

X-Learner: Learning Cross Sources and Tasks for Universal Visual Representation

Yinan He^{1*} Gengshi Huang^{1*} Siyu Chen^{1*} Jianing Teng^{1*}
Wang Kun¹ Zhenfei Yin¹ Lu Sheng² Ziwei Liu³ Yu Qiao⁴ Jing Shao^{1†}

¹SenseTime Research ²College of Software, Beihang University

³S-Lab, Nanyang Technological University ⁴Shanghai AI Laboratory

{heyinan, huanggengshi, chensiyu, tengjianing, wangkun, yinzhenfei, shaojing}@senseauto.com

lsheng@buaa.edu.cn ziwei.liu@ntu.edu.sg qiaoyu@pjlab.org.cn

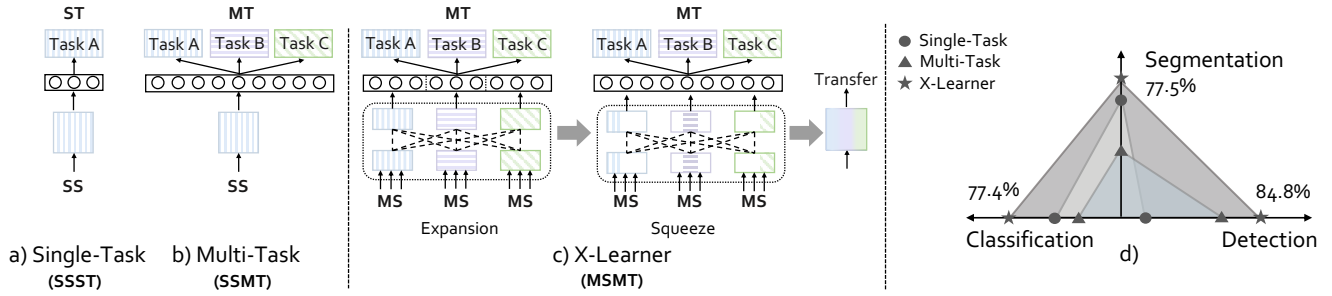


Figure 1. a) Single-Source Single-Task; b) Single-Source Multi-Task; c) X-Learner: Multi-Source Multi-Task; d) Our proposed X-Learner achieves the best performances in Classification (average linear probe results across 10 classification datasets), Detection (Pascal VOC Detection [15]) and Segmentation (Pascal VOC Semantic Segmentation [15]).

Abstract

In computer vision, pre-training models based on large-scale supervised learning have been proven effective over the past few years. However, existing works mostly focus on learning from individual task with single data source (e.g., ImageNet for classification or COCO for detection). This restricted form limits their generalizability and usability due to the lack of vast semantic information from various tasks and data sources. Here, we demonstrate that jointly learning from heterogeneous tasks and multiple data sources contributes to universal visual representation, leading to better transferring results of various downstream tasks. Thus, learning how to bridge the gaps among different tasks and data sources is the key, but it still remains an open question. In this work, we propose a representation learning framework called **X-Learner**, which learns the universal feature of multiple vision tasks supervised by various sources, with expansion and squeeze stage: **1) Expansion Stage:** X-Learner learns the task-specific feature to alleviate task interference and enrich the

representation by reconciliation layer. **2) Squeeze Stage:** X-Learner condenses the model to a reasonable size and learns the universal and generalizable representation for various tasks transferring. Extensive experiments demonstrate that X-Learner achieves strong performance on different tasks without extra annotations, modalities and computational costs compared to existing representation learning methods. Notably, a single X-Learner model shows remarkable gains of 3.0%, 3.3% and 1.8% over current pre-trained models on 12 downstream datasets for classification, object detection and semantic segmentation.

1. Introduction

Substantial advances have been achieved in visual representation learning, such as those based on curated large-scale image datasets with supervised [30, 59], weakly-supervised [29, 41], semi-supervised [65, 66], as well as self-supervised [7, 11, 12, 21, 25] pre-training. These visual representations show promising abilities in improving the performance on downstream tasks.

Among these pre-training techniques, supervised pre-training is widely adopted for its clear objective and steady

*Equal contribution.

†Corresponding author.

training process. Nevertheless, existing works in this direction only consider individual upstream task¹ (e.g., classification or detection) and most of them solely utilize one single data source (e.g., ImageNet [13] or COCO [39]). We argue this single-source single-task (SSST, Fig. 1 (a)) paradigm has several drawbacks: 1) The learned representation in SSST is specialized for one given task and is likely to have inferior performance on other tasks [19, 26, 44, 55, 56]. 2) It misses the potentials of a more robust representation by integrating characteristic semantic information from different tasks. Intuitively, we can opt to a simple hard-sharing method, i.e. single-source, multi-task (SSMT) paradigm, as described in Fig. 1 (b), by building many heads, each of which is specific for one task [24, 55]. However, this over-simplified algorithm usually encounters task interference [43, 73], especially for heterogeneous tasks, leading to a significant drop in performance. Besides, it requires the same image with a variety of labels [71, 72], which is not scalable easily due to the high annotation cost. A recent self-training work [19] attempts to create a pseudo multi-task dataset to alleviate the data-scarcity issue of multi-task learning, which follows a similar spirit to other SSMT works.

In light of issues with previous settings, we focus on utilizing numerous data sources of multiple tasks to learn a universal visual representation which should transfer well to various downstream tasks like classification, object detection and semantic segmentation. To leverage cross-source, cross-task information and mitigate undesired task interference, we propose a new pre-training paradigm *X-Learner*, as shown in Fig. 1 (c). The X-Learner contains two dedicated stages: **1) Expansion Stage:** It first trains a set of sub-backbones, each of which specifically exploits one task enriched with multiple sources. It then joins together these sub-backbones and combine their representational knowledge via our proposed *reconciliation layer*, forming an expanded backbone with enhanced modeling capacity. **2) Squeeze Stage:** Given the expanded backbone, this stage reduces the model complexity back to sub-backbone level and produces a unified and compact multi-task-aware representation. This new paradigm has two main advantages: **1)** It can effectively consolidate diverse knowledge from our new multi-source multi-task learning and avoid task conflicts. The resulting representation generalizes well to different types of tasks simultaneously. **2)** Compared to traditional multi-task methods, it is highly extensible with new tasks and sources, since we only require data sources annotated with single-task labels.

Our contributions are summarized as follows:

- We propose a new **multi-source multi-task learning** set-

¹To avoid ambiguity, we refer to a *task* as a general vision problem such as classification, detection or segmentation, and a *source* as a specific dataset or context within a certain *task*.

ting that only requires single-task label per datum, and is highly scalable with more tasks and sources without requiring any extra annotation effort.

- We present **X-Learner**, a general framework for learning a universal representation from supervised multi-source multi-task learning, with Expansion Stage and Squeeze Stage. Task interference can be well mitigated by Expansion Stage, while a compact and generalizable model is produced by Squeeze Stage. With X-Learner, heterogeneous tasks can be jointly learned, and the resulting single model renders a universal visual representation suitable for various tasks.
- We show the **strong transfer ability** of feature representations learned by our X-Learner. In terms of transfer learning performance, multi-source multi-task learning with our two-stage design outperforms traditional supervised single/multi-task training, self-supervised learning and self-training methods. As illustrated in Fig. 1 (d), a model pre-trained with X-Learner exhibits significant gains (3.0%, 3.3% and 1.8%) over the ImageNet supervised counterpart on downstream image classification, object detection and semantic segmentation.
- We offer **several new insights** into representation learning and the framework design for multi-task and multi-source learning through extensive experiments.

2. Related Work

Visual Representation Learning. Significant progress has been made in the field of visual representation learning, including unsupervised method [10, 11, 14, 25, 47, 49], supervised training [30, 59], weakly-supervised learning [29, 41], and semi-supervised learning [65, 66]. A large quantity of prior works use supervised datasets, including ImageNet1k [31], ImageNet-21K [52], IG-3.5B-17k [41] and JFT [30], for learning visual representations. In supervised pre-training, labeled training data provide significant improvement for transfer performance in the same task as the one for which the data are annotated. However, the ability of transferring across different tasks is not good enough [57]. In unsupervised learning, [49] focuses on multi-modal vision language pre-training to achieve strong performances in classification, but not do well in other visual tasks like detection [22]. In order to obtain uniformly high transfer performance on diverse task types, it is important to improve the task diversity of training data, justifying the necessity of multi-task pre-training.

Multi-Task Learning. There has been substantial interest in multi-task learning [4, 8, 23, 40, 50, 62, 72, 74, 77] in the community. A common practice for multi-task learning is to share the hidden layers of a backbone model

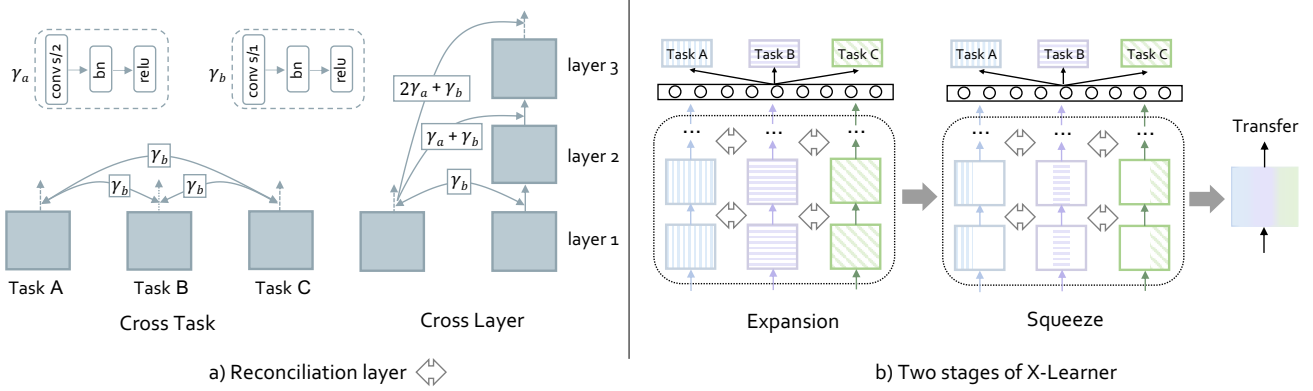


Figure 2. **Structure of X-Learner.** a) illustrates how reconciliation layers make the features from different tasks interact with each other. We use γ to represent the reconciliation layer. We present two typical ways of connection by reconciliation layer: cross different tasks and cross multiple layers; b) Features for different tasks are learned in Expansion Stage and unified in Squeeze Stage. After the two stages, X-Learner obtains a general representation for transferring to downstream tasks.

across different tasks, which is called “hard-sharing” in the literature. However, such sharing is not always beneficial, in many times hurting performance [23, 63, 69, 70]. To alleviate this, there are several lines of works to solve the problem in different ways. One of them is the use of a split architecture with parallel backbones for different tasks [18, 40, 45]. [45] proposes a cross-stitch module, which intelligently combines task-specific networks, avoiding the need to brute-force search through numerous architectures. Another line of works is improving optimization during learning [35, 63, 69, 70]. For example, [70] mitigates gradient interference by altering the gradients directly, i.e., performing “gradient surgery”. [63] addresses interference by de-conflicting gradients via projection. [35, 36] use distillation to avoid interference, but they are limited to a retrained setting, either single-task multi-source or single-source multi-task. Other works attempt to develop systematic techniques to determine which tasks should be trained together in a multi-task neural network to avoid harmful conflicts between non-affinitive tasks [1–3, 17, 34]. These methods perform multi-task learning to improve the performances of tasks involved, but they are not concerned with the transfer performance on downstream tasks. [37] applies vision transformer on multiple modalities and achieves impressive performance. For the image modality, it deals with the classification task only, and learns in a simple hard-sharing way. The problem of multi-task learning remains. A recent work [19] turns to semi-supervised learning and constructs cross-task pseudo labels with task-specific teachers, creating a complete multi-task dataset for pre-training. Yet it only considers the single-source setting, and its student training still follows a hard-sharing regime.

3. X-Learner

In this section, we introduce X-Learner, which leverages multiple vision tasks and various data sources to learn a unified representation that transfers well to a wide range of downstream tasks. It combines the superior modelling capacity of a split architecture design with the simplicity of hard parameter sharing. The whole two-stage framework is shown in Fig. 2. In Expansion Stage, we learn individual sub-backbones for different tasks with multi-source data in parallel. We further interconnect them to an expanded backbone that effectively alleviates interference among tasks. We then condense the expanded backbone to a normal-sized one in Squeeze Stage, producing the final general representation for downstream transfer.

3.1. Multi-Task and Multi-Source Learning

As illustrated in Fig. 1 (a), the most common supervised learning setting involves only one task with a single source, i.e., a datum from the source has one label or annotation corresponding to the only task (SSST). There is no task interference during optimization, yet the generated representation is weak in terms of transferability to other tasks.

Traditional multi-task approaches in previous works concurrently learn multiple tasks within a single data source (SSMT), which is shown in Fig. 1 (b). The single data source should have multiple sets of labels, each for one task. Such a data source is hardly scalable due to the high annotation cost.

To fix the drawbacks of previous setups, we propose our multi-source multi-task setting (MSMT), which is displayed in Fig. 1 (c). More concretely, let T be the number of tasks, then for each task $t \in \{1, 2, \dots, T\}$, there are N_t data sources $\mathcal{S}^t = \{(X_n^t, Y_n^t)\}_{n=1}^{N_t}$ with labels of the task. In this way, we only require $N = \sum_{t=1}^T N_t$ single-task data

Algorithm 1 Expansion Stage

Input: Data sources of T tasks $\{\mathcal{S}^t\}_{t=1}^T$, where $\mathcal{S}^t = \{(X_n^t, Y_n^t)\}_{n=1}^{N_t}$; Sub-backbones $\{\mathcal{E}^t\}_{t=1}^T$; Task losses $\{\ell_t\}_{t=1}^T$; Set of reconciliation layers γ ; Total step number K ; Step threshold τ

Output: pre-trained expanded backbone \mathcal{E}

```
1: Initialize  $\{\mathcal{E}^t\}_{t=1}^T$  and  $\gamma$ 
2: for  $k \leftarrow 1$  to  $K$  do
3:   for  $t \leftarrow 1$  to  $T$  do
4:     Sample a batch  $\mathcal{B}^t$  from  $\mathcal{S}^t$  with  $N_t$  sources
5:     if  $k \leq \tau$  then
6:       Forward with data  $\mathcal{B}^t$  on sub-backbone  $\mathcal{E}^t$ 
7:       Compute task loss  $\ell_t$ 
8:       Update  $\mathcal{E}^t$  separately with gradients from  $\ell_t$ 
9:     end if
10:   end for
11:   if  $k > \tau$  then
12:     Forward with multi-task data  $\{\mathcal{B}^t\}_{t=1}^T$  on ex-
        panded backbone  $\{\mathcal{E}^t\}_{t=1}^T \cup \gamma$ 
13:     Compute averaged loss  $L$  with Eq. (1)
14:     Jointly update  $\{\mathcal{E}^t\}_{t=1}^T \cup \gamma$  with gradients from  $L$ 
15:   end if
16: end for
17: return  $\{\mathcal{E}^t\}_{t=1}^T \cup \gamma$ 
```

sources which are easily attainable, avoiding the difficulty of multi-task annotation. Our setting is also highly extensible since adding new tasks or data sources becomes an effortless process. During training, the optimization objective of our multi-task and multi-source paradigm is to simply minimize the average loss over all the N data sources consisting of T different tasks:

$$\min_{\theta} L(\theta, \{\mathcal{S}^t\}_{t=1}^T) = \frac{1}{N} \sum_{t=1}^T \sum_{n=1}^{N_t} \ell_t(\theta, (X_n^t, Y_n^t)) \quad (1)$$

where θ denotes model parameters, and ℓ_t refers to the loss function for task t .

3.2. Expansion Stage

We aim to learn general representation from heterogeneous tasks while being least affected by the harmful interference among tasks. This motivates us to design this Expansion Stage to learn a split architecture combining multiple single-task networks. We first train T sub-backbones individually for the T tasks, leveraging their own data sources. We then join all T sub-backbones into one holistic architecture, integrating information learned from all tasks to form a general representation. Specifically, we introduce an expanded backbone composed of multiple sub-backbones corresponding to T tasks, along with several reconciliation layers for connecting them, which we describe

in detail below. The expanded backbone learned in this pipeline largely 1) preserves the high precision of single-task training, and 2) combines advantages of all tasks to achieve better generalizability on downstream tasks. The full training process is summarized in Algorithm 1.

Reconciliation Layer. As shown in Fig. 2 (a), each reconciliation layer is a link between two sub-backbones of two tasks. It obtains features from one task, transforms them with a few operations, and then fuses them into the features of another task at the same or a deeper layer.

Suppose each sub-backbone has D output layers, and we denote the original output of layer $i \in \{1, 2, \dots, D\}$ from the sub-backbone for task $t \in \{1, 2, \dots, T\}$ by \mathcal{E}_i^t . Let $\gamma_{j \rightarrow i}^{k \rightarrow t}$ ($j \leq i, k \neq t$) refer to the reconciliation layer taking \mathcal{E}_j^k as input and providing its output to the i^{th} layer of another task t . According to Fig. 2 (a), $\gamma_{j \rightarrow i}^{k \rightarrow t}$ can be expressed as the composition of one γ_b and $i - j$ times of γ_a . Receiving all cross-task and cross-layer features, we take a summation to compute the final fused output F_i^t at layer i of the sub-backbone for task t :

$$F_i^t = \mathcal{E}_i^t + \sum_{\substack{k=1 \\ k \neq t}}^T \sum_{j=1}^i \gamma_{j \rightarrow i}^{k \rightarrow t} (\mathcal{E}_j^k). \quad (2)$$

Adding reconciliation layers directly facilitates interactions among information from different tasks. Thus it closely unifies all sub-backbones into one expanded backbone expressing an integrated and general representation. In practical implementation, to avoid task interference introduced by such cross-task communication, we detach inputs to all reconciliation layers from the computational graph to cut off further gradient propagation.

3.3. Squeeze Stage

The previous Expansion Stage gives a concerted representation provided by the expanded backbone uniting all T sub-backbones of T tasks. However, it also introduces an undesirable T times increase in the number of model parameters and computational complexity. To maintain performance while reducing the expanded parameters, we present the Squeeze Stage. The final squeezed model remains highly generalizable for downstream transfer while sharing the same number of parameters with a single-task sub-backbone.

In Squeeze Stage, given an expanded backbone, we adopt distillation to consolidate the model. We employ the FitNets [53] approach, but with multiple targets (hints) from the expanded backbone as the student's supervision. Formally, given multiple outputs from the expanded teacher indexed by $t \in \{1, 2, \dots, T\}$, we refer to F^t as the output feature of task t , and \hat{F} as the feature of the student network. We perform distillation between the student model and the

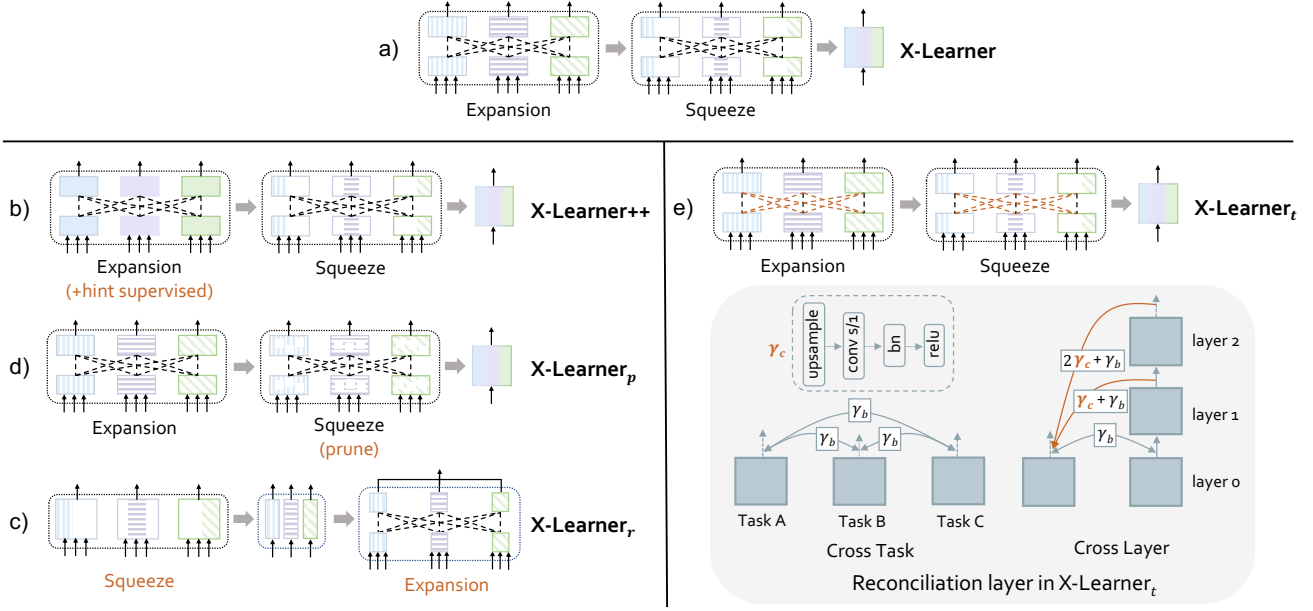


Figure 3. **Variants of X-Learner.** (a) is the default form of **X-Learner**. (b) The expansion stage of **X-Learner++** is supervised by extra hints from single-task single-source pre-trained models. (c) **X-Learner_r** is a Squeeze-Expansion version. (d) **X-Learner_p** replace the distillation with pruning in the squeeze stage. (e) We switch to a new reconciliation layer in **X-Learner_t**. Differences between variants and the default X-Learner are highlighted in red.

bunch of teacher outputs. Specifically, we project the single student feature \hat{F} through a task-specific guidance layer \mathcal{G}^t , and expect the outcome to match the teacher’s version F^t . Therefore, our distillation loss L_{squeeze} is simply the sum over squared L_2 losses of all teacher-student pairs:

$$L_{\text{squeeze}} = \sum_{t=1}^T \left\| F^t - \mathcal{G}^t(\hat{F}) \right\|_2^2 \quad (3)$$

The guidance layer \mathcal{G}^t is composed of a convolutional layer and a normalization layer:

$$\mathcal{G}^t(x) = \text{Norm}(\text{Conv}(x)). \quad (4)$$

We adopt an 1×1 convolution which transforms the student’s feature to have the same number of channels as the teacher’s output. For the normalization function, we simply choose Batch Normalization [28] as in [53].

3.4. Variants of X-Learner

X-Learner is a highly flexible multi-task pre-training framework, and many variants can be designed from the default setting. In this section, we describe several possibilities, which are illustrated in Fig. 3. More detailed differences among those variants are listed in Fig. 4.

X-Learner_r. We notice that the number of parameters in each individual model is first rising and then declining in our default X-Learner. It is natural to also study the reversed order, i.e., Squeeze-Expansion. In the new squeeze

stage, we use T task-specific teachers trained with multiple sources to distill T more light-weight sub-backbones. They are then combined into one network with normal computational complexity via reconciliation layers in the following expansion stage.

X-Learner_t. We make a modification on the reconciliation layers and let them take features from deeper layers of other sub-backbones as input and fuse to low-level features of a task. We also replace γ_a in cross-layer reconciliation layers with γ_c which is composed of an up-sampling layer and a convolutional layer.

X-Learner_p. We replace the distillation operation with unstructured pruning in Squeeze Stage. It is another way to reduce computation consumption while maintaining the performance of a network. We adopt a simple unstructured pruning method referencing [78].

X-Learner++. Inspired by [36], in the Expansion Stage, we add extra supervisions from single-task single-source pre-trained model in the form of hints besides the original supervision from labels of multiple data sources. This can be viewed as adding a pre-distillation process with multiple SSST teachers prior to training the expanded backbone.

Table 1. **Datasets used for X-Learner pre-training.** We grouped them into manually defined image domains according to [44].

Dataset	Task	Domain	Train Size
ImageNet [54]	General CLS.	Websearch	1.3M
Places365 [75]	General CLS.	Websearch	8.0M
iNat2021 [61]	Fine-Grained CLS.	Consumer	2.7M
CompCars [67]	Fine-Grained CLS.	Close-ups	120k
Tsinghua Dogs [79]	Fine-Grained CLS.	Close-ups	65k
COCO [39]	General DET.	Consumer	118k
Objects365 [56]	General DET.	Consumer	609k
WIDER FACE [68]	Face DET.	Websearch	13k
ADE20K [76]	Semantic SEG.	Consumer	20k
COCO-Stuff [6]	Semantic SEG.	Consumer	164k

Table 2. **Comparison with supervised and self-supervised methods on classification, detection and segmentation.** * represents the model is not pre-trained with semantic segmentation. We compare X-Learner to supervised pre-training, self-supervised learning, and a simple hard-sharing multi-task learning baseline. Relative gains are computed with respect to the ImageNet supervised baseline.

Method	AVG CIs	PASCAL Det	PASCAL Seg
ImageNet [54] Supervised	74.4	81.5	75.7*
SimCLR [10]	74.6	82.9	74.1*
Hard-sharing	73.2	83.7	70.5*
X-Learner	77.1 (+2.7)	84.4 (+2.9)	77.1* (+1.4)
X-Learner++	77.4 (+3.0)	84.8 (+3.3)	77.5* (+1.8)
X-Learner w/ seg	77.7 (+3.3)	84.3 (+2.8)	77.6 (+1.9)

4. Experiments

4.1. Pre-Training Settings

Pre-Training Sources (Datasets). Tab. 1 summarizes the sources we use for experiments. Most of our experiments are conducted in a base setting, where we pre-train models with 2 tasks: classification and object detection. We use 3 sources for image classification: ImageNet [54], iNat2021 [61] and Places365 [75] (Challenge version), and 2 sources for object detection: COCO [39] and Objects365 [56]. We also consider two extended settings: 1) to investigate the effect of more sources on X-Learner, we add CompCars [67] as well as Tsinghua Dogs [79] as two extra classification sources, and select WIDER FACE [68] as a new object detection source; 2) we study the impact of adding a new task, which is semantic segmentation, with ADE20K [76] and COCO-Stuff [6] as its sources.

Implementation Details. We implement X-Learner and its variants described in Sec. 3.4 using ResNet-50 [27] as the basic backbone throughout our experiments unless otherwise specified. The weights of reconciliation layers are initialized with [20]. We use SGD optimizer with a momentum of 0.9 [60], 10^{-4} weight decay and a base learning rate of 0.2. We decay the learning rate three times by a multi-step schedule with factors 0.5, 0.2 and 0.1 at 50%, 70% and 90% of the total iterations respectively.

4.2. Downstream Task Settings

Classification. We select 10 datasets from the well-studied evaluation suite introduced by [31], including general object classification (CIFAR-10 [33], CIFAR-100 [33]); fine-grained object classification (Food-101 [5], Stanford Cars [32], FGVC-Aircraft [42], Oxford-IIIT Pets [48], Oxford 102 Flower [46], Caltech-101 [16]), and scene classification (SUN397 [64]). We follow the linear probe evaluation setting used in [49]. We use the average accuracy of 10 classification datasets (AVG CIs) to represent the overall performance on the classification task. We train a logistic regression classifier using the L-BFGS optimizer, with a maximum of 1,000 iterations. We search the value for the L2 regularization strength λ over a set which distributes evenly over the range between 10^{-1} and 10^{-5} . We use images of resolution 224×224 for both training and evaluation.

Detection. We fine-tune our pre-trained model on PASCAL VOC07+12 (PASCAL Det) [15] for the detection task. We use Faster-RCNN [51] architecture in our experiments and run 24,000 iterations with a batch size of 16. We use SGD as the optimizer and search the best learning rate between 0.001 and 0.05. Weight decay is set to 10^{-4} , and momentum is set to 0.9. Evaluation is performed on the PASCAL VOC 2007 test set, with the shorter edges of images scaled to 800 pixels.

Semantic Segmentation. We evaluate models on PASCAL VOC 2012 (PASCAL Seg) [15]. We run 33,000 iterations with a batch size of 16. The architecture is based on Deeplab v3 [9]. We use SGD as the optimizer with a learning rate between 0.001 and 0.07. Weight decay is set to 10^{-4} , and momentum is set to 0.9. Images are scaled to 513×513 .

4.3. Main Results

Pre-Training Paradigm Comparison. Tab. 2 compares our pre-training scheme X-Learner with supervised training and self-supervised learning (SimCLR [10]) on ImageNet [54], as well as a simple hard-parameter-sharing baseline (named as “Hard-sharing”) on our multi-task and multi-source setting. We report performances on all three types of downstream tasks. Under the base setting, X-Learner uniformly outperforms all compared methods in terms of all evaluated metrics, especially AVG CIs. We also observe that the Hard-sharing model has better performance than the ImageNet-supervised model on PASCAL Det, but suffers a performance drop of 1.2% in AVG CIs. This suggests that the hard-sharing model benefits from multi-task pre-training with object detection sources included, but is harmed by task interference. In contrast, our X-Learner clearly overcomes the shortcoming and alleviates undesirable interference, leading to performance boosts on all considered tasks. Moreover, compared with training solely on

Experiment	Sub-Backbone	Expansion	Squeeze	Pre-Distillation	Parameters
Hard-sharing	ResNet-50	×	×	×	→
X-Learner	ResNet-50	✓	D	×	↗ ↘
X-Learner _r	HalfResNet-50	✓	D	×	↗ ↘
X-Learner _t	ResNet-50	✓	D	×	↗ ↘
X-Learner _p	ResNet-50	✓	P	×	↗ ↘
X-Learner++	ResNet-50	✓	D	✓	↗ ↘
X-Learner w/o Rec.	ResNet-50	×	D	×	↗ ↘

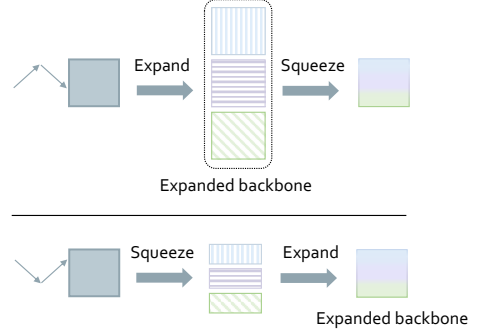


Figure 4. **Differences among X-learner variants.** We conduct different ablation study of X-Learner. Pre-distillation refers to applying extra supervisions from single-task single-source pre-trained models as is introduced in X-Learner++. In the Squeeze column, we denote distillation by D and pruning by P if there is a squeeze stage present in the pipeline. The change of the parameter can refer to the figure on the right.

Table 3. **Comparison on extended settings with extra pre-training sources.** By adding sources in different tasks (marked in bold italic), Hard-sharing suffers performance drops on both upstream and downstream tasks, while our X-Learner is stable across different settings, benefiting from the proposed *Expansion Stage*.

Experiments	Methods	Pre-train								Transfer	
		ImageNet	iNat2021	Places	<i>Cars</i>	<i>Dogs</i>	COCO	Objects365	<i>FACE</i>	AVG CIs	PASCAL Det
Base	Hard-sharing	75.0	75.3	53.0	–	–	35.5	17.4	–	73.2	83.7
	X-Learner	77.3	79.7	54.4	–	–	39.9	22.2	–	77.1	84.4
+ CIs Sources	Hard-sharing	73.7	73.6	52.3	98.5	85.3	35.4	17.6	–	77.5	83.1
	X-Learner	77.3	77.9	54.4	98.4	86.9	40.5	22.6	–	80.6	84.3
+ CIs & Det Sources	Hard-sharing	73.6	73.6	52.0	98.4	85.4	34.9	16.5	31.5	77.1	83.2
	X-Learner	76.9	78.6	54.6	98.6	85.9	40.1	22.1	33.6	80.5	84.3

ImageNet which is already specialized for classification, our approach still enjoys a 2.5% increase on AVG CIs. This result demonstrates that our setting of learning with multiple tasks simultaneously is beneficial for all involved pre-training tasks, such as classification here.

In addition, our X-Learner++ mentioned in Sec. 3.4 further enhances performance by means of its extra distillation process during sub-backbone training in the Expansion Stage, and achieves the best performance on all three downstream tasks.

We also compare our X-Learner++ with the multi-task self-training method MuST [19] in Tab. 4. For fair comparison, we fine-tune on the CIFAR-100 dataset instead of applying our default linear probe setting, evaluate PASCAL Det with pre-trained FPN [38], and set output stride to 8 in segmentation.

Our model surpasses MuST on classification and detection tasks despite using ResNet-50 instead of the more advanced ResNet-152 applied by MuST. To better show the effectiveness of our setting, we also conduct an experiment with the ResNet-152 backbone. Tab. 4 shows the performance of X-Learner_{R152} as well as MuST on four different tasks. We observe that our framework outperforms the self-training method by significant margins on all evaluated downstream tasks. Moreover, it is worth mentioning that

on NYU-Depth V2, our X-Learner, without any depth estimation pre-training, surpasses MuST which is learned with MiDaS, a mixture dataset with 10 depth-wise datasets. This zero-shot result further demonstrates the strong generalization capability of X-Learner.

We also compare our X-Learner_{R152} with a stronger version of MuST model pre-trained with JFT-300M, which is much larger than our datasets. As our X-Learner achieves 89.7 and 88.6 in downstream classification and detection tasks. This comparison proves that the dataset size is not an important factor, and our design has its superiority.

Cross-Task Generalization and Scalability. In Tab. 2, among methods that are not pre-trained on semantic segmentation, our X-Learner++ has the highest result on PASCAL Seg. This validates that our models produce more generalizable representations in terms of unseen tasks.

In addition to generalizability, our framework is also highly scalable and can incorporate extra tasks or sources effortlessly. As a demonstration, we add a semantic segmentation task according to the extended setting with ADE20K and COCO-Stuff. Results of “X-Learner w/ seg” in Tab. 2 show improvement on PASCAL Seg by 0.5 mIoU compared to the basic X-Learner. Classification performance is also benefitted from the new task introduced,

Table 4. **Comparison with self-training.** PASCAL Seg is an unseen task for X-Learner++, which is marked with *. NYU-Depth V2 is an unseen task for X-Learner_{R152}, which is marked with *.

Method	Backbone	Pre-training Settings	CIFAR-100 [33]	PASCAL Det [15]	PASCAL Seg [15]	NYU-Depth V2 [58]
MuST [19]	ResNet-152	ImageNet + DET. + SEG. + DEP.	86.3	85.1	80.6	87.8
MuST [19]	ResNet-152	JFT300M + DET. + SEG. + DEP.	88.3	87.9	82.9	89.5
X-Learner++	ResNet-50	ImageNet + DET.	87.0 (+0.7)	87.3 (+2.2)	78.8* (-1.8)	89.0 (+1.2)
X-Learner _{R152}	ResNet-152	ImageNet + OBJ365 + COCO	88.7 (+2.4)	88.5 (+3.4)	81.4 (+0.8)	91.2* (+3.4)
X-Learner _{R152}	ResNet-152	ImageNet + DET. + SEG.	89.7 (+3.4)	88.6 (+3.5)	82.6 (+2.0)	91.3* (+3.5)

Table 5. **The effect of applying reconciliation layers in the Expand Stage.** The reconciliation layer can significantly improve the performance in multi-task learning.

	AVG CIs	PASCAL Det
X-Learner w/o Rec	74.8	83.9
X-Learner	77.1	84.4

demonstrating the effectiveness of our multi-task learning approach.

Necessity of Reconciliation Layers. As shown in Tab. 5, we train an X-Learner without reconciliation layer to study the importance of the component. Compared to the default setting, removing reconciliation layers leads to significant performance drops at downstream transfer learning, especially on fine-grained datasets. We find that the feature from detection sub-backbone contains more detail, and it can be enhanced to a universal feature by the reconciliation layer. This phenomenon also verifies that reconciliation layers play a crucial role in coordinating multiple tasks towards the common goal of general representation learning.

4.4. In-Depth Studies

4.4.1 Multi-Task and Multi-Source Pre-Training

Observation 1: Proper multi-task learning promotes collaboration instead of bringing interference. As is discussed in Sec. 4.3, X-Learner not only resolves the task interference issue encountered by the hard-sharing model, but also surpasses single-task pre-trained models such as the ImageNet baseline in terms of downstream results. This shows that with an appropriately designed learning scheme, multi-task training is able to collaboratively enhance performances on all pre-training tasks. This conclusion is again corroborated by the results of X-Learner++ in Tab. 2. With a more elaborated design, performances on all tasks are again consistently boosted.

Observation 2: Additional sources further improve multi-task and multi-source representation learning if task conflicts are well-mitigated. We experiment on the extended setting with extra classification and detection sources. The added sources, such as CompCars [67] and WIDER FACE [68], have data in domains very different from exist-

ing sources. Ideally, including sources of complementary nature should help the overall multi-task and multi-source learning, since information available for pre-training is enriched and is more likely to cover downstream domains. However, this may also increase conflicts among tasks if not dealt with properly. In Tab. 3, we can see that the over-simplified hard-sharing baseline has considerably inferior results at both upstream and downstream if more sources are added. In pre-training stage, there is slight decrease after adding classification sources. This is due to the increase in task conflict when introducing new data domains. Nonetheless, we can find that additional sources becomes beneficial to transfer learning tasks both in hard-sharing and X-Learner. Compared to hard-sharing, X-Learner has mitigated such detrimental conflict to a certain extent with the aid of our two-stage design. This suggests that when task interference is properly alleviated, new data sources can be fully utilized by the model to learn more diverse knowledge and enhance the final representation.

4.4.2 Design of X-Learner Framework

Observation 3: Expansion-Squeeze is better than Squeeze-Expansion. In Sec. 3.4, we have described the X-Learner_r variant in which the order of the two stages within X-Learner is reversed. Performing squeezing first would result in smaller single-task sub-backbones with $1/T$ of the original size. Since $T = 2$ in our base setting, we should get two halved ResNet-50 models, corresponding to HalfResNet-50 in Fig. 4, which are to be joined in the further expansion process. HalfResNet-50 is a sub-backbone with only $1/\sqrt{2}$ of the original ResNet-50 channels. As shown in Tab. 6, X-Learner_r has lower performance on most pre-training tasks and all downstream tasks than the default X-Learner. This finding is reasonable since by intuition, shrinking sub-backbones first is likely to cause unrecoverable information

Table 6. **Comparison of various X-Learner variants.** Pre-training tasks and downstream tasks are evaluated on X-Learner variants. Our framework always performs better than Hard-sharing.

Method	Pre-train					Transfer	
	ImageNet	iNat2021	Places	COCO	Objects365	AVG Cls	PASCAL Det
Hard-sharing	75.0	75.3	53.0	35.5	17.4	73.2	83.7
X-Learner	77.3	79.7	54.4	39.9	22.2	77.1	84.4
X-Learner _r	73.9	76.6	52.5	41.1	21.7	73.9	84.1
X-Learner _t	76.3	79.9	53.3	42.5	22.0	74.5	83.5
X-Learner _p	76.1	78.6	53.5	42.4	23.4	77.2	83.1
X-Learner++	77.2	80.4	54.6	40.1	22.4	77.4	84.8

loss. It also validates our choice of Expansion-Squeeze for the default setup. Note that X-Learner_r is still better than the hard-sharing model, which again highlights the importance of a two-stage paradigm to mitigate task interference.

Observation 4: Reconciliation layers should receive information from lower levels. We also evaluate the alternative design of X-Learner_t, where reconciliation layers take features from deeper layers instead of shallower ones. Experiments in Tab. 6 show that the modified and original setups are both competitive at upstream pre-training. However, X-Learner_t is not as good as X-Learner in terms of downstream tasks. In conclusion, low-level features are more suitable to serve as complementary information among heterogeneous tasks.

Observation 5: Pruning may replace distillation in Squeeze Stage. In Tab. 6, X-Learner_p achieves results similar to those of X-Learner. This shows that pruning is also a valid choice for squeezing the expanded backbone, and thus is able to substitute distillation in Squeeze Stage.

5. Discussion and Conclusion

In this paper, we propose a flexible multi-task and multi-source pre-training paradigm called X-Learner, the general framework for representation learning by supervised multi-task learning. Heterogeneous tasks and diverse sources can be jointly learned with the help of the Expansion Stage and Squeeze Stage. We validate that X-Learner mitigates the well-known task interference problem and learns unified general representation that generalizes well to multiple seen and unseen tasks. We also show that X-Learner is superior to traditional supervised and self-supervised learning methods, as well as self-training approaches. In addition, We also demonstrate that our framework is highly flexible and additional tasks or sources can be integrated in a “plug-and-play” manner. Moreover, we offer several insightful observations through our experiments. One possible limitation is that the representation capability of our current pre-training is confined by the scale of publicly available datasets. It is possible to study with larger sources and more tasks in our framework. We hope this work will encourage further

researches towards creating general representations by performing multi-task and multi-source learning at scale.

References

- [1] Alessandro Achille, Giovanni Paolini, Glen Mbeng, and Stefano Soatto. The information complexity of learning tasks, their structure and their distance. *Information and Inference: A Journal of the IMA*, 10(1):51–72, 2021. 3
- [2] Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000. 3
- [3] Shai Ben-David and Reba Schuller. Exploiting task relatedness for multiple task learning. In *Learning theory and kernel machines*, pages 567–580. Springer, 2003. 3
- [4] Hakan Bilen and Andrea Vedaldi. Universal representations: The missing link between faces, text, planktons, and cat breeds. *arXiv preprint arXiv:1701.07275*, 2017. 2
- [5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014. 6
- [6] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 6
- [7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 1
- [8] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 2
- [9] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 6
- [10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 6
- [11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 2
- [12] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 1

- [13] 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
- [14] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27:766–774, 2014. 2
- [15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 1, 6, 8
- [16] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR workshop*, pages 178–178. IEEE, 2004. 6
- [17] Christopher Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *arXiv preprint arXiv:2109.04617*, 2021. 3
- [18] Yuan Gao, Jiayi Ma, Mingbo Zhao, Wei Liu, and Alan L. Yuille. Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3205–3214, 2019. 3
- [19] Golnaz Ghiasi, Barret Zoph, Ekin D Cubuk, Quoc V Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8856–8865, 2021. 2, 3, 7, 8
- [20] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 6
- [21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Dohersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 1
- [22] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Zero-shot detection via vision and language knowledge distillation. *arXiv e-prints*, pages arXiv–2104, 2021. 2
- [23] Yunhui Guo, Yandong Li, Liqiang Wang, and Tajana Rosing. Depthwise convolution is all you need for learning multiple visual domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8368–8375, 2019. 2, 3
- [24] Hu Han, Anil K Jain, Fang Wang, Shiguang Shan, and Xilin Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2597–2609, 2017. 2
- [25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 2
- [26] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019. 2
- [27] 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. 6
- [28] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015. 5
- [29] Armand Joulin, Laurens Van Der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer, 2016. 1, 2
- [30] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020. 1, 2
- [31] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2661–2671, 2019. 2, 6
- [32] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 6
- [33] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6, 8
- [34] Abhishek Kumar and Hal Daume III. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*, 2012. 3
- [35] Wei-Hong Li and Hakan Bilen. Knowledge distillation for multi-task learning. In *European Conference on Computer Vision*, pages 163–176. Springer, 2020. 3
- [36] Zhizhong Li, Avinash Ravichandran, Charles Fowlkes, Marzia Polito, Rahul Bhotika, and Stefano Soatto. Representation consolidation for training expert students. *arXiv preprint arXiv:2107.08039*, 2021. 3, 5
- [37] Valerii Likhoshesterov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, Yi Tay, Adrian Weller, and Mostafa Dehghani. Polyvit: Co-training vision transformers on images, videos and audio. *arXiv preprint arXiv:2111.12993*, 2021. 3
- [38] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 7
- [39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 6

- [40] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1871–1880, 2019. 2, 3
- [41] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision*, pages 181–196, 2018. 1, 2
- [42] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6
- [43] Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. Attentive single-tasking of multiple tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1851–1860, 2019. 2
- [44] Thomas Mensink, Jasper Uijlings, Alina Kuznetsova, Michael Gygli, and Vittorio Ferrari. Factors of influence for transfer learning across diverse appearance domains and task types. *arXiv preprint arXiv:2103.13318*, 2021. 2, 6
- [45] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3994–4003, 2016. 3
- [46] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *CVPR*, volume 2, pages 1447–1454. IEEE, 2006. 6
- [47] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2
- [48] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE, 2012. 6
- [49] 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. *arXiv preprint arXiv:2103.00020*, 2021. 2, 6
- [50] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *arXiv preprint arXiv:1705.08045*, 2017. 2
- [51] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 6
- [52] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 2
- [53] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 4, 5
- [54] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 6
- [55] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013. 2
- [56] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8430–8439, 2019. 2, 6
- [57] Zhiqiang Shen, Zhuang Liu, Jianguo Li, Yu-Gang Jiang, Yurong Chen, and Xiangyang Xue. Object detection from scratch with deep supervision. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):398–412, 2019. 2
- [58] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 8
- [59] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 1, 2
- [60] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013. 6
- [61] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisín Mac Aodha. Benchmarking representation learning for natural world image collections. In *CVPR*, pages 12884–12893, 2021. 6
- [62] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. Towards universal object detection by domain attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7289–7298, 2019. 2
- [63] Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. *arXiv preprint arXiv:2010.05874*, 2020. 3
- [64] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *IJCV*, 119(1):3–22, 2016. 6
- [65] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. 1, 2
- [66] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6509–6518, 2020. 1, 2
- [67] Linjie Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3973–3981, 2015. 6, 8
- [68] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of*

the IEEE conference on computer vision and pattern recognition, pages 5525–5533, 2016. 6, 8

- [69] Yilin Yang, Akiko Eriguchi, Alexandre Muzio, Prasad Tadepalli, Stefan Lee, and Hany Hassan. Improving multilingual translation by representation and gradient regularization. *arXiv preprint arXiv:2109.04778*, 2021. 3
- [70] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *arXiv preprint arXiv:2001.06782*, 2020. 3
- [71] Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11197–11206, 2020. 2
- [72] 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
- [73] Xiangyun Zhao, Haoxiang Li, Xiaohui Shen, Xiaodan Liang, and Ying Wu. A modulation module for multi-task learning with applications in image retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 401–416, 2018. 2
- [74] Xiangyun Zhao, Samuel Schulter, Gaurav Sharma, Yi-Hsuan Tsai, Manmohan Chandraker, and Ying Wu. Object detection with a unified label space from multiple datasets. In *European Conference on Computer Vision*, pages 178–193. Springer, 2020. 2
- [75] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 6
- [76] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 6
- [77] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. *arXiv preprint arXiv:2102.13086*, 2021. 2
- [78] L. Zhuang, M. Sun, T. Zhou, H. Gao, and T. Darrell. Rethinking the value of network pruning. 2018. 5
- [79] Ding-Nan Zou, Song-Hai Zhang, Tai-Jiang Mu, and Min Zhang. A new dataset of dog breed images and a benchmark for finegrained classification. *Computational Visual Media*, 6(4):477–487, 2020. 6