TASK2BOX: Box Embeddings for Modeling Asymmetric Task Relationships

Rangel Daroya Aaron Sun Subhransu Maji
University of Massachusetts, Amherst
{rdaroya, aaronsun, smaji}@umass.edu

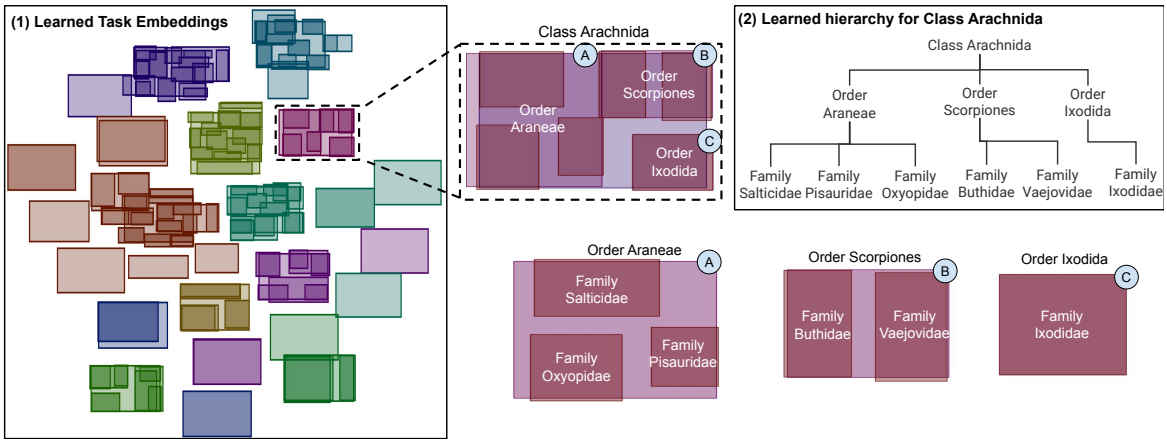

Figure 1. **Box Embeddings of 150 Datasets of iNaturalist + CUB and Corresponding Learned Hierarchy for Class Arachnida**. Each taxonomic category is treated as a separate dataset for which TASK2BOX embeddings are learned. (1) Shows the learned box embeddings where datasets from the same group (taxonomic class) have the same color. Datasets naturally cluster to their ground truth groups. (2) Shows the hierarchy learned through TASK2BOX for a specific class. The hierarchy matches the ground truth relationships based on biological classification. Orders that belong to class Arachnida are learned as boxes (Ⓐ, Ⓑ, Ⓒ) contained by the larger box for Arachnida; families under each of the orders are learned as smaller boxes contained by the corresponding orders they belong to.

Abstract

Modeling and visualizing relationships between tasks or datasets is an important step towards solving various meta-tasks such as dataset discovery, multi-tasking, and transfer learning. However, many relationships, such as containment and transferability, are naturally asymmetric and current approaches for representation and visualization (e.g., t-SNE [44]) do not readily support this. We propose TASK2BOX*, an approach to represent tasks using box embeddings—axis-aligned hyperrectangles in low dimensional spaces—that can capture asymmetric relationships between them through volumetric overlaps. We show that* TASK2BOX *accurately predicts unseen hierarchical relationships between nodes in ImageNet and iNaturalist datasets, as well as transferability between tasks in the Taskonomy benchmark. We also show that box embeddings estimated from task representations (e.g., CLIP [36], Task2Vec [4], or attribute based [15]) can be used to predict relationships between unseen tasks more accurately than classifiers trained on the same representations, as well as handcrafted asymmetric distances (e.g., KL divergence). This suggests that low-dimensional box embeddings can effectively capture these task relationships and have the added advantage of being interpretable. We use the approach to visualize relationships among publicly available image classification datasets on popular dataset hosting platform called Hugging Face.*

1. Introduction

The success of deep learning has led to the proliferation of datasets for solving a wide range of computer vision problems. Yet, there are few tools available to enable practitioners to find datasets related to the task at hand, and to solve various meta-tasks related to it. We present TASK2BOX, a method to represent tasks using axis-aligned hyperrectangles (or box embeddings). TASK2BOX is framed as a learnable mapping from dataset representation to boxes, and can be trained to predict various relationships between novel tasks such as transferability, hierarchy, and overlap.

Box embeddings [48] extend order embeddings [46] by using volumetric relationships between axis-aligned hyper-

rectangles to encode pairwise relationships. Prior work in natural language processing has utilized box embeddings to represent the WordNet [30] hierarchy and to model conditional distributions. To model relationships between novel datasets, we develop a technique to map from Euclidean representations of datasets into the space of boxes. We explore simple image and label embedding from large vision-language models such as CLIP [36], TASK2VEC [4], and attribute-based vectors [15] as base representations of tasks.

We test our framework to model asymmetric relationships between nodes in iNaturalist [45] and Caltech-UCSD Birds (CUB) [50] and ImageNet [11] datasets, as well as to predict transferability on the Taskonomy benchmark [52]. Table 1 and 2 show that low-dimensional box embeddings accurately predict novel relationships between datasets seen during training, as well as relationships with novel datasets. Remarkably, TASK2BOX outperforms classifiers trained to directly predict the relationships on the same representations, suggesting that the box embedding provides a strong inductive bias for learning hierarchical relationships. We also outperform simple asymmetric distances proposed in prior work such as Kullback-Leibler (KL) divergence [4]. To model the heterogeneous tasks in the Taskonomy benchmark [52] we map each task to a set of attributes from which a box embedding is learned. Once again, we obtain significantly higher correlation between the true and predicted transferability for both existing and novel datasets compared to standard classifiers (Table 3). Such attribute-based representations can be readily derived from datasheets [15] and modelcards [31].

Finally, the low-dimensional box embeddings have the added advantage of being interpretable. Fig. 1 and Fig. 3 show relationships on the iNaturalist+CUB and ImageNet categories, respectively. The 2D box representation allows us to readily visualize the strength and direction of task relationships based on the overlapping volumes, which is not possible using symmetric distances with Euclidean representations (e.g., t-SNE [44]). At the same time, new datasets can be embedded in constant time without needing to retrain or re-optimize. Fig. 5 uses TASK2BOX to visualize relationships among 131 publicly available datasets on Hugging Face [1], a popular platform for hosting datasets. Our main contributions are as follows:

- We introduce a novel method (TASK2BOX) that uses box embeddings to learn asymmetric (e.g., hierarchical, transfer learning) dataset relationships.
- We demonstrate that TASK2BOX can predict the relationships of *new tasks* with a collection of existing tasks.
- We illustrate the interpretability of our model, and the ability to visualize public classification datasets on Hugging Face.

The code for this project is publicly available at https://github.com/cvl-umass/task2box.

2. Related Work

Task Representations. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, consisting of images $x_i \in \mathcal{X}$ and labels $y_i \in \mathcal{Y}$, a range of approaches have been proposed for dataset representation. The most straightforward approach involves modeling the distribution of either the images x or the labels y within the dataset independently, using embeddings referred to in prior work as "domain" and "label" embeddings [4]. To capture the joint dependency between images and labels, [4] proposed the use of the Fisher Information Matrix (FIM) derived from a "probe network" trained to minimize a loss function $\ell(\hat{y}, y)$ over the dataset [22, 25, 35]. This approach leverages the similarity of FIMs to predict task transferability and for model selection. However, the utility of the FIM critically depends on the choice of the probe network and a pre-defined similarity may not accurately represent the various relationships between datasets.

We also investigate the use of vision-language models (VLMs) such as CLIP [36]. This model, trained on a wide range of visual domains, can generalize to tasks involving vision and language data. CLIP features have been effective for image classification [3, 9], semantic segmentation [18, 24, 26], object detection [16, 47, 51], and even closing domain gaps for performance improvement [23, 53]. Both images $x \in \mathcal{X}$ and labels $y \in \mathcal{Y}$ represented as text, can be mapped into a shared space using the vision encoder (ϕ) and text encoder (ψ) of CLIP, allowing us to model the dataset as a set of image and label embeddings $\{(\phi(x_i), \psi(y_i))\}_{i=1}^n$.

We compare FIMs with representations derived from CLIP as base representations for tasks and *learn* box embeddings to model a variety of relations among tasks.

Task Relations in Computer Vision. Understanding the relationships between tasks can lead to efficient solutions to new tasks. Previous work has measured task similarity by using model gradients [14] or based on their learned features [21] for grouping tasks for efficient multi-tasking. Similarly, predicting which pre-trained models will generalize the best on a new dataset could streamline model selection. Taskonomy [52] investigates transfer learning across vision tasks, varying from segmentation to pose estimation, by computing pairwise transfer distances or task affinities. These affinities are calculated by evaluating the extent to which a model trained on a source task generalizes to a target task [13, 42], though this process is computationally expensive.

Dataset Visualization. Low-dimensional Euclidean embeddings derived from UMAP [28], t-SNE [44], and LargeVis [43] are widely used to visualize relationships between datasets. They have been shown to successfully recover clusters of various data modalities by preserving the

relationship of each data point with its neighbors [4, 5, 40]. In low-dimension space, relationships with other data points are defined by their Euclidean distances. However, these are commonly used to represent symmetric relations.

Visualizations using tidy trees [37], circle packing [49], or cone trees [29] organize asymmetric relations as tree-structured hierarchies in low dimension. However, cyclic data relationships cannot be properly represented for these methods (e.g., when a node has two or more parents).

Asymmetric Distances over Datasets. Kullback-Leibler (KL) divergence between image or label distributions provides a natural way to represent asymmetric distances between datasets. TASK2VEC [4] computes the similarity between two tasks (e.g., cosine distance), and introduces asymmetry by using the complexity of the first task as a reference. The complexity is measured by the similarity of the task embedding to a "trivial embedding" (embedding of a task that is easy or has no examples).

Order embeddings on images were first proposed in [46] to capture tree-structured relationships. Given an dataset of $P = (u, v)$ drawn from an partially ordered set $(X, \preceq_X)$, they frame the problem as learning a mapping $f : (X, \preceq_X) \rightarrow (Y, \preceq_Y)$ that is order preserving, i.e., $u \preceq_X v \iff f(u) \preceq_Y f(v)$. The reserved product order was used for $\preceq_Y$, i.e., $x \preceq y \iff x_i > y_i, \forall i$. Box embeddings [8] generalized this framework by representing points as axis-aligned hyper-rectangles and using volumetric relationships (e.g., intersection over union) to represent asymmetric relations. They used the framework to model conditional distributions and hypernymy relations (e.g., dog "is a" mammal) on the WordNet graph [2, 19, 33, 38, 48].

Hyperbolic spaces provide yet another way to model asymmetric relationships. Examples include the Poincare disk model which uses hyperbolic cosine (cosh) to measure distance between points in a disk. Poincare embeddings have been similarly used to represent WordNet hierarchies [32] and other relations in general graphs. Hyperbolic representations have also been proposed for representing images for efficient zero-shot learning given a taxonomic structure of the labels [27].

To the best of our knowledge, no prior work has explored the use of these spaces for representing entire datasets and their effectiveness in capturing various task relationships. We adopt box embeddings in this work due to the effectiveness over alternatives in previous work [2, 6, 34, 38], ease of visualization, as well as due to open-source libraries for robust learning. However, instead of learning box embeddings directly, we learn mappings from task representations.

3. TASK2BOX Framework

We define the problem as follows: given a collection of datasets $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m\}$, and an asymmetric relationship given the pairwise relationship between datasets $d(\mathcal{D}_i, \mathcal{D}_j) \in [0, 1]$, we aim to encode each dataset into a low-dimension space that preserves the relationships between datasets and is interpretable.

To achieve this, we propose using box embeddings for encoding each of the datasets. This process involves two main steps: (1) deriving the base representation e of each dataset, and (2) learning a model $f_\theta : e \rightarrow z$, where $z \in \mathbb{R}^{2\times k}$ represents a k-dimensional axis-aligned hyper-rectangle (i.e., box), denoted by its lower left and upper right coordinates. These steps are detailed further below.

3.1. Base Task Representations

Each dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ is defined as a collection of pairs of images $x_i \in \mathcal{X}$ and labels $y_i \in \mathcal{Y}$. For obtaining a base embedding e for each dataset, we utilize methods such as CLIP [36, 41], TASK2VEC [4], or attribute-based approaches [15].

CLIP. Using a pre-trained CLIP model [7, 20, 36, 41], we compute the mean and variance of the individual sample embeddings within each dataset. For each data sample, the image embedding is concatenated with the label embedding, the latter generated from text prompts (e.g., "A photo of [CLS]"). This concatenation models the joint distribution of images and labels. Eq. 1 and 2 detail how the mean and variance embeddings are derived, where $[i, j]$ represents the concatenation of vectors i and j, ϕ is the vision encoder, and ψ is the text encoder. The covariance is approximated as diagonal for tractability.

$$\mu_{CLIP} = \frac{1}{N}\sum_{i=1}^{N}[\phi(x_i), \psi(y_i)] \tag{1}$$

$$\sigma^2_{CLIP} = \frac{1}{N}\sum_{i=1}^{N}\left([\phi(x_i), \psi(y_i)] - \mu_{CLIP}\right)^2 \tag{2}$$

The base representation is defined as $e := \mu_{CLIP} \in \mathbb{R}^{2048}$ or $e := [\mu_{CLIP}, \sigma^2_{CLIP}] \in \mathbb{R}^{4096}$. A ViT-H/14 [12] pretrained on LAION-2B [41] was used to extract the embeddings.

TASK2VEC [4] encodes a dataset using the approximate Fisher Information Matrix (FIM) of a network trained on the given dataset. The FIM represents the importance of parameters in the feature extractor by perturbing the weights $\hat{w}$ of a given probe network with Gaussian noise $\mathcal{N}(0, \Lambda)$. The precision matrix Λ is estimated to be close to an isotropic prior $\mathcal{N}(\hat{w}, \lambda^2 I)$ while having a good expected error. Eq. 3 is minimized to find Λ where H is the cross entropy loss, $\hat{w}$ are the weights of the network, β is the magnitude of the prior, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$.

$$\mathcal{L}(\hat{w};\Lambda) = \mathbb{E}_{w\sim\mathcal{N}(\hat{w},\Lambda)}[H_{p_w,\hat{p}}p(y|x)] + \beta KL\left(\mathcal{N}(0,\Lambda)\|\mathcal{N}(0,\lambda^2 I)\right) \quad (3)$$

The matrix Λ provides an estimate of the FIM and is approximated as a diagonal matrix. The diagonal components are used as a base representation of a dataset ($e := \text{FIM} \in \mathbb{R}^{17024}$). ResNet-34 [17] pretrained on ImageNet [11] is used as the probe network for all datasets.

Attribute-based. A task can be characterized by a set of t binary attributes [15] represented as a vector of dimension t. Some of the attributes explored for representing tasks are: (1) Is the task generative? (2) Is the task output in 2D? (3) Does the task involve camera pose estimation? Taking these 3 characteristics, for example, we can represent a 2D segmentation task as the vector $e = [0, 1, 0]$ as a discriminative 2D task that does not need camera poses.

Tasks in Taskonomy [52] involving multiple modalities benefit from this attribute-based representation due to its model independence. This approach also enables generalization to unseen tasks by identifying the presence or absence of various characteristics. Each vision task in Taskonomy is represented with 15 attributes, resulting in vectors $e \in \mathbb{R}^{15}$. The full list of attributes is in Appendix **??**.

3.2. Learning Box Embeddings

From a base task representation e, we learn the parameters θ of a model $f_\theta : e \to z$ that preserves the asymmetric similarity function $d(\mathcal{D}_i, \mathcal{D}_j)$ between any two datasets $\mathcal{D}_i$ and $\mathcal{D}_j$. In Eq. 4, the learning objective is shown where $\mathcal{L}_E$ is a loss function (mean squared error), $\mathcal{L}_D$ is a distance function, $\mathcal{L}_R$ is a regularization term, and λ is a hyperparameter. The embeddings $z_i, z_j \in \mathbb{R}^{2\times k}$ represent the coordinates of the lower left and the upper right corners of the respective k dimemsional boxes, with $f_\theta(e_i) = z_i$ and $f_\theta(e_j) = z_j$.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i,j} \left(\mathcal{L}_E\left(d(\mathcal{D}_i, \mathcal{D}_j), d_{box}\left(z_i, z_j\right)\right) + \lambda \mathcal{L}_D\left(z_i, z_j\right) + \mathcal{L}_R\right) \quad (4)$$

The asymmetric relationship between box embeddings, denoted as $d_{box}(z_i, z_j)$, is computed in Eq. 5 by calculating the volume of the intersection between z_i and z_j, normalized by the volume of z_i. For z_i to be fully contained inside z_j ($z_i \subset z_j$), is it required that $d_{box}(z_i, z_j) = 1$. Conversely, for z_j to only partially contain z_i, $d_{box}(z_j, z_i)$ must fall within the range $(0, 1)$. Fig. 2 illustrates this through a 2-dimensional example, where z_1 represents the box embedding for Canidae, and z_2 for Mammalia.

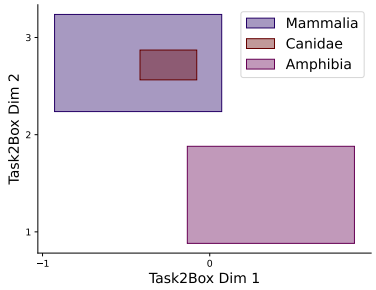

Figure 2. **TASK2BOX Embeddings in 2D for Mammalia, Canidae, and Amphibia Datasets from iNaturalist**. Each embedding represents the coordinates of the lower left and upper right corners of each box/rectangle. Since Canidae (z_1) is a proper subset of Mammalia (z_2): $d_{box}(z_1, z_2) = 1$ and $d_{box}(z_2, z_1) = 0.1$.

$$d_{box}\left(z_i, z_j\right) = \frac{\text{vol}(z_i \cap z_j)}{\text{vol}(z_i)} \quad (5)$$

$\mathcal{L}_D$ is applied to datasets where $d(\mathcal{D}_i, \mathcal{D}_j) > 0$ such that the Euclidean distance between the center coordinates of z_i and z_j is minimized when starting from a non-overlapping state. This allows non-overlapping embeddings to move closer to each other, complementing $\mathcal{L}_E$ to learn relationships. $\mathcal{L}_R$ encourages solutions with regular-shaped boxes for better interpretability. The formulation of $\mathcal{L}_R$, given in Eq. 6, applies to a k-dimensional box embedding z_i, where s_a represents the size of the a-th dimension (for example, width), and α, β are hyperparameters. To prevent the trivial solution of minimizing box volume to zero, the inverse of the box volume is included. It's crucial to normalize the terms with respect to the embedding dimension, as the first and second terms scale quadratically and exponentially with dimension increase, respectively.

$$\mathcal{L}_R = \left(\frac{\alpha}{k^2}\sum_{a=1}^{k}\sum_{b=a+1}^{k}|s_a - s_b|\right) + \beta\left(\text{vol}(z_i)\right)^{-1/k} \quad (6)$$

The architecture of f consists of three fully-connected layers followed by two linear heads: one predicts a k-dimensional lower-left coordinate, and the other predicts the k-dimensional sizes of each box dimension.

4. Experiments

We consider the following goals to evaluate the capability of TASK2BOX to represent datasets:

1. Given a collection of **existing datasets** $\mathcal{D}_E$ and a subset of pairwise relationships $\mathcal{R}$: can the model generalize on unseen relationships $\mathcal{R}'$ within $\mathcal{D}_E$ where $\mathcal{R}' \cap \mathcal{R} = \emptyset$?
2. Given a collection of **novel datasets** $\mathcal{D}_N$ not seen during training: can the model accurately identify the relationships with the existing datasets $\mathcal{D}_E$?

iNaturalist + CUB						
Method	**Existing Datasets**			**Novel Datasets**		
	μ_{CLIP}	$[\mu, \sigma^2]_{CLIP}$	FIM	μ_{CLIP}	$[\mu, \sigma^2]_{CLIP}$	FIM
TASK2BOX (2D)	69.23%	67.84%	39.61%	50.07%	39.66%	10.06%
TASK2BOX (3D)	<u>79.66%</u>	<u>79.35%</u>	<u>57.63%</u>	<u>70.04%</u>	<u>64.53%</u>	<u>20.65%</u>
TASK2BOX (5D)	**84.67%**	**82.41%**	**79.72%**	**73.79%**	**72.11%**	**34.88%**
MLP Classifier	45.25%	61.45%	26.34%	39.06%	44.54%	19.90%
Linear Classifier	4.40%	3.11%	7.06%	4.77%	5.87%	15.92%
KL Divergence	-	6.58%	7.94%	-	5.90%	0.00%
Asymmetric Cosine	9.29%	11.54%	2.83%	1.47%	1.47%	1.47%
Asymmetric Euclidean	1.71%	1.71%	8.53%	1.47%	1.47%	1.91%
Random		2.06%			1.49%	

Table 1. **Average F1 Score for Predicting Hierarchical Relationships on iNaturalist + CUB Dataset.** For all feature types, TASK2BOX outperforms other methods by more than 20% on existing datasets and more than 10% on novel datasets. The best-performing model is in **bold**, and the second-best is <u>underlined</u>. Results on **Novel Datasets** demonstrate that our model can generalize beyond the seen tasks it has previously seen. The dimensions such as 2D, 3D, and 5D refer to different box dimensionalities. Results using KL divergence for μ_{CLIP} are not shown since it is not a distribution.

The goals are evaluated through two experimental setups, demonstrating our model's ability to predict various types of relationships. The configurations, and their corresponding datasets, relationships, and metrics, are detailed below. Baseline methods for performance comparison are also reviewed. Implementation details are in the Appendix.

4.1. Experimental Setup

4.1.1 Hierarchical Task Relationships

We use a combination of iNaturalist [45] and Caltech-UCSD Birds (CUB) [50], and instrument-related classes in ImageNet [11] to evaluate the ability of TASK2BOX to represent hierarchical relationships. The first two are composed of images of various species, and the third is composed of images of various objects. The classes naturally follow a hierarchical form based on biological classification (taxonomy of iNaturalist+CUB species), and on semantic relations (hypernymy of objects in ImageNet).

Datasets. For iNaturalist+CUB, the datasets are defined as the classes, the orders, and the families in the taxonomy that contain a significant number of samples per dataset as in [4]. There are 47 classes, 202 orders, and 589 families for a total of 838 datasets. For ImageNet, the instrument-related objects were processed using WordNet [30] to obtain hierarchical information between classes. This resulted in 131 datasets for training and evaluation.

Dataset Relationships. For any two datasets, their relationship is captured as $d(\mathcal{D}_i, \mathcal{D}_j) \in \{0, 1\}$, where $d(\mathcal{D}_i, \mathcal{D}_j) = 1$ if and only if $\mathcal{D}_i \subset \mathcal{D}_j$, and $d(\mathcal{D}_i, \mathcal{D}_j) = 0$ otherwise. Fig. 2 provides an example: Canidae ($\mathcal{D}_1$) is a family within the class Mammalia ($\mathcal{D}_2$); thus, $d(\mathcal{D}_1, \mathcal{D}_2) = 1$ and $d(\mathcal{D}_2, \mathcal{D}_1) = 0$. Meanwhile, Amphibia ($\mathcal{D}_3$) is unrelated to either dataset, resulting in $d(\mathcal{D}_1, \mathcal{D}_3) = 0$

Evaluation Metrics. We evaluate the ability of the model to classify the presence of containment relationships between datasets using the F1 score due to the imbalance between positive and negative relationships. To obtain F1 score, precision and recall are first calculated. Precision is computed as the ratio of true positive pairs predicted to the total number of positive predictions. Recall is the ratio of true positive pairs predicted to the total number of true positive pairs. F1 score is then reported as the harmonic mean between the precision and recall. These calculations are done on both (1) the unseen relationships $\mathcal{R}'$ within existing datasets $\mathcal{D}_E$ and (2) the relationships between novel datasets $\mathcal{D}_N$ and existing datasets $\mathcal{D}_E$. For the latter, we evaluate the relationships in both directions, i.e., $\{(\mathcal{D}_e, \mathcal{D}_n) \; \forall \; \mathcal{D}_e \in \mathcal{D}_E\} \cup \{(\mathcal{D}_n, \mathcal{D}_e) \; \forall \; \mathcal{D}_e \in \mathcal{D}_E\}$.

4.1.2 Transfer Learning Between Datasets

The Taskonomy [52] benchmark is used to evaluate the ability of TASK2BOX to predict task affinities.

Datasets. Taskonomy defines a set of 25 visual tasks with corresponding pairwise task affinities. These tasks range from object detection, semantic segmentation, pose estimation, and more. We treat each visual task as a dataset. The tasks are described in Appendix **??**.

Dataset Relationships. In contrast to the containment relationship defined in § 4.1.1, Taskonomy quantifies relationships with task affinity measured by the performance gain achieved by transfer learning from a source dataset $\mathcal{D}_j$ to a target dataset $\mathcal{D}_i$. The values are computed using Analytic Hierarchy Process [39, 52]. Dataset relationships are computed and normalized based on an ordinal approach and determined by the percentage of images that transfer well to a target task given a set of source tasks as detailed in [52].

Method	ImageNet					
	Existing Datasets			Novel Datasets		
	μ_{CLIP}	$[\mu, \sigma^2]_{CLIP}$	FIM	μ_{CLIP}	$[\mu, \sigma^2]_{CLIP}$	FIM
TASK2BOX (2D)	62.72%	63.33%	31.32%	47.76%	48.35%	9.46%
TASK2BOX (3D)	83.12%	82.79%	58.16%	73.06%	66.48%	24.70%
TASK2BOX (5D)	**90.58%**	**88.48%**	**64.91%**	**76.95%**	**78.84%**	37.39%
MLP Classifier	54.20%	60.43%	44.85%	62.24%	64.56%	**41.22%**
Linear Classifier	9.84%	8.33%	25.46%	11.89%	11.87%	22.98%
KL	-	5.53%	9.90%	-	10.41%	0.00%
Asymmetric Cosine	4.28%	4.28%	6.92%	4.52%	4.52%	0.00%
Asymmetric Euclidean	3.73%	3.73%	7.16%	4.52%	4.52%	4.73%
Random		3.64%			5.02%	

Table 2. **Average F1 Score for Predicting Hierarchical Relationships on ImageNet Instruments**. TASK2BOX is shown to generalize well on both predicting relationships between existing datasets (**Existing Datasets**), and predicting box embeddings of new datasets and their relationships with existing datasets (**Novel Datasets**).

For a pair of datasets $(\mathcal{D}_i, \mathcal{D}_j)$, the task affinity is defined as $d(\mathcal{D}_i, \mathcal{D}_j) \in [0, 1]$. The higher the task affinity to a target dataset $\mathcal{D}_i$ from a source dataset $\mathcal{D}_j$, then $d(\mathcal{D}_i, \mathcal{D}_j)$ gets closer to 1, where a value of 1 would show $\mathcal{D}_i \subset \mathcal{D}_j$.

Evaluation Metrics. We evaluate the prediction of the model based on the Spearman correlation between the ground truth task affinity values, and the predicted values based on the box distances in Eq. 5. Similar to § 4.1.1, we evaluate on both (1) unseen relationships $\mathcal{R}'$ for existing datasets, and (2) on relationships between unseen novel datasets $\mathcal{D}_N$ and existing datasets $\mathcal{D}_E$.

4.2. Baseline Methods

We compare the performance of TASK2BOX with alternative models and simple asymmetric distances proposed in prior work. For hierarchical relationships, given two datasets $\mathcal{D}_i, \mathcal{D}_j$, the models predict $\hat{d}(\mathcal{D}_i, \mathcal{D}_j) \in \{0, 1\}$. For task affinity, the models predict $\hat{d}(\mathcal{D}_i, \mathcal{D}_j) \in [0, 1]$. The structure for the various methods are discussed below.

Linear Model. A linear model is trained to predict the relationship value between two datasets. The input is the concatenation of the base representation e for the two datasets.

MLP Model. A 4-layer MLP is used instead for prediction on the same inputs as the linear model.

KL Divergence. Each dataset is treated as a multivariate Gaussian using the mean and variance of the image and label features (CLIP) or directly as the FIM. The optimal threshold t is selected for the minimum distance between two dataset distributions $\text{KL}(\mathcal{D}_i||\mathcal{D}_j) < t$ for a prediction of $\hat{d}(\mathcal{D}_i, \mathcal{D}_j) = 1$, and $\hat{d}(\mathcal{D}_i, \mathcal{D}_j) = 0$ otherwise. The F1 Score (hierarchical) or the correlation (task affinity) is the objective for selecting t on the train set.

Asymmetric Cosine Similarity. The cosine similarity d_{cos} is a symmetric measure between two embeddings e_i, e_j. An asymmetric variant was proposed in [4] and shown in Eq. 7

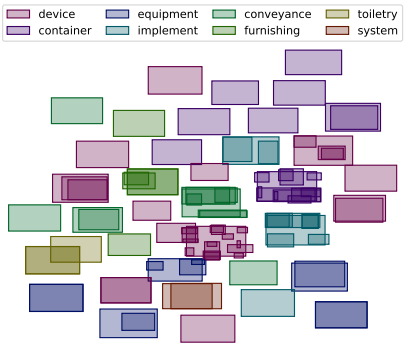

Figure 3. **Visualization of Instrument-related Datasets in ImageNet**. Datasets that belong to the same superset are shaded in the same color. TASK2BOX learns the hierarchy of various groups, and clusters similar datasets closer.

by considering the similarity of e_i and e_j relative to the complexity of e_i. The complexity of is measured as the distance to the trivial embedding e_o, and α is a hyperparameter.

$$d_{asym}(e_i, e_j) = d_{cos}(e_i, e_j) - \alpha d_{cos}(e_i, e_o) \quad (7)$$

An optimal threshold t is found on the train set where $d_{asym}(e_i, e_j) < t$ results to a prediction $\hat{d}(\mathcal{D}_i, \mathcal{D}_j) = 1$. The same threshold t is used for test set evaluation.

Asymmetric Euclidean Similarity. The asymmetric similarity is computed as in Eq. 7 but uses the Euclidean distance instead of cosine.

Random. The probability of containment (for hierarchical) or the value of the task affinity is uniformly random.

5. Results and Discussion

5.1. Hierarchical Relationships

Tables 1 and 2 present the average F1 score for predicting hierarchical relationships on iNaturalist+CUB and Ima-

geNet datasets. Various methods of extracting dataset embeddings were evaluated. TASK2BOX outperforms baseline methods, indicating that TASK2BOX can both generalize across unseen relationships (Existing Datasets), and to accurately represent relationships with existing datasets for unseen datasets (Novel Datasets). CLIP features also generalize better than FIM features, perhaps because the CLIP multi-modal embedding of images and labels was trained across a broad set of domains[36]. However, while CLIP is confined to image and text modalities, FIM can accommodate any modality by adapting the probe network's last layer for different prediction tasks.

Beyond predicting relationships, our method allows the visualization of datasets. Fig. 1 illustrates this for a subset of datasets in iNaturalist and CUB using TASK2BOX (2D), and Fig. 3 for ImageNet. The hierarchical organization of the datasets is apparent in the TASK2BOX representation – datasets that contain others appear as larger boxes, while more specialized datasets are depicted as smaller boxes neted with borader, more general dataset boxes. However, while 2D visualization offers insights, the flat surface may restrict the representation of complex relationships. Representations in higher dimensions is explored in § 5.4.

5.2. Task Affinity

Table 3 presents the results of task representations in Taskonomy. Our method show that, even with attribute-based embeddings, it can learn box embeddings through relationship supervision between datasets. TASK2BOX is shown to correlate highly with the ground truth task affinities compared to other methods. Given that only 25 tasks are available in Taskonomy, with 3 held out of training as novel datasets, there is a higher uncertainty in predicting unseen datasets. We expect that performance on novel datasets will improve with more datasets available during training.

Fig. 4 displays the learned representation with TASK2BOX (2D). Each subfigure in (a)-(c) represents a subset of tasks identified as having strong transfer relationships. TASK2BOX not only identifies but also visually represents related tasks suitable for fine-tuning. The small highlighted boxes indicate target tasks, while the larger enclosing boxes represent source tasks. Although other tasks may not be proper subsets, the box distance in Eq. 5 provides an estimate for task affinity. A small box distance between two datasets, $d_{box}(z_1, z_2)$, suggests lower transfer performance from a source task z_2 to a target task z_1.

5.3. Visualizing Public Datasets

Apart from predicting both hierarchical relationships and quantified task affinities, our method can also be used to visualize public datasets that lack available ground truth relationships. Using a set of vision tasks from Hugging Face [1], μ_{CLIP} + TASK2BOX (2D) were utilized to predict dataset embeddings. A constraint on the box sizes was added to reflect information about the dataset sizes. Fig. 5 displays the results of TASK2BOX on 131 datasets from Hugging Face, where similar datasets, such as sentiments and documents, are shown to overlap. This approach allows for the analysis and visualization of dataset variations and similarities even with only samples of images and labels. Simultaneously, the sizes of various datasets can be visualized as box sizes – with larger datasets containing more samples depicted as larger boxes. This tool enables computer vision practitioners to see how their datasets compare with other existing datasets. Beyond visualizing the variation of available datasets (based on the number of clusters), it can expedite the process of finding suitable data sources by examining embedding overlaps.

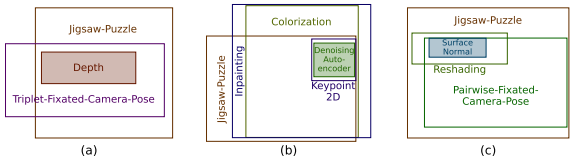

Figure 4. **Visualization of Tasks in Taskonomy showing Source Tasks that Transfer Well to Target Tasks (shaded)**. (a) Jigsaw and Triplet Fixated Camera Pose estimation are source tasks that transfer well to Depth estimation. (b) and (c) show different source tasks (larger boxes) that transfer well to the shaded boxes of Denoising Autoencoder and Surface Normal, respectively.

Method	Existing Datasets	Novel Datasets
	Spearman's ρ	Spearman's ρ
TASK2BOX (2D)	0.85 ± 0.06	0.12 ± 0.21
TASK2BOX (3D)	$\underline{0.93 \pm 0.02}$	$\mathbf{0.48 \pm 0.24}$
TASK2BOX (5D)	$\mathbf{0.94 \pm 0.03}$	0.39 ± 0.22
MLP	0.88 ± 0.06	0.31 ± 0.18
Linear	0.75 ± 0.11	$\underline{0.40 \pm 0.24}$
Random	0.05 ± 0.14	0.15 ± 0.07

Table 3. **Spearman Correlation and Standard Deviation between Predicted and Ground Truth Task Affinities on Taskonomy**. Our method shows higher correlation with the task affinities compared to the baseline. Attribute-based embeddings were used.

5.4. Analysis of TASK2BOX

We discuss properties of TASK2BOX below. Additional insights are also included in Appendix **??**.

Using a Box Prior Improves Performance. While TASK2BOX was trained with a 3-layer MLP, it achieved significantly better performance than the MLP baseline without a box prior. Tables 1, 2, and 3 demonstrate our method outperforming the baselines. Representing each dataset as an entity with shape and volume (such as a box), instead of solely learning relationships from embeddings, proves effective for generalization on unseen relationships. This

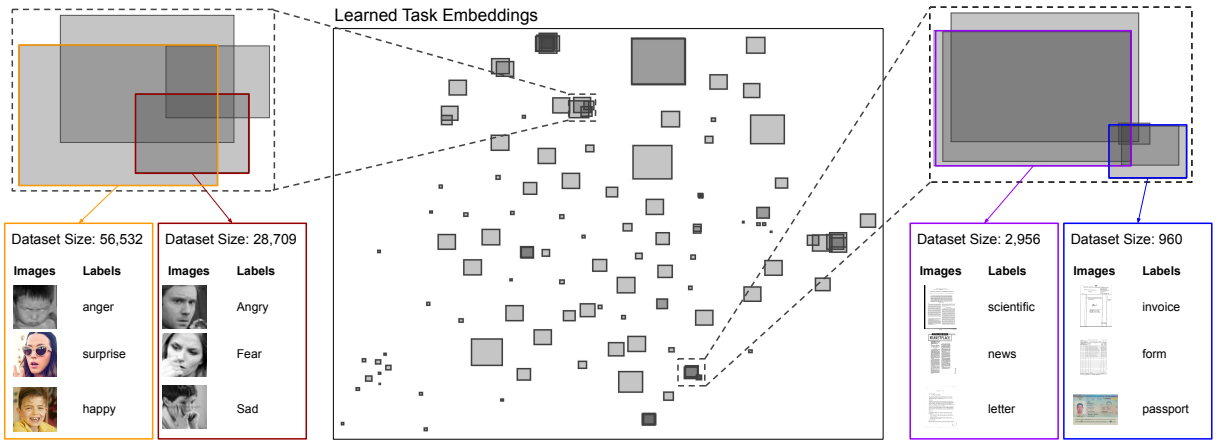

Figure 5. **Visualizing Image Classification Datasets in Hugging Face**. The sample data points annotated on the highlighted datasets show that common tasks overlap with each other (e.g., sentiment classification and document classification datasets). Although labels could slightly differ between datasets, TASK2BOX can infer the level of similarity and represent it as the amount of overlap. The embedding size (box area) also shows the number of available data.

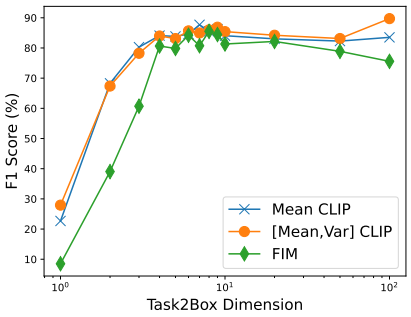

Figure 6. **Effect of TASK2BOX Embedding Dimension on the Accuracy of Predicting Task Relations.** As the dimension increases, the performance generally increases.

improvement could be attributed to the explicit modeling of relationships through physical shapes, which enforces consistency: if $z_i \subset z_j$ and z_j has no overlap with z_k, then having a physical representation ensures that $z_i \cap z_k = \emptyset$.

TASK2BOX can Represent Relations in Varying Dimensions. Fig. 6 illustrates the performance of TASK2BOX as the box dimension increases. Representing relationships in two dimensions results in easily interpretable embeddings; however, modeling complex task relationships could benefit from expansion to higher dimensions. Projecting to higher dimensions may yield even better performance, as relationships between datasets become more accurately represented, given the model's increased representation capacity. Nonetheless, as the dimensionality further increases, learning and visualizing the embedding space also becomes more challenging.

Using Boxes to Represent Tasks Enables Effective Calculations on Embeddings. Boxes offer the advantage of being closed under intersection, meaning the intersection of two boxes results in another box, a property not shared by circles or ellipses. Among region-based embeddings that utilize geometric objects (e.g., cones, disks, and boxes) to represent entities, operations involving boxes are the most straightforward to calculate [10, 38].

6. Conclusion

We present a novel method that learns low-dimensional, interpretable embeddings of various task relationships. With box representations, we demonstrate that asymmetric relationships, such as task affinities and hierarchies, can be accurately modeled using various base representations. While CLIP embeddings have been shown to outperform FIM when integrated with our model, future work could investigate how CLIP might be adapted to modalities beyond text and images. Attribute-based features offer a viable alternative, and can be extracted from from datasheets using natural language processing techniques.

The distinct properties of TASK2BOX were analyzed, revealing its ability to perform effectively across varying dimensions, and to model and visualize overlaps among public classification datasets on Hugging Face. This could enable computer vision practitioners to assess dataset utility for a task in hand. Although our model successfully represents task relationships, it does not incorporate information about optimal training procedures and model architecture. Future work could explore the inclusion of additional information for a more detailed understanding of the task space.

Acknowledgements. This work was supported by awards from the National Science Foundation (2329927 and 1749833) and the NASA AIST program. The experiments were performed on the University of Massachusetts GPU cluster funded by the Mass. Technology Collaborative.

References

[1] Hugging Face datasets. https://huggingface.co/datasets?task_categories=task_categories:image-classification. 2, 7

[2] Ralph Abboud, Ismail Ceylan, Thomas Lukasiewicz, and Tommaso Salvatori. Boxe: A box embedding model for knowledge base completion. *Advances in Neural Information Processing Systems*, 33:9649–9661, 2020. 3

[3] Rabab Abdelfattah, Qing Guo, Xiaoguang Li, Xiaofeng Wang, and Song Wang. Cdul: Clip-driven unsupervised learning for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1348–1357, 2023. 2

[4] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6430–6439, 2019. 1, 2, 3, 5, 6

[5] Sanjeev Arora, Wei Hu, and Pravesh K. Kothari. An analysis of the t-sne algorithm for data visualization. In *Proceedings of the 31st Conference On Learning Theory*, pages 1455–1462. PMLR, 2018. 3

[6] Michael Boratko, Dongxu Zhang, Nicholas Monath, Luke Vilnis, Kenneth L Clarkson, and Andrew McCallum. Capacity and bias of learned geometric embeddings for directed graphs. *Advances in Neural Information Processing Systems*, 34:16423–16436, 2021. 3

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 3

[8] Tejas Chheda, Purujit Goyal, Trang Tran, Dhruvesh Patel, Michael Boratko, Shib Sankar Dasgupta, and Andrew McCallum. Box embeddings: An open-source library for representation learning using geometric structures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 203–211, 2021. 3

[9] Marcos V Conde and Kerem Turgutlu. Clip-art: Contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3956–3960, 2021. 2

[10] Shib Dasgupta, Michael Boratko, Siddhartha Mishra, Shriya Atmakuri, Dhruvesh Patel, Xiang Li, and Andrew McCallum. Word2box: Capturing set-theoretic semantics of words using box embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022. 8

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 4, 5

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021. 3

[13] Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy & transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12387–12396, 2019. 2

[14] Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516, 2021. 2

[15] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. 1, 2, 3, 4

[16] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2021. 2

[17] 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. 4

[18] Wenbin He, Suphanut Jamonnak, Liang Gou, and Liu Ren. Clip-s4: Language-guided self-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11207–11216, 2023. 2

[19] EunJeong Hwang, Jay-Yoon Lee, Tianyi Yang, Dhruvesh Patel, Dongxu Zhang, and Andrew McCallum. Event-event relation extraction using probabilistic box embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 235–244, 2022. 3

[20] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 3

[21] Zhuoliang Kang, Kristen Grauman, and Fei Sha. Learning with whom to share in multi-task feature learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 521–528, 2011. 2

[22] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of fisher information in deep neural networks: Mean field approach. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1032–1041. PMLR, 2019. 2

[23] Zhengfeng Lai, Noranart Vesdapunt, Ning Zhou, Jun Wu, Cong Phuoc Huynh, Xuelu Li, Kah Kuen Fu, and Chen-Nee Chuah. Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16155–16165, 2023. 2

[24] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7061–7070, 2023. 2

[25] Zhibin Liao, Tom Drummond, Ian Reid, and Gustavo Carneiro. Approximate fisher information matrix to characterize the training of deep neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 42(1): 15–26, 2018. 2

[26] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15305–15314, 2023. 2

[27] Shaoteng Liu, Jingjing Chen, Liangming Pan, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. Hyperbolic visual embedding learning for zero-shot recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[28] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. 2

[29] Guy Melancon and Ivan Herman. *Circular drawings of rooted trees*. CWI (Centre for Mathematics and Computer Science), 1998. 3

[30] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2, 5

[31] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019. 2

[32] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017. 3

[33] Dhruvesh Patel, Shib Sankar Dasgupta, Michael Boratko, Xiang Li, Luke Vilnis, and Andrew McCallum. Representing joint hierarchies with box embeddings. In *Automated Knowledge Base Construction*, 2020. 3

[34] Dhruvesh Patel, Pavitra Dangati, Jay-Yoon Lee, Michael Boratko, and Andrew McCallum. Modeling label space interactions in multi-label classification using box embeddings. In *International Conference on Learning Representations*, 2022. 3

[35] Jeffrey Pennington and Pratik Worah. The spectrum of the fisher information matrix of a single-hidden-layer neural network. *Advances in neural information processing systems*, 31, 2018. 2

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 7

[37] Edward M. Reingold and John S. Tilford. Tidier drawings of trees. *IEEE Transactions on software Engineering*, (2): 223–228, 1981. 3

[38] Hongyu Ren, Weihua Hu, and Jure Leskovec. Query2box: Reasoning over knowledge graphs in vector space using box embeddings. In *International Conference on Learning Representations*, 2019. 3, 8

[39] Roseanna W Saaty. The analytic hierarchy process—what it is and how it is used. *Mathematical modelling*, 9(3-5):161–176, 1987. 5

[40] Saquib Sarfraz, Marios Koulakis, Constantin Seibold, and Rainer Stiefelhagen. Hierarchical nearest neighbor graph embedding for efficient dimensionality reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 336–345, 2022. 3

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3

[42] Astuti Sharma, Tarun Kalluri, and Manmohan Chandraker. Instance level affinity-based transfer for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5361–5371, 2021. 2

[43] Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th international conference on world wide web*, pages 287–297, 2016. 2

[44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 1, 2

[45] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 2, 5

[46] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *International Conference on Learning Representations (ICLR)*, 2015. 1, 3

[47] Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3219–3229, 2023. 2

[48] Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. Probabilistic embedding of knowledge graphs with box lattice measures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 263–272, 2018. 1, 3

[49] Weixin Wang, Hui Wang, Guozhong Dai, and Hongan Wang. Visualization of large hierarchical data by circle packing. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 517–520, 2006. 3

[50] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. Technical Report CNS-TR-201, Caltech, 2010. 2, 5

[51] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7031–7040, 2023. 2

[52] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 2, 4, 5

[53] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. In *International Conference on Learning Representations*, 2021. 2