

# Transferable and Principled Efficiency for Open-Vocabulary Segmentation

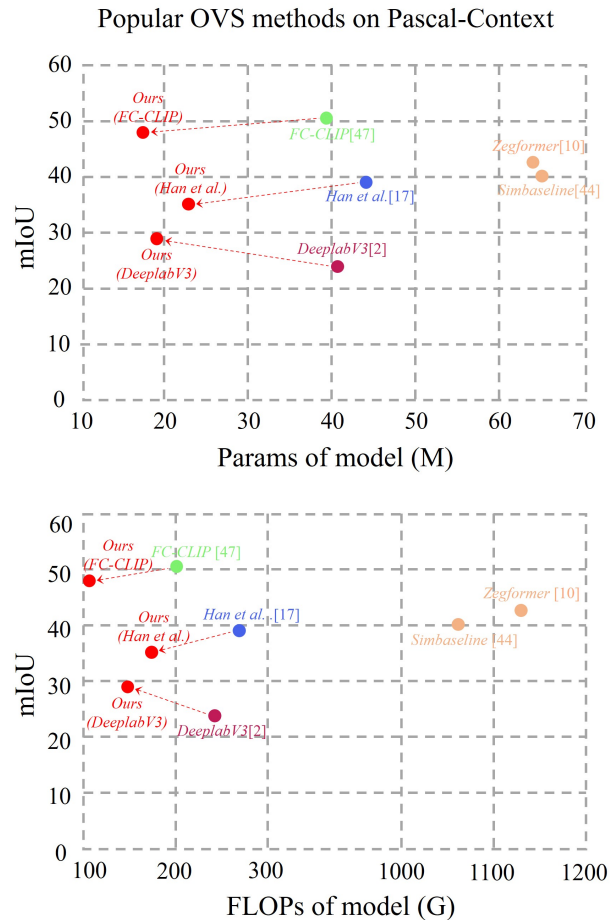
Jingxuan Xu<sup>1</sup>, Wuyang Chen<sup>2</sup>, Yao Zhao<sup>1</sup>, Yunchao Wei<sup>1</sup>  
<sup>1</sup>Beijing Jiaotong University <sup>2</sup>Simon Fraser University

## Abstract

Recent success of pre-trained foundation vision-language models makes Open-Vocabulary Segmentation (OVS) possible. Despite the promising performance, this approach introduces heavy computational overheads for two challenges: 1) large model sizes of the backbone; 2) expensive costs during the fine-tuning. These challenges hinder this OVS strategy from being widely applicable and affordable in real-world scenarios. Although traditional methods such as model compression and efficient fine-tuning can address these challenges, they often rely on heuristics. This means that their solutions cannot be easily transferred and necessitate re-training on different models, which comes at a cost. In the context of efficient OVS, we target achieving performance that is comparable to or even better than prior OVS works based on large vision-language foundation models, by utilizing **smaller models that incur lower training costs**. The core strategy is to make our efficiency **principled** and thus seamlessly **transferable** from one OVS framework to others without further customization. Comprehensive experiments on diverse OVS benchmarks demonstrate our superior trade-off between segmentation accuracy and computation costs over previous works. Our code is available on <https://github.com/Xujxyang/OpenTrans>

## 1. Introduction

Open-vocabulary segmentation has been developed to address the limitations of traditional semantic segmentation [2, 22, 42, 50], which cannot identify categories beyond its training set, by leveraging human-like recognition of novel categories using arbitrary text descriptions. This method enables the segmentation of arbitrary categories from text inputs, expanding its use in areas like image editing and human-robot interaction. Early models used cross-modal alignment between pixel and text embeddings [25, 43, 48], while recent methods have improved performance with region-level alignment [10, 15, 19, 28, 44]. These methods learn from base class embeddings and are often paired with a frozen pre-trained vision-language foundation model, such as CLIP [37], to assist in re-classification, with strate-



**Figure 1.** Comparison with popular open-vocabulary segmentation works on Pascal-Context [36] (with Resnet50 as backbone). The red points represent our works, the orange points stand for traditional two-stage open-vocabulary segmentation [10, 44], the blue point stands for single-stage open-vocabulary segmentation [17], the green denotes a novel dual-channel prediction model with frozen backbone [47], the purple represent traditional convolutional segmentation networks [2].

gies that incorporate knowledge distillation or a frozen CLIP backbone to maintain a generalizable representation while fine-tuning on known classes.

Despite the promising performance of recent OVS frameworks, these methods introduce heavy computational over-

heads. The core bottleneck to the efficiency of recent OVS frameworks is rooted in the challenging balance between the preservation of the generalization ability of pretrained CLIP model versus the overhead of model size and training costs. **First**, as the pretrained CLIP model supports the extraction of generalizable features of unseen images, the entire heavy image encoder has to be inherited for the OVS task. This introduces the requirement for the large model size of the CLIP image encoder. **Second**, after inheriting from CLIP, one needs to fine-tune together with the segmentation head. This further requires expensive training computations, and careful and delicate fine-tuning to preserve the pretrained generalizable knowledge in the CLIP image encoder. These challenges hinder this OVS strategy from being more widely applicable and affordable in real-world scenarios.

Traditional methods, including model compression [3, 4, 13] and efficient fine-tuning [20, 21, 39], can mitigate these challenges. For example, iterative magnitude pruning can find semantic-aware lottery tickets that introduce highly sparse networks adapted to downstream tasks. Adapter-based fine-tuning can partially update pretrained models by introducing lightweight modules. Despite the efficiency, however, these methods come with a price of **heuristics**: their solutions cannot be easily transferred and have to be re-trained on different models. One has to re-run the pruning to find task-specific and model-specific sparse masks that can *hardly be transferred* to different frameworks. During the fine-tuning, adapter layers have to be *empirically customized* and inserted into pretrained models, which introduces extra computation costs. Therefore, we are motivated to ask our core question:

*Can we design **principled** methods to make OVS efficient, and seamlessly **transferable** to different frameworks?*

This work targets achieving performance as competitive as large vision-language foundation models but using smaller models with less training costs. We propose a Transferable Open-vocabulary segmentation technic (dubbed OpenTrans), to establish a principled and seamlessly transferable efficiency across various OVS frameworks, eliminating the need for additional customization. This enables us to achieve the “one design, transferable to all” fashion. **First**, to address the large model size of the CLIP image encoder, we explore small subnetwork by iterative magnitude pruning. To avoid overfitting the seen OVS classes and finding sparse masks specific to OVS frameworks, we prune the CLIP image encoder without semantic supervision. This strategy enables our pruning approach to discover highly transferable subnetwork that seamlessly adapt to the latest OVS frameworks without requiring any modifications, avoiding any further pruning costs. **Second**, to address the heavy OVS fine-tuning computation, we propose to only select partial layers to update. To enable a “plug-and-play” strategy for layer-wise selective fine-tuning, we target analyzing the quality of pre-

trained layers without any data or any forward/backward process. We investigate the spectrum of pretrained weights, freeze layers with heavy tails, and only update those with light tails. This principled analysis not only further accelerates the training with fewer computations, but also introduces zero overhead in the pretrained model and can be directly adopted in different OVS frameworks. We conduct comprehensive experiments on diverse OVS benchmarks, our method can employ 21.2M smaller networks (Parameters), 59.11P less training costs (Training FLOPs).

We summarize our contributions below:

- Transferable sparse model: We reduce the backbone for 21.2M smaller in parameters and reduce 94.6G FLOPs. We avoid extra pruning costs on diverse latest OVS frameworks by directly transferring our subnetwork.
- Principled efficient fine-tuning: We design a plug-and-play layer-wise selection approach to enable efficient fine-tuning that reduces the training costs in OVS.
- Superior mIoU-efficiency trade-off: Extensive experiments demonstrate that we achieve comparable or even improved mIoU over diverse OVS benchmarks with significantly reduced computation costs for both training and inference.

## 2. Related work

### 2.1. Vision-Language Pre-training Model

Vision-language pre-training is a pre-training approach that integrates language and vision domains. In its early stages [26, 27], research in this field was initially limited to smaller datasets, requiring subsequent fine-tuning on downstream tasks. Recent studies emphasize the advantages of leveraging the extensive pool of web data accessible to researchers. Through the implementation of contrastive learning, CLIP [37] establishes meaningful connections between images and their corresponding captions, achieving notable performance in cross-modal alignment. Alin [23] adopts a straightforward dual-encoder architecture and employs a contrastive loss to align the visual and linguistic representations of image and text pairs. The advent of these methods has played a substantial role in advancing Open-Vocabulary Segmentation. Despite the promising performance, vision-language pre-trained models still impose significant computational overheads, primarily attributed to the large sizes of the backbone. This poses challenges in terms of wide applicability and affordability. In contrast, we utilize CLIP as a source of open-domain capabilities, but propose efficiency to its image encoder with transferable sparse masks.

### 2.2. Open-Vocabulary Segmentation

The purpose of OVS is to segment diverse classes, including those that are inaccessible during the training process. The developmental trajectory of OVS can be delineated into two

phases: pre- and post-introduction of vision-language pre-training models. Research conducted before the advent of vision-language pre-training models, as observed in works such as [1, 43], often yield suboptimal results. Therefore, integrating these models into OVS emerges as a highly advantageous strategy [11, 24, 28]. Just after CLIP came out, Simbaseline [44] adopts proposal masks for cropping the input image and leverages CLIP to extract region-level features. Zegformer [10] utilizes CLIP as its encoder and incorporates MaskFormer [5] for extracting mask proposals, both demonstrating improved performance in OVS. Recently, CAT-Seg [7] introduces a spatial and class aggregation network, integrating multi-modality guidance features to enhance open-vocabulary pixel classification effectiveness, achieving state-of-the-art performance on part datasets. FC-CLIP [47] proposes a structure with two classifiers and freezes the CLIP backbone to uphold open-vocabulary recognition, also attaining state-of-the-art performance on certain datasets. Han et al. [17] develop an efficient framework that avoids relying on the extra computational burden of the CLIP model. In our research, our focus is on advancing principled efficiency for OVS frameworks. This is accomplished through the implementation of transferable sparse masks and efficient fine-tuning, employing spectrum analysis of pretrained weights.

### 2.3. Parameter-Efficient Transfer Learning

Parameter-Efficient Transfer Learning (PETL) is a research direction focused on minimizing the computational burden associated with adapting large pretrained models to new tasks. This is achieved by selectively updating specific parameters, eliminating the necessity to update the entire model. Adapters [20, 39] are compact bottleneck modules that are inserted within transformer layers. Empirical studies have demonstrated that using layer normalization layers alone for training adapters is adequate to achieve optimal performance during the fine-tuning process. Side-Adapters [40, 45] partitions the trainable parameters of the backbone model and forms a distinct side network. The designated task for this side network is to adapt the entire model to novel tasks, leading to its specific design tailored to different structures and tasks. Following a similar pattern, LoRA [21] introduces trainable rank decomposition matrices into a pre-trained model that remains fixed. Although the majority of advancements in PETL have been observed in the NLP domain, researchers also extend the application of this technique to the fields of vision-language (VL) [41, 45, 49, 52]. However, the adapter-based efficient fine-tuning methods still have to introduce extra layers or modules to the pretrained layers, which require extra design and customization efforts. Instead, we aim to propose a principled and plug-and-play efficient fine-tuning method that can be seamlessly adopted in different OVS frameworks.

## 3. Method

### 3.1. Preliminaries

**General OVS Pipeline.** OVS aims at training a segmentation model capable of segmenting arbitrary classes using text descriptions. Most of the existing methods decoupled OVS into two steps: mask proposals generation and mask proposals classification. They first involve MaskFormer[6] to generate class-agnostic mask proposals, followed by leveraging CLIP’s transferability to classify these proposals. The mask proposals classification process can be defined as:

$$C_i = v_i \otimes T = \text{softmax}((v_i * t_1), (v_i * t_2) \dots (v_i * t_{|C|})), \quad (1)$$

where  $C$  is the classification score.  $v_i$  is the vision embeddings for  $i$  mask proposals, which executed with either masked crops [10, 44] or masked attention [45].  $T$  is the text embeddings from CLIP-pretrained text encoder. Building upon this inspiration, Our framework maintains the utilization of the Mask2Former architecture..

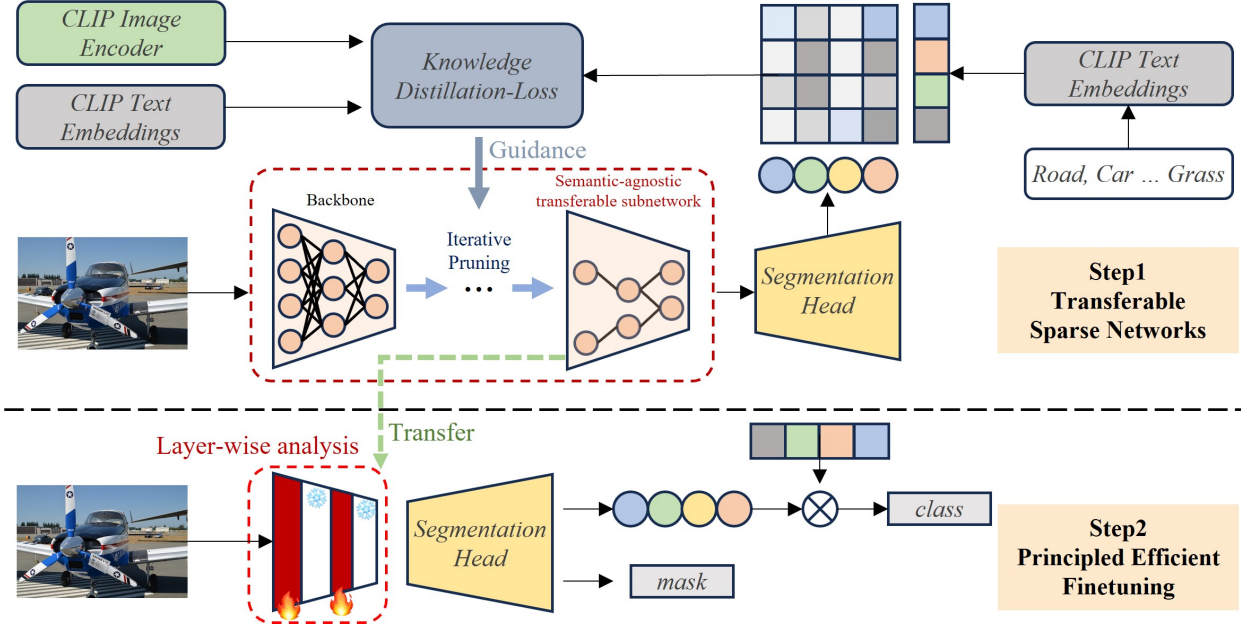
During training, in order to preserve the generalizability of CLIP, we employ a distillation loss [17] (“Knowledge Distillation Loss” in Fig. 2, denoted as  $\mathcal{L}_{KD}$ ) to distill the CLIP’s knowledge to our backbone. The distillation loss guarantees that the distance between text embeddings and visual embeddings remains consistent across all pairwise categories. The alignment of the CLIP text space with the image feature space enhances the performance of open vocabulary, while being agnostic to any semantic class. Refer to the supplementary material for details. In our work, we introduced this loss to enhance the open-domain characteristics of our pruning method.

**Out Target.** In this work, we target introducing two levels of efficiency to current OVS works: both model and training efficiency. More importantly, our core contribution lies in not only establishing these efficiency principles but also ensuring their transferability across various OVS frameworks. We will introduce two steps (see Fig. 2):

1. **Model efficiency** (Sec. 3.2): We prune the heavy CLIP image encoder without semantic awareness to make the backbone transferable to different OVS frameworks.
2. **Training efficiency** (Sec. 3.3): During fine-tuning, we only partially update selective layers of the backbone to further reduce training costs.

### 3.2. Model Efficiency

In this section, our main objective is to improve model efficiency through model pruning. Specifically, we aim to answer the question: *Can we find efficient subnetwork that can be seamlessly transferred to different OVS frameworks?* We focus on finding a subnetwork within the model’s backbone that allows for the seamless transfer of its sparse masks across different OVS works.



**Figure 2.** Overview of OpenTrans(Ours). We introduce principled and transferable efficiency in two folds. **Step1:** we prune the heavy CLIP image encoder without semantic awareness, this turns the backbone into a **semantic-agnostic transferable subnetwork**. This allows us to seamlessly transfer subnetwork to other OVS Frameworks, such as Deeplabv3 [2], Mask2former [6] and FC-CLIP [47]. **Step2:** during fine-tuning, we further explore and prioritize principled efficiency by introducing **layer-wise heavy-tail spectrum analysis**. This method involves selectively updating layers with light-tail spectra in their pretrained weights, while keeping layers with heavy-tail spectra frozen.

**Pruning method.** To identify the pruning mask for the backbone, we utilize the classical Iterative Magnitude Pruning (IMP) approach, as described in the LTH (Lottery Ticket Hypothesis) literature [3, 4, 13, 14]. The pruning of the model’s backbone is carried out in a two-step process. First, we train an unpruned network as the pre-training model. Then, we remove a portion of weights with the globally smallest magnitudes [4, 18, 38].

We acknowledge the crucial role of the knowledge distillation loss  $\mathcal{L}_{KD}$  in aligning feature spaces between CLIP text and vision features, which significantly impacts the model’s open-vocabulary performance. Hence, during the pruning process to find the pruning mask, we solely utilize the knowledge distillation loss and exclude any semantic supervision. This approach also allows us to remain agnostic to any specific OVS framework. More importantly, it helps us discover a pruning mask on one OVS framework that can be transferable to other frameworks without any further customization, while still preserving the open-domain capabilities in CLIP without any risk of overfitting to specific semantic classes.

We denote our backbone as  $f(x; \theta)$ , where  $x$  is the input and  $\theta \in \mathbb{R}^d$  as its parameters. The subnetwork of the backbone can be described as  $f(x; m \odot \theta)$  with a pruned binary mask  $m \in \{0, 1\}^d$ , where  $\odot$  is the element-wise product. We depict our pruning outlines in Algorithm 1, the sparsity level indicates the remaining weights of the model.

**Subnetwork Transfer.** Upon obtaining the pruning mask

---

**Algorithm 1** Semantic-agnostic Model Pruning.

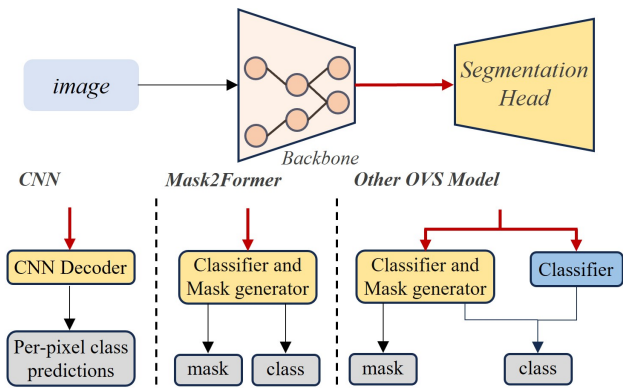
---

- 1: Set the initial mask to  $m = 1^d$ , with pre-trained parameters  $\theta$ .
  - 2: **repeat**
  - 3:   Train  $f(x; m \odot \theta)$  for  $t$  iterations with only the distillation loss  $\mathcal{L}_{KD}$ .
  - 4:   Prune  $p\%$  of remaining weights in  $\theta$  and update  $m$  accordingly.
  - 5: **until** the sparsity of  $m$  reaches the desired sparsity level  $s$ .
  - 6: Return  $f(x; m \odot \theta)$ .
- 

through IMP, we proceed to transfer it to other OVS frameworks (Fig. 3) by applying it to different segmentation backbones [2, 6, 47]. Importantly, no fine-tuning or additional adjustments are required when incorporating our subnetwork into these alternative frameworks, as it seamlessly integrates with the model.

To demonstrate that our transferable subnetwork can be widely adopted, we choose to study three representative and diverse OVS frameworks:

- **Traditional CNN-based OVS:** In the traditional convolutional segmentation model, CLIP text embeddings can be utilized as the classifier. Thus, in Eq. (1), the  $v_i$  represents the vision embeddings for the  $i$ -th pixel. We choose DeeplabV3 [2] as our transfer target, which incorporates the ASPP module to enhance its performance.
- **Mask2Former-based OVS:** This line of work separates the process of acquiring masks and predicting mask cat-



**Figure 3.** We transfer our semantic-agnostic sparse masks to different OVS pipelines. **Left:** Traditional CNN-based frameworks: these frameworks commonly employ simple upsampling techniques. **Middle:** Mask2Former-based frameworks, the approach utilizes a decoupled mask and class segmentation head. **Right:** OVS with extra classifiers: some frameworks propose to leverage a frozen CLIP classifier to further utilize the pretrained visual knowledge (with the blue box indicating frozen classifier layers of the model).

egories, resulting in higher quality masks compared to CNN methods. We choose Han et al. [17] as our transfer target.

- OVS with extra classifiers: Beyond Mask2Former, many recent OVS frameworks additionally utilize an extra pre-trained classifier obtained from CLIP. We choose FC-CLIP [47] as our target, which uses an out-of-vocabulary classifier during inference and greatly helps in OVS.

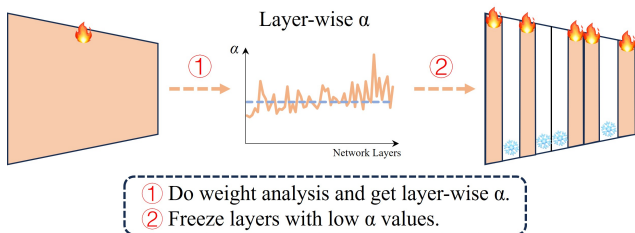
In our experiments (Sec. 4.2), we will show that by adopting our transferable subnetwork, we can facilitate the efficiency of all three types of OVS frameworks. After transferring the subnetwork, the sparsified pretrained backbone will be fine-tuned with CLIP text embeddings in the next section.

### 3.3. Training Efficiency

In current OVS frameworks, heavy training costs are a common issue. The fine-tuning stage typically updates all layers without considering the quality of different pretrained weights, resulting in inefficiencies. In this section, our objective is to address the question: *Can we make the fine-tuning of OVS frameworks principled and efficient?* To achieve this, we propose a principled and explainable metric that guides the fine-tuning of pretrained OVS models. Importantly, our approach does not introduce any extra parameters and does not incur any overhead during training.

#### 3.3.1 Fine-tuning with Selective Layers

We target a plug-and-play efficient fine-tuning, with minimal extra computation costs and without introducing any extra layers. Our strategy is simple and principled: during fine-tuning process, we selectively calculate gradients and



**Figure 4.** We partially fine-tune layers of our pretrained backbone by analyzing heavy-tail spectrum of pretrained weights. We first fine-tune the whole model for  $10^4$  iterations. Then, we compute  $\alpha$  values for all layers of the model. During fine-tuning, we freeze layers with small  $\alpha$  values (indicating good pretrained quality) and only fine-tune layers with large  $\alpha$  (bad pretrained quality).

update layers that are still under-trained (weights of worse pretrained quality), while freeze and skip calculating gradients of layers that are already well-trained (weights of good pretrained quality), thus saving computation costs during fine-tuning. We will defer how to determine the quality of pretrained layers in Sec. 3.3.2. We depict our selective layer-wise fine-tuning in Fig. 4.

#### 3.3.2 Layer Selections via Heavy-tail Analysis

We assess the quality of pretrained weights and determine under-trained/well-trained layers by analyzing their heavy-tail behaviors in the weights spectrums.

The fundamental concept of heavy-tail theory [31] is that in the empirical spectral density (ESD) of weight matrices, heavy-tail structures emerge as these matrices become more correlated and effectively capture diverse correlations in data during the optimization process [30–33, 35]. A key practical implication of this theory is the ability to forecast model performance by determining the power-law coefficient from the ESDs, which only requires the weights of the model. Studies, such as Yang et al. [46], have shown that lower coefficients typically indicate greater test accuracy.

Imagine a neural network composed of  $L$  layers, each associated with weight matrices  $\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_L$ . For every weight matrix  $\mathbf{W}_i$ , which has a dimension of  $N \times M$ , assuming  $N \geq M$ , the correlation matrix is defined as  $\Sigma_i = \mathbf{W}_i^T \mathbf{W}_i$ . The eigenvalues of  $\Sigma_i$  are represented as  $\{\lambda_j\}_{j=1}^M$ , where  $\lambda_j$  equals the square of the singular values of  $\mathbf{W}_i$ , denoted as  $\{\sigma_j\}_{j=1}^M$ . The largest eigenvalue of the correlation matrix  $\Sigma_i$  is referred to as  $\lambda_{i,\max}$ . The term ESD (empirical spectral density) for the weight matrix  $\mathbf{W}_i$  describes the empirical distribution of the eigenvalues of  $\Sigma_i$ , often visualized through a histogram. The density function that models the ESD, denoted as  $p(\lambda)$ , is considered within the range  $(\lambda_{\min}, \lambda_{\max})$ . For a power law,  $p$  satisfies

$$p(\lambda) \propto \lambda^{-\alpha}, \quad \lambda_{\min} < \lambda < \lambda_{\max}. \quad (2)$$

**Table 1. Model efficiency via subnetwork transfer.** By discovering semantic-agnostic sparse masks, we can directly transfer the subnetwork to different OVS frameworks, significantly reducing their model sizes and computation costs, while preserving their OVS performance after fine-tuning. *Random* means transferring random subnetworks.

Model	COCO	Cityscapes	ADE20K-150	ADE20K-847	PAS-20	PC-59	PC-459	Params.	FLOPs
Han et al. [17]	46.0	33.9	16.6	2.5	71.2	39.0	7.1	44.1M	268.2G
<i>Random</i>	37.4	28.7	13.2	2.2	60.1	33.2	5.8	22.9M	173.3G
Ours	42.5	31.7	15.8	2.6	64.6	35.1	6.4	22.9M <span style="color:red">48.1%↓</span>	173.3G <span style="color:red">35.4%↓</span>
DeeplabV3[2]	26.3	20.3	8.8	-	44.1	23.9	4.1	40.3M	241.3G
<i>Random</i>	17.9	16.3	6.4	-	30.2	16.5	2.7	19.1M	146.8G
Ours	34.8	24.3	10.8	-	55.2	28.9	5.2	19.1M <span style="color:red">52.6%↓</span>	146.8G <span style="color:red">39.2%↓</span>
FC-CLIP [47]	58.7	53.2	23.3	7.1	89.5	50.5	12.9	39.0M	200.1G
<i>Random</i>	52.8	50.0	17.2	3.2	85.5	44.8	8.7	17.8M	105.6G
Ours	56.8	52.1	19.1	4.2	87.6	47.4	9.9	17.8M <span style="color:red">54.4%↓</span>	105.6G <span style="color:red">47.2%↓</span>

$\alpha$  is essentially the slope of the tail of the ESD of the weights on a log-log scale.

A **smaller**  $\alpha$  indicates a slow decay of large eigenvalues in the ESD, which tends to lead to a **better** quality of pretrained weights [33]. Therefore, during fine-tuning, we only update layers with large  $\alpha$  values (i.e. weights that are not well-trained), and will freeze layers with small  $\alpha$  values (i.e. pretrained weights of good quality). We update 50% layers of top (larger)  $\alpha$  values, and freeze the other 50% layers of smaller  $\alpha$  values, which helps prevent overfitting.

## 4. Experiments

Sufficient experiments are conducted to verify the effectiveness of the model. We first separately study the contribution from our transferable subnetworks (in Sec. 4.2) and our principled layer-selective fine-tuning (in Sec. 4.3). We then show our final results (in Sec. 4.4) to demonstrate our superior balance between OVS accuracy and efficiency.

### 4.1. Implementation Details

**Datasets.** We conduct experiments in the following popular OVS datasets. **COCO-Panoptic** [29] is a large-scale semantic segmentation dataset with 133 classes. **ADE20K** [51] is a large-scale dataset for semantic segmentation, it contains more than 20k training images and 2k validation images. This dataset is divided into two splits. ADE20K-150 comprises 150 semantic classes, while ADE20K-857 encompasses 857 classes. **Cityscapes** [8] is a dataset for urban scene understanding, its validation set contains 500 high-resolution images. **PASCAL** datasets [12, 36] contains two splits of datasets. Pascal VOC 2012 contains 1,449 validation images from 20 classes. Pascal Context is an extensive of Pascal VOC 2010, and it has 5005 validation images. We utilize both the commonly used PC-59 version and the more challenging PC-459 version.

We evaluate our OpenTrans in a cross dataset setting following previous works [7, 17, 45, 47]. Our subnetwork is trained on COCO and evaluated on the others.

**Training Strategy.** Our implementation is inspired by the

work of Han et al. [17]. Throughout the model training process, we consistently utilize four NVIDIA RTX 4090 Ti GPUs. The input image is resized to  $512 \times 512$ . During the IMP stage outlined in Algorithm 1, we set the learning rate to 0.0001 without implementing any specific step schedule. The parameters  $s$  and  $p$  in the algorithm indicate a sparsity level and pruning ratio of 10%, respectively. When transferring the subnetwork to different architectures, we perform fine-tuning for several iterations. The experimental settings are adjusted accordingly, depending on the specific architectures being used. For a thorough understanding of the experimental details, please refer to the supplementary material.

**Parameters and FLOPs.** For precise and equitable computation, we employed a uniform calculation methodology to assess parameters and FLOPs across all tasks. Our evaluation encompasses all model parameters during testing. FLOPs were consistently computed using the initial graph from the COCO validation set with dimensions  $800 \times 1216$ .

### 4.2. Model Efficiency via Subnetwork Transfer

In this section, we employ three exemplary methodologies [2, 17, 47] to evaluate the efficiency of the transferable subnetwork transfer strategy. We compare these three representative methods with original approaches and random subnetwork transfer strategies.

**Comparisons with original OVS approaches.** We prune once but contribute to the efficiency of peer works via subnetwork transfer. Application of our subnetwork yields notable reductions in Params and FLOPs for models such as Han et al., DeeplabV3, and FC-CLIP, as delineated in Tab. 1. This means that all our three models share the same subnetwork, which validates the transferability of our method. Specifically, for Han et al., we achieve a reduction in Params from 44.1M to 22.9M, and FLOPs from 268.2G to 173.3G. Despite being nearly half the size, the mIoU only saw a marginal 2.2% reduction in Cityscapes and 0.8% in ADE20K-150. In the case of DeeplabV3, we successfully halved the model size while surpassing its original performance. The improvements amounted to a 4% increase in mIoU for Cityscapes

**Table 2. Efficient Fine-tuning.** By analyzing the heavy-tail spectrum in pretrained weights, our layer-selective strategy can reduce training costs in OVS frameworks in a principled manner. Here we choose Han et al. [17] as our test-bed. “random” indicates randomly freezing 50% layers, “ $\alpha^*$ ” means freezing 50% layers with top (greater)  $\alpha$  values, and “ $\alpha$ ” means freezing 50% layers with bottom (smaller)  $\alpha$  values.

Training method	COCO	Cityscapes	ADE20K-150	ADE20K-847	PAS-20	PC-59	PC-459	Training FLOPs
Han et al.	46.0	33.9	16.6	2.5	71.2	39.0	7.1	181.4 P
Han et al.(random)	44.5	33.5	16.4	2.5	70.1	38.5	7.2	164.5 P 9.3%↓
Han et al.+ $\alpha^*$	45.3	33.6	16.7	2.7	73.2	39.2	7.3	172.3 P 5.0%↓
Han et al.+ $\alpha$	<b>47.2</b>	<b>34.0</b>	<b>17.3</b>	<b>2.9</b>	<b>74.0</b>	<b>39.9</b>	<b>7.7</b>	<b>159.6 P 12.0%↓</b>

**Table 3. Balance: OpenSeg vs. Efficiency.** We compare OVS performance on popular image segmentation datasets. By adopting our methods, we can significantly boost the efficiency of popular OVS frameworks while preserving their OVS performance.

Model	Backbone	Training Set	COCO	PAS-20	PC-59	ADE20K-150	Params.	FLOPs
ZS3Net [1]	R-101	PASCAL-15	-	38.3	19.4	-	-	-
SPNet [43]	R-101	PASCAL-15	-	18.3	-	24.3	-	-
LSeg [25]	R-101	PASCAL-15	-	47.4	-	-	-	-
ZegFormer [10]	R-50	COCO Stuff	-	80.7	42.8	16.4	63.9M	1128.7G
Cat-Seg [7]	R-101	COCO Stuff	-	93.7	57.5	27.2	60.8M	3847.8G
OVSeg [28]	R-101	COCO Stuff	-	89.2	53.3	24.0	61.1M	1116.2G
LSeg+ [16]	R-101	COCO Panoptic	-	59.0	36.0	13.0	-	-
OpenSeg [16]	R-101	COCO Panoptic	36.9	-	36.9	15.3	-	-
Simbaseline [44]	R-50	COCO Panoptic	39.5	-	40.1	14.4	64.6M	1060.6G
DeeplabV3 [2]	R-50	COCO Panoptic	26.3	44.1	23.9	8.8	40.3M	241.3G
Han et al. [17]	R-50	COCO Panoptic	46.0	71.2	39.0	16.6	44.1M	268.2G
FC-CLIP [47]	R-50	COCO Panoptic	58.7	89.5	50.5	23.3	39.0M	200.1G
Ours (Han et al.)	R-50	COCO Panoptic	42.3	64.8	35.7	15.3	22.9M	173.3G
Ours (DeeplabV3)	R-50	COCO Panoptic	34.2	54.6	28.8	10.7	19.1M	146.8G
Ours (FC-CLIP)	R-50	COCO Panoptic	56.8	87.3	47.4	18.8	17.8M	105.6G

and a 2% boost for ADE20K-150. Compared with the original FC-CLIP model, we are able to compress it to 17.95M Params and 105.6G FLOPs, surpassing even the compactness achieved by Han et al. Although FC-CLIP emphasizes the significance of freezing the CLIP backbone, it is noteworthy that the subnetwork we utilized exerted a minimal impact on the model’s performance in the task of OVS.

**Comparisons with random subnetwork transfer.** The results for the *Random*, characterized by equivalent sparsity to our transferable subnetwork, are omitted in Tab. 1. It is evident that ours significantly outperforms *Random* across all datasets, This highlights the effectiveness of our proposed method in preserving the transferability of performance from previous OVS approaches.

In a word, the proposed transferable subnetwork significantly decreases computational redundancy with minimal impacts on performance.

### 4.3. Efficient Fine-tuning with Layer Selections

Based on the computation method outlined in Sec. 3.3, we can compute the  $\alpha$  value for each layer of the backbone. The model training is initiated for  $10^4$  iterations before performing the  $\alpha$  derivation and freezing specific layers.

In Tab. 2, we adopt diverse approaches to freeze half of

the layers within the backbone. Through the identification of layers with small  $\alpha$  values as high-quality layers, excluding them from training results in enhanced performance not only on the training dataset but also significant improvements compared to the original model across all evaluated OVS datasets. Notably, we achieve 0.7%, 2.8%, and 0.9% improvement on ADE20K-150, PAS-20 and PC-59 in terms of mIoU. In comparison to randomly freezing half of the layers without training involvement, our proposed method consistently surpasses the performance of the original model on all datasets. The freezing method allows for the conservation of training FLOPs, leading to more efficient resource utilization, as detailed in Tab. 4. It becomes apparent that through fine-tuning the identified poorly performing layers, the network can attain improved training results while utilizing fewer parameters.

### 4.4. Stronger Balance: OpenSeg vs. Efficiency

To assess the open-vocabulary performance of our proposed method, we compare its effectiveness against state-of-the-art techniques on various well-known image segmentation datasets [8, 12, 29, 36, 51]. The results presented in Tab. 3 illustrate the outcomes of employing our two previously proposed methods for enhancing model and training efficiency.

**Table 4.** Reduction of training FLOPs with our methods (ResNet-50). “*pruning*” means our transferable subnetworks in Sec 3.2. “ $\alpha$ ” means selective layer-wise fine-tuning in Sec 3.3.

Model	Training FLOPs
Han et al.	181.4 P
Han et al.+ $\alpha$	159.6 P 12.0%↓
Han et al.+ $\alpha$ + <i>pruning</i>	122.2 P 32.6%↓

These findings showcase that our model not only attains impressive speed but also demonstrates robust OVS capabilities. To elaborate, our method achieves a notable 56.8% mIoU on the training set COCO, encompassing 133 categories. It further attains an 18.8% mIoU on the ADE20K dataset with 150 categories and secures 47.4% and 9.9% mIoU on Pascal Context with 49 categories and 459 categories, respectively.

In comparison to the state-of-the-art method Cat-Seg across multiple datasets, our approach reveals a performance gap of only 6.4% mIoU on the PAS-20 dataset and 8.4% mIoU on the ADE20k-150 dataset. Notably, we achieve a remarkable reduction of over 240% in model parameters, coupled with a nearly 40-fold improvement in computational speed. Han et al. pioneer the elimination of the two-stage process, demonstrating commendable performance and efficiency. In contrast, our approach showcases advantages of 2.2% mIoU on ADE20K-150 and 8.4% mIoU on PC-59. Moreover, we achieve more than a twofold improvement in both parameter efficiency and operational speed.

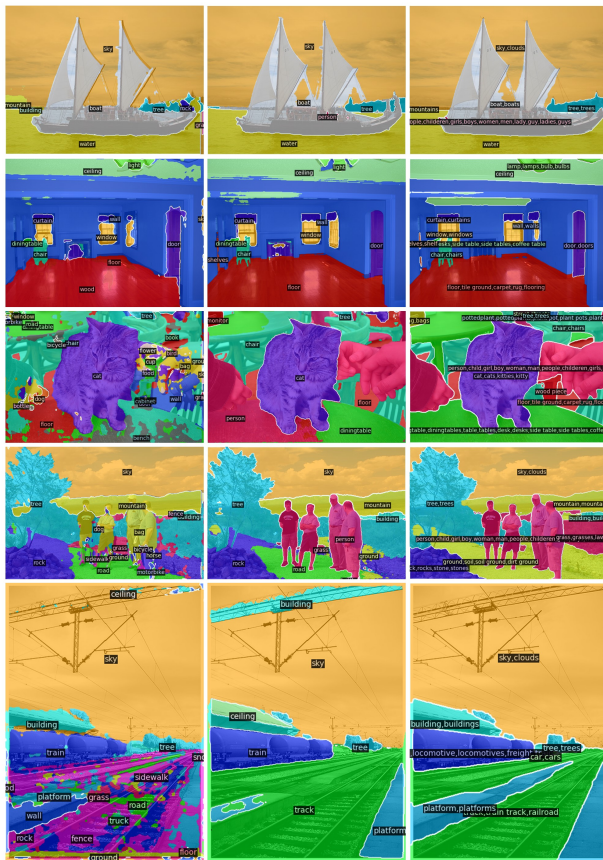
#### 4.5. Training cost

The layer-wise fine-tuning method we proposed, has proven to be effective across OVS works. Moreover, the combination of this layer-wise fine-tuning method with our subnetwork further significantly amplifies training efficiency. In Tab. 4, our focus is exclusively on the training FLOPs of the entire model, given that our subnetwork and layer-wise fine-tuning methods are specifically applied to the backbone component. The incorporation of  $\alpha$  results in a substantial reduction of 12.0% in the training parameters for the model. When coupled with subnetwork, the training parameters are further diminished to only 32.6% of their original value. A comprehensive breakdown of the training FLOPs calculation will be provided in the appendix section.

#### 4.6. Qualitative Results

Fig. 5 displays several visualization results obtained from our method. There exists a notable disparity in the quality of masks obtained by various baselines. In intricate scenes, the conventional convolution segmentation head exhibits subpar performance. Moreover, the unique structural design of FC-CLIP facilitates a more precise assignment of unseen category labels to the generated masks, enhancing its capacity to predict both foreground and background categories.

Ours (DeeplabV3)    Ours (Han et al.)    Ours (FC-CLIP)



**Figure 5.** Visualizations of examples on the PC-59 validation set by our models (trained on COCO panoptic training set, zero-shot evaluated on the PC-59 validation set).

## 5. Conclusion

In this paper, we introduce a transferable subnetwork approach that markedly reduces the size and computational demands of popular OVS models. Additionally, we propose a principled layer-selective fine-tuning method, enabling efficient fine-tuning of OVS models by only updating under-trained layers. Through extensive experiments on various OVS datasets, we illustrate that our model strikes a commendable balance between performance and efficiency. We hope our method can serve as a solid baseline and help advance the future research of OVS. Some current limitations and potential future works: 1) adaptation of our method on larger convolutional backbones (e.g. ConvNeXt-Large) and ViT-based backbones; 2) more fine-grained selection of weight elements or channels for fine-tuning, going beyond the layer level; 3) extension to other open-vocabulary tasks, such as open-set object detection.

**Acknowledgement.** This research was funded by the Fundamental Research Funds for the Central Universities (2023JBZD003) and the National NSF of China (No.U23A20314).



## References

- [1] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 3, 7
- [2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1, 4, 6, 7
- [3] Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. The lottery ticket hypothesis for pre-trained bert networks. *Advances in neural information processing systems*, 33:15834–15846, 2020. 2, 4
- [4] Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Michael Carbin, and Zhangyang Wang. The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16306–16316, 2021. 2, 4
- [5] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 3
- [6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 3, 4
- [7] Seokju Cho, Heeseong Shin, Sunghwan Hong, Seungjun An, Seungjun Lee, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation, 2023. 3, 6, 7
- [8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6, 7
- [9] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13(1):795–828, 2012. 4
- [10] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 1, 3, 7
- [11] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022. 3
- [12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 6, 7
- [13] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018. 2, 4
- [14] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pages 3259–3269. PMLR, 2020. 4
- [15] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021. 1
- [16] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 7
- [17] Kunyang Han, Yong Liu, Jun Hao Liew, Henghui Ding, Jiajun Liu, Yitong Wang, Yansong Tang, Yujia Yang, Jiashi Feng, Yao Zhao, et al. Global knowledge calibration for fast open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 797–807, 2023. 1, 3, 5, 6, 7
- [18] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015. 4
- [19] Shuting He, Henghui Ding, and Wei Jiang. Semantic-promoted debiasing and background disambiguation for zero-shot instance segmentation. In *CVPR*, 2023. 1
- [20] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 2, 3
- [21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 3
- [22] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019. 1
- [23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2
- [24] Siyu Jiao, Yunchao Wei, Yaowei Wang, Yao Zhao, and Humphrey Shi. Learning mask-aware clip representations for zero-shot segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [25] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 1, 7
- [26] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11336–11344, 2020. 2

- [27] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. **2**
- [28] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. **1, 3, 7**
- [29] 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 *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. **6, 7**
- [30] Charles H Martin and Michael W Mahoney. Traditional and heavy tailed self regularization in neural network models. In *International Conference on Machine Learning*, pages 4284–4293, 2019. **5**
- [31] Charles H Martin and Michael W Mahoney. Heavy-tailed universality predicts trends in test accuracies for very large pre-trained deep neural networks. In *SIAM International Conference on Data Mining*, pages 505–513. SIAM, 2020. **5**
- [32] Charles H Martin and Michael W Mahoney. Post-mortem on a deep learning contest: a Simpson’s paradox and the complementary roles of scale metrics versus shape metrics. Technical Report Preprint: arXiv:2106.00734, 2021.
- [33] Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021. **5, 6**
- [34] Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021. **4**
- [35] Charles H Martin, Tongsu Serena Peng, and Michael W Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(1):1–13, 2021. **5**
- [36] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. **1, 6, 7**
- [37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. **1, 2**
- [38] Alex Renda, Jonathan Frankle, and Michael Carbin. Comparing rewinding and fine-tuning in neural network pruning. *arXiv preprint arXiv:2003.02389*, 2020. **4**
- [39] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. **2, 3**
- [40] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. *Advances in Neural Information Processing Systems*, 35: 12991–13005, 2022. **3**
- [41] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. VI-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022. **3**
- [42] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7268–7277, 2018. **1**
- [43] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019. **1, 3, 7**
- [44] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. **1, 3, 7**
- [45] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. **3, 6**
- [46] Yaoqing Yang, Ryan Theisen, Liam Hodgkinson, Joseph E Gonzalez, Kannan Ramchandran, Charles H Martin, and Michael W Mahoney. Evaluating natural language processing models with generalization metrics that do not need access to any training or testing data. *arXiv preprint arXiv:2202.02842*, 2022. **5**
- [47] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36, 2024. **1, 3, 4, 5, 6, 7**
- [48] Hui Zhang and Henghui Ding. Prototypical matching and open set rejection for zero-shot semantic segmentation. In *ICCV*, pages 6974–6983, 2021. **1**
- [49] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. **3**
- [50] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. **1**
- [51] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k

dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. [6](#), [7](#)

- [52] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. [3](#)