taxonomies from existing datasets, we test alignment quality directly. We measure relation discovery with average precision (AP) and graph edit distance (GED), and report accuracy on the unified taxonomy to assess multi-dataset training.

Our contributions are as follows:

- We formalize relation discovery as a standalone task, introducing synthetic dataset collections with controllable ground-truth relations and evaluation via AP and GED.
- We present SLAMDUNKS, a model that simultaneously aligns taxonomies and trains classifiers across datasets.
- We validate SLAMDUNKS extensively, showing that it outperforms existing methods and uniquely handles standalone classes.

2 RELATED WORK

We discuss the two tasks performed by SLAMDUNKS: multi-dataset training and semantic relation discovery in dataset collections.

2.1 MULTI-DATASET TRAINING

Multi-dataset training continues to be a powerful paradigm for improving generalization and mitigating dataset-specific biases. Initial strategies either used dataset-specific prediction heads on shared backbones Kalluri et al. (2019), or merged datasets without resolving label inconsistencies Masaki et al. (2021). Subsequent refinement introduced cross-logit interaction Fourure et al. (2017) and open-set recognition techniques to detect unknown or novel classes Chan et al. (2021); Biase et al. (2021); Uhlemeyer et al. (2022), treating non-primary datasets as anomaly or negative domains Tian et al. (2021).

108 Label taxonomy alignment has also evolved. Approaches using hierarchical structures Liang et al.
 109 (2018); Meletis & Dubbelman (2018) often required complex pipelines, while manual relabeling
 110 collapsed labels at the expense of specificity Lambert et al. (2020); Zendel et al. (2022). A more
 111 scalable approach constructs a universal taxonomy, representing dataset-specific labels as unions
 112 of universal classes—supporting consistent multi-dataset training without altering annotations Be-
 113 vandić et al. (2024); Cour et al. (2011).

114 Recent works have further advanced the field: DaTaSeg Gu et al. (2023) and LMSeg Zhou et al.
 115 (2023) leverage modular architectures and language-driven objectives for cross-dataset alignment.
 116 LMSeg, in particular, employs textual supervision (e.g., CLIP) to link heterogeneous annotations
 117 dynamically. HTTS Meletis & Dubbelman (2023) integrates a taxonomy-aware transformer archi-
 118 tecture, exploiting hierarchical label priors via network design. MultiTalent Ulrich et al. (2023)
 119 demonstrates improved results by co-training across varied medical segmentation datasets. TMT-
 120 VIS Zheng et al. (2023) applies taxonomy-aware joint training within video instance segmentation.
 121 Plain-Det Shi et al. (2024) delivers competitive performance through architectural simplicity and
 122 carefully designed multi-dataset training strategies.

123

124 2.2 AUTOMATED CROSS-DATASET CLASS RELATION DISCOVERY

125

126 Manual label alignment becomes increasingly impractical at scale. Semantic text embeddings (e.g.,
 127 CLIP-based) offer automation Li et al. (2022); Yin et al. (2022), but are unreliable due to label
 128 ambiguity Uijlings et al. (2022). Visual cues, in contrast, often yield more accurate alignments.

129 One line of work uses post-hoc optimization (e.g., linear programming) on output logits to match
 130 classes, yet these methods assume one-to-one class correspondence and fail to handle hierarchi-
 131 cal relations—leading to competition and reduced performance Zhou et al. (2022); Bevandić et al.
 132 (2022).

133

134 Newer, more flexible approaches include: DynAlign Sun et al. (2025) performs dynamic, unsuper-
 135 vised alignment during training using foundation model priors—allowing detection of fine-grained
 136 and overlapping categories via joint optimization HTTS Meletis & Dubbelman (2023) enforces hier-
 137 archical consistency through its taxonomy-aware transformer architecture. AutoUniSeg Rong et al.
 138 (2024) employs graph neural networks to infer cross-dataset class relations by jointly considering
 139 semantic and visual feature spaces. RESI Zhangli et al. (2024) targets inconsistencies in label se-
 140 mantics through structured, task-specific label alignment.

141

142 Variants of language-vision approaches continue to emerge. LMSeg Zhou et al. (2023) integrates
 143 text-based label representations into training, replacing hard labels with text-guided supervision and
 144 enabling seamless adaptation across datasets and tasks.

145

146 Our work differentiates itself by embedding taxonomy discovery directly into the training loop,
 147 jointly optimizing segmentation performance and label alignment within a universal taxonomy. This
 148 unified, data-driven process supports taxonomic overlap, label granularity variation, and semantic
 149 inconsistency—without relying on manual mappings or fixed label embeddings.

150

151 3 METHOD

152

153 We introduce **SLAMDUNKS**, a model designed for simultaneous multi-dataset training and se-
 154 mantic relation discovery between class taxonomies. The key idea is to embed the discovery of
 155 cross-dataset relations directly into the training loop. After training, the model can classify objects
 156 according to the induced universal taxonomy, without requiring any additional post-processing or
 157 manual label mappings.

158

159 3.1 PRELIMINARIES

160

161 We assume that two classes are related if they share at least one common visual concept. For
 162 example, the classes *Vistas-car* and *ADE-van* are related because both include pickup trucks.

163 We represent such inter-dataset class relations in two equivalent ways:

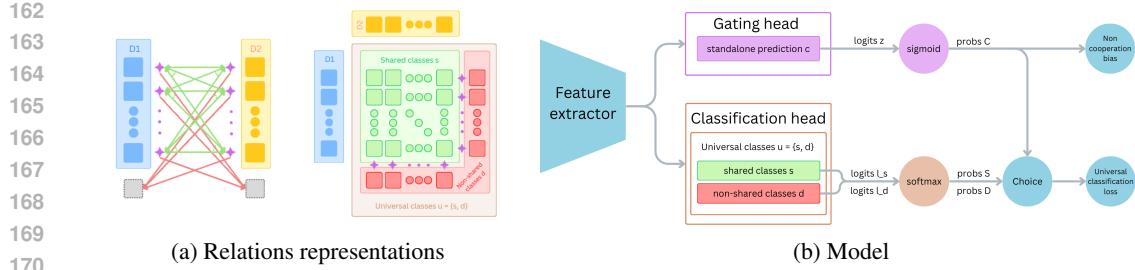


Figure 2: Method

1. **Bipartite Graph.** Vertices correspond to dataset-specific classes, and we introduce an additional dummy vertex for each dataset to represent the *rest of the world* (*RoW*)—concepts outside the given taxonomy. Edges connect related classes. If a class has no valid relation to the other dataset, it connects to its dataset’s *RoW* vertex, marking it as a *standalone class*.
2. **Universal Class Decomposition.** Alternatively, we can define a set of primitive visual concepts (universal classes) that appear across datasets, and express dataset-specific classes as unions of these universal classes. The two views are interchangeable, since each bipartite edge encodes a potential universal class.

When relating two datasets D_1 and D_2 , we make no assumptions about prior alignments. Initially, any class in D_1 may be related to any class in D_2 —including the dummy *RoW* class. Multiple relations per class are permitted, except for *RoW* connections, which indicate that the class is entirely *standalone*. This setup defines an upper bound on the number of recoverable universal classes:

$$(|D_1| + 1)(|D_2| + 1) - 1.$$

Figure 2a illustrates this framework. Green edges denote potential inter-class relations, red edges connect classes to the *RoW* vertex (standalone classes), and purple gates indicate the model’s choice between shared and standalone assignments. The bipartite graph can also be represented as a matrix of universal classes, where each dataset-specific class corresponds to a union of rows or columns in the matrix.

3.2 SLAMDUNKS

Our model builds on a shared feature extractor with two task-specific heads that jointly determine class relations and classify samples in the universal taxonomy.

Gating Head. The first head produces $|D_1| + |D_2|$ logits, denoted by z , with a sigmoid applied to each. This head can be interpreted as a collection of binary classifiers, one per dataset-specific class, that predicts whether the class is *shared* with the other dataset or *standalone*. These predictions correspond to the purple gates in Figure 2a.

$$P(C_y = 1 | \mathbf{x}) = \frac{1}{1 + \exp -z_y}. \quad (1)$$

Classification Head. The second head outputs $(|D_1| + 1)(|D_2| + 1) - 1$ logits, followed by a softmax. This head assigns each sample to one of the universal classes, covering all theoretically possible relations. Specifically: - $|D_1||D_2|$ logits l_s represent all candidate *shared classes*, and - $|D_1| + |D_2|$ logits l_d represent all possible *standalone classes*.

$$P(S = s' | \mathbf{x}) = \frac{\exp l_{s'}}{\sum_{s \in S} \exp l_s + \sum_{d \in D} \exp l_d}. \quad (2)$$

$$P(D = d_y | \mathbf{x}) = \frac{\exp l_{d_y}}{\sum_{s \in S} \exp l_s + \sum_{d \in D} \exp l_d}. \quad (3)$$

216 3.3 TRAINING OBJECTIVE
217218 The two heads of SLAMDUNKS interact to define the posterior probability of each dataset-specific
219 class. Intuitively, the cooperation (gating) head decides whether a class is *shared* or *standalone*,
220 while the classification head provides the distribution over all candidate universal classes.
221222 **Shared classes.** If a dataset-specific class y is shared, its posterior probability equals the product
223 of (i) the probability that it is not standalone, and (ii) the probability mass assigned to the set of
224 related universal classes:
225

226
$$P(Y = y | x, s) = P(C_y = 0 | x) \sum_{s' \in m_{S_d}(y)} P(S = s' | x). \quad (4)$$

227

228 **Standalone classes.** If y is not shared, the posterior is given by the product of (i) the probability
229 that it is standalone, and (ii) the probability of its associated standalone universal class:
230

231
$$P(Y = y | x, ns) = P(C_y = 1 | x) P(D = d_y | x). \quad (5)$$

232

233 The two remaining combinations of outputs are invalid and ignored.
234236 **Total posterior.** The overall posterior probability of class y is obtained by summing the two valid
237 cases:
238

239
$$P(Y = y | x) = P(C_y = 0 | x) \sum_{s' \in m_{S_d}(y)} P(S = s' | x) + P(C_y = 1 | x) P(D = d_y | x). \quad (6)$$

240

242 Expanding this expression highlights the role of the cooperation logits z :
243

244
$$P(Y = y | x) = \frac{\sum_{s' \in m_{S_d}(y)} \exp(l_{s'} - z_y) + \exp(l_{d_y})}{(1 + \exp(-z_y)) (\sum_{s \in S} \exp(l_s) + \sum_{d \in D} \exp(l_d))}. \quad (7)$$

245

247 This formulation shows that z_y dampens shared logits l_s whenever the model favors a standalone
248 interpretation, allowing the standalone logit l_{d_y} to dominate. Without this mechanism, standalone
249 logits would not be learnable, as they always co-occur with shared ones—an issue observed in
250 weakly supervised settings Cour et al. (2011); Bevandić et al. (2024).
251252 **Loss function.** We train SLAMDUNKS by minimizing the negative log-likelihood of the dataset-
253 specific posterior, augmented with a bias toward the standalone prediction:
254

255
$$\mathcal{L}(x, y | m_{S_d}) = -\ln P(Y = y | x) - \lambda \ln P(C_y = 1 | x). \quad (8)$$

256

257 This objective creates a competition between the two heads: the cooperation head pushes classes
258 toward standalone assignments, while the classification head pulls them toward shared relations
259 when evidence supports it. If two classes are easily distinguishable, the bias toward standalone
260 wins; if they are visually similar, the classification head overrides this bias to link them.
261262 **Inference.** At test time, we discard the cooperation head and use only the classification head to
263 predict labels in the universal taxonomy.
264265 4 EXPERIMENTS
266268 We evaluate SLAMDUNKS on the task of semantic relation discovery across datasets. To this end,
269 we describe our experimental setup, the construction of synthetic dataset collections, evaluation
metrics, baselines, and implementation details.
270

270 4.1 EXPERIMENTAL SETUP
271272 Most prior work has concentrated on semantic segmentation benchmarks. While realistic, these
273 require very large models and long training cycles, making systematic evaluation difficult. To enable
274 controlled and reproducible experiments, we instead focus on the image classification setting.
275276 Specifically, we create *synthetic dataset collections* by splitting and re-labeling existing classifica-
277 tion datasets. This approach has two advantages: (i) it allows us to define *ground-truth relations*
278 between classes explicitly, and (ii) it provides a corresponding *universal taxonomy* against which
279 recovered relations can be validated.
280281 With ground truth available, we can move beyond the common—but potentially misleading—metric
282 of per-dataset accuracy. Instead, we directly evaluate (i) the correctness of extracted relations and (ii)
283 performance in the unified label space, which better reflects the objectives of multi-dataset training.
284285 4.2 DATASETS
286287 We create 5 synthetic dataset pairs with MNIST Deng (2012), CIFAR-10 Krizhevsky (2009) and
288 SVHN Netzer et al. (2011). All of these datasets have 10 classes in the original label space. For four
289 of the five pairs we create two custom taxonomies by merging some classes and omitting others. We
290 present the two splits by listing the concepts represented by their indices in the original label space.
291 Dataset 1 has six classes: $\{0\}$, $\{1, 2\}$, $\{3\}$, $\{4\}$, $\{5, 6, 7\}$ and $\{8\}$. Dataset 2 has eight classes: $\{1\}$,
292 $\{2, 3\}$, $\{4\}$, $\{5\}$, $\{6\}$, $\{7\}$, $\{8\}$, $\{9\}$. These splits cover all interesting scenarios between classes:
293 equality between classes, subset-superset relation, overlapping classes and not related classes. We
denote the custom taxonomy with the extension "-C" in the pairs name
294295 **MNIST-C** and **CIFAR-C** are made by splitting MNIST and CIFAR10 into two subsets of equal size.
These datasets represent the simplest experimental setting, as the dataset pairs share the domain.
296297 **MNIST-SVHN** uses MNIST and SVHN as the dataset pair. These datasets share a taxonomy (dig-
298 its), which means that the only relation that exists between classes is equality. Still, this setup is
challenging, as there is a significant domain shift between the two datasets.
299300 Finally, **MNIST-SVHN-C** and **SVHN-MNIST-C** apply the custom taxonomies to MNIST and
301 SVHN, where the ordering indicates which dataset has taxonomy 1, and which taxonomy 2. This
302 experimental setup is the most challenging as there is both a domain shift and complex relations
303 between classes
304305 4.3 METRICS
306307 A common way to evaluate multi-dataset training is to report performance on each individual dataset.
308 However, this can be misleading: a large model may simply memorize and separate datasets without
309 truly sharing knowledge or discovering meaningful relations.
310311 Since our synthetic dataset collections provide ground-truth relations and a known universal taxon-
312 omy, we evaluate *relation discovery* directly as a standalone task, alongside classification perfor-
313 mance in the shared label space.
314315 **Relation discovery as classification.** As described in Subsection 3.1, inter-dataset relations can
316 be represented as a bipartite graph connecting classes across datasets. The total number of candidate
317 relations is $(|D_1|+1)(|D_2|+1)-1$. Thus, relation discovery can be viewed as a binary classification
318 problem, where the model must decide whether each potential relation is valid. If a method outputs
319 continuous confidence scores for relations, we measure performance using **Average Precision (AP)**,
320 which evaluates the precision–recall tradeoff without requiring a threshold.
321322 **Graph-based evaluation.** If a method instead outputs a discrete set of relations, we compare
323 the resulting bipartite graph to the ground truth using **Graph Edit Distance (GED)**. We count the
number of edge additions and deletions required to transform the predicted graph into the true one.
324

324	Method	MNIST-C AP	CIFAR-10-C AP	MNIST-SVHN AP	MNIST-SVHN-C AP	SVHN-MNIST-C AP
325	Missing link	0.8 ± 0.0	0.8 ± 0.0	0.6 ± 0.1	0.6 ± 0.1	0.5 ± 0.1
326	SLAMDUNKS	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	0.7 ± 0.1	0.7 ± 0.1
327	Method	MNIST-C GED	CIFAR-10-C GED	MNIST-SVHN GED	MNIST-SVHN-C GED	SVHN-MNIST-C GED
328	AUT	5.4 ± 1.0	4.2 ± 0.4	4.6 ± 7.8	6.2 ± 4.0	7.0 ± 2.7
329	SLAMDUNKS	0.0 ± 0.0	0.4 ± 0.8	4.0 ± 2.2	8.8 ± 1.3	14.4 ± 2.6

332 Table 1: Quantitative results of relation discovery on custom classification dataset pairs. We measure
 333 average precision in the top part of the table to compare our approach to Missing link, and graph
 334 edit distance to compare our approach to Automatic Universal Taxonomies.

337 **Classification accuracy.** Finally, because SLAMDUNKS simultaneously learns to classify samples into the universal taxonomy, we also report **accuracy in the shared label space**. This provides
 338 an additional measure of how well the recovered relations support effective multi-dataset training.

341 4.4 BASELINES

343 We compare SLAMDUNKS against two existing approaches, adapted from semantic segmentation to the image classification setting: (i) **Missing Link** Uijlings et al. (2022), and (ii) **Automatic
 344 Universal Taxonomies (AUT)** Bevandić & Šegvić (2022).

347 **Missing Link.** This method trains a model on dataset A and applies it to dataset B to estimate
 348 correlations between predictions and ground-truth labels. We extend the approach to open-set recogni-
 349 tion by using max-softmax to detect outliers. Repeating this procedure in both directions yields
 350 two correlation matrices, which can be interpreted as weighted bipartite graphs. The final graph is
 351 obtained by merging edges and taking the larger of the two weights. Since this method does not
 352 output a conclusive relation graph, we evaluate it using **Average Precision (AP)** only.

353 **Automatic Universal Taxonomies (AUT).** AUT trains a model whose output label space is
 354 formed by concatenating the classes from both datasets. Two co-occurrence matrices are then com-
 355 puted, one for dataset A and one for dataset B . We adapt the approach to include open-set prediction
 356 via max-softmax. Only the strongest connections are retained, followed by pruning of contradic-
 357 tory edges that degrade individual dataset performance. Because this method outputs only a discrete set
 358 of accepted relations, we evaluate it using **Graph Edit Distance (GED)**.

360 4.5 IMPLEMENTATION DETAILS

362 We use a ResNet-18 backbone He et al. (2016) initialized with ImageNet pretraining. Each head
 363 consists of a single linear layer with weights and biases initialized to zero. Training is performed
 364 with mini-batches corresponding to 0.5% of the training set (but never fewer than 10 samples). Each
 365 batch is balanced to contain an equal number of samples from both datasets. To further stabilize
 366 training, we applied weighted sampling that favors rarer classes within each dataset. For MNIST
 367 and CIFAR experiments, we train for 5 epochs with a learning rate of 1×10^{-4} with Adam optimizer.
 368 For MNIST–SVHN experiments, we train for 20 epochs with a learning rate of 5×10^{-4} and resize
 369 the images to 32×32 pixels. We do not apply weight decay. The coefficient λ in Equation 8 is
 370 selected via a simple linear search in $[0, 1]$. All reported results are the mean and variance over five
 371 independent runs. Experiments are conducted on a single NVIDIA GTX 1080 GPU.

373 5 RESULTS

375 We organize our results into three parts. First, we evaluate the relation discovery ability of SLAM-
 376 DUNKS by comparing it with baseline methods. Second, we present ablation studies that assess
 377 the impact of key design choices. Finally, we analyze the feature spaces learned by different multi-
 dataset training methods.

378	pretrain	zero init	weighted sampling	MNIST-C		CIFAR-10-C		MNIST-SVHN	
				AP	GED	AP	GED	AP	GED
380		✓		0.6 ± 0.2	7.6 ± 4.3	0.3 ± 0.1	13.4 ± 2.7	0.1 ± 0.0	26.6 ± 2.1
381			✓	0.4 ± 0.1	11.4 ± 3.8	0.3 ± 0.1	12.8 ± 3.2	0.1 ± 0.0	28.8 ± 2.0
382	✓	✓		0.8 ± 0.0	4.0 ± 0.0	0.8 ± 0.1	6.0 ± 0.0	0.6 ± 0.2	16.6 ± 3.1
383			✓	1.0 ± 0.0	0.0 ± 0.0	0.8 ± 0.2	2.0 ± 2.4	0.8 ± 0.1	11.0 ± 3.0
384	✓		✓	0.7 ± 0.1	5.6 ± 2.7	0.6 ± 0.3	10.0 ± 6.1	0.2 ± 0.1	25.8 ± 2.1
385		✓	✓	0.5 ± 0.2	10.8 ± 2.8	0.2 ± 0.1	16.6 ± 2.2	0.1 ± 0.1	28.8 ± 2.1
			✓	0.7 ± 0.1	4.0 ± 0.0	0.7 ± 0.0	6.0 ± 0.0	0.6 ± 0.2	12.2 ± 6.0
	✓	✓	✓	1.0 ± 0.0	0.0 ± 0.0	1.0 ± 0.0	0.4 ± 0.8	1.0 ± 0.0	4.0 ± 2.2

Table 2: Quantitative results on custom classification dataset pairs

5.1 RELATION RECOVERY

Our first set of experiments compares the proposed *SLAMDUNKS* with two baselines: (i) Missing Link Uijlings et al. (2022), and (ii) Automatic Universal Taxonomies Bevandić & Šegvić (2022).

For Missing Link, we report only AP, since producing a conclusive relation graph would require thresholding its outputs. For AUT, we report only GED, as the method does not yield confidence scores for relations. Table 1 summarizes the results.

SLAMDUNKS clearly outperforms Missing Link across all dataset pairs. Missing Link relies on training two separate models, each of which tends to overfit to its respective dataset. This limits cross-dataset generalization and produces inferior correlation matrices for relation discovery.

On same-domain dataset pairs, SLAMDUNKS also surpasses AUT. AUT frequently introduces spurious relations for outlier classes, despite having a pruning step to mitigate them. On cross-domain dataset pairs, results are more mixed: while AUT continues to suffer from spurious connections, SLAMDUNKS sometimes separates classes that should be connected. This behavior stems from the gating head, which more readily favors standalone assignments under domain shift.

5.2 ABLATION STUDIES

We next investigate how different design choices affect the performance of *SLAMDUNKS*. Our analysis covers initialization strategies, sampling, pretraining, parameter λ , and dataset size.

Two considerations are particularly important: (i) initializing the weights and biases of both heads to zero, and (ii) applying weighted sampling across classes. Zero initialization avoids unintended bias at the start of training: in the classification head, all potential relations are equally likely, and in the gating head, each class is equally likely to be shared or standalone. Weighted sampling reduces the natural tendency of the gating head to default to predicting frequent classes as standalone.

We also study the effect of initializing the backbone with pretrained ImageNet features. Pretraining provides limited benefit on its own, but it improves performance when combined with zero initialization and weighted sampling. This suggests that pretrained features alone do not capture the semantics required for aligning heterogeneous datasets.

Our approach combines two losses 8 and balanced by the parameter λ . Figure 3 shows the effect of varying λ in terms of AP, GED, and classification accuracy. For same-domain datasets, higher values of λ yield the best results, as strong visual similarity makes it easy for the model to connect classes. For cross-domain datasets, lower λ performs better, as domain shift makes it easier for the

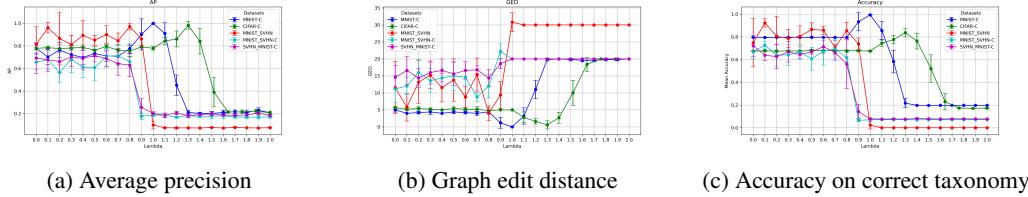


Figure 3: Impact of lambda on taxonomy construction quality

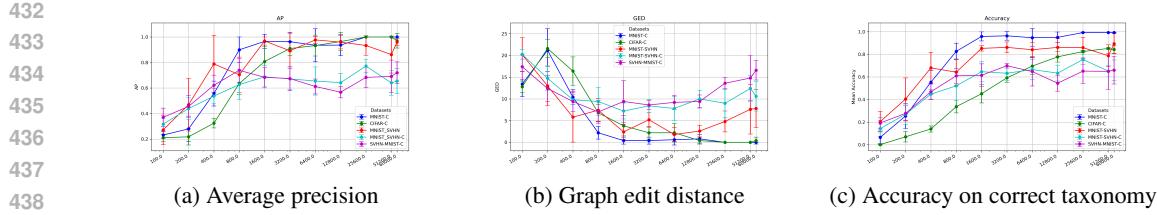


Figure 4: Impact of training set size on taxonomy construction quality

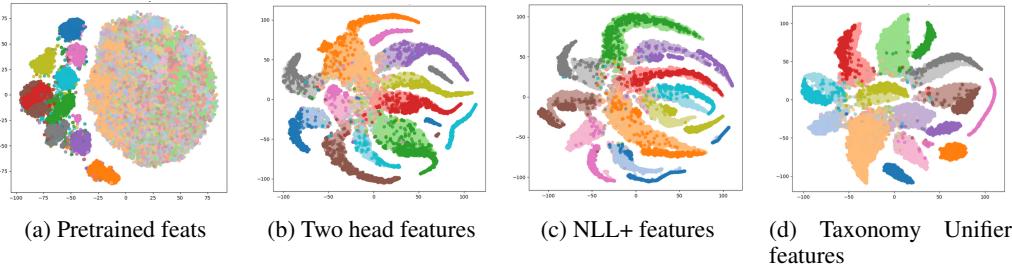


Figure 5: Visualization of feature spaces for SVHN and MNIST sample embeddings for different models

gating head to treat each class as standalone. Interestingly, classification accuracy does not always correspond to GED, indicating incorrect relations may affect relatively few samples in practice.

Finally, we analyze how dataset size influences performance (Figure 4). Larger datasets consistently improve relation discovery, but even with only 25% of the training data, the model already achieves strong results. This suggests that reliable relation discovery can be obtained using subsets of available data, offering a path toward more efficient training.

5.3 MULTI-DATASET TRAINING

We next analyze how different training strategies affect the latent space learned by the feature extractor. We compare SLAMDUNKS against: (i) a pretrained model, (ii) training with dataset-specific heads Kalluri et al. (2019), and (iii) the NLL+ loss Bevandić et al. (2024). All experiments are conducted on MNIST and SVHN, and the resulting feature spaces are visualized in Figure 5.

Using a pretrained ResNet-18 backbone without further adaptation produces feature spaces that clearly separate MNIST from SVHN, with no semantic alignment between related classes. Training with NLL+ in the known universal label space achieves stronger alignment: universal classes are cleanly separated, and samples from both datasets are well mixed within each class. Our approach seems to further improve representation quality. It aligns semantically related classes across datasets, while at the same time preserving a distinction between domains. This demonstrates that SLAMDUNKS simultaneously supports cross-dataset semantic consistency and domain awareness.

6 CONCLUSION

We addressed multi-dataset training and relation discovery in settings where datasets share concepts but have unaligned taxonomies. Our proposed SLAMDUNKS learns both classification and class relations through two competing heads on a shared backbone: one predicting whether classes are standalone, and the other mapping samples into a universal taxonomy. This design enables discovery of shared concepts while preserving standalone categories.

Using synthetic dataset pairs with controllable ground-truth relations, we showed that our method outperforms prior approaches, particularly in handling standalone classes, and yields more informative feature spaces. Future work will extend this framework to more complex tasks.

486 REFERENCES
487

488 Petra Bevandić and Siniša Šegvić. Automatic universal taxonomies for multi-domain semantic seg-
489 mentation. In *BMVC*, 2022.

490 Petra Bevandić, Marin Oršić, Ivan Grubišić, Josip Šarić, and Siniša Šegvić. Multi-domain semantic
491 segmentation with overlapping labels. In *Proceedings of the IEEE/CVF Winter Conference on*
492 *Applications of Computer Vision (WACV)*, pp. 2615–2624, January 2022.

493 Petra Bevandić, Marin Oršić, Ivan Grubišić, Josip Šarić, and Siniša Šegvić. Weakly supervised train-
494 ing of universal visual concepts for multi-domain semantic segmentation. *International Journal*
495 *of Computer Vision*, 132:2450 – 2472, 2024.

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

498 Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classifi-
500 cation for out-of-distribution detection in semantic segmentation. In *International Conference on*
501 *Computer Vision, ICCV*, 2021.

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

504 DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning,
505 2025. URL <https://arxiv.org/abs/2501.12948>.

506 Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE*
507 *Signal Processing Magazine*, 29(6):141–142, 2012.

508 Veronika Eyring, William Collins, Pierre Gentine, Elizabeth Barnes, Marcelo Barreiro, Tom Beu-
509 cler, Marc Bocquet, Christopher Bretherton, Hannah Christensen, Katherine Dagon, David
510 Gagne, David Hall, Dorit Hammerling, Stephan Hoyer, Fernando Iglesias-Suarez, Ignacio Lopez-
511 Gomez, Marie McGraw, Gerald Meehl, Maria Molina, and Laure Zanna. Pushing the frontiers in
512 climate modelling and analysis with machine learning. *Nature Climate Change*, 14, 08 2024. doi:
513 10.1038/s41558-024-02095-y.

514 Damien Fourure, Rémi Emonet, Elisa Fromont, Damien Muselet, Natalia Neverova, Alain Tréneau,
515 and Christian Wolf. Multi-task, Multi-domain Learning: application to semantic segmentation
516 and pose regression. *Neurocomputing*, 2017. URL <https://hal.archives-ouvertes.fr/hal-01507132>.

517 Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation
518 with image-level labels. In *ECCV*, pp. 540–557, 2022.

519 Xiuye Gu, Yin Cui, Jonathan Huang, Abdullah Rashwan, Xuan Yang, Xingyi Zhou, Golnaz Ghiasi,
520 Weicheng Kuo, Huizhong Chen, Liang-Chieh Chen, et al. Daseg: Taming a universal multi-
521 dataset multi-task segmentation model. In *NeurIPS*, 2023.

522 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recog-
523 nition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.
524 770–778, 2016.

525 Tarun Kalluri, Girish Varma, Manmohan Chandraker, and CV Jawahar. Universal semi-supervised
526 semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vi-
527 sion*, pp. 5259–5270, 2019.

528 Ivan Krešo, Marin Oršić, Petra Bevandić, and Siniša Šegvić. Robust semantic segmentation with
529 ladder-densenet models. *arXiv preprint arXiv:1806.03465*, 2018.

530 Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL <https://api.semanticscholar.org/CorpusID:18268744>.

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

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

542

543 Xiaodan Liang, Hongfei Zhou, and Eric Xing. Dynamic-structured semantic propagation network.
 544 In *CVPR*, pp. 752–761, 2018.

545

546 Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr
 547 Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pp.
 548 740–755, 2014.

549

550 Shota Masaki, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Multi-domain
 551 semantic-segmentation using multi-head model. In *2021 IEEE International Intelligent Trans-
 552 portation Systems Conference (ITSC)*, pp. 2802–2807, 2021. doi: 10.1109/ITSC48978.2021.
 553 9564940.

554

555 Panagiotis Meletis and Gijs Dubbelman. Training of convolutional networks on multiple hetero-
 556 geneous datasets for street scene semantic segmentation. In *Intelligent Vehicles Symposium*, pp.
 557 1045–1050, 2018.

558

559 Panagiotis Meletis and Gijs Dubbelman. Training semantic segmentation on heterogeneous datasets.
 560 *arXiv preprint arXiv:2301.07634*, 2023.

561

562 Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Ng. Reading
 563 digits in natural images with unsupervised feature learning. *NeurIPS*, 01 2011.

564

565 Ma Rong, Jie Chen, Xiangyang Xue, and Jian Pu. Automated label unification for multi-dataset
 566 semantic segmentation with gnns. In *NeurIPS*, 2024.

567

568 Adam Schmidt, Omid Moharer, Simon DiMaio, Michael C. Yip, and Septimiu E. Salcudean. Track-
 569 ing and mapping in medical computer vision: A review. *Medical Image Analysis*, 94:103131,
 570 2024. ISSN 1361-8415. doi: <https://doi.org/10.1016/j.media.2024.103131>. URL <https://www.sciencedirect.com/science/article/pii/S1361841524000562>.

571

572 Cheng Shi, Yuchen Zhu, and Sibei Yang. Plain-det: A plain multi-dataset object detector. In *ECCV*,
 573 2024.

574

575 Han Sun, Rui Gong, Ismail Nejjar, and Olga Fink. Dynalign: Unsupervised dynamic taxonomy
 576 alignment for cross-domain segmentation. In *ICLR*, 2025.

577

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

581

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

584

585 Jasper R. R. Uijlings, Thomas Mensink, and Vittorio Ferrari. The missing link: Finding label
 586 relations across datasets. In *ECCV*, pp. 540–556, 2022.

587

588 Constantin Ulrich, Fabian Isensee, Tassilo Wald, Maximilian Zenk, Michael Baumgartner, and
 589 Klaus H Maier-Hein. Multitalent: A multi-dataset approach to medical image segmentation.
 590 In *MICCAI*, 2023.

591

592 Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello.
 593 Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. *arXiv preprint
 594 arXiv:2303.04803*, 2023.

595

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

598

599 Oliver Zendel, Matthias Schörghuber, Bernhard Rainer, Markus Murschitz, and Csaba Beleznai.
 600 Unifying panoptic segmentation for autonomous driving. In *Proceedings of the IEEE/CVF Con-
 601 ference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21351–21360, June 2022.

594 Qilong Zhangli, Di Liu, Abhishek Aich, Dimitris Metaxas, and Samuel Schulter. Resolving inco-
595 sistent semantics in multi-dataset image segmentation. *arXiv preprint arXiv:2409.09893*, 2024.
596

597 Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, Yu Qiao, and Hengshuang Zhao. Tmt-vis:
598 Taxonomy-aware multi-dataset joint training for video instance segmentation. In *NeurIPS*, 2023.

599 Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene
600 parsing through ade20k dataset. In *CVPR*, pp. 633–641, 2017.

601

602 Qiang Zhou, Yuang Liu, Chaohui Yu, Jingliang Li, Zhibin Wang, and Fan Wang. Lmseg: Language-
603 guided multi-dataset segmentation. In *ICLR*, 2023.

604 Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *CVPR*,
605 2022.

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647