[Paper Review] Surgical Fine-Tuning Improves Adaptation to Distribution Shifts

Stanford University, *ICLR 2023*

Jaechul (Harry) Roh

Table of Content

SURGICAL FINE-TUNING IMPROVES ADAPTATION TO DISTRIBUTION SHIFTS

Yoonho Lee* Annie S. Chen* Fahim Tajwar Ananya Kumar
Huaxiu Yao Percy Liang Chelsea Finn

Stanford University

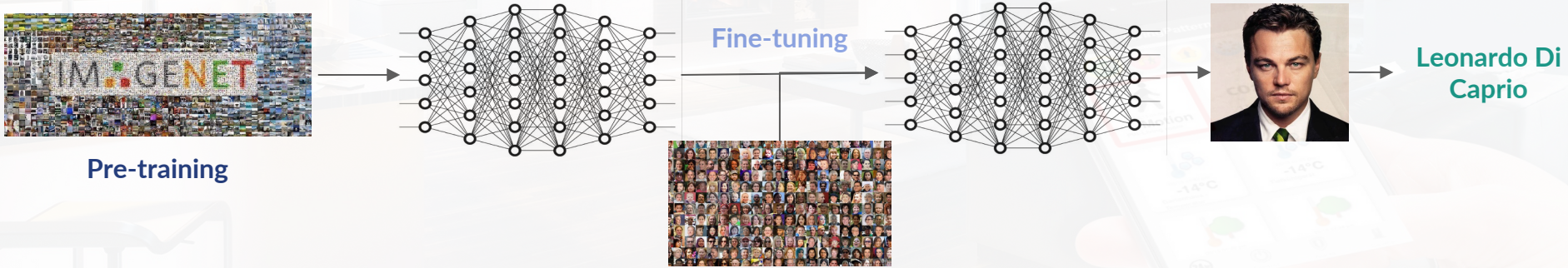
Section 1: Introduction

Problems Existing in Distribution Shift

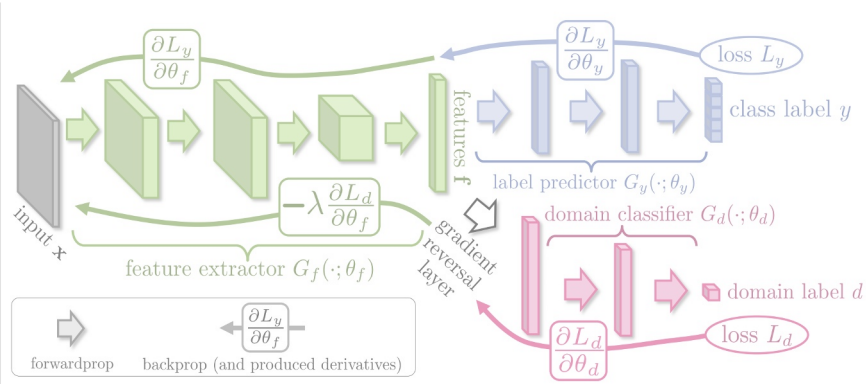
- **Distribution Shift**
  - Change in the distribution of input data between the training phase and testing phase
- **Potential Problems:**
  - Overfitting
  - Model Collapse
  - Performance Degradation
  - Inconsistent performance

Transfer Learning

- **Model** trained on one task
- **Used** as starting point for another related task
- **Advantage:** when there is limited amount of data available for target task



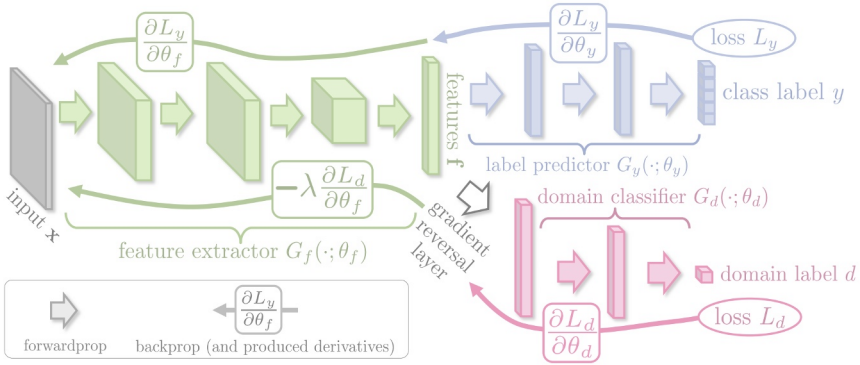
Unsupervised Adaptation (1)



- **Unsupervised Adaptation:**
  - Type of transfer learning approach
  - The target dataset does not have labeled data
  - Adaptation of the model is done without using any labeled data from the target domain
- Model is fine-tuned ***to perform well on the target distribution***
  - Without the use of **supervised information**

Unsupervised Adaptation (2)

- Unsupervised adaptation for fine-tuning (in the presence of distribution shift)
- **Aim of the paper:**
  - Surgical fine-tuning outperforms traditional unsupervised adaptation methods
  - Surgically removing parts of the model that are not useful
  - Fine-tuning only the ***useful parts***
- **Result:**
  - Improved performance on target dataset
  - Even when labeled data is not available



Brief Overview of the Paper (1)

- **General Summary:**
  - Explores novel approach to address the challenge of distribution shift in deep learning
  - Proposes **Surgical Fine-Tuning**
    - Selectively fine-tunes only a portion of the network in response to distribution shift
    - Instead of fine-tuning the entire network
  - **Relative Gradient Norm (RGN)**
    - Identify critical parts of the network
    - That need to be fine-tuned in response to distribution shifts
    - More explained later…

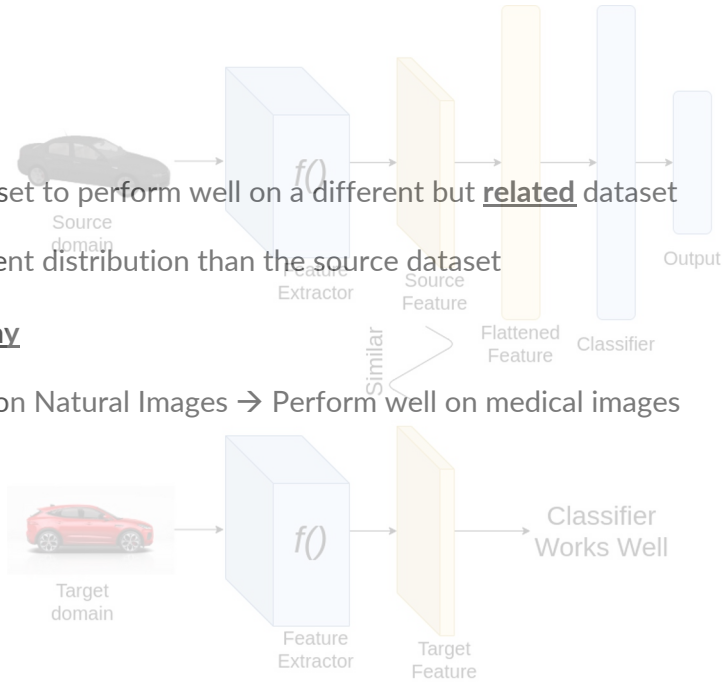
Brief Overview of the Paper (2)

- **Evaluation:**
  - Evaluate surgical fine-tuning on a range of benchmark datasets
  - Compare its performance against the traditional fine-tuning method
- **Results:**
  - Compare its performance against the traditional fine-tuning method
  - **Surgical fine-tuning** improves performance compared to traditional fine-tuning

Section 2: Background Information

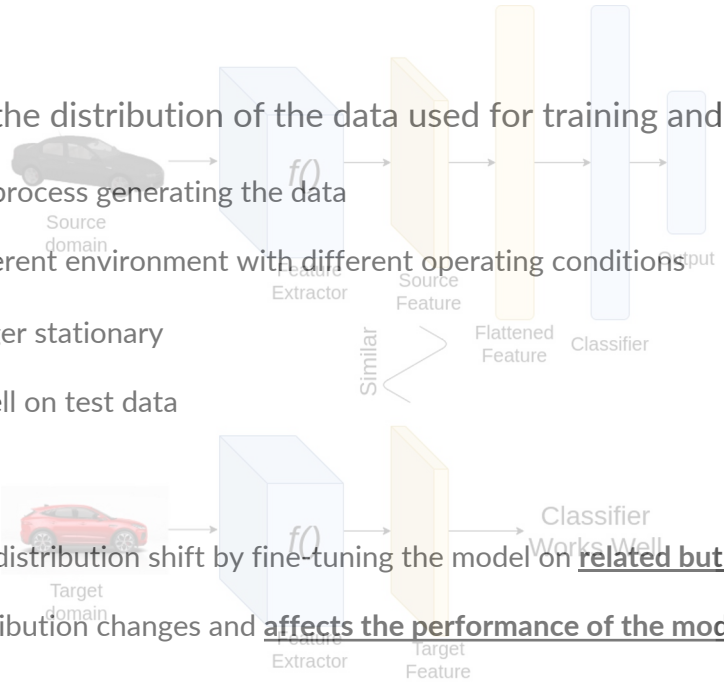
Domain Adaptation vs. Distribution Shift (1)

- **Domain Adaptation:**
  - Model trained on one dataset to perform well on a different but **related** dataset
  - Target dataset has a different distribution than the source dataset
    - **Related in some way**
    - E.g. Model trained on Natural Images → Perform well on medical images



Domain Adaptation vs. Distribution Shift (2)

- **Distribution Shift:** change in the distribution of the data used for training and testing a model
  - Changes in the underlying process generating the data
  - Model is deployed in a different environment with different operating conditions
  - **Data distribution** is no longer stationary
  - **Model** may not perform well on test data
- **Summary**:
  - ***Domain Adaptation***: tackle distribution shift by fine-tuning the model on **related but different datasets**
  - ***Distribution Shift***: data distribution changes and **affects the performance of the model**

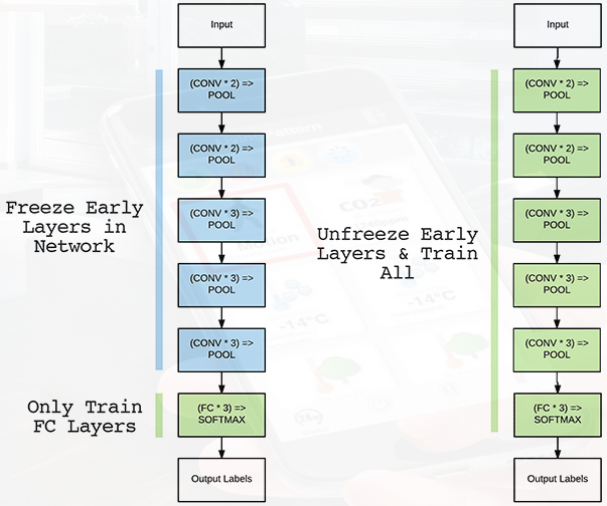


Prior Works (1)

- **Domain adaptation:**
  - Aims to adapt a model to a new target domain
  - By leveraging the knowledge learned from the source domain
- **Domain generalization:**
  - Aims to learn a model that is robust to unseen domains
- **Domain-invariant feature learning:**
  - Aims to learn a feature representation
  - Invariant to the distribution shifts across different domain

Prior Works (2)

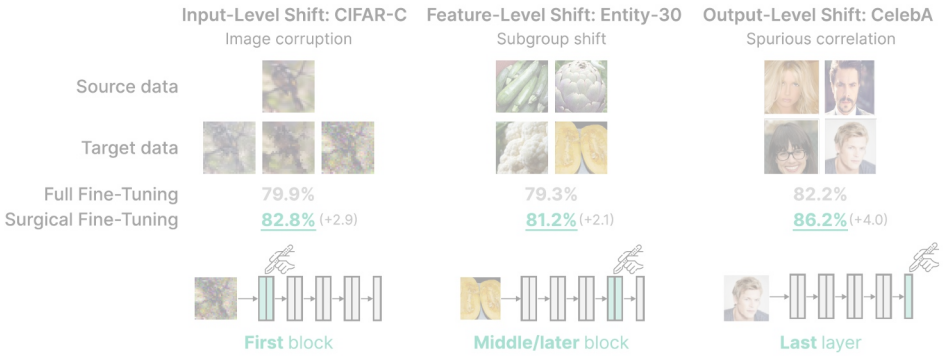
- **From paper:**
  - Using different architectures or pre-training the model on a wider range of data to make it more robust
  - **Fine-tuning**: pre-trained model is adapted to a new task or dataset
  - **Challenges of fine-tuning**:
    - Overfitting to the new task
    - Difficulty in selecting an appropriate strategy
    - Computational cost



Section 3: Methodology

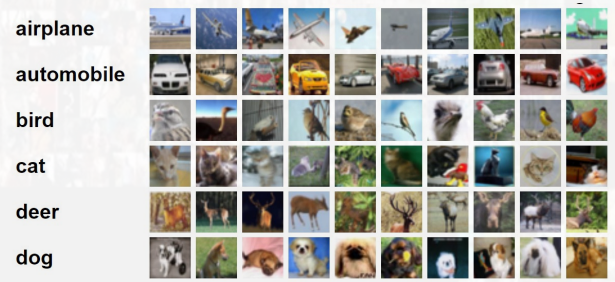
Definition: Surgical Fine-Tuning

- Problems of full fine-tuning:
  - Cause network to forget the knowledge learned from the source domain
  - Poor performance
- Surgical fine-tuning:
  - Fine-tuning only the **parts** of the network
  - Keeping other parts fixed
  - Achieved by ***Auto-RGN***
    - *Measures the gradient norm of each layer*
    - *Determine which layers should be fine-tuned*



Datasets

- CIFAR-10-C
  - CIFAR-10 Task: classify images into 10 classes
  - CIFAR-10-C: target distribution contains 14 types of corrupted images
- ImageNet-C
  - ImageNet Task: classify images into 1000 classes
  - ImageNet-C: target distribution contains 15 types of corrupted images
- Living 17 and Entity-30
  - Classify to 17 animal categories or one of 30 entities
  - Tune on 850 images from the target distribution
- Waterbirds
  - Classify into either “waterbird” (water background) or “landbird” (land background)
  - Tune on 400 images from the target distribution
- CelebA
  - Classify into “blond” or “not blond”
  - Tune on 400 images from the target distribution



Experimental Procedure (Datasets) (1)

- Datasets
  - Input-level shift
    - CIFAR-C, ImageNet-C
  - Feature-level shift
    - Living-17, Entity-30
  - Output-level shift (Challenging)
    - CIFAR-FLIP, Waterbirds, CelebA

“Model can all result in degraded performance, as the model may not be able to generalize well to the new data"

“Authors propose the surgical fine-tuning approach to alleviate such issues”

Experimental Procedure (Datasets) (2)

- **Pre-training & Model Architectures**
  - ResNet-50 pre-trained on ImageNet
  - ResNet-26 for CIFAR-C and CIFAR-Flip
    - Adam optimizer
    - Tune over 3 learning rates
- **CIFAR-10-C & CIFAR-FLIP**:
  - Fine-tune on the labeled data for 15 total epochs
  - 3 learning rates: 1e-3, 1e-4, 1e-5
  - Last-layer fine-tuning: 1e-1, 1e-2, 1e-3
  - Weight decay: 0.0001

Experimental Procedure (Datasets) (3)

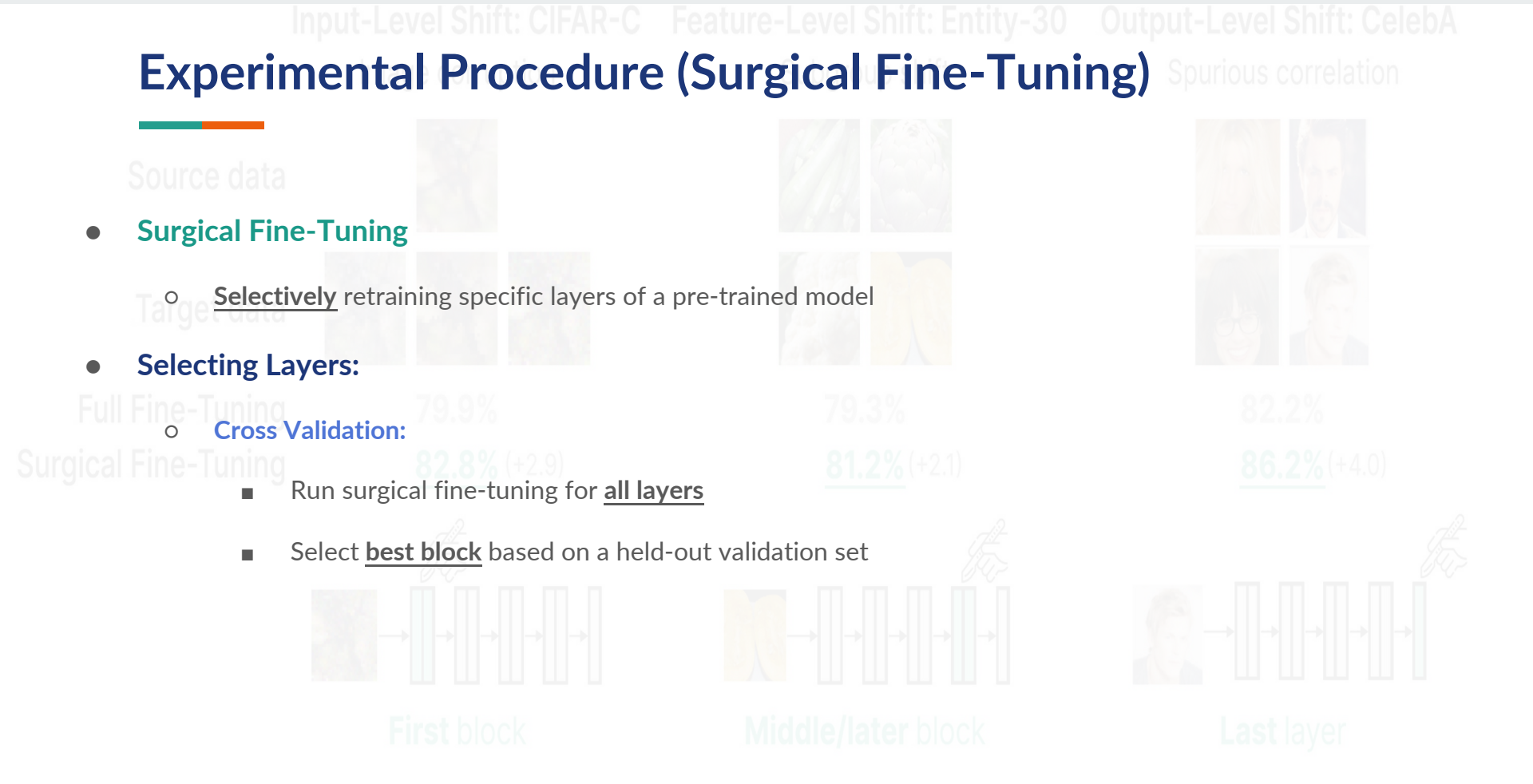
- **ImageNet-C**
  - Fine-tune on the labeled target data for 10 total epochs
  - 3 learning rates: [1e-3, 1e-4, 1e-5]
  - Weight decay: 0.0001
- **Living-17 & Entity-30**
  - Train source data for 5 epochs
  - Fine-tune for 15 epochs
  - 3 learning rates: [0.0005, 0.0001, 0.00001]
  - No Weight Decay

Experimental Procedure (Datasets) (4)

- **Waterbirds**
  - ResNet-50 pretrained on ImageNet (train on source distribution for 300 epochs)
  - Fine-tune on the labeled target data for 100 total epochs
  - 3 learning rates: [0.005, 0.001, 0.0005]
  - Weight Decay: 0.0001
- **CelebA**
  - ResNet-50 pretrained on ImageNet (train on source distribution for 50 epochs)
  - Fine-tune for 50 epochs
  - 3 learning rates: [0.001, 0.0005, 0.0001]
  - Weight decay: 0.0001

Experimental Procedure (Surgical Fine-Tuning)

- **Surgical Fine-Tuning**
  - **Selectively** retraining specific layers of a pre-trained model
- **Selecting Layers:**
  - **Cross Validation:**
    - Run surgical fine-tuning for **all layers**
    - Select **best block** based on a held-out validation set



Experimental Procedure (Selecting layers) (1)

- **Selecting Layers:**
  - **Relative Gradient Norm (Auto-RGN):**
    - Measured by the magnitude of change in the gradient during fine-tuning
    - *Relative* magnitude of change in the gradient during fine-tuning

$$\mathrm{RGN}(\theta_i) = \frac{(g_i)}{||\theta_i||}$$

Gradient norm in the fine-tuning stage

Parameter Norm

Experimental Procedure (Selecting layers) (2)

- **Selecting Layers:**
  - **Signal-to-Noise Ratio (Auto-SNR):**
    - Objective of surgical fine-tuning:
      - Increase auto-SNR in the layer of interest
      - Selectively emphasize the activations in the layer of interest

$$\mathrm{SNR}(g_i) = \frac{\mathrm{Avg}(g_i)^2}{\mathrm{Var}(g_i)}$$

Average of gradient norm

Variance of gradient norm

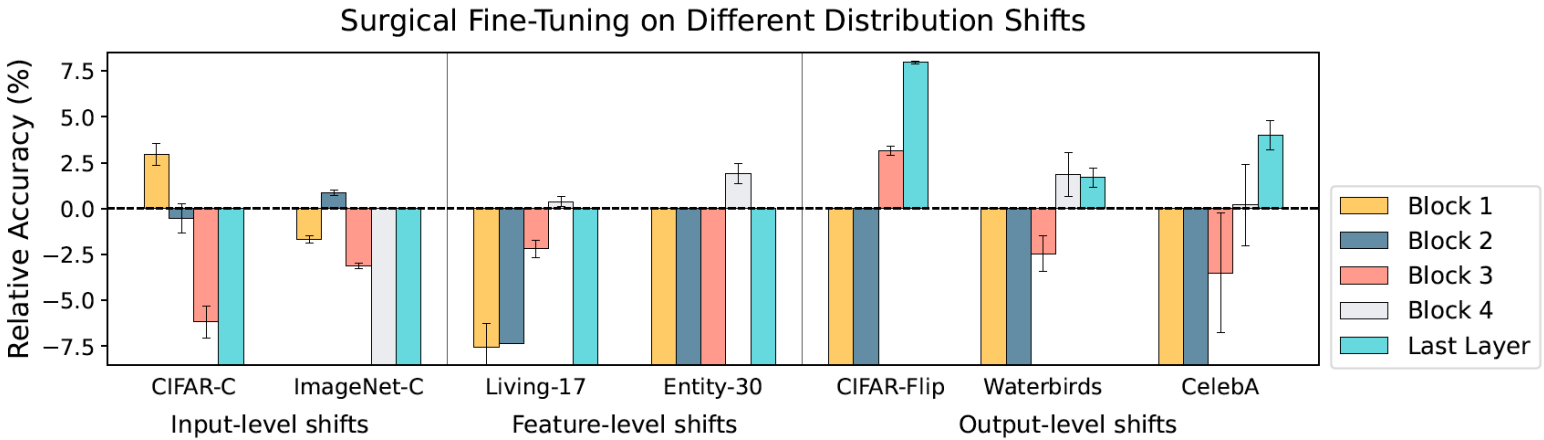
Section 4: Results & Analysis

Results Based on Criteria

Method	Input-level Shifts		Feature-level Shifts		Output-level Shifts			Avg Rank
	CIFAR-C	IN-C	Living-17	Entity-30	CIFAR-F	Waterbirds	CelebA	
No Adaptation	52.6 (0)	18.2 (0)	80.7 (1.8)	58.6 (1.1)	0 (0)	31.7 (0.3)	27.8 (1.9)	-
Cross-Val	82.8 (0.6)	51.6 (0.1)	93.2 (0.3)	81.2 (0.6)	93.8 (0.1)	89.9 (1.2)	86.2 (0.8)	-
Full Fine-Tuning (All)	79.9 (0.7)	50.7 (0.1)	92.8 (0.7)	79.3 (0.6)	85.9 (0.4)	**88.0 (1.2)**	**82.2 (1.3)**	2.29
L_1 Regularize (Xuhong et al., 2018)	**81.7 (0.6)**	48.8 (0.3)	93.4 (0.5)	78.4 (0.1)	84.2 (1.2)	87.6 (1.9)	82.6 (1.8)	2.57
Auto-SNR	80.9 (0.7)	49.9 (0.2)	**93.5 (0.2)**	77.3 (0.3)	17.3 (0.7)	86.3 (0.7)	78.5 (1.8)	3.14
Auto-RGN	81.4 (0.6)	**51.2 (0.2)**	**93.5 (0.3)**	**80.6 (1.2)**	**87.7 (2.8)**	**88.0 (0.7)**	**82.2 (2.7)**	**1.29**

- **Highest accuracies** are based on **cross-validation**
- **Satisfactory results** demonstrated using **Auto-RGN**

Relative Accuracy of Surgical Fine-tuning



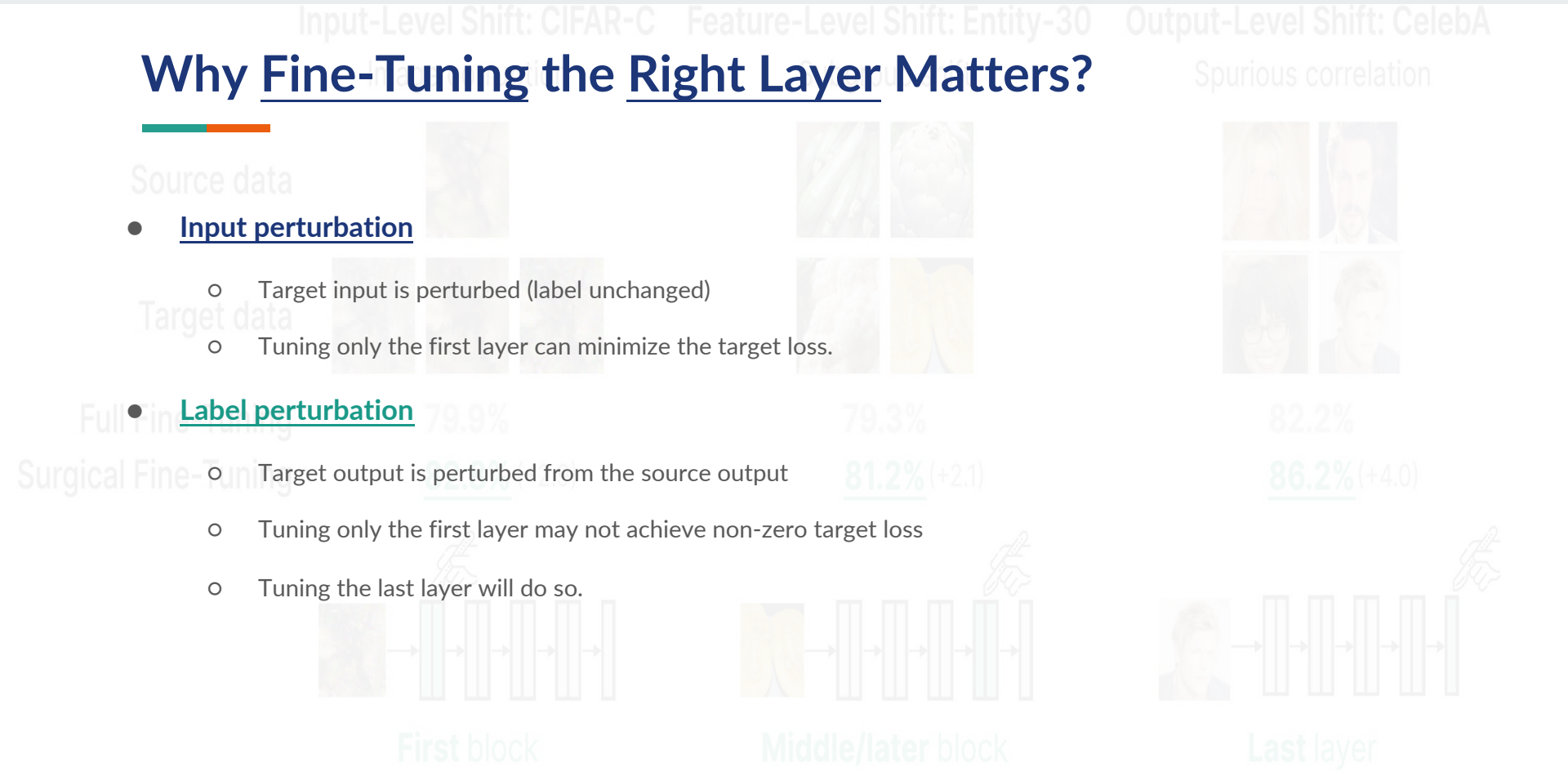
- **Relative Accuracy: (Surgical fine-tuning accuracy) – (Full fine-tuning accuracy)**
- Tuning **earlier** layers performs best for **input-level shifts**
- Tuning **later** layers performs best for **output-level shifts**

Why Surgical Fine-Tuning Works?

- **Surgical fine-tuning vs. Full fine-tuning:**
  - It outperforms full fine-tuning since it addresses the challenge of **catastrophic forgetting**
- **Catastrophic forgetting**
  - Fine-tuned neural net **forget** how to perform well on its original task
- **How surgical fine-tuning solves catastrophic forgetting?**
  - Preserve the network's knowledge of the original task
  - While still allowing it to adapt to the new task

Why Fine-Tuning the Right Layer Matters?

- **Input perturbation**
  - Target input is perturbed (label unchanged)
  - Tuning only the first layer can minimize the target loss.
- **Label perturbation**
  - Target output is perturbed from the source output
  - Tuning only the first layer may not achieve non-zero target loss
  - Tuning the last layer will do so.



Section 5: Evaluation & Summary

Key Findings (1)

- Surgical Fine-Tuning
  - New fine-tuning method to resolve the issue of distribution shift
- **Outperforms** conventional full fine-tuning
  - By selectively fine-tuning a smaller subset of layers in the network
- Demonstrates **effectiveness** of surgical fine-tuning
  - Reduces the relative gradient norm
  - Increases signal-to-noise ratio

Key Findings (2)

- **Robust** to various types of distribution shifts
  - Input-level shift
  - Feature-level shift
  - Output-level shift
- Results demonstrate the **usefulness** of surgical fine-tuning
  - Improves robustness
  - Improves generalization

Personal Comments

- **Well-designed experiments**
  - Effective choice of model architecture (ResNet)
  - Wide range of datasets and experimental types
- **Question towards applicability**
  - What if the model gets bigger?
  - How about large language models (Transformer, BERT, ...)?
- From **OpenReview**:
  - Pros: “paper can be cited as a justification of the experimental design for future researchers”
  - Cons: “Could have proposed more unique perspective"

Interesting Catch

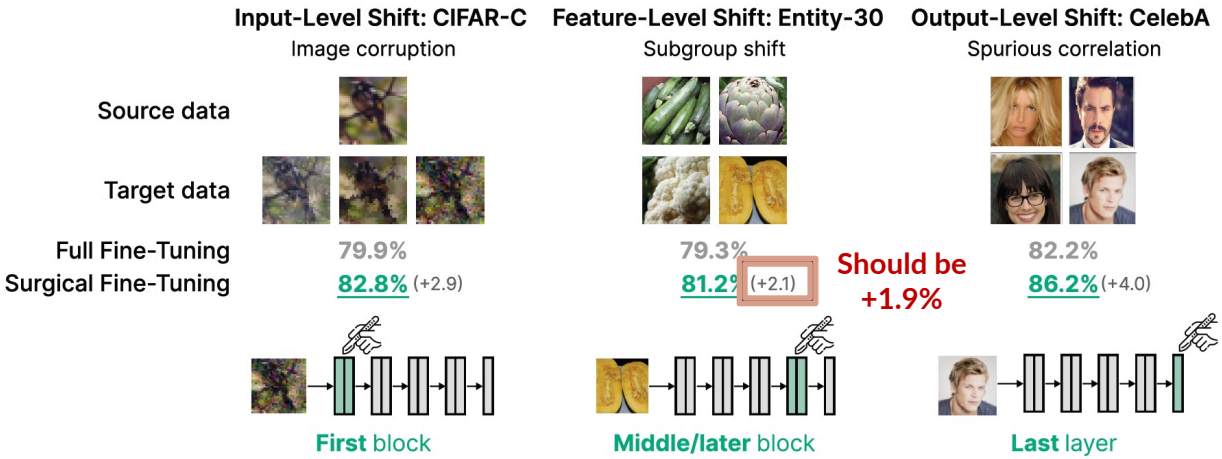


Figure 1: Surgical fine-tuning, where we tune only one block of parameters and freeze the remaining parameters, outperforms full fine-tuning on a range of distribution shifts. Moreover, we find that tuning different blocks performs best for different types of distribution shifts. Fine-tuning the first block works best for input-level shifts such as CIFAR-C (image corruption), later blocks work best for feature-level shifts such as Entity-30 (shift in entity subgroup), and tuning the last layer works best for output-level shifts such as CelebA (spurious correlation between gender and hair color).

Thank You!