A Self-explaining Neural Architecture for Generalizable Concept Learning

Sanchit Sinha[1] , **Guangzhi Xiong**[1] and **Aidong Zhang**[1]
[1]University of Virginia
Charlottesville, VA, USA
{sanchit, hhu4zu, aidong}@virginia.edu

Abstract

With the wide proliferation of Deep Neural Networks in high-stake applications, there is a growing demand for explainability behind their decision-making process. Concept learning models attempt to learn high-level 'concepts' - abstract entities that align with human understanding, and thus provide interpretability to DNN architectures. However, in this paper, we demonstrate that present SOTA concept learning approaches suffer from two major problems - lack of **concept fidelity** wherein the models fail to learn consistent concepts among similar classes and limited **concept interoperability** wherein the models fail to generalize learned concepts to new domains for the same task. Keeping these in mind, we propose a novel self-explaining architecture for concept learning across domains which - i) incorporates a new *concept saliency network* for representative concept selection, ii) utilizes *contrastive learning* to capture representative domain invariant concepts, and iii) uses a novel *prototype-based concept grounding* regularization to improve concept alignment across domains. We demonstrate the efficacy of our proposed approach over current SOTA concept learning approaches on four widely used real-world datasets. Empirical results show that our method improves both concept fidelity measured through concept overlap and concept interoperability measured through domain adaptation performance. An appendix of the paper with more comprehensive results can also be viewed at https://arxiv.org/abs/2405.00349.

1 Introduction

Deep Neural Networks (DNNs) have revolutionized a variety of human endeavors from vision to language domains. Increasingly complex architectures provide state-of-the-art performance which, in some cases has surpassed even human-level performance. Even though these methods have incredible potential in saving valuable man-hours and minimizing inadvertent human mistakes, their adoption has been met with rightful skepticism and extreme circumspection in critical applications like medical diagnosis [Liu *et al.*, 2021; Aggarwal *et al.*, 2021], credit risk analysis [Szepannek and Lübke, 2021], etc.

With the recent surge in interest in Artificial General Intelligence (AGI) through DNNs, the broad discussion around the lack of rationale behind DNN predictions and their opaque decision-making process has made them notoriously **black-box** in nature [Rudin, 2019; Varoquaux and Cheplygina, 2022; D'Amour *et al.*, 2020; Weller, 2019]. In extreme cases, this can lead to a lack of *alignment* between the designer's intended behavior and the model's actual performance. For example, a model designed to analyze and predict creditworthiness might look at features that should not play a role in the decision such as race or gender [Bracke *et al.*, 2019]. This, in turn, reduces the trustworthiness and reliability of model predictions (even if they are correct) which defeats the purpose of their usage in critical applications [Hutchinson and Mitchell, 2019; Raji *et al.*, 2020].

In an ideal world, DNNs would be inherently explainable by their *inductive biases*, as it is designed keeping stakeholders in account. However, such an expectation is gradually relaxed with the increasing complexity of the data which in itself drives up the complexity of the architectures of DNNs to fit said data. Several approaches to interpreting DNNs have been proposed. Some approaches assign relative importance scores to features deemed important like LIME [Ribeiro *et al.*, 2016], Integrated Gradients [Sundararajan *et al.*, 2017], etc. Other approaches rank training samples by their importance to prediction like influence functions [Koh and Liang, 2017], data shapley [Ghorbani and Zou, 2019], etc.

However, the aforementioned methods only provide a post-hoc solution and to truly provide interpretability, a more *accesible* approach is required. Recently, there have been multiple concept-based models incorporate concepts during model training [Kim *et al.*, 2018; Zhou *et al.*, 2018]. It is believed that explaining model predictions using abstract human-understandable "concepts" better **aligns** model's internal working with **human thought process**. Concepts can be thought of as abstract entities - shared across multiple samples providing a general model understanding. The general approach to train such models is to first map inputs to a concept space. Subsequently, alignment with the concepts is performed in the concept space and a separate model is learned on the concept space to perform the downstream task.

The ideal method to extract concepts from a dataset would be to manually curate and define what concepts best align with the requirements of stakeholders/end-users using extensive domain knowledge. This approach requires manual annotation of datasets and forces models to extract and encode only the pre-defined concepts as Concept Bottleneck Models [Koh *et al.*, 2020; Zaeem and Komeili, 2021] do. However, with increasing dataset sizes, it becomes difficult to manually annotate each data sample, thus limiting the efficiency and practicality of such approaches [Yuksekgonul *et al.*, 2022].

As a result, many approaches incorporate unsupervised **concept discovery** for concept-based prediction models. One such architecture is Self Explaining Neural Networks (SENN) proposed in [Alvarez-Melis and Jaakkola, 2018]. The concepts are extracted using a bottleneck architecture, and appropriate relevance scores to weigh each concept are computed in tandem using a standard feedforward network. The concepts and relevance scores are then combined using a network to perform downstream tasks (e.g. classification). Even though such concept-based explanations provide a clear explanation to understand neural machine intelligence, concept-based approaches are not without their faults. One critical problem we observed is that concepts learned across multiple domains using concept-based models are not consistent among samples from the same class, implying low **concept fidelity**. In addition, concepts are unable to generalize to new domains implying a lack of **concept-interoperability**.

In this paper, we propose a concept-learning framework with a focus on generalizable concept learning which improves concept interoperability across domains while maintaining high concept fidelity. Firstly, we propose a salient concept selection network that enforces representative concept extraction. Secondly, our framework utilizes self-supervised contrastive learning to learn domain invariant concepts for better interoperability. Lastly, we utilize prototype-based concept grounding regularization to minimize concept shifts across domains. Our novel methodology not only improves concept fidelity but also achieves superior concept interoperability, demonstrated through improved domain adaptation performance compared to SOTA self-explainable concept learning approaches. Our contributions are - (1) We analyze the current SOTA self-explainable approaches for concept interoperability and concept fidelity when trained across domains - problems that have not been studied in detail by recent works. (2) We propose a novel framework that utilizes a *salient concept selection network* to extract representative concepts and a self-supervised contrastive learning paradigm for enforcing domain-invariance among learned concepts. (3) We propose a prototype-based concept grounding regularizer to mitigate the problem of concept shift across domains. (4) Our evaluation methodology is the first to quantitatively evaluate the domain adaptation performance of self-explainable architectures and comprehensively compare existing SOTA self-explainable approaches.

2 Related Work

Related work on concept-level explanations. Recent research has focused on designing concept-based deep learning methods to interpret how deep learning models can use high-level human-understandable concepts in arriving at decisions [Ghorbani *et al.*, 2019; Chen *et al.*, 2019; Wu *et al.*, 2020; Koh *et al.*, 2020; Yeh *et al.*, 2019; Huang *et al.*, 2022; Sinha *et al.*, 2021; Sinha *et al.*, 2023]. Such concept-based deep learning models aim to incorporate high-level concepts into the learning procedure. Concept priors have been utilized to align model concepts with human-understandable concepts [Zhou *et al.*, 2018; Murty *et al.*, 2020; Chen *et al.*, 2019] and bottleneck models were generalized wherein any prediction model architecture can be transformed [Koh *et al.*, 2020; Zaeem and Komeili, 2021] by integrating an intermediate layer to represent a human-understandable concept representation. Similar work on utilizing CBMs for downstream tasks include [Jeyakumar *et al.*, 2021; Pittino *et al.*, 2021].

Related work on self-supervised learning with images. Self-supervised learning [Xu *et al.*, 2019; Saito *et al.*, 2020] via pretext tasks has been demonstrated to learn high-quality domain invariant representations from images using a variety of transformations such as rotations [Xu *et al.*, 2019; Gidaris *et al.*, 2018]. The transformations are usually small enough to not cause a significant shift in the intended and actual features in the latent space and are trained using a form of contrastive loss [Wang and Liu, 2021].

Related work on automatic interpretable concept learning. Supervised concept learning requires the concepts of each training sample to be manually annotated, which is impossible with a moderately large dataset and the concepts are restricted to what humans can conceptualize. To alleviate such bottlenecks, automatic concept learning is becoming increasingly appealing. One dominant architecture is Self Explaining Neural Networks (SENN) proposed in [Alvarez-Melis and Jaakkola, 2018]. Several other popular methods have been proposed which automatically learn concepts are detailed [Kim *et al.*, 2018; Ghorbani *et al.*, 2019; Yeh *et al.*, 2019; Wu *et al.*, 2020; Goyal *et al.*, 2019].

Comparision with existing work. Our work aims to address a challenge existing approaches face, concepts learned by self-explaining models may not be able to generalize well across domains, as the learned concepts are mixed with domain-dependent noise and less robust to light transformations due to a lack of supervision and regularization. Our proposed approach tackles this largely unsolved problem by designing a novel representative concept extraction framework and regularizes it using self-supervised contrastive concept learning and prototype-based grounding.

Concurrent to our work, BotCL [Wang, 2023] also proposes to utilize self-supervised learning to learn interpretable concepts. However, our approach is significantly different in both training and evaluation. We utilize multiple SOTA *transformations* to learn distinct concepts. Our evaluation framework comprises concept interoperability by evaluating performance across domains in addition to accuracy. Another work related to ours [Sawada, 2022b] incorporates unsupervised concepts in the bottleneck layer of CBMs which differs from our approach as we learn all concepts in a self-supervised manner, without supervision. Another concurrent work [Sawada, 2022a] utilizes an autoencoder setup with a discriminator and weak supervision using an object-detecting network (Faster RCNN) with limited generalization.

3 Methodology

In this section, we first provide a detailed description of our proposed learning pipeline, including (a) the Representative Concept Extraction (RCE) framework which incorporates a novel Salient Concept Selection Network in addition to the Concept and Relevance Networks, (b) Self-Supervised Contrastive Concept Learning (CCL) which enforces domain invariance among learned concepts, and (c) a Prototype-based Concept Grounding (PCG) regularizer that mitigates the problem of concept-shift among domains. We then provide details for the end-to-end training procedure with additional Concept Fidelity regularization which ensures concept consistency among similar samples.

3.1 Representative Concept Extraction

Figure 1 presents the proposed Representative Concept Extraction framework. For a given input sample $x \in \mathbb{R}^n$, the self-explainable concept learning framework learns a set of K representative d-dimensional concepts $C = \{c_1, \cdots, c_K\} \in \mathbb{R}^d$ and relevance scores associated with the concepts $\mathcal{S} = \{s_1, \cdots, s_K\} \in \mathbb{R}^d$ for the downstream task.

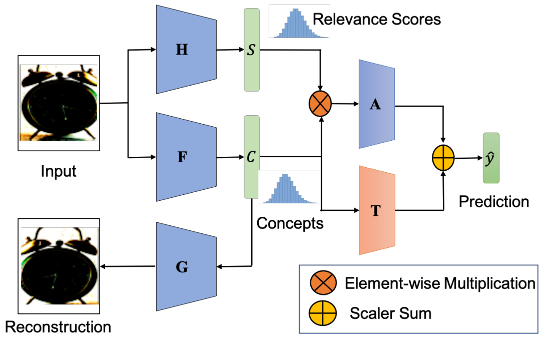

Figure 1: The proposed Representative Concept Extraction (RCE) framework. The networks **F** and **H** respectively extract concepts and associated relevance scores and **A** aggregates them. Network **G** reconstructs original input from the concepts while **T** selects the most representative concepts to the prediction.

Concept Network. The Concept Extraction Network consists of an encoder function **F**, which maps from the input space to the concept representation space ($\mathbb{R}^n \to \mathbb{R}^d$). To preserve the maximum amount of information content in the concept representation, the entire network is modeled as an autoencoder with the decoder function **G** which maps from the concept representation space to the input space ($\mathbb{R}^d \to \mathbb{R}^n$).

Relevance Networks. The Relevance Network function **H** is modeled similarly to the function **F**, which maps from the input space to the concept representation space ($\mathbb{R}^n \to \mathbb{R}^d$). The relevance network outputs a score associated with each concept - encapsulating each concept's relevance to the prediction. Mathematically, the relevance network **H** ($\mathbb{R}^n \to \mathbb{R}^d$) outputs a set of score vectors $\mathcal{S} = \{s_1, ..s_k\}$ for an input sample x.

Salient Concept Selection Network. Note that approaches like [Alvarez-Melis and Jaakkola, 2018] employ simple sparsity regularizations on the concept space to increase diversity and select representative concepts. However, we utilize a novel strategy that conditions the concept selection on the prediction performance. Effectively, utilizing a shallow network **T**, which maps from the concept space to the prediction space ($\mathbb{R}^d \to \mathbb{R}$) *selects* only those concepts that are *most responsible* or *salient* for prediction.

Aggregation and Prediction. Subsequently, the concepts and the relevance scores are aggregated to perform the final prediction using the aggregation function **A** which maps from the concept space to the output prediction space ($\mathbb{R}^d \to \mathbb{R}$). Mathematically, the function **A** aggregates a given concept vector $\mathbf{F}(x)$ and relevance score vector $\mathbf{H}(x)$ respectively for a given sample x. A shallow fully connected network models the function **A**. Note that the function **A** should be as shallow as possible to maximize interpretability.

The final prediction is computed using a weighted sum of outputs from the Aggregation Network **A** and the Salient Concept Selection Network **T**. Mathematically,

$$\hat{y} = \omega_1 * \mathbf{A}(\mathbf{F}(x) \odot \mathbf{H}(x)) + \omega_2 * \mathbf{T}(\mathbf{F}(x)) \qquad (1)$$

where $\odot$ is the element-wise product of the concept and relevance vectors. This weighted prediction strategy with tunable parameters ω_1 and ω_2 exerts greater control over concept selection. Note that higher values of ω_2 enforce representative concept selection.

Training Objective. As the Concept Network is modeled as an autoencoder, the training objective can be mathematically given by:

$$\mathcal{L}_{rec} = L(x, \mathbf{G}(\mathbf{F}(x))) + \lambda|\mathbf{F}(x)|_1 \qquad (2)$$

Note that λ is the strength of L_1 norm in Equation 2 - that regularizes the concept space and prevents degenerate concept learning (such as all concepts being a unit vector). The reconstruction loss $\mathcal{L}_{rec}$ is composed of L which quantifies the difference between an input sample x and its reconstruction $\mathbf{G}(\mathbf{F}(x))$.

Note that as the network **F** is responsible for extracting representative concepts and the **H** is responsible for calculating the relevance of the concepts extracted by **F**, they must be modeled by networks with similar complexity to avoid overfitting and learning of degenerate concepts.

The complete training objective of the Concept Extraction Framework where $\mathcal{L}$ is any prediction loss (such as Cross Entropy) is as follows:

$$\mathcal{L}_{CE} = \mathcal{L}_{rec} + \mathcal{L}(y, \hat{y}) \qquad (3)$$

3.2 Self-supervised Contrastive Concept Learning

Even though the RCE framework generates representative concepts, the concepts extracted are adulterated with *domain noise* thus limiting their generalization. In addition, with limited training data, the concept extraction process is not robust. Self-supervised learning contrastive training objectives are the most commonly used paradigm [Thota and Leontidis, 2021] for learning robust visual features in images. We incorporate self-supervised contrastive learning to learn domain invariant concepts, termed CCL.

Contrastive Sampling Procedure. The underlying idea revolves around utilizing multiple strong transformations of

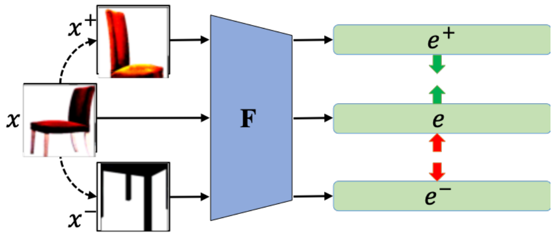

Figure 2: Self-supervised contrastive concept learning. Images sampled from a set of positive X^+ and negative samples X^- associated with an anchor image x. Green arrows depict direction of maximizing similarity, red arrows depict direction of minimizing similarity.

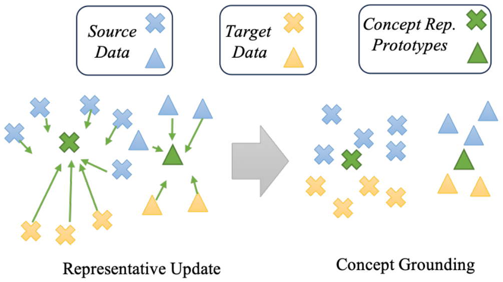

Figure 3: Prototype-based concept grounding (PCG). Concept grounding ensures the concept representations learned from both source and target domains are *grounded* to a representative concept representation prototype (Green).

an input sample x_i (anchor) and maximizing the similarity between their representations and minimizing the similarity between non-related transformations in the concept space, as shown in Figure 2. Mathematically, given an image sample x_i and the Concept Network $\mathbf{F}$, a set of transformations $T = \{t_1, t_2, ...t_n\}$, Contrastive learning begins by imputing a set of positive samples wrt x_i $X^+ = \{t_1(x_i), t_2(x_i)...t_n(x_i)\}$ and negative samples wrt x_j $X^- = \{t_1(x_j), t_2(x_j)...t_n(x_j)\}$. Note that the negative samples are not sampled from transformations of x_i but another sample x_j such that $i \neq j$. The concept representations wrt the positive and negative sets given a Concept Network $\mathbf{F}$ are $E^+ = \{e_i^+ = \mathbf{F}(x_i) \ \ \forall x_i \in X^+\}$ and $E^- = \{e_i^- = \mathbf{F}(x_i) \ \forall x_i \in X^-\}$ respectively.

Self-Supervised Training Objective. The extent of similarity is adjusted using a tunable hyperparameter τ (*temperature*), which controls the penalty on both positive and negative samples. Formally, the self-supervised loss $\mathcal{L}_{ssl}$ parameterized by an anchor image sample's concept representation $e = \mathbf{F}(x)$), its associated positive and negative sets' concept representations E^+ and E^- can be formulated as Equation 4.

$$\mathcal{L}_{ssl} = -\log\left(\frac{\exp(s(e, e^+)/\tau)}{\sum^{|E^-|}\exp(s(e, e^-)/\tau))}\right) \tag{4}$$

where $e^+ \in E^+, e^- \in E^-$ and s is any similarity function.

3.3 Prototype-based Concept Grounding

Ideally, concepts should be invariant entities shared among samples from similar classes and aligned across domains. However, due to imbalanced data and domain noise, models without explicit regularization learn significantly divergent concept representations for similar samples across domains, a phenomenon termed *concept-shift*. For proper concept alignment, it is important to ensure concept representations associated with samples of the same class from different domains are as close as possible. To achieve this, we utilize a prototype as an anchor, which *grounds* concept representations from multiple domains and reduces *concept-shift* during training. An illustration of concept grounding is presented in Figure 3. The blue and yellow data points correspond to concept representations for a class in the source and target domains respectively while the crosses and triangles represent different types of concepts. Our objective is to ground the source and target concept representations using a *concept representation prototype* (shown in green). Note that the training data X in our setting is a set of abundant samples from a source domain, X^s, and non-abundant samples from a target domain, X^t, i.e., $X = X^s \cup X^t$. Our prototype-based concept grounding method (PCG) utilizes a dynamically updated bank of concept representation prototypes to enforce concept alignment during training. The *concept bank* is constructed with concept representations of randomly sampled data points for each class from both source and target domains. Let N be the set of classes in the task. We sample a set of samples $S^s \subset X^s$ such that $S^s = \cup_{c=1}^{N} \mathcal{S}_c^s$ where S_c^s is a set of randomly selected samples belonging to class c from the source domain. Similarly, set $S^t \subset X^t$ is sampled from the target domain such that $S^t = \cup_{c=1}^{N} \mathcal{S}_c^t$. The representative concept prototype corresponding to a class c, $\mathcal{C}_c$, is updated after every training step with a weighted sum of the source and target concept prototypes associated with S^t and S^s:

$$\mathcal{C}_c \leftarrow \frac{\mu}{|S_c^s|}\sum_{x \in S_c^s} \mathbf{F}(x) + \frac{(1-\mu)}{|S_c^t|}\sum_{x \in S_c^t} \mathbf{F}(x) \tag{5}$$

where μ is a tunable hyperparameter used to control the extent of concept shift. Note that the higher the μ, the more concepts will be grounded to the source domain. The grounding concept code bank $\mathcal{C} = \{\mathcal{C}_c, \ \forall\ c \in \ N\}$ is used to supervise the concept representation learning as follows:

$$\mathcal{L}_{grnd} = L(\mathbf{F}(\mathbf{x}), \mathcal{C}) \tag{6}$$

where L is the same loss function as shown in Formula 2, which can be implemented as Mean Square Error.

Concept Fidelity Regularization. Concept fidelity attempts to enforce the similarity of concepts through a similarity measure $s(\cdot, \cdot)$ of data instances from the same class in the same domain. Formally,

$$\mathcal{L}_{fid} = s(\mathbf{F}(x_i), \mathbf{F}(x_j)) \quad \text{for} \ \ y_i = y_j \tag{7}$$

3.4 End-to-end Composite Training

Overall, the training objective can be formalized as a weighted sum of CCL and PCG objectives:

$$\mathcal{L}_{CL} = \mathcal{L}_{ssl} + \lambda_1 * \mathcal{L}_{grnd} + \lambda_2 * \mathcal{L}_{fid} \tag{8}$$

where λ_1 and λ_2 are tunable hyperparameters controlling the strength of contrastive learning and prototype grounding regularization. The end-to-end training objective can be represented as:

$$\mathcal{L}_{CE} + \beta * \mathcal{L}_{CL} \tag{9}$$

The tunable hyperparameter β controls the effect of generalization and robustness on the RCE framework. Note that a higher value of β makes the concept learning procedure brittle and unable to adapt to target domains. However, a very low value of β makes the concept learning procedure overfit on the source domain, implying a tradeoff between concept generalization and performance.

4 Experiments

Method	Explainable	Prototypes	Interoperability	Fidelity
S+T	✗	✗	✗	✗
SENN	✓	✓	✗	✗
DiSENN	✓	✓	✓	✗
BotCL	✓	✓	✗	✓
UnsupCBM	✓	✗	✗	✗
Ours	✓	✓	✓	✓

Table 1: A summary of salient features of our method as compared to the baselines considered. The column 'Explainable' shows whether the method is inherently explainable without any post-hoc methodologies. Column 'Prototypes' depicts if a method can explain predictions by selecting prototypes from the train set, 'Interoperability' shows if learned concepts maintain consistency across domains and 'Fidelity' depicts if the method maintains intra-class consistency among learned concepts.

4.1 Datasets and Networks

We consider four widely used task settings commonly utilized for domain adaptation. The task in each of the following settings is classification.

- **Digits**: This setting utilizes MNIST and USPS [LeCun *et al.*, 1998; Hull, 1994] with Hand-written images of digits and Street View House Number Dataset (SVHN) [Netzer *et al.*, 2011] with cropped house number photos.
- **VisDA-2017** [Peng *et al.*, 2017]: contains 12 classes of vehicles sampled from Real (R) and 3D domains.
- **DomainNet** [Venkateswara *et al.*, 2017]: contains 126 classes of objects (clocks, bags, etc.) sampled from 4 domains - Real (R), Clipart (C), Painting (P), and Sketch (S).
- **Office-Home** [Peng *et al.*, 2019]: Office-Home contains 65 classes of office objects like calculators, staplers, etc. sampled from 4 different domains - Art (A), Clipart (C), Product (P), and Real (R).

Network Choice: For Digits, we utilize a modified version of LeNet [LeCun *et al.*, 1998] which consists of 3 convolutional layers for digit classification with ReLU activation functions and a dropout probability of 0.1 during training. For all other datasets we utilize a ResNet34 architecture similar to [Yu and Lin, 2023] and initialize it with pre-trained weights from Imagenet1k. For details, refer Appendix.

Baselines. We start by comparing against standard non-explainable NN architectures - the S+T setting as described in [Yu and Lin, 2023]. Next, we compare our proposed method against 5 different self-explaining approaches. As none of the approaches specifically evaluate concept generalization in the form of domain adaptation, we replicate all approaches. **SENN** and **DiSENN** utilize a robustness loss calculated on the Jacobians of the relevance networks with DiSENN utilizing a VAE as the concept extractor. **BotCL** [Wang, 2023] also proposes to utilize contrastive loss but uses it for position grounding. Similar to BotCL, Ante-hoc concept learning [Sarkar *et al.*, 2022] uses contrastive loss on datasets with known concepts, hence we do not explicitly compare against it. Lastly, **UnsupervisedCBM** [Sawada, 2022b] uses a mixture of known and unknown concepts and requires a small set of known concepts. For our purpose, we provide the one-hot class labels as known concepts in addition to unknown. A visual summary of the salient features of each baseline is depicted in Table 1.

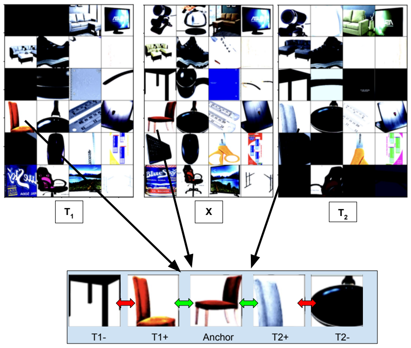

Figure 4: Schematic overview of proposed SimCLR transformations for OfficeHome dataset from the Product(P) domain. Note that green arrows depict maximizing similarity while red arrows depict minimizing similarity in concept space. Transformation sets T_1+ and T_2+ comprise images transformed from chair while T_1- and T_2- consist of images transformed from non-chair classes.

4.2 Hyperparameter Settings

RCE Framework: We utilize the Mean Square Error as the reconstruction loss and set sparsity regularizer λ to 1e-5 for all datasets. The weights $\omega_1 = \omega_2 = 0.5$ are utilized for digit, while they are set at $\omega_1 = 0.8$ and $\omega_2 = 0.2$ for object tasks.

Learning: We utilize the *lightly*[1] library for implementing SimCLR transformations [Chen, 2020]. We set the temperature parameter (τ) to 0.5 by default [Xu *et al.*, 2019] for all datasets. The hyperparameters for each transformation are defaults utilized from SimCLR. The training objective is Contrastive Cross Entropy (NTXent) [Chen, 2020]. Figure 4 depicts an example of various transformations along with the adjudged positive and negative transformations. For the training procedure, we utilize the SGD optimizer with momentum set to 0.9 and a cosine decay scheduler with an initial learning rate set to 0.01. We train each dataset for 10000 iterations with early stopping. The regularization parameters of λ_1 and λ_2 are set to 0.1 respectively. For Digits, β is set to 1 while it is set to 0.5 for objects. For further details, refer to Appendix.

4.3 Evaluation Metrics

We consider the following evaluation metrics to evaluate each component of the concept discovery framework.

- **Generalization**: We start by quantitatively evaluating the quality of concepts learned by measuring how well the learned concepts can generalize to new domains. To

[1] https://github.com/lightly-ai/lightly

	A → C	A → P	A → R	C → A	C → P	C → R	P → A	P → C	P → R	R → A	R → C	R → P
S+T	54.0	73.1	74.2	57.6	72.3	68.3	63.5	53.8	73.1	67.8	55.7	80.8
SENN	52.5	73.1	74.2	57.6	72.3	68.3	59.5	53.8	73.1	66.3	55.7	80.8
DiSENN	48.5	69.2	70.1	52.5	69.1	66.1	58.8	51.2	70.3	64.9	52.3	77.0
BotCL	53.1	72.8	74.0	**58.2**	70.4	67.9	58.4	52.1	72.6	65.3	56.3	78.2
UnsupCBM	54.0	73.1	74.2	57.6	72.3	68.3	63.5	53.8	73.1	**67.8**	55.7	80.8
RCE	52.5	73.1	74.2	57.6	72.3	68.3	**63.5**	53.8	73.1	67.8	55.7	80.8
RCE+PCG	55.2	73.1	74.0	57.9	71.2	68.1	58.1	53.6	73.2	66.9	56.1	80.3
RCE+PCG+CCL	**58.7**	**73.7**	**75.0**	58.0	**71.9**	**68.9**	62.1	**55.4**	**74.8**	67.2	**60.2**	**81.3**

Table 2: Domain generalization performance for the Office-Home Dataset with domains Art (A), Clipart (C), Product (P) and Real (R).

	DomainNet							VisDA	
	R → C	R → P	P → C	C → S	S → P	R → S	P → R	R → 3D	3D → R
S+T	60.0	62.2	59.4	55.0	59.5	50.1	73.9	79.8	49.4
SENN	59.2	60.1	57.2	53.8	56.1	49.0	72.4	79.6	49.2
DiSENN	57.3	58.1	55.3	51.2	55.1	47.4	71.0	78.1	48.1
BotCL	60.0	60.1	57.2	53.8	56.1	49.0	72.4	80.2	49.8
UnsupCBM	60.0	62.2	59.4	55.0	**59.5**	50.1	73.9	80.3	49.9
RCE	59.2	60.1	57.2	53.8	56.1	49.0	72.4	79.6	49.2
RCE+PCG	60.8	59.9	59.9	54.6	58.9	51.6	73.6	81.3	49.5
RCE+PCG+CCL	**61.2**	**60.5**	**62.9**	**55.0**	59.1	**52.1**	**74.2**	**82.4**	**53.4**

Table 3: Domain generalization performance for the [Left] DomainNet dataset with domains Real (R), Clipart (C), Picture (P), and Sketch (S) and [Right] VisDA dataset with domains Real (R) and 3-Dimensional visualizations (3D).

	M → U	M → S	U → M	U → S	S → M	S → U
S+T	0.54	0.16	0.74	0.13	0.92	0.65
SENN	0.43	0.11	0.73	0.09	0.92	0.64
DiSENN	0.43	0.11	0.73	0.09	0.92	0.64
BotCL	0.58	0.14	0.17	0.12	0.38	0.51
UnsupCBM	0.54	0.16	0.74	0.13	0.92	0.65
RCE	0.43	0.11	0.73	0.09	0.92	0.64
RCE+PCG	0.58	0.23	0.79	0.19	0.94	0.71
RCE+PCG+CCL	**0.60**	**0.23**	**0.81**	**0.20**	**0.95**	**0.71**

Table 4: Domain generalization performance for the Digit datasets with domains MNIST (M), USPS (U) and SVHN (S). In addition, the results of multiple source adaptation are in the Appendix.

achieve this, we compare our proposed method against the aforementioned baselines on domain adaptation settings.

- **Concept Fidelity**: To evaluate consistency in the learned concepts, we compute the intersection over union of the concept sets associated with for two data points x_i and x_j from same class as defined in Equation 10:

$$\text{Fidelity score} = |C^{x_i} \cap C^{x_j}| \, / \, |C^{x_i} \cup C^{x_j}| \qquad (10)$$

4.4 Genenralization Results

Tables 2, 3, and 4 report the domain adaptation results on the OfficeHome, DomainNet, VisDA and the Digit datasets, respectively. The notation X→Y represents models trained on X as the source domain (with abundant data) and Y as the target domain (with limited data) and evaluated on the test set of domain Y. The best statistically significant accuracy is reported in bold. The last three rows in all the tables list the performance of the **RCE** framework, RCE trained with regularization (**RCE+PCG**), and RCE trained with both regularization and contrastive learning paradigm (**RCE+PCG+CCL**).

Comparision with baselines. The first row in each table lists the performance of a standard Neural Network trained using the setting described in [Yu and Lin, 2023] (S+T). As a standard NN is not inherently explainable, we consider this setting as a baseline to understand the upper bound of the performance-explainability tradeoff.

The second and third rows in each table lists the performance of SENN and DiSENN respectively. SENN performs worse than S+T setting in almost all settings, except in a handful of settings where the performance matches S+T. This is expected, as SENN is formulated as an overparameterized version of a standard NN with regularization. Recall that DiSENN replaces the autoencoder in SENN with a VAE, and as such is not generalizable to bigger datasets without domain engineering. DiSENN performs the worst among all approaches for all datasets due to poor VAE generalization.

Recall that UnsupervisedCBM is an improved version of SENN architecture with a discriminator in addition to the aggregation function. In most cases, it performs slightly better than SENN and is at par with S+T. However, in particular cases in OfficeHome data (R→A) and DomainNet (S→P), UnsupCBM performs the best. We attribute this result to two factors: first, the Art (A) and Sketch (S) domains are significantly different from Real (R) and Picture (P) domains due to both of the former being hand-drawn while the latter being photographed as mentioned in [Yu and Lin, 2023]. Second, the use of a discriminator as proposed in UnsupervisedCBM helps enforce domain invariance in those particular cases.

BotCL explicitly attempts to improve concept fidelity and applies contrastive learning to *discover* concepts. However, the contrastive loss formulation is rather basic and they never focuses on domain invariance. BotCL's performance is similar to S+T for the most part except in OfficeHome data (C→A), where it just outperforms all other approaches. One possible reason is that Clipart domain is significantly less noisy, and hence basic transformations in BotCL work well.

As the last row demonstrates, our proposed framework RCE+PCG+CCL outperforms all baselines on a vast majority of the settings across all four datasets and is comparable to SOTA baselines in the other settings.

Ablation studies. We also report the performance corresponding to various components of our proposed approach. We observe that the performance of RCE is almost identical to SENN, which is expected as there is very weak regularization in both cases. In almost all cases, adding prototype-based grounding regularization (RCE+PCG) improves performance over RCE while models trained with both PCG regularization and contrastive learning (RCE+PCG+CCL) outperform all approaches on a vast majority of settings across all datasets. Note that the setting RCE+CCL is not reported, as it defeats the fundamental motivation of maintaining concept fidelity.

Effect of number of concepts and dimensions. We observe that there are no significant differences in performance over varying number of concepts or dimensions. For all results reported, the number of concepts is set as number of classes and are unidimensional. Refer Appendix.

4.5 Concept Fidelity

As RCE framework is explicitly regularized with a concept fidelity regularizer and grounded using prototypes, we would expect high fidelity scores. Table 5 lists the fidelity scores for the aforementioned baselines and our proposed method. Fidelity scores are averaged for each domain when taken as target (e.g. for domain (A) in DomainNet, the score is average of C→A, P→A and R→A). As expected, our method and BotCL, both with specific fidelity regularization outperform all other baseline approaches. Our method outperforms BotCL on most settings, except when the target domains are Art in DomainNet and Clipart in OfficHome due to siginificant domain dissonance.

	Digit			**DomainNet**				**OfficeHome**			
	M	U	S	A	C	P	R	C	P	R	S
SENN	0.81	0.74	0.61	0.21	0.24	0.26	0.30	0.31	0.27	0.29	0.30
DSENN	0.79	0.71	0.63	0.14	0,22	0.21	0.27	0.29	0.23	0.29	0.32
BotCL	0.93	**0.94**	0.89	**0.49**	0.55	0.51	0.58	**0.73**	0.66	0.61	0.64
U-CBM	0.79	0.74	0.63	0.21	0.24	0.26	0.30	0.31	0.27	0.29	0.30
RCE	0.86	0.80	0.73	0.39	0.50	0.45	0.42	0.54	0.49	0.50	0.49
R+P	0.94	0.94	0.89	0.47	0.56	0.51	0.59	0.70	0.67	0.61	0.63
R+P+C	**0.94**	0.94	**0.89**	0.47	**0.55**	**0.52**	**0.59**	0.71	**0.68**	**0.63**	**0.64**

Table 5: Average Intra-class Concept Fidelity scores for each domain for all settings where the domain is target. The columns show the domains in each dataset. For the complete table, refer Appendix.

4.6 Qualitative Visualization

Domain Alignment. We consider the extent to which the models trained using both concept grounding and contrastive learning maintain concept consistency not only within the source domain but also across the target domain as well. To understand what discriminative information is captured by a particular concept, Figure 5 shows the most important prototypes selected from the training set of both the source and target domains corresponding to five randomly selected concepts. We observe that prototypes explaining each concept are visually similar. For more results, refer Appendix.

Explanation using prototypes. For a given input sample, we also plot the prototypes associated to the highest activated concept, i.e., the important concept. Figure 6 shows the prototypes associated with the concepts most responsible for prediction (highest relevance scores). As can be seen, the prototypes possess distinct features, for eg., they capture round face of alarm clock. More results are reported in Appendix.

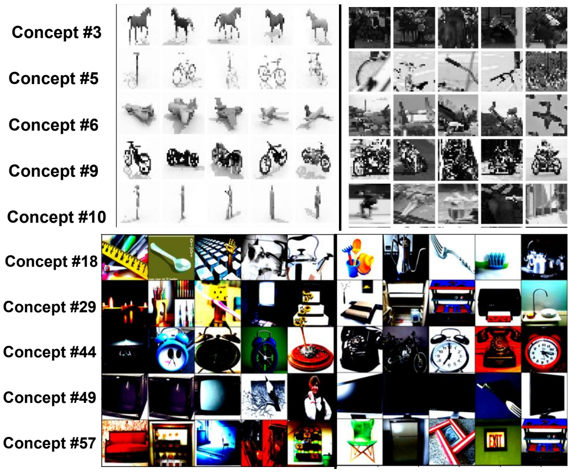

Figure 5: Top-5 most important prototypes for chosen concepts on models trained using our methodology on the VisDA [TOP] and OfficeHome [BOTTOM] dataset for the 3D → Real and Art (A) → Real (R) domains respectively. As can be seen, in the VisDA dataset Concept #6 captures samples with wings - namely airplanes while in OfficeHome, Concept #44 captures training samples with rounded faces in both domains - including alarm clocks, rotary telephones, etc. Similarly, Concept #29 captures flat screens - TVs/monitors.

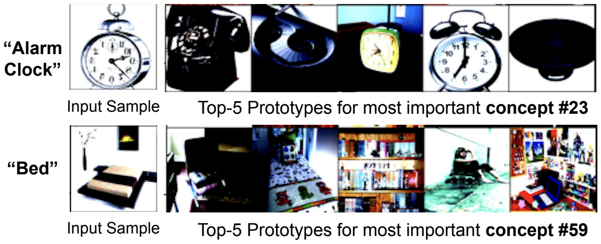
"Alarm Clock"
Input Sample
Top-5 Prototypes for most important **concept #23**
"Bed"
Input Sample
Top-5 Prototypes for most important **concept #59**

Figure 6: Top-5 most important prototypes associated with the highest activated concept. Prototypes associated with 'Alarm Clock' are distinctly circular objects while those associated with bed class are mostly flat.

5 Conclusion

In this paper, we discuss a fairly less-studied problem of *concept interoperability* which involves learning domain invariant concepts that can be generalized to similar tasks across domains. Next, we introduce a novel Representative Concept Extraction framework that improves on present self-explaining neural architectures by incorporating a Salient Concept Selection Network. We propose a Self-Supervised Contrastive Learning-based training paradigm to learn domain invariant concepts and subsequently propose a Concept Prototype-based regularization to minimize concept shift and maintain high fidelity. Empirical results on domain adaptation performance and fidelity scores show the efficacy of our approach in learning generalizable concepts and improving concept interoperability. Additionally, qualitative analysis demonstrates that our methodology learns domain-aligned concepts and can explain samples from both domains equally well. We hope our research helps the community utilize self-explaining models in domain alignment in the future.

Acknowledgements

This work is supported in part by the US National Science Foundation under grants 2217071, 2213700, 2106913, 2008208, 1955151. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[Aggarwal *et al.*, 2021] Ravi Aggarwal, Viknesh Sounderajah, Guy Martin, Daniel SW Ting, Alan Karthikesalingam, Dominic King, Hutan Ashrafian, and Ara Darzi. Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. *NPJ digital medicine*, 4(1):1–23, 2021.

[Alvarez-Melis and Jaakkola, 2018] David Alvarez-Melis and Tommi S Jaakkola. Towards robust interpretability with self-explaining neural networks. *arXiv preprint arXiv:1806.07538*, 2018.

[Bracke *et al.*, 2019] Philippe Bracke, Anupam Datta, Carsten Jung, and Shayak Sen. Machine learning explainability in finance: an application to default risk analysis. 2019.

[Chen *et al.*, 2019] Runjin Chen, Hao Chen, Jie Ren, Ge Huang, and Quanshi Zhang. Explaining neural networks semantically and quantitatively. In *ICCV*, pages 9187–9196, 2019.

[Chen, 2020] Ting Chen. A simple framework for contrastive learning of visual representations. In *ICML*. PMLR, 2020.

[D'Amour *et al.*, 2020] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *JMLR*, 2020.

[Ghorbani and Zou, 2019] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *ICML*, pages 2242–2251. PMLR, 2019.

[Ghorbani *et al.*, 2019] Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. *NeurIPS*, 2019.

[Gidaris *et al.*, 2018] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[Goyal *et al.*, 2019] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019.

[Huang *et al.*, 2022] Jinbin Huang, Aditi Mishra, Bum-Chul Kwon, and Chris Bryan. Conceptexplainer: Understanding the mental model of deep learning algorithms via interactive concept-based explanations. *arXiv preprint arXiv:2204.01888*, 2022.

[Hull, 1994] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

[Hutchinson and Mitchell, 2019] Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *FAccT*, pages 49–58, 2019.

[Jeyakumar *et al.*, 2021] Jeya Vikranth Jeyakumar, Luke Dickens, Yu-Hsi Cheng, Joseph Noor, Luis Antonio Garcia, Diego Ramirez Echavarria, Alessandra Russo, Lance M Kaplan, and Mani Srivastava. Automatic concept extraction for concept bottleneck-based video classification. 2021.

[Kim *et al.*, 2018] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, pages 2668–2677. PMLR, 2018.

[Koh and Liang, 2017] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *ICML*. PMLR, 2017.

[Koh *et al.*, 2020] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, pages 5338–5348. PMLR, 2020.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Liu *et al.*, 2021] Xiaoqing Liu, Kunlun Gao, Bo Liu, Chengwei Pan, Kongming Liang, Lifeng Yan, Jiechao Ma, Fujin He, Shu Zhang, Siyuan Pan, et al. Advances in deep learning-based medical image analysis. *Health Data Science*, 2021, 2021.

[Murty *et al.*, 2020] Shikhar Murty, Pang Wei Koh, and Percy Liang. Expbert: Representation engineering with natural language explanations. In *ACL*, pages 2106–2113, 2020.

[Netzer *et al.*, 2011] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[Peng *et al.*, 2017] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[Peng *et al.*, 2019] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pages 1406–1415, 2019.

[Pittino *et al.*, 2021] Federico Pittino, Vesna Dimitrievska, and Rudolf Heer. Hierarchical concept bottleneck models for explainable images segmentation, objects fine classification and tracking. *Objects Fine Classification and Tracking*, 2021.

[Raji *et al.*, 2020] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 33–44, 2020.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.

[Rudin, 2019] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[Saito *et al.*, 2020] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. *NeurIPS*, 33:16282–16292, 2020.

[Sarkar *et al.*, 2022] Anirban Sarkar, Deepak Vijaykeerthy, Anindya Sarkar, and Vineeth N Balasubramanian. A framework for learning ante-hoc explainable models via concepts. In *CVPR*, pages 10286–10295, 2022.

[Sawada, 2022a] Yoshihide Sawada. C-senn: Contrastive senn. 2022.

[Sawada, 2022b] Yoshihide Sawada. Cbm with add. unsup concepts. 2022.

[Sinha *et al.*, 2021] Sanchit Sinha, Hanjie Chen, Arshdeep Sekhon, Yangfeng Ji, and Yanjun Qi. Perturbing inputs for fragile interpretations in deep natural language processing. In *Proceedings of the Fourth BlackboxNLP Workshop at EMNLP*, pages 420–434, 2021.

[Sinha *et al.*, 2023] Sanchit Sinha, Mengdi Huai, Jianhui Sun, and Aidong Zhang. Understanding and enhancing robustness of concept-based models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15127–15135, 2023.

[Sundararajan *et al.*, 2017] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, pages 3319–3328. PMLR, 2017.

[Szepannek and Lübke, 2021] Gero Szepannek and Karsten Lübke. Facing the challenges of developing fair risk scoring models. *Frontiers in artificial intelligence*, 4, 2021.

[Thota and Leontidis, 2021] Mamatha Thota and Georgios Leontidis. Contrastive domain adaptation. In *CVPR*, pages 2209–2218, 2021.

[Varoquaux and Cheplygina, 2022] Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1):1–8, 2022.

[Venkateswara *et al.*, 2017] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017.

[Wang and Liu, 2021] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *CVPR*, pages 2495–2504, 2021.

[Wang, 2023] Bowen Wang. Learning bottleneck concepts in image classification. In *CVPR*, pages 10962–10971, 2023.

[Weller, 2019] Adrian Weller. Transparency: motivations and challenges. In *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 23–40. Springer, 2019.

[Wu *et al.*, 2020] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. Towards global explanations of convolutional neural networks with concept attribution. In *CVPR*, pages 8652–8661, 2020.

[Xu *et al.*, 2019] Jiaolong Xu, Liang Xiao, and Antonio M López. Self-supervised domain adaptation for computer vision tasks. *IEEE Access*, 7:156694–156706, 2019.

[Yeh *et al.*, 2019] Chih-Kuan Yeh, Been Kim, Sercan O Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *arXiv preprint arXiv:1910.07969*, 2019.

[Yu and Lin, 2023] Yu-Chu Yu and Hsuan-Tien Lin. Semi-supervised domain adaptation with source label adaptation. In *CVPR*, pages 24100–24109, 2023.

[Yuksekgonul *et al.*, 2022] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

[Zaeem and Komeili, 2021] Mohammad Nokhbeh Zaeem and Majid Komeili. Cause and effect: Concept-based explanation of neural networks. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2730–2736. IEEE, 2021.

[Zhou *et al.*, 2018] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *ECCV*, pages 119–134, 2018.