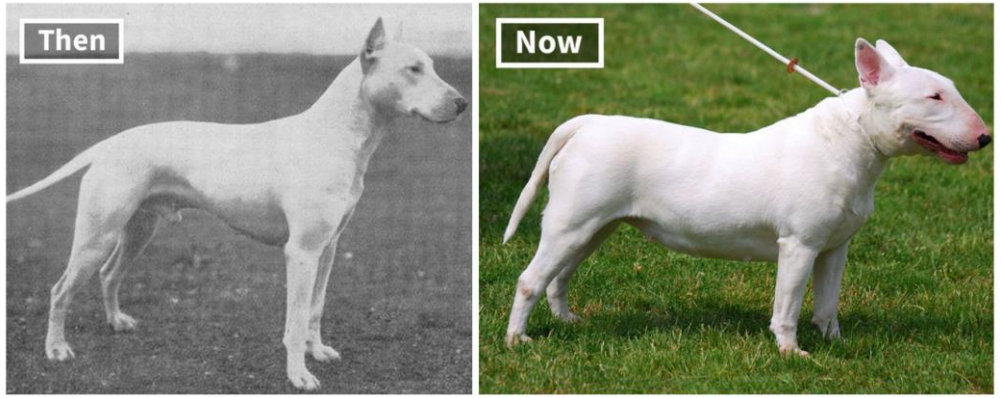
Robust fine-tuning of zero-shot models

Tae-hwi Kim

CVPR, 2022

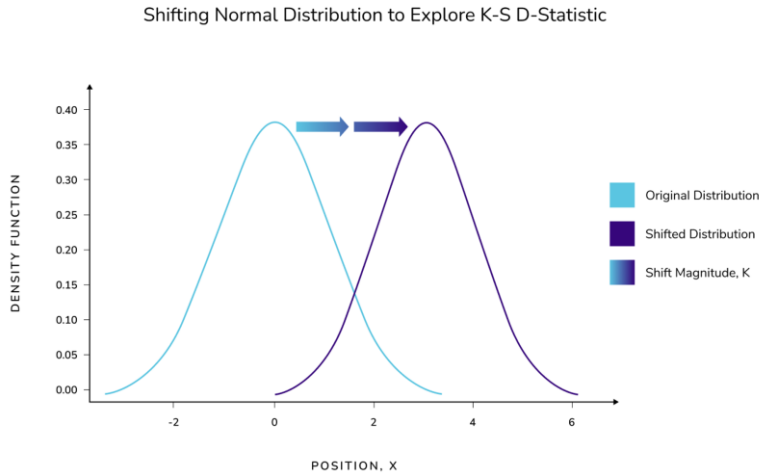
WiSE-FT

- *Introduction*



다양한 데이터 분포(Data Shift)에서도
잘 작동하는 모델 필요
> Robustness가 필요하다.

최근 CLIP, ALGIN과 같은 큰 규모의
사전 학습 모델에서 큰 Robustness를 가짐이 증명됨



특히, CLIP에서는 모델을 Fine-tuning한 것이
아닌 Zero-shot으로 사용한 것이 더 높은
Robustness 성능을 보였다.

WiSE-FT

- *Introduction*

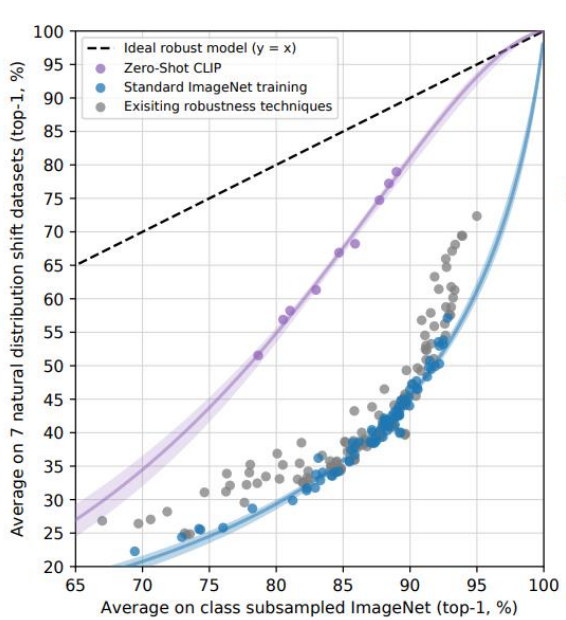


Figure 13. **Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models.** (Left) An ideal robust model

Target Data Distribution에 대해 모델을 Fine-tuning하면 보통 성능이 더 오르는데,
이는 **모델의 Robustness를 희생**하면서 나타나는 결과임.

WiSE-FT

- *Introduction*

Robustness vs. Accuracy

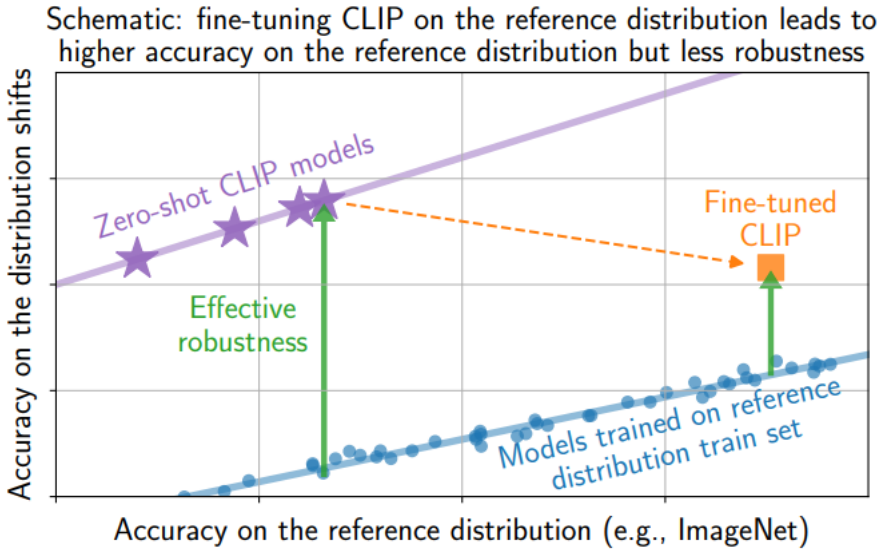
> 모델이 가질 수 있는 두 성능이 Trade-off 관계에 있음을 확인함.
하지만, Fine-tuning도 하면서 Distribution shift에 잘 적응하게 할 수 있지 않을까?

WiSE-FT

- *Methods*

WiSE-FT를 제안,
두 가지 Step을 통해 모델을 Fine-tuning한다.

First step. Target Distribution(D_ref, no shift)
에 대해 zero-shot 모델을 Fine-tuning 한다.
> Strandard fine-tuning on CLIP-like models



$$\arg\min_\theta \left\{ \sum_{(x_i,y_i)\in \mathcal{S}_{\text{ref}}^{\text{tr}}} \ell(f(x_i,\theta),\ y_i) + \lambda R(\theta) \right\}$$

$$f(x,\theta) = g\left(x, \mathbf{V}_{\text{enc}}\right)^\top \mathbf{W}_{\text{classifier}}$$

WiSE-FT

- *Methods*

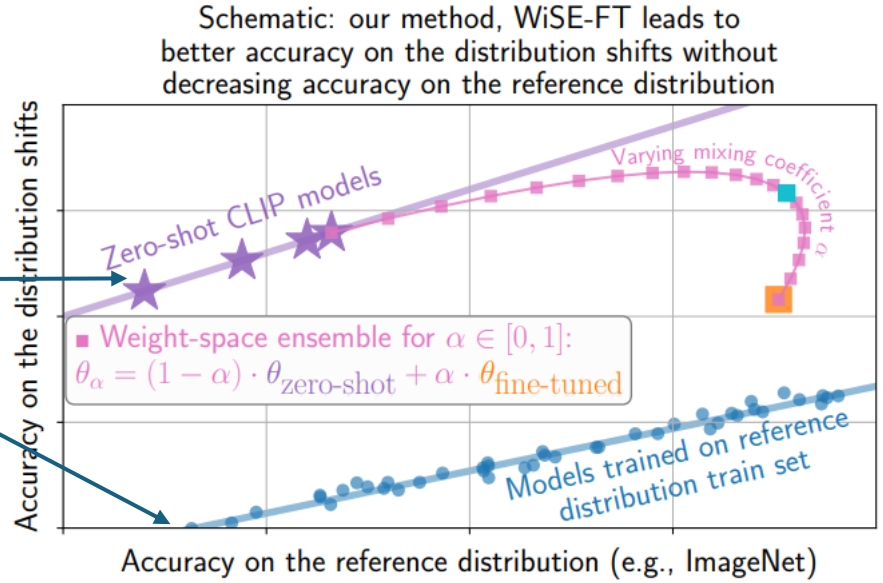
Second step. Weight-space ensembles for fine-tuning (WISE-FT)

> Zero-shot 모델(θ_0)과 Standard fine-tuned 모델(θ_1)의 **Weight에 대한 element-wise interpolation**을 hyperparameter α를 통해 수행한다.

$$\text{wse}(x, \alpha) = f(x, (1-\alpha)\cdot\theta_0 + \alpha\cdot\theta_1)\,,$$

$$f(x,\theta) = g\left(x, \mathbf{V}_{\text{enc}}\right)^\top \mathbf{W}_{\text{classifier}}$$

$$g(x, \mathbf{V}_{\text{enc}})^\top \mathbf{W}_{\text{zero-shot}}$$



WiSE-FT

- *Experimental Setting*

CLIP : ViT-L/14 @ 336px 사용.

사용된 데이터셋 : 악의적이거나 인공적인 조작이 없는 **Natural Distribution Shift** (e.g. lighting, geographic location)

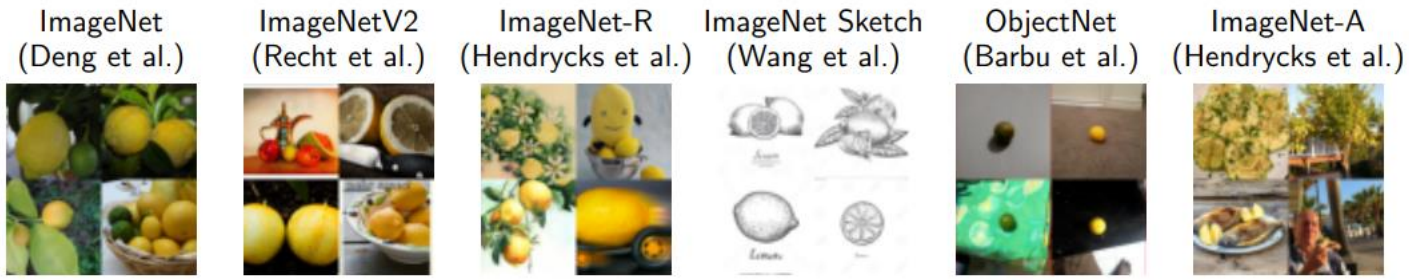
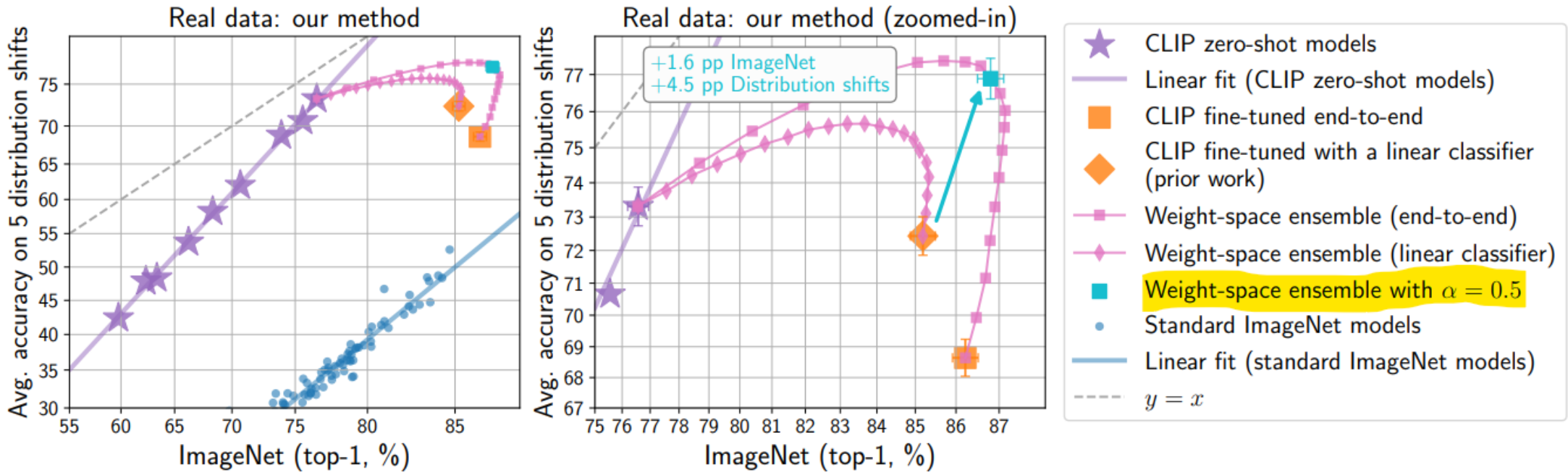


Figure 2. Samples of the class *lemon*, from the reference distribution ImageNet [17] and the derived distribution shifts considered in our main experiments: ImageNet-V2 [81], ImageNet-R [35], ImageNet Sketch [98], ObjectNet [4], and ImageNet-A [38].

WiSE-FT

- *Experimental Results*

5개의 Distribution Shift에 대한 평균 Acc (y), ImageNet에 대한 Fine-tuning Acc (x)



WiSE-FT

- *Experimental Results*

	IN (reference)	Distribution shifts					Avg shifts	Avg ref., shifts
		IN-V2	IN-R	IN-Sketch	ObjectNet*	IN-A		
CLIP `ViT-L/14@336px`								
Zero-shot [79]	76.2	70.1	88.9	60.2	70.0	77.2	73.3	74.8
Fine-tuned LC [79]	85.4	75.9	84.2	57.4	66.2	75.3	71.8	78.6
Zero-shot (PyTorch)	76.6	70.5	89.0	60.9	69.1	77.7	73.4	75.0
Fine-tuned LC (ours)	85.2	75.8	85.3	58.7	67.2	76.1	72.6	78.9
Fine-tuned E2E (ours)	86.2	76.8	79.8	57.9	63.3	65.4	68.6	77.4
WiSE-FT (ours)								
LC, α=0.5	83.7	76.3	89.6	63.0	70.7	79.7	75.9	79.8
LC, optimal α	85.3	76.9	89.8	63.0	70.7	79.7	75.9	80.2
E2E, α=0.5	86.8	**79.5**	89.4	64.7	71.1	79.9	76.9	81.8
E2E, optimal α	**87.1**	**79.5**	**90.3**	**65.0**	**72.1**	**81.0**	**77.4**	**81.9**

Table 1. Accuracy of various methods on ImageNet and derived distribution shifts for CLIP `ViT-L/14@336px` [79]. E2E: end-to-end; LC: linear classifier. *Avg shifts* displays the mean performance among the five distribution shifts, while *Avg reference, shifts* shows the average of ImageNet (reference) and Avg shifts. For optimal α, we choose the single mixing coefficient that maximizes the column. Results for additional models are provided in Appendix E.7.

WiSE-FT

- *Experimental Results*

BASIC, ALIGN 과 같은 다른 Zero-shot 모델에도 적용 가능,

+) 보다 큰 모델인 ViT-H/14 모델로 실험.

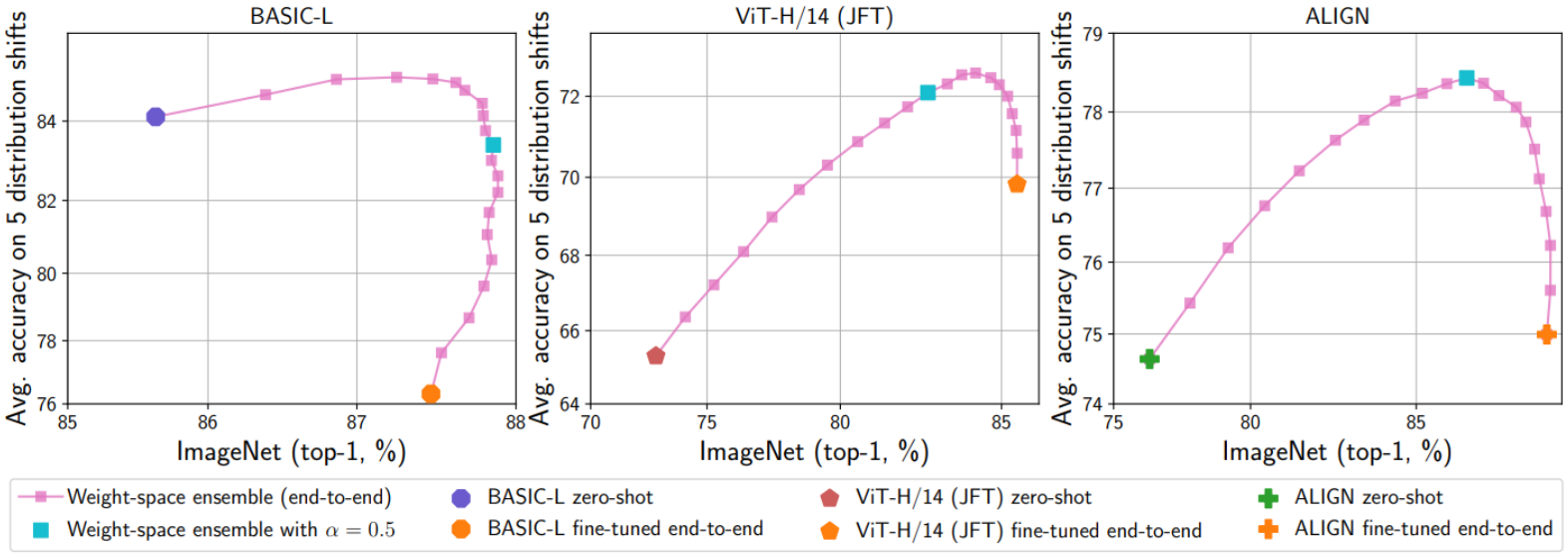


Figure 4. WiSE-FT applied to BASIC-L [75], a ViT-H/14 [21] model pre-trained on JFT-300M [91] and ALIGN [44].

WiSE-FT

- *Summary*

- Large pre-trained 모델의 Robustness와 Fine-tuning accuracy 간의 Trade-off 해결.
- 이를 통해 보다 신뢰 가능한 Zero-shot 모델을 구축할 수 있음.
- 다만 Task가 Image Classification에 한정, 최적의 α값이 0.5인 점도 의문점으로 남아 있음.