Side-Tuning: A Baseline for Network Adaptation via Additive Side Networks

Berkeley University of California

Stanford University

EPFL Swiss Federal Institute of Technology

Jeffrey O. Zhang

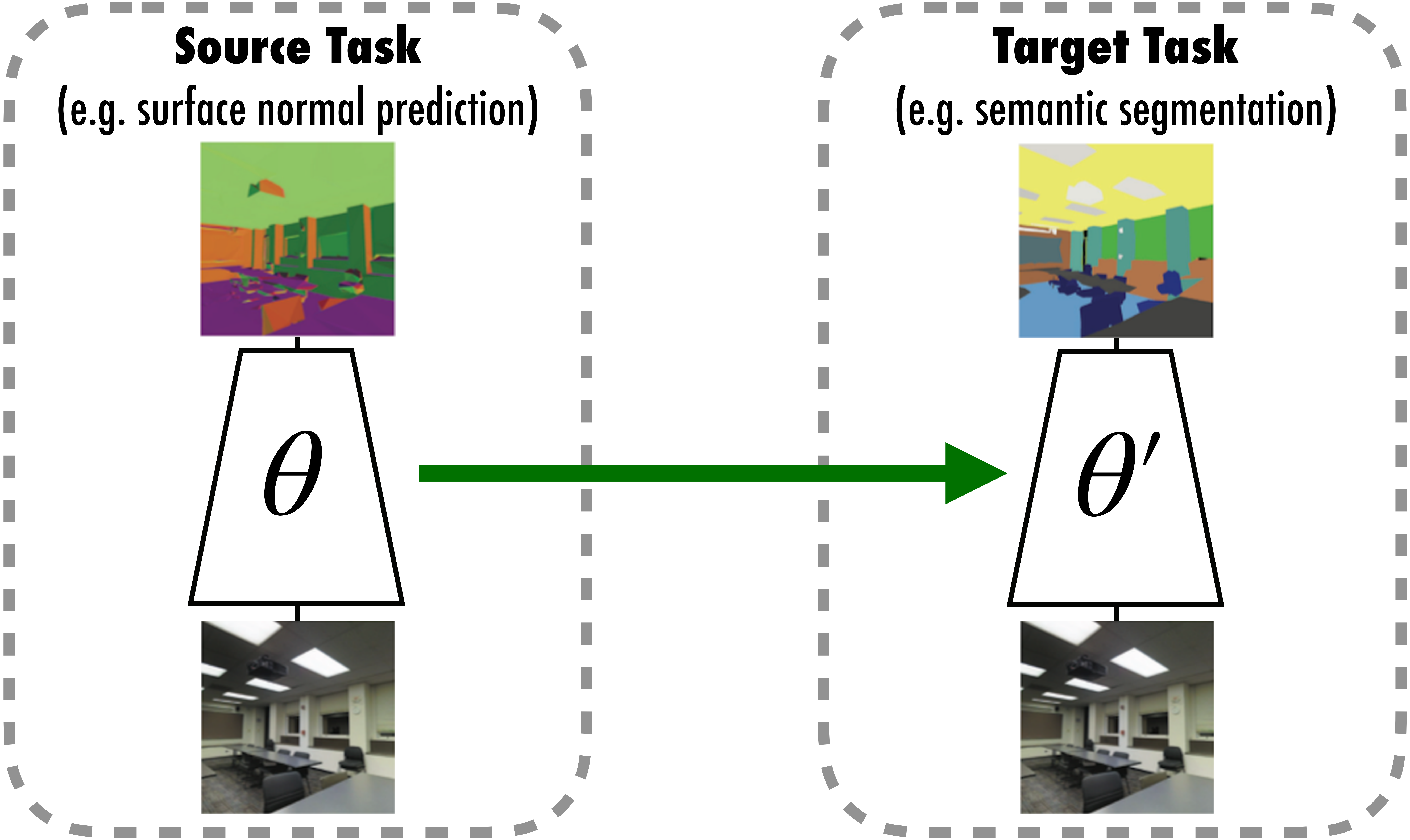
Alexander Sax

Amir Zamir

Leonidas Guibas

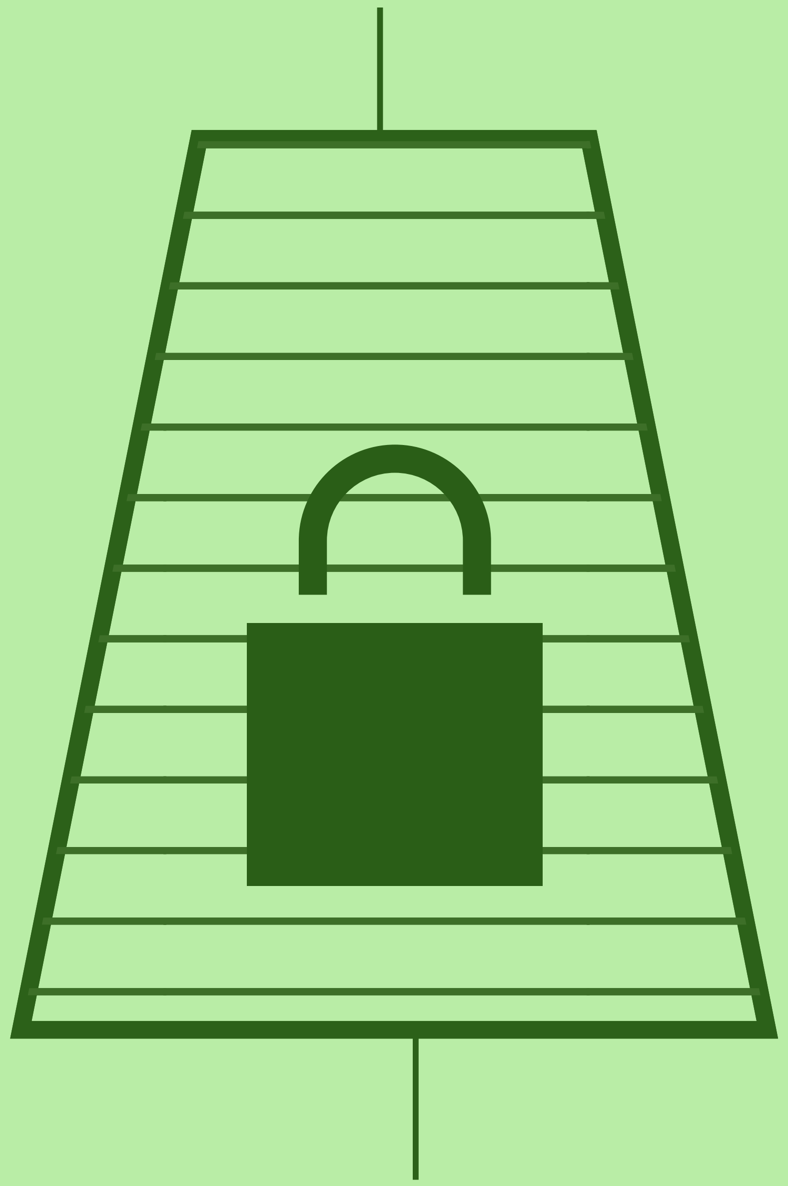
Jitendra Malik

Network Adaptation



Approaches for Network Adaptation

Freeze + Add



Simplest Approach:
Fixed Features

- CNN features Off the Shelf (Sharif et al.)
- Piggyback (Mallya et al.)
- PackNet (Mallya et al.)
- Deep Adaptation (Rosenfeld et al.)
- Hard Attention (Serra et al.)
- Residual Continual Learning (Lee et al.)
- Progressive NN (Rusu et al.)
- Efficient parameterization (Rebuffi et al.)

Update

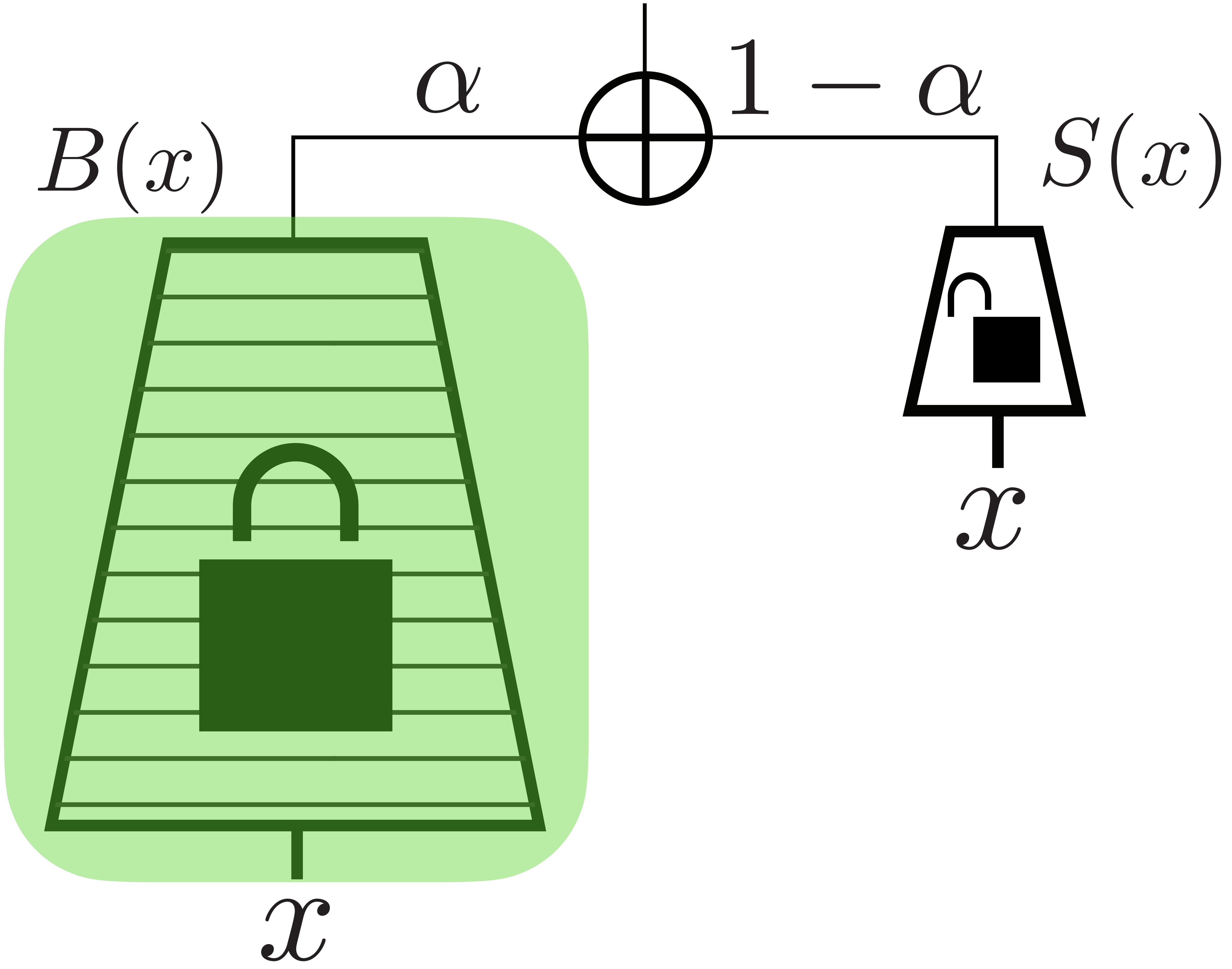
Simplest Approach:
Fine-Tuning

- Elastic Weight Consolidation (Kirkpatrick et al.)
- Learning without Forgetting (Li et al.)
- Superposition of Many Models into One (Cheung et al.)

Features vs. Fine-Tuning

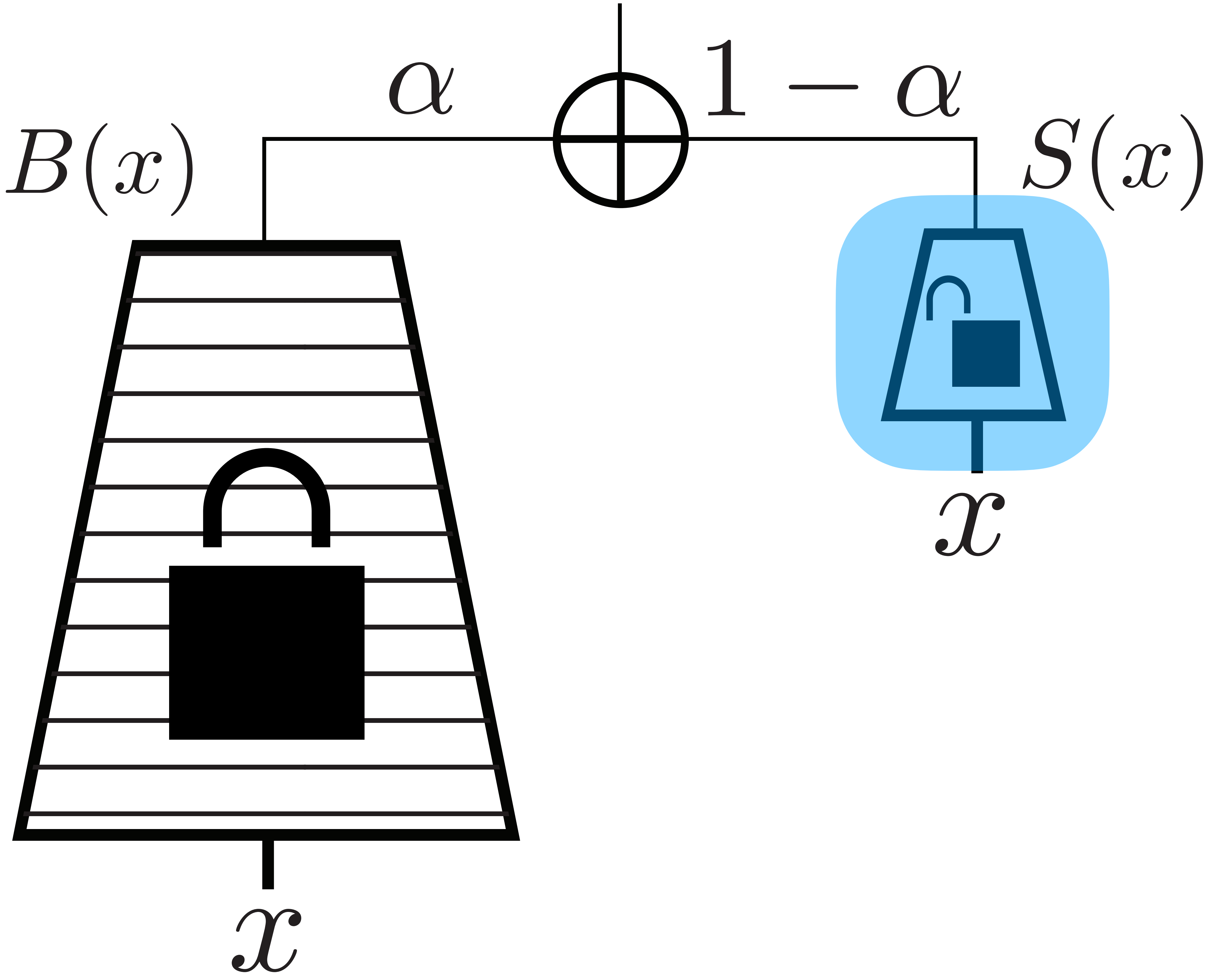
Method	**1 Target Task** Low Data	High Data	**> 1 Target Tasks** (incremental)
Fixed features	✓	✗ (Info Loss)	✗ (Info Loss)
Fine-tuning	✗ (Overfit)	✓	✗ (Forgetting)

Side-tuning: a straightforward freeze+add approach



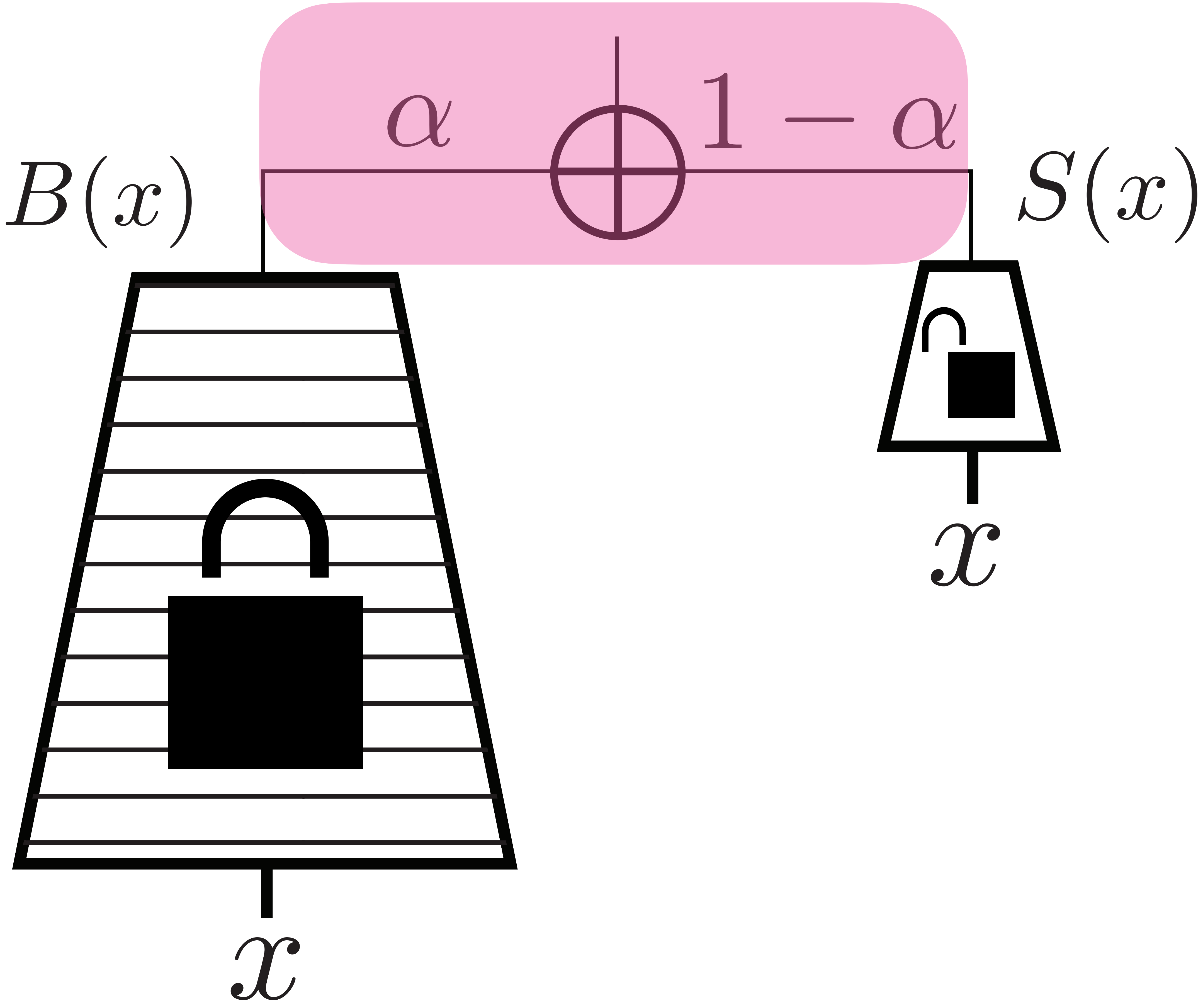
- Base network, B(x) → pre-trained

Side-tuning: a straightforward freeze+add approach



- Base network, B(x) → pre-trained
- Side network, S(x) → updated for target task

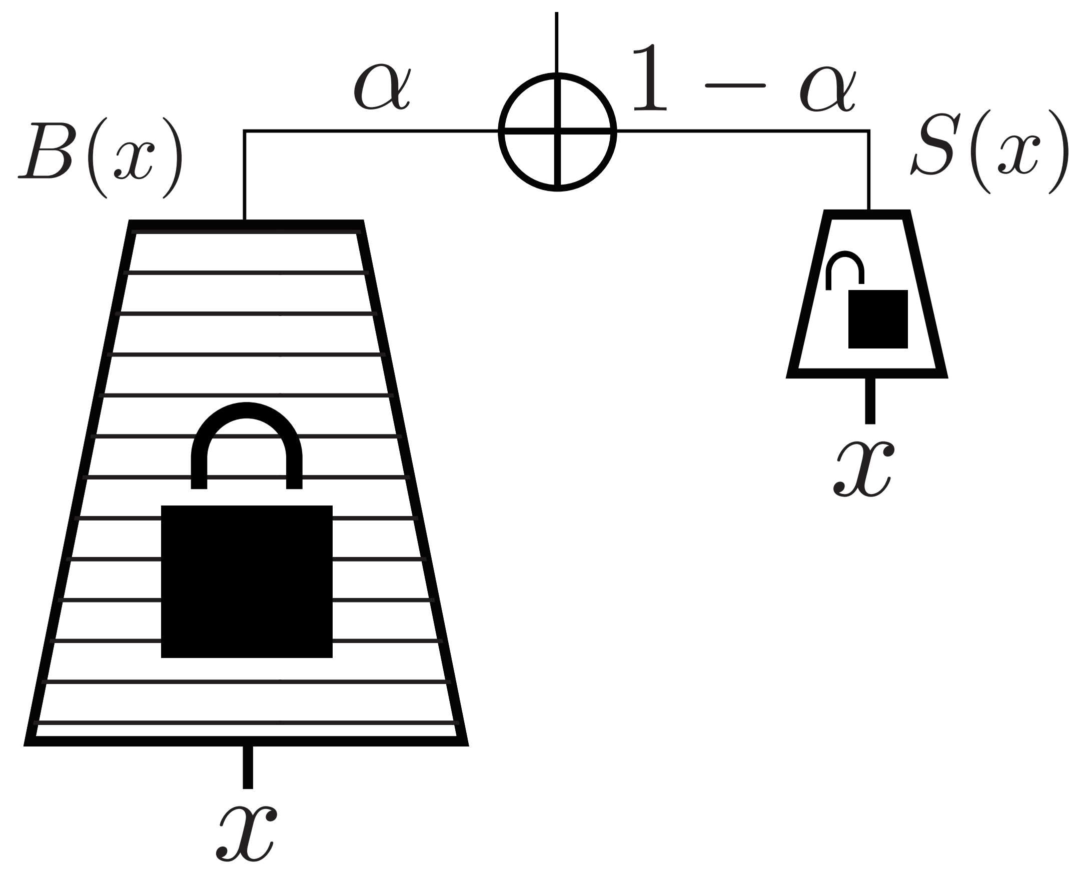
Side-tuning: a straightforward freeze+add approach

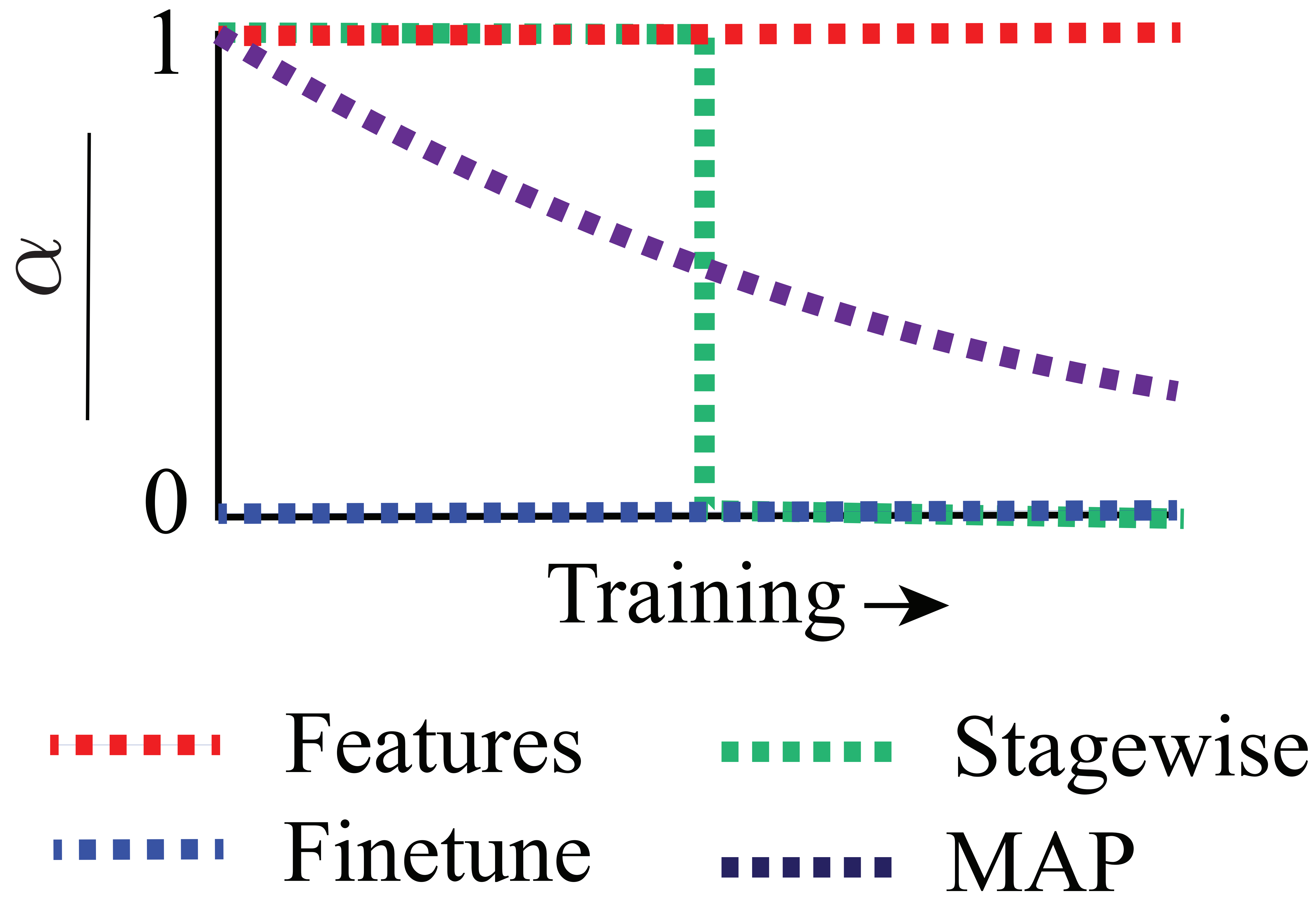


- Base network, B(x) → pre-trained
- Side network, S(x) → updated for target task
- Combined via alpha-blending

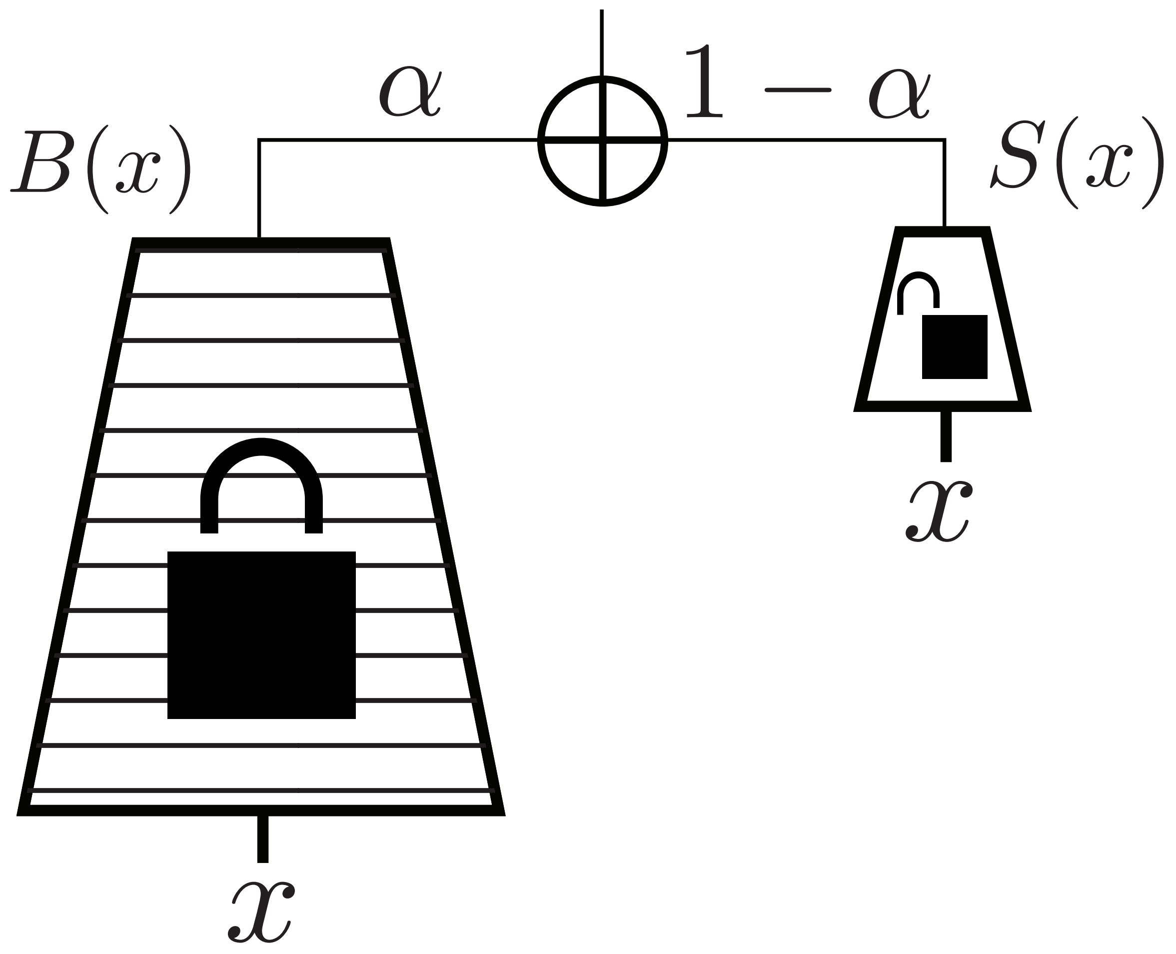
Side-Tuning: A Baseline for Network Adaptation via Additive Side Networks, ECCV 2020

Side-Tuning: Learning α

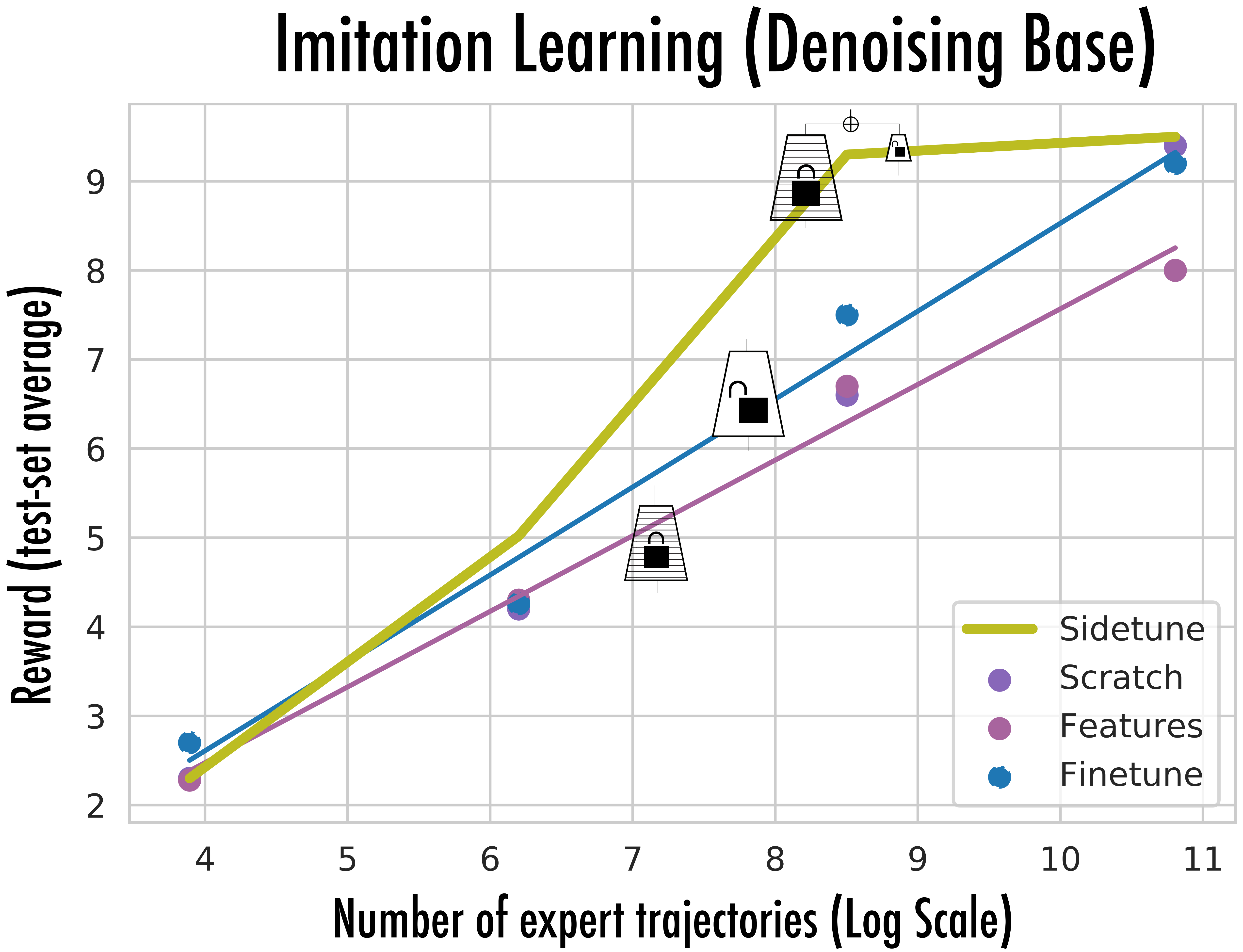




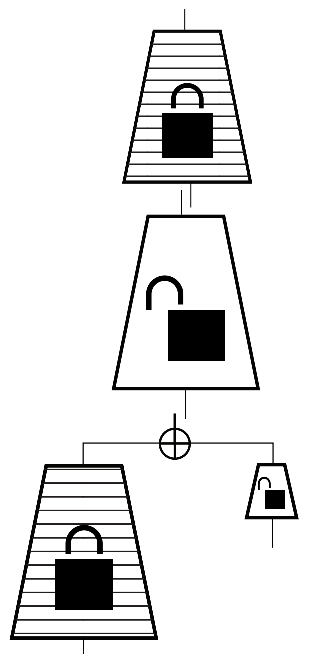
Side-tuning for intermediate amounts of data



- Base network → useful bias
- Side network → consistency



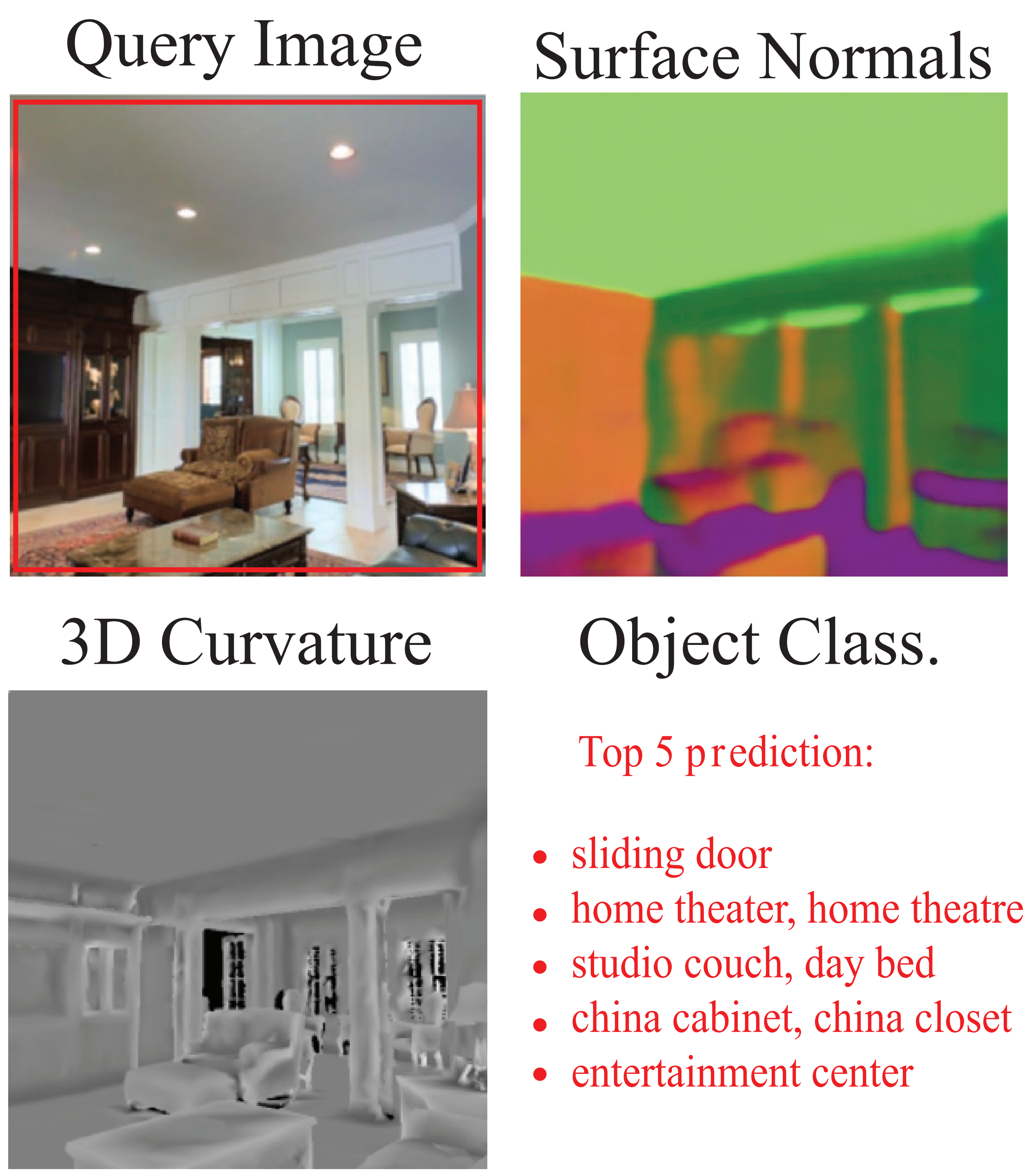
Features vs. Fine-Tuning



	1 Target Task		**> 1 Target Tasks**
Method	Low Data	High Data	(incremental)
Fixed features	✓	✗ (Info Loss)	✗ (Info Loss)
Fine-tuning	✗ (Overfit)	✓	✗ (Forgetting)
Side-tuning	✓	✓	✓

Side-tuning in varied settings

Computer Vision



Taskonomy (Zamir et al.)

NLP

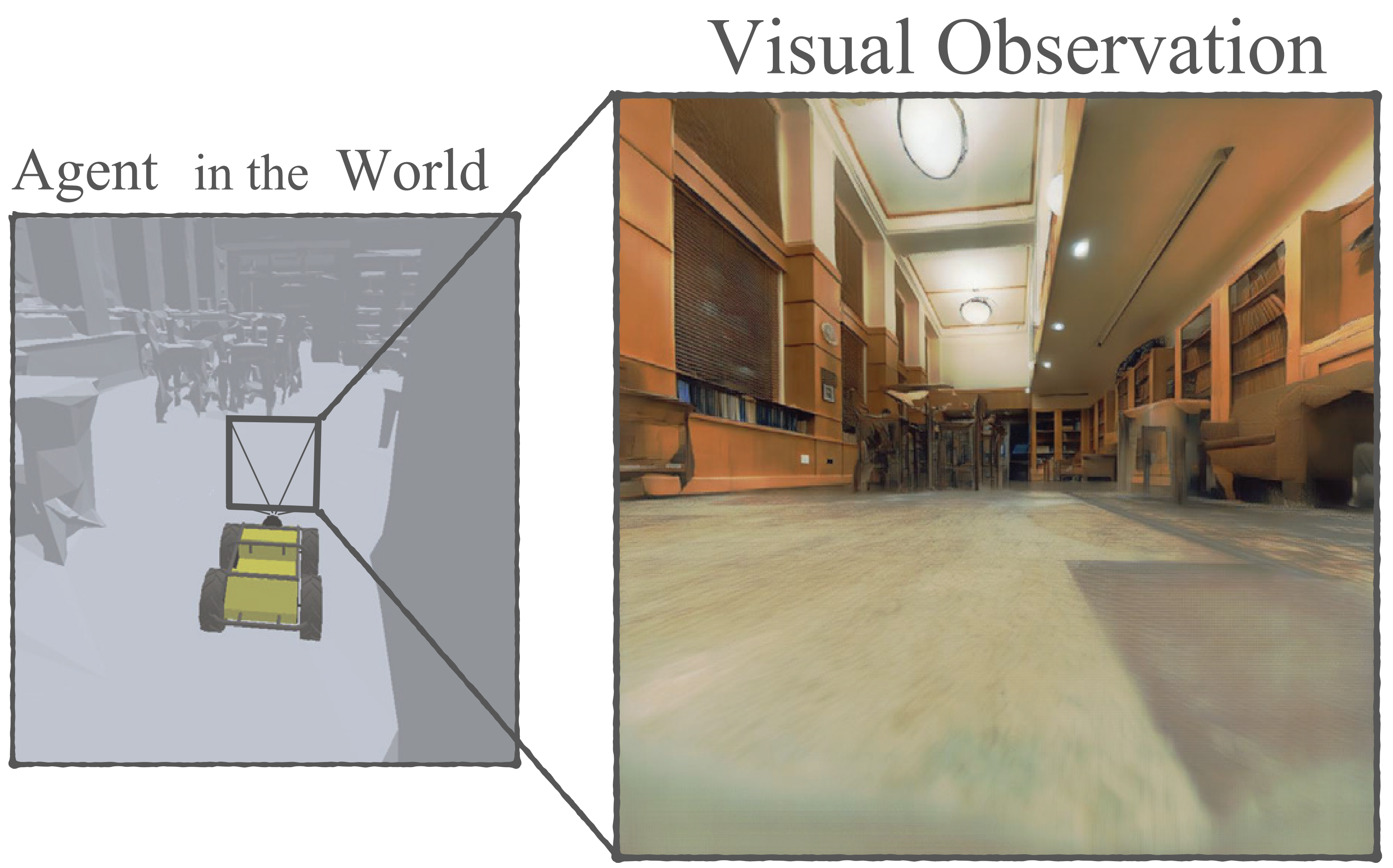
Article: Endangered Species Act
Paragraph: " …Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised."

Question 1: "Which laws faced significant opposition?"
Plausible Answer: later laws

Question 2: "What was the name of the 1937 treaty?"
Plausible Answer: Bald Eagle Protection Act

SQUAD v2 (Rajpurkar et al.)

Robotics (Tested in Gibson)



Habitat (Savva et al.)
Gibson (Xia et al.)

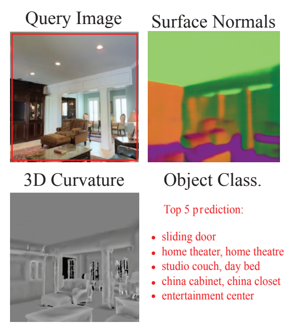
Solid performance across multiple domains and settings

Solid performance across multiple domains and settings

Transfer Learning in Taskonomy

Method	From Curvature (100/4M ims.)	
	Normals (MSE ↓)	Obj. Cls. (Acc. ↑)
Fine-tune	**0.200** / **0.094**	**24.6** / **62.8**
Features	**0.204** / 0.117	24.4 / 45.4
Scratch	0.323 / **0.095**	19.1 / 62.3
Side-tune	**0.199** / **0.095**	**24.8** / **63.3**

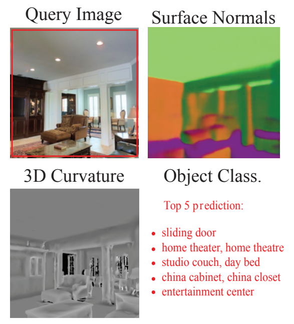
- **Low-dimensional prediction tasks**
- **High-dimensional pix-to-pix tasks**
- **Low-data (100 images)**
- **High-data (4M images)**



Solid performance across multiple domains and settings

Method	Transfer Learning in Taskonomy: From Curvature (100/4M ims.) Normals (MSE ↓)	Transfer Learning in Taskonomy: From Curvature (100/4M ims.) Obj. Cls. (Acc. ↑)	QA on SQuAD: Match (↑) Exact	QA on SQuAD: Match (↑) F1
Fine-tune	**0.200** / **0.094**	**24.6** / **62.8**	**79.0**	**82.2**
Features	**0.204** / 0.117	24.4 / 45.4	49.4	49.5
Scratch	0.323 / **0.095**	19.1 / 62.3	0.98	4.65
Side-tune	**0.199** / **0.095**	**24.8** / **63.3**	**79.6**	**82.7**

- Low-dimensional prediction tasks
- High-dimensional pix-to-pix tasks
- Low-data (100 images)
- High-data (4M images)



- **NLP domain**
- **Different architecture (transformer)**

Article: Endangered Species Act
Paragraph: " …Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised."

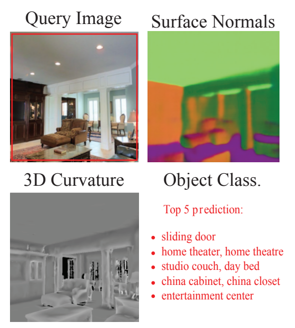
Question 1: "Which laws faced significant opposition?"
Plausible Answer: later laws

Question 2: "What was the name of the 1937 treaty?"
Plausible Answer: Bald Eagle Protection Act

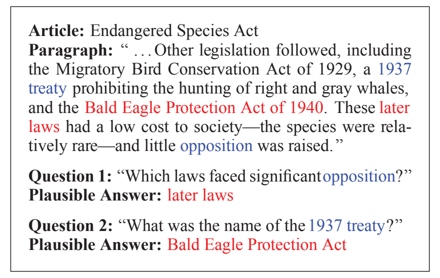
Solid performance across multiple domains and settings

Method	Transfer Learning in Taskonomy: From Curvature (100/4M ims.) Normals (MSE ↓)	Transfer Learning in Taskonomy: From Curvature (100/4M ims.) Obj. Cls. (Acc. ↑)	QA on SQuAD: Match (↑) Exact	QA on SQuAD: Match (↑) F1	Navigation (IL): Nav. Rew. (↑) Curv.	Navigation (IL): Nav. Rew. (↑) Denoise
Fine-tune	**0.200** / **0.094**	**24.6** / **62.8**	**79.0**	**82.2**	10.5	9.2
Features	**0.204** / 0.117	24.4 / 45.4	49.4	49.5	**11.2**	8.2
Scratch	0.323 / **0.095**	19.1 / 62.3	0.98	4.65	9.4	**9.4**
Side-tune	**0.199** / **0.095**	**24.8** / **63.3**	**79.6**	**82.7**	**11.1**	**9.5**

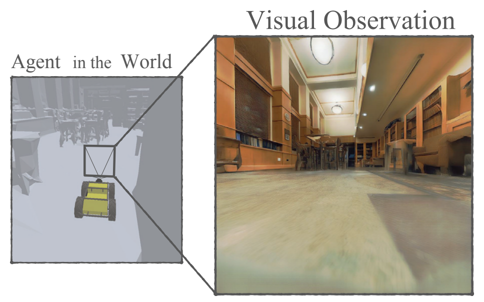
- Low-dimensional prediction tasks
- High-dimensional pix-to-pix tasks
- Low-data (100 images)
- High-data (4M images)



- NLP domain
- Different architecture (transformer)



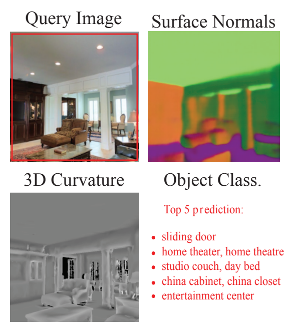
- **Active POMDP settings**



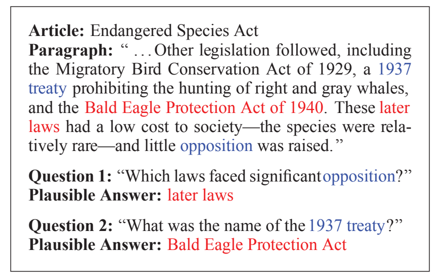
Solid performance across multiple domains and settings

	Transfer Learning in Taskonomy		QA on SQuAD		Navigation (IL)		Navigation (RL)	
	From Curvature (100/4M ims.)		Match (↑)		Nav. Rew. (↑)		Nav. Rew. (↑)	
Method	Normals (MSE ↓)	Obj. Cls. (Acc. ↑)	Exact	F1	Curv.	Denoise	Curv.	Denoise
Fine-tune	**0.200 / 0.094**	**24.6 / 62.8**	**79.0**	**82.2**	10.5	9.2	10.7	**10.0**
Features	**0.204** / 0.117	24.4 / 45.4	49.4	49.5	**11.2**	8.2	**11.9**	8.3
Scratch	0.323 / **0.095**	19.1 / 62.3	0.98	4.65	9.4	**9.4**	7.5	7.5
Side-tune	**0.199 / 0.095**	**24.8 / 63.3**	**79.6**	**82.7**	**11.1**	**9.5**	**11.8**	**10.4**

- Low-dimensional prediction tasks
- High-dimensional pix-to-pix tasks
- Low-data (100 images)
- High-data (4M images)



- NLP domain
- Different architecture (transformer)

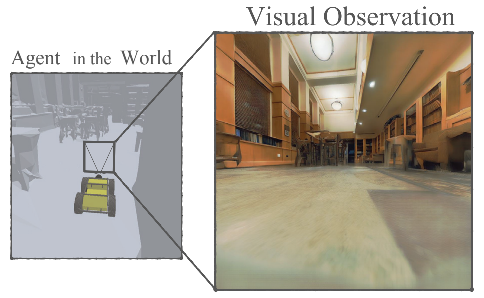


Article: Endangered Species Act
Paragraph: " …Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised."

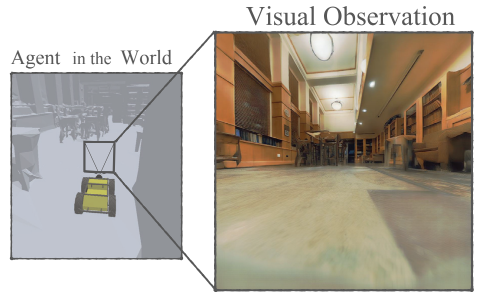
Question 1: "Which laws faced significant opposition?"
Plausible Answer: later laws

Question 2: "What was the name of the 1937 treaty?"
Plausible Answer: Bald Eagle Protection Act

- Active POMDP settings



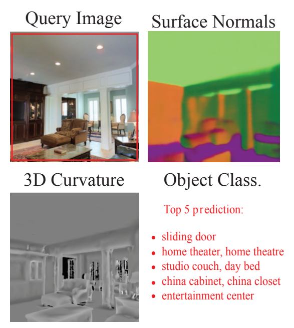
- Different learning algorithms (PPO instead of supervised learning)



Solid performance across multiple domains and settings

Method	Transfer Learning in Taskonomy: From Curvature (100/4M ims.)		QA on SQuAD: Match (↑)		Navigation (IL): Nav. Rew. (↑)		Navigation (RL): Nav. Rew. (↑)	
	Normals (MSE ↓)	Obj. Cls. (Acc. ↑)	Exact	F1	Curv.	Denoise	Curv.	Denoise
Fine-tune	**0.200** / **0.094**	**24.6** / **62.8**	**79.0**	**82.2**	10.5	9.2	10.7	**10.0**
Features	**0.204** / 0.117	24.4 / 45.4	49.4	49.5	**11.2**	8.2	**11.9**	8.3
Scratch	0.323 / **0.095**	19.1 / 62.3	0.98	4.65	9.4	**9.4**	7.5	7.5
Side-tune	**0.199** / **0.095**	**24.8** / **63.3**	**79.6**	**82.7**	**11.1**	**9.5**	**11.8**	**10.4**

- Low-dimensional prediction tasks
- High-dimensional pix-to-pix tasks
- Low-data (100 images)
- High-data (4M images)



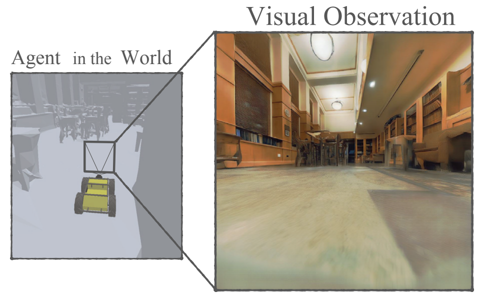
- NLP domain
- Different architecture (transformer)

Article: Endangered Species Act
Paragraph: " ...Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised."

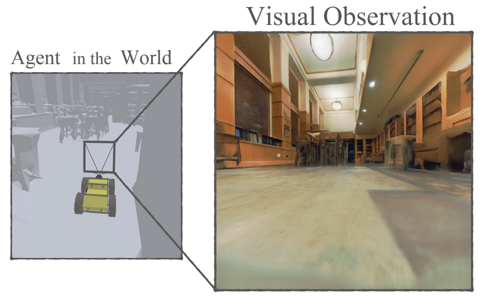
Question 1: "Which laws faced significant opposition?"
Plausible Answer: later laws

Question 2: "What was the name of the 1937 treaty?"
Plausible Answer: Bald Eagle Protection Act

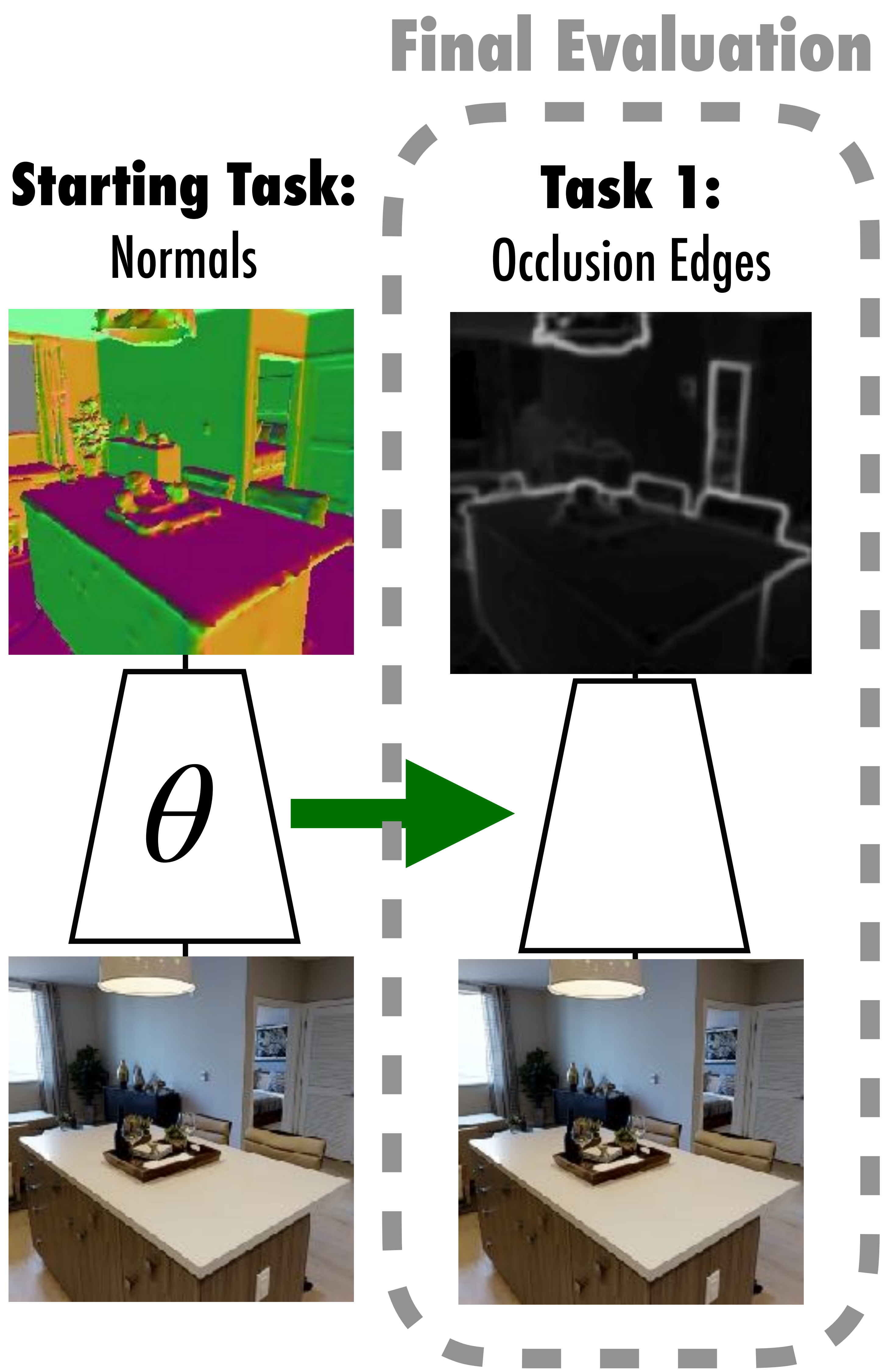
- Active POMDP settings



- Different learning algorithms (PPO instead of supervised learning)

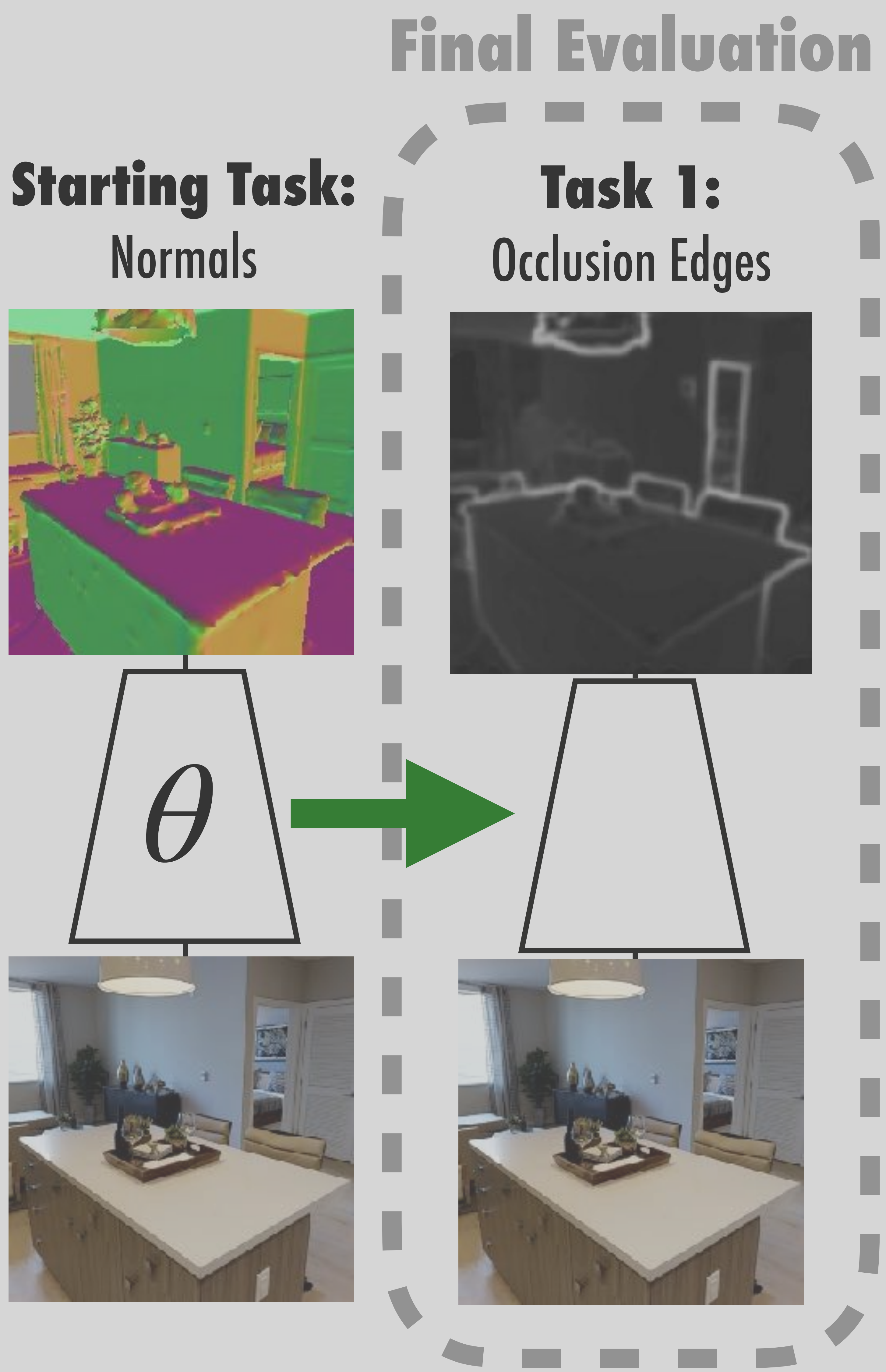


Adaptation

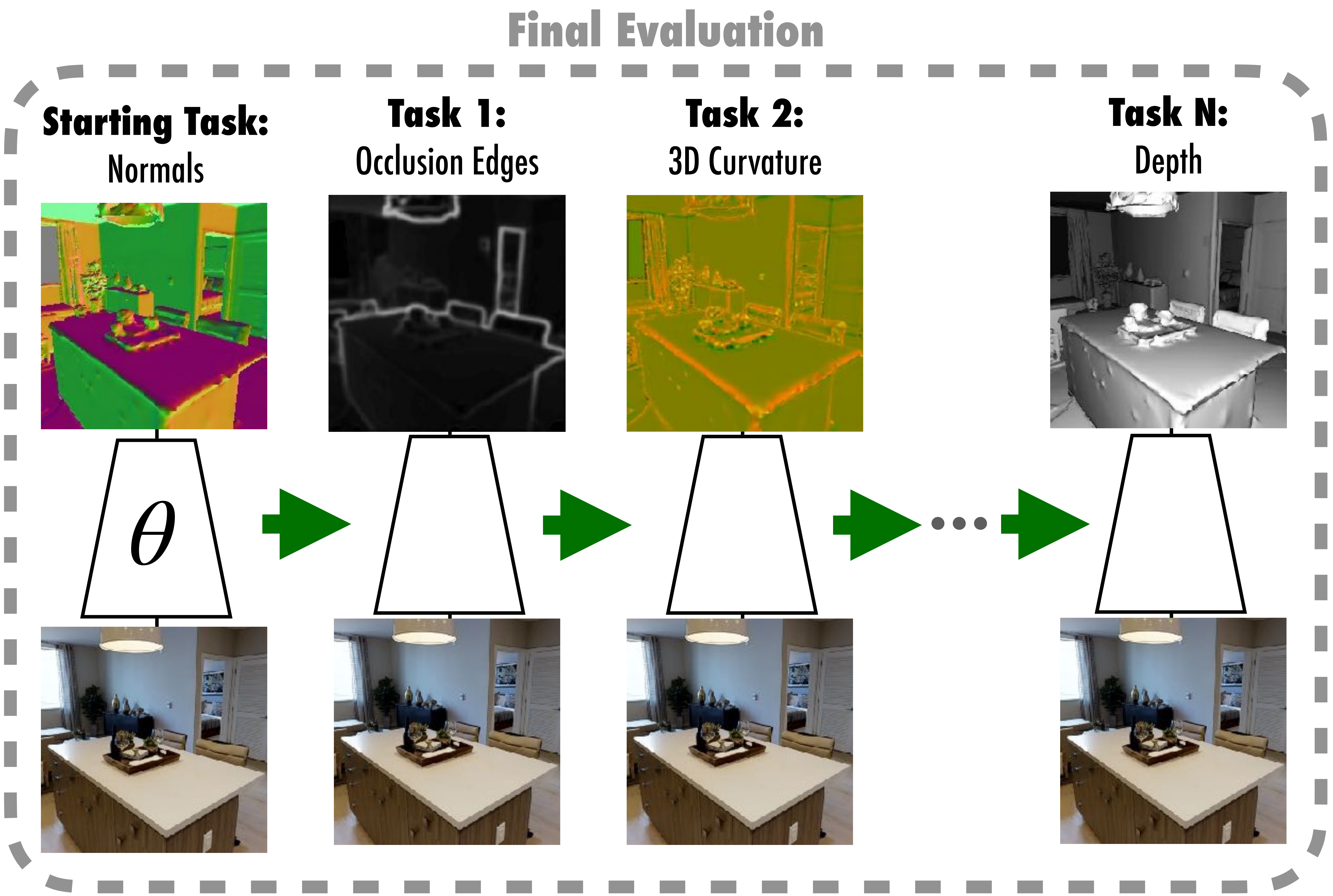


Final Evaluation
Starting Task:
Normals
θ
Task 1:
Occlusion Edges

Adaptation



Incremental Learning



Incremental learning: forgetting and rigidity

Catastrophic Forgetting

Tendency of a network to lose previously learned knowledge upon learning new information.

Catastrophic interference in connectionist networks: the sequential learning problem (McCloskey + Cohen, 1989)

Incremental learning: forgetting and rigidity

Catastrophic Forgetting

Tendency of a network to lose previously learned knowledge upon learning new information.

Catastrophic interference in connectionist networks: the sequential learning problem (McCloskey + Cohen, 1989)

Rigidity (Intransigence)

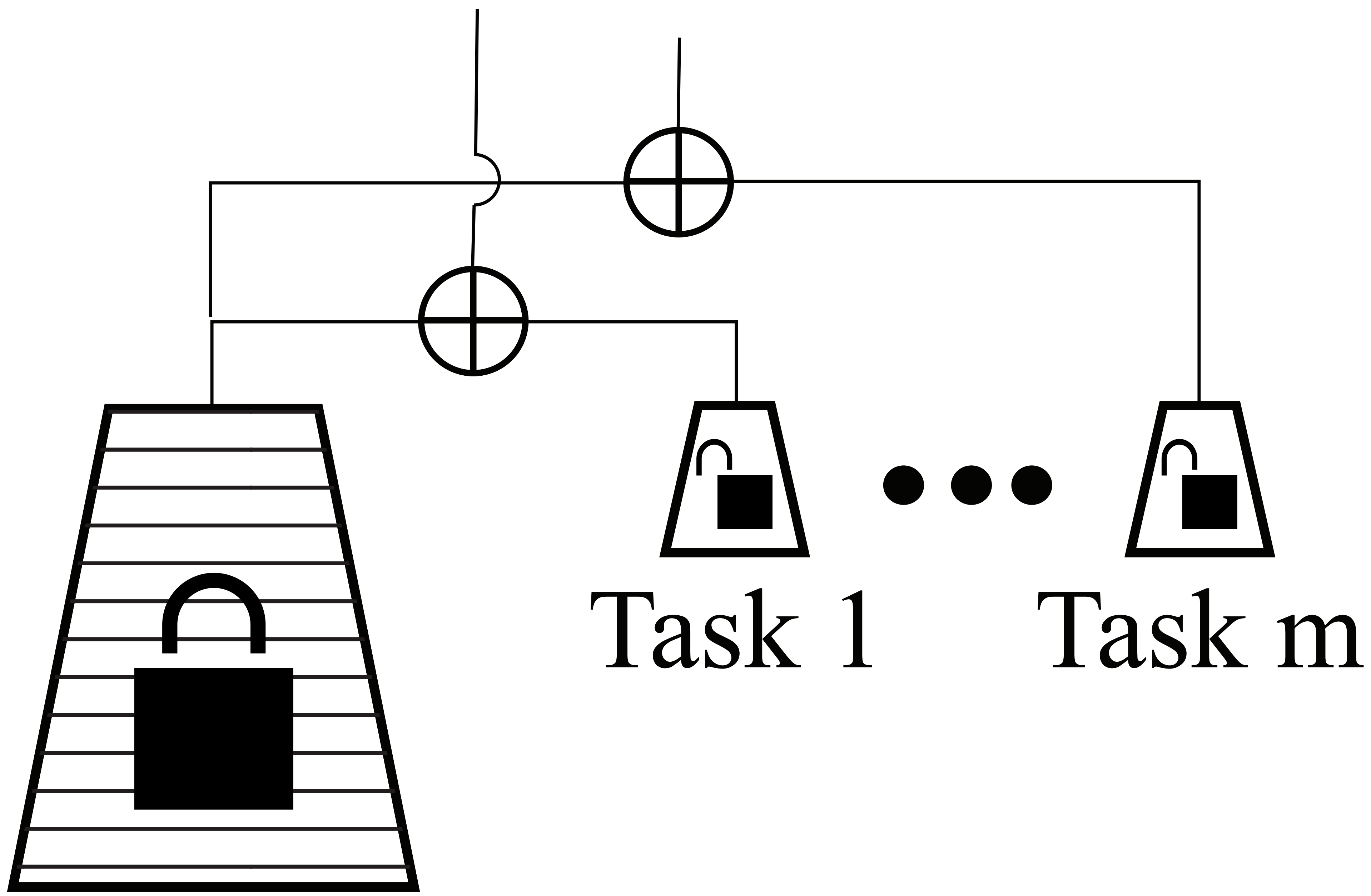
Increasing inability of a network to adapt to new problems as it accrues constraints from previous problems.

The concept of rigidity: an enigma (Len, 1983)

Riemannian walk for incremental learning: understanding forgetting and intransigence (Chaudhry et al 2018)

Side-tuning for incremental learning

Architecture



Incremental learning: forgetting and rigidity

Catastrophic Forgetting

Tendency of a network to lose previously learned knowledge upon learning new information.

Catastrophic interference in connectionist networks: the sequential learning problem (McCloskey + Cohen, 1989)

Rigidity (Intransigence)

Increasing inability of a network to adapt to new problems as it accrues constraints from previous problems.

The concept of rigidity: an enigma (Len, 1983)

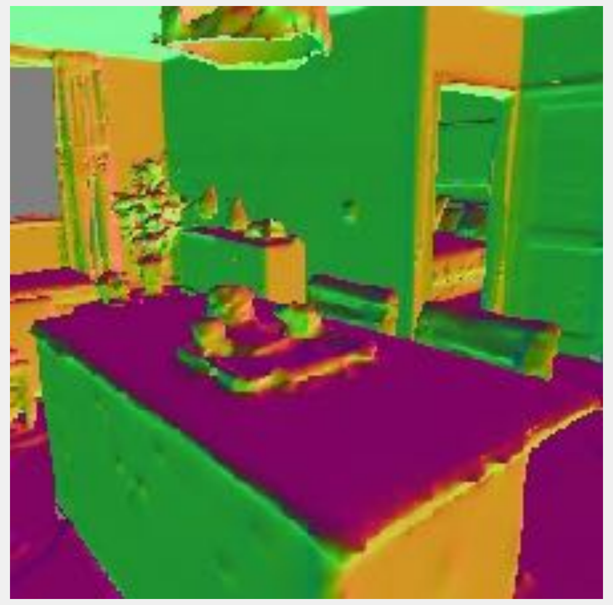
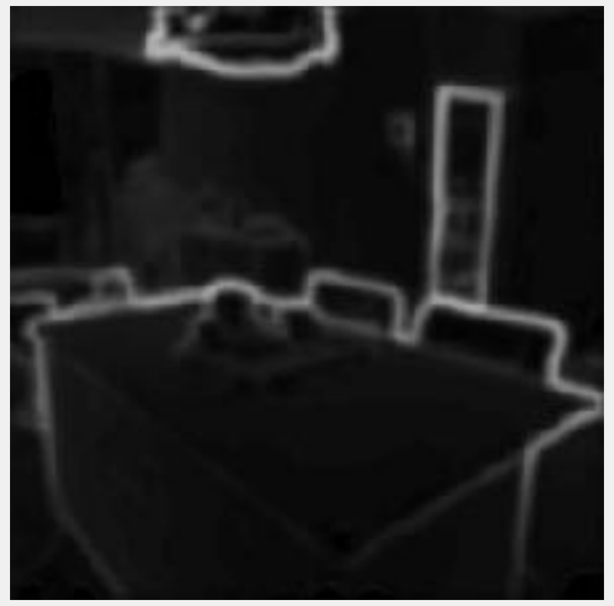
Riemannian walk for incremental learning: understanding forgetting and intransigence (Chaudhry et al 2018)

Side-tuning: **no** forgetting + **no** rigidity

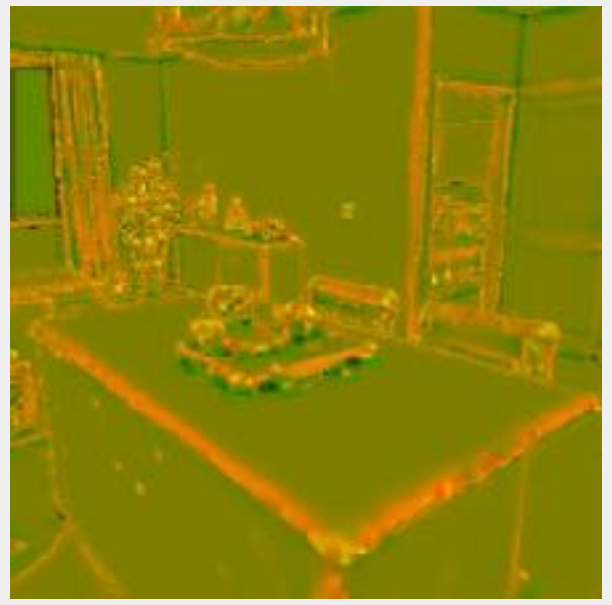
Incremental learning datasets: tasks

iTaskonomy
Zamir et al.
—
512x512 images
12 tasks

Autoencoding → **Normals** Estimation → **Occlusion Edges** → **100-way** object classification: → **Curvature** → **63-way** scene classification: ••• ▸ **Depth** Estimation

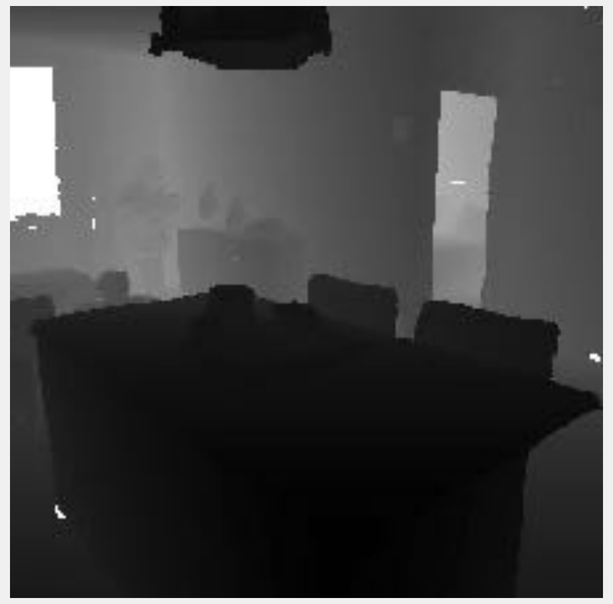




Answer: Table



Answer: Office

•••

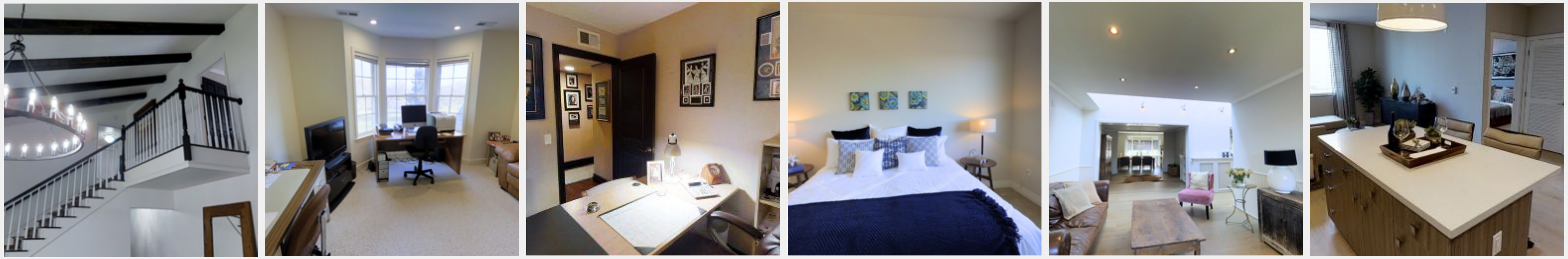


iCIFAR
Rebuffi et al.
—
32x32 images
20 tasks

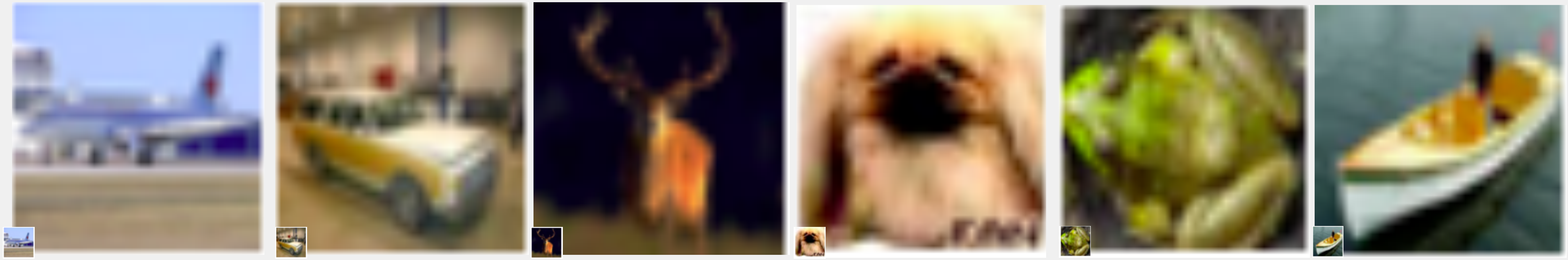
5-way object classification: Answer: Cup → **5-way** object classification: Answer: Cow → **5-way** object classification: Answer: Cloud → **5-way** object classification: Answer: Lion → **5-way** object classification: Answer: Train → **5-way** object classification: Answer: Squirrel ••• ▸ ••• **5-way** object classification: Answer: Butterfly

Incremental learning datasets: query images

iTaskonomy
Zamir et al.
—
512x512 images
12 tasks

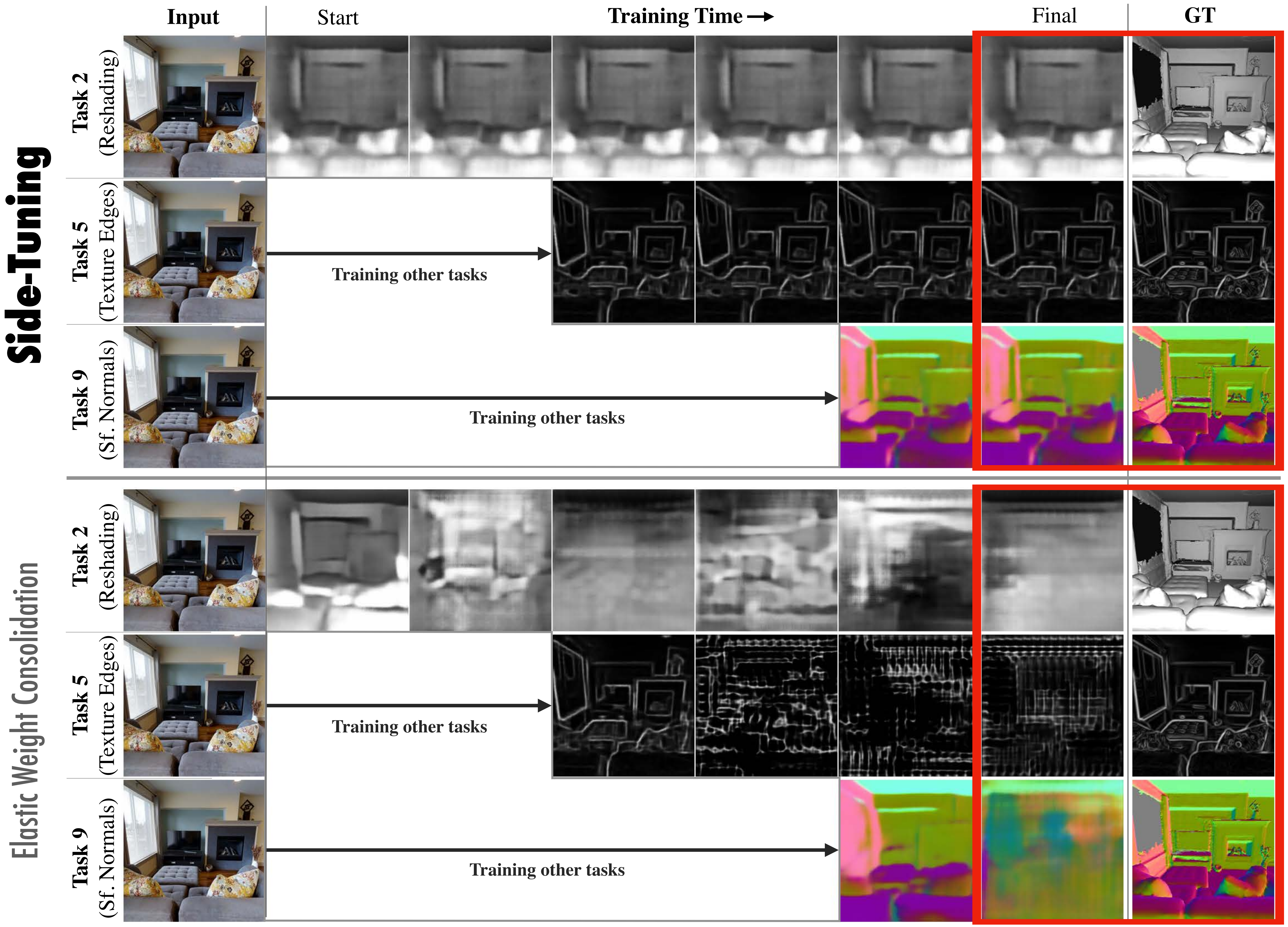


iCIFAR
Rebuffi et al.
—
32x32 images
20 tasks



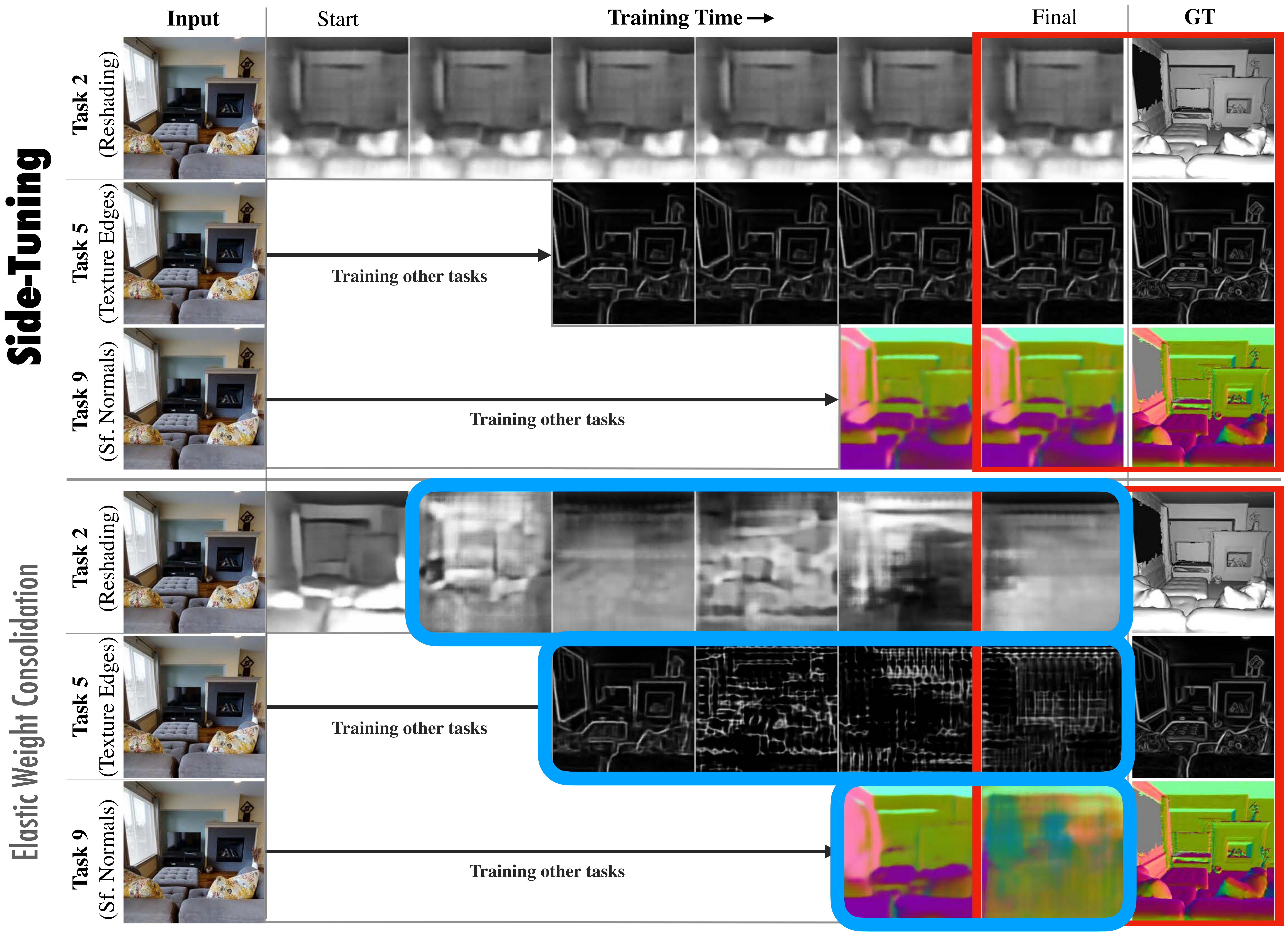
Trends shift on harder datasets: “Which Tasks Should Be Learned Together in Multi-task Learning?” Standley et al (ICML 2020)

Evaluation on iTaskonomy



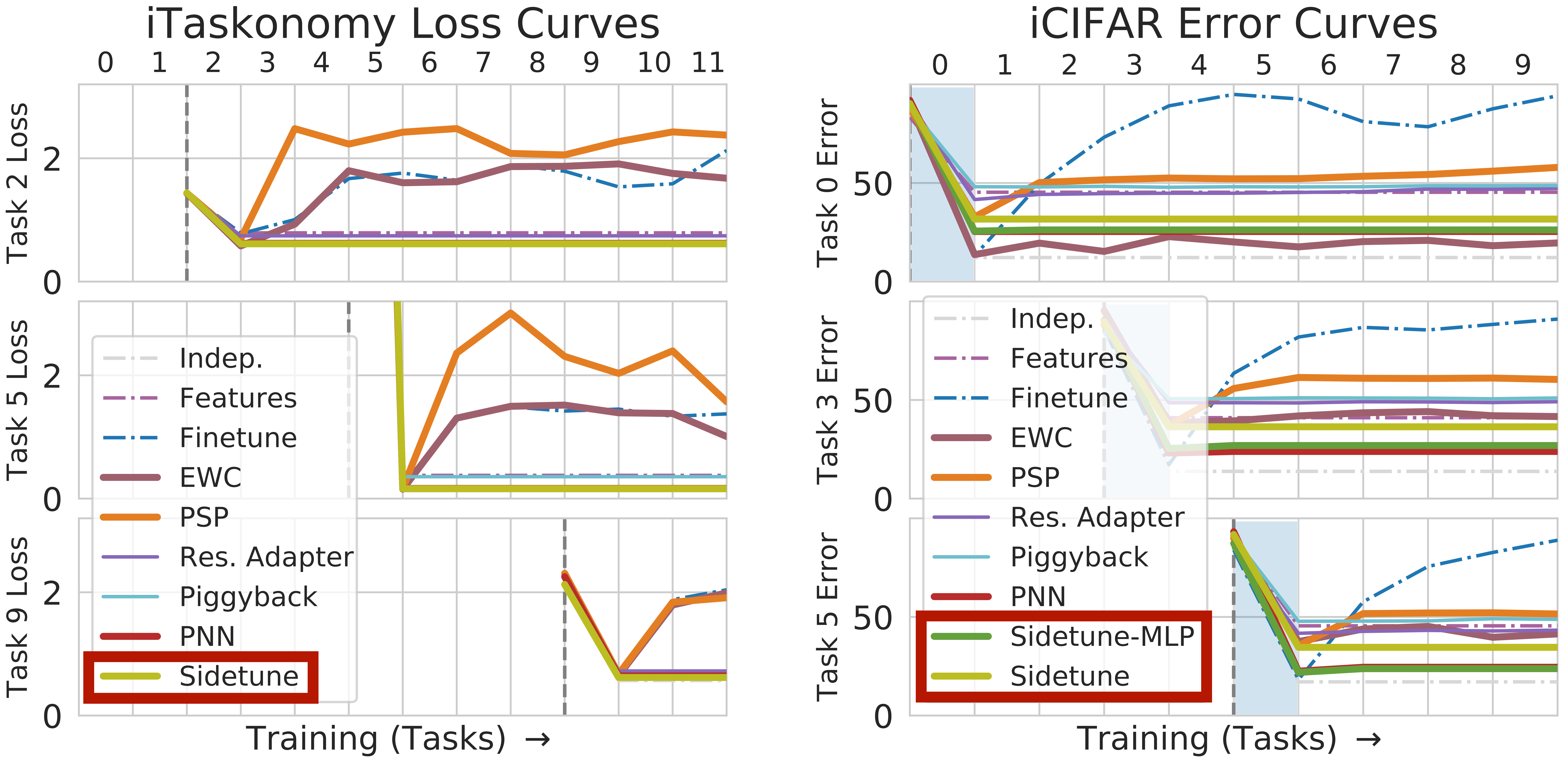
Side-Tuning: A Baseline for Network Adaptation via Additive Side Networks, ECCV 2020

Evaluation on iTaskonomy

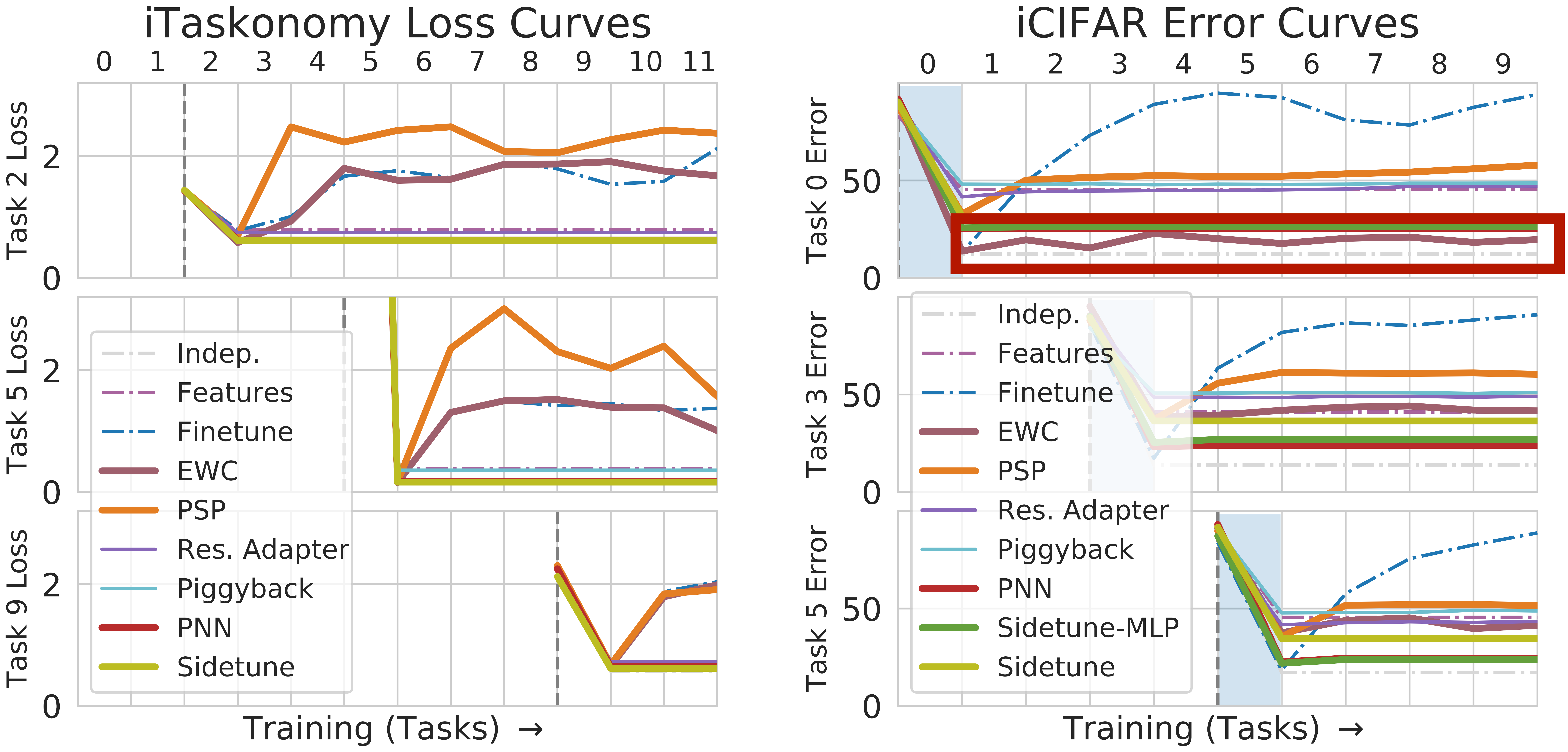


Side-Tuning: A Baseline for Network Adaptation via Additive Side Networks, ECCV 2020

Results: catastrophic forgetting in incremental learning

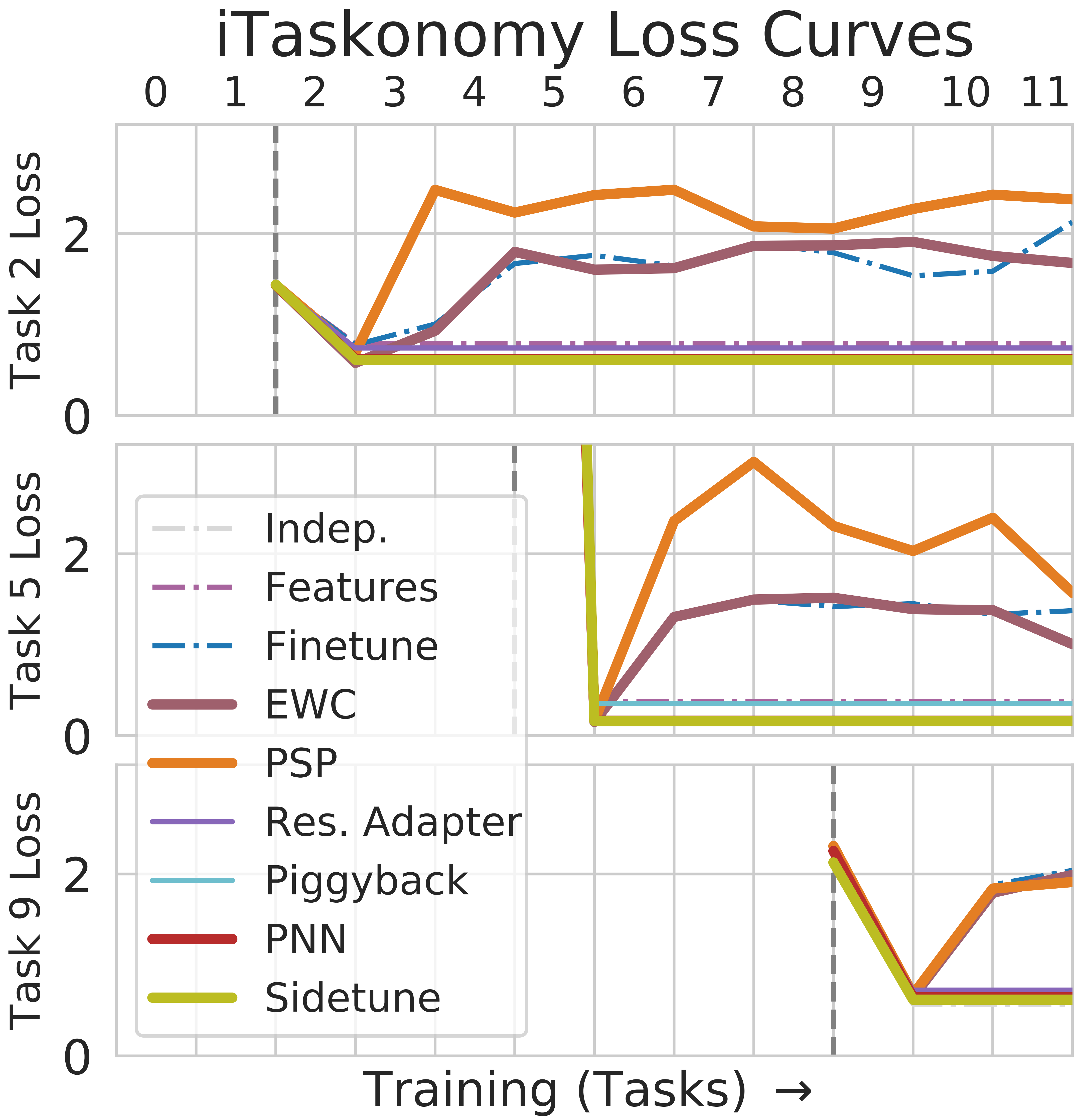


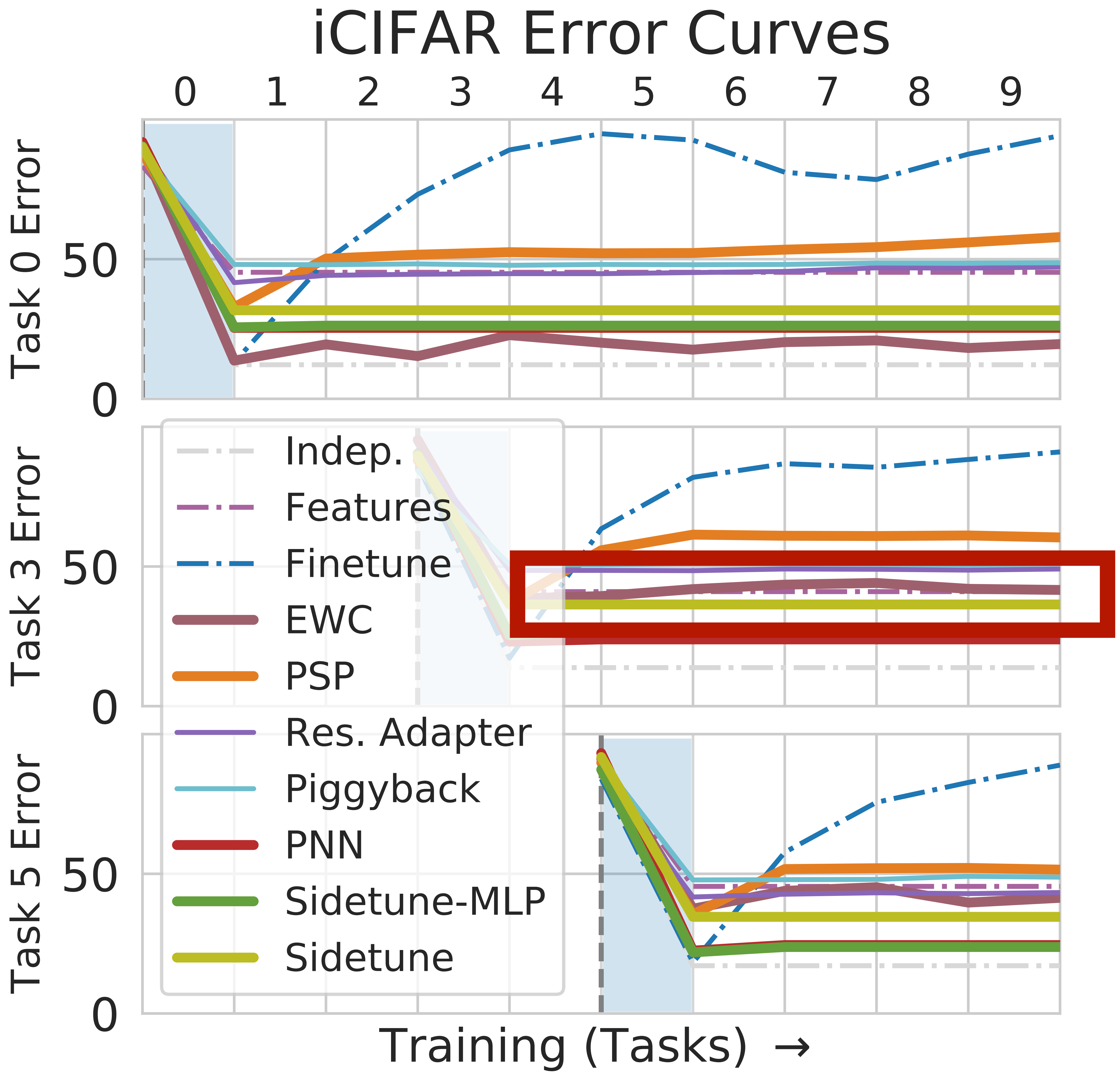
Results: catastrophic forgetting in incremental learning



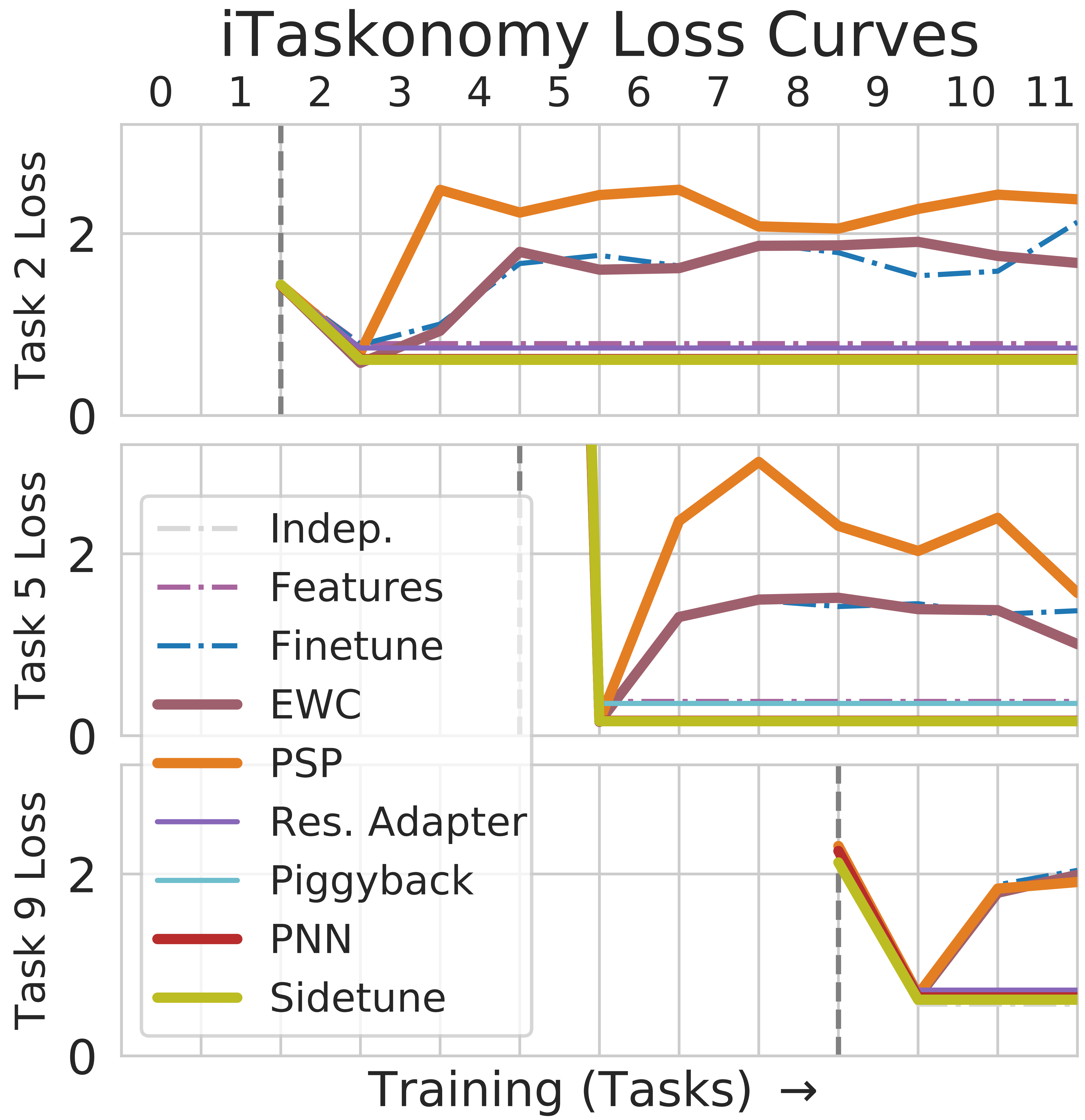
Side-Tuning: A Baseline for Network Adaptation via Additive Side Networks, ECCV 2020

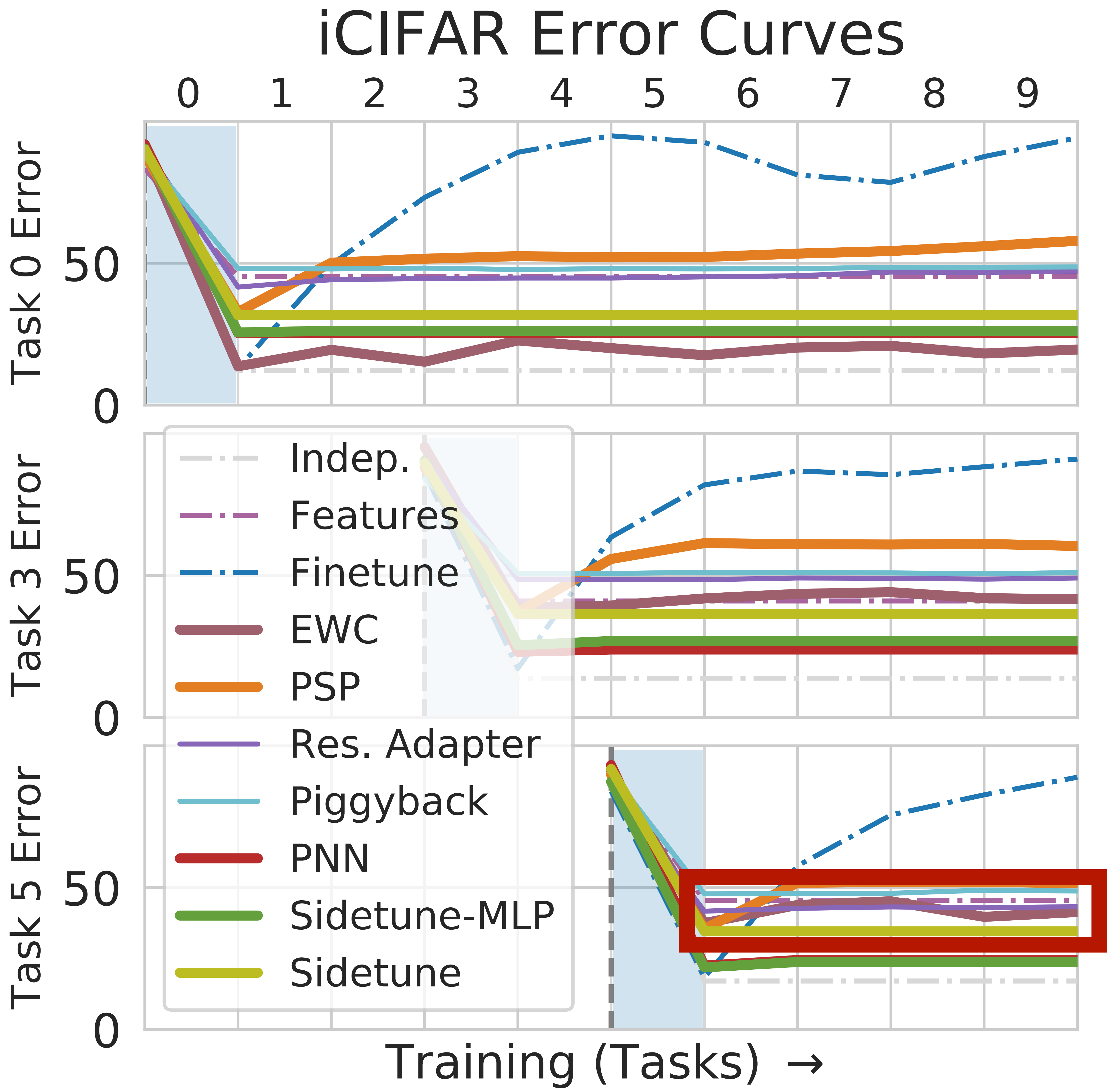
Results: catastrophic forgetting in incremental learning



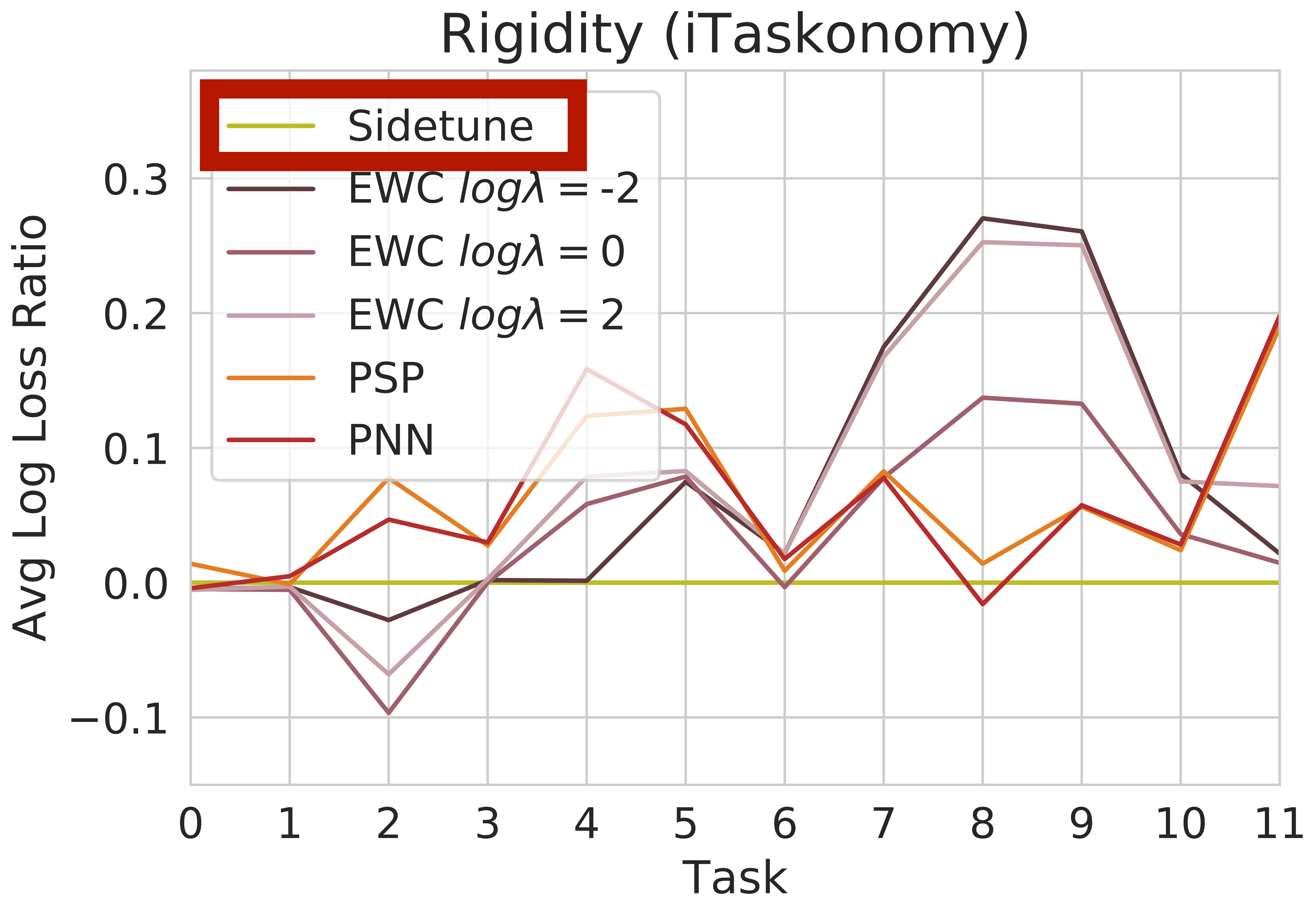


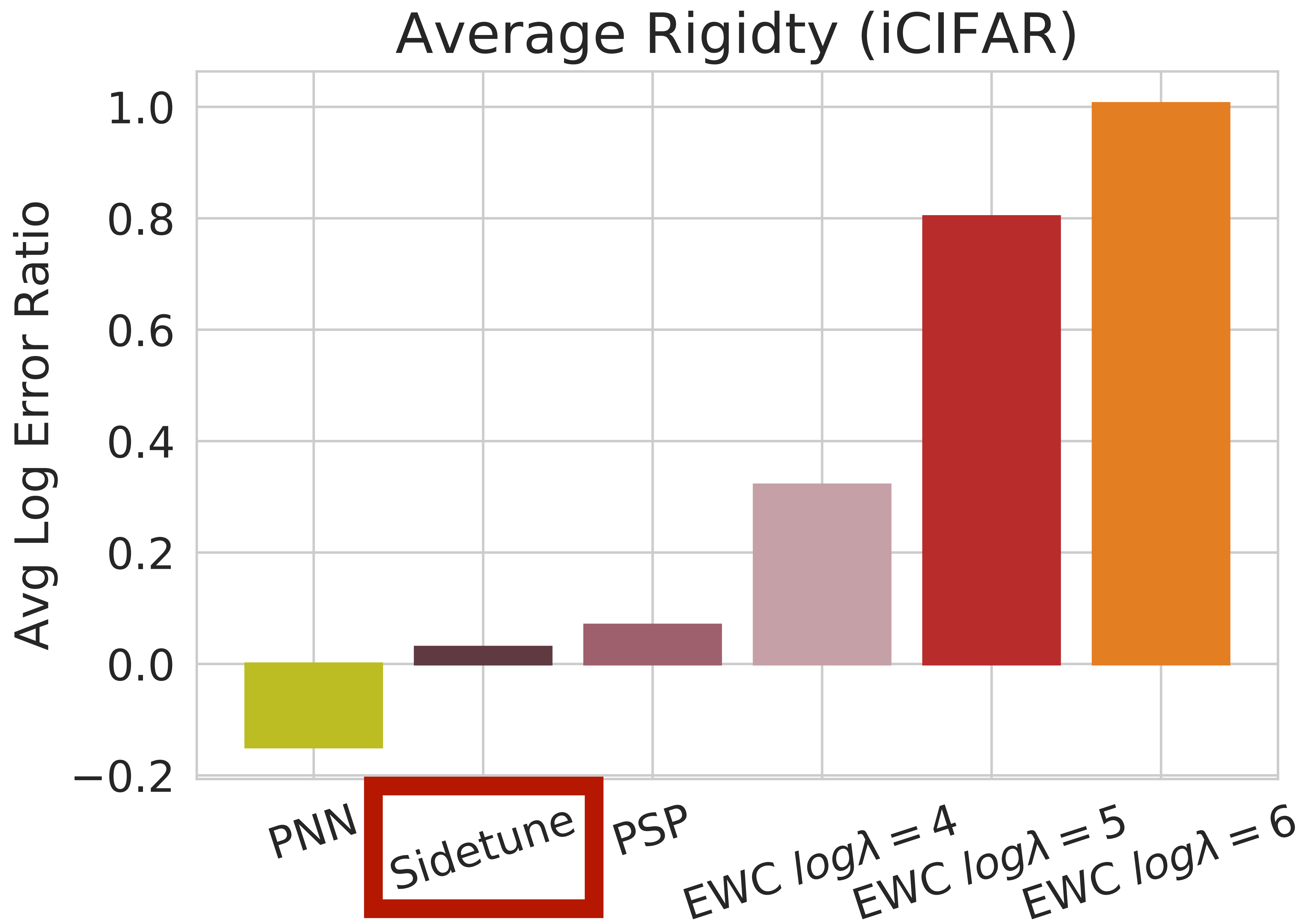
Results: catastrophic forgetting in incremental learning



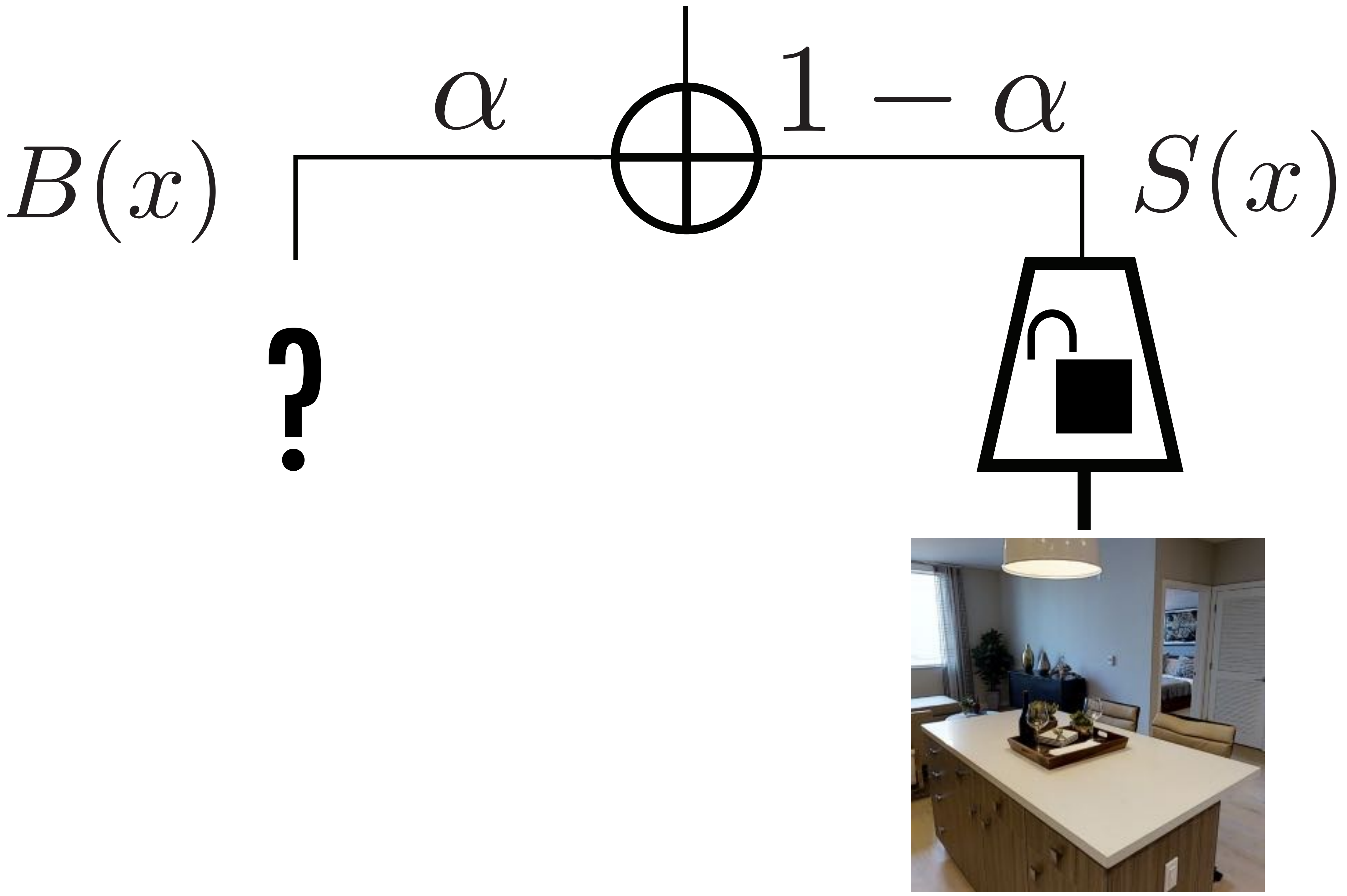


Results: rigidity in incremental learning



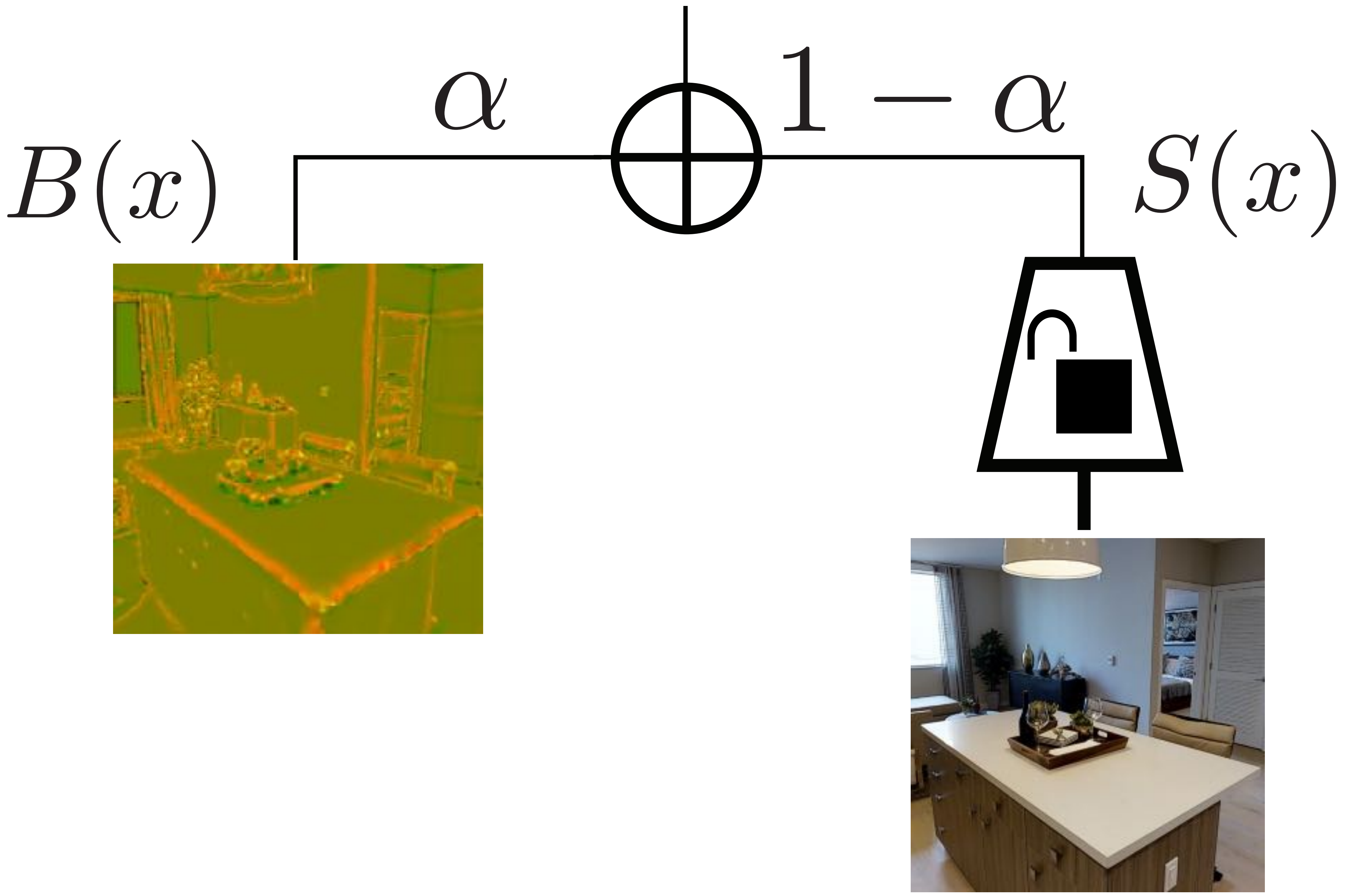


Side-Tuning: Beyond Network Adaptation



- Base model needn’t be a network
- Decision tree, or oracle for some other task

Side-Tuning: Beyond Network Adaptation

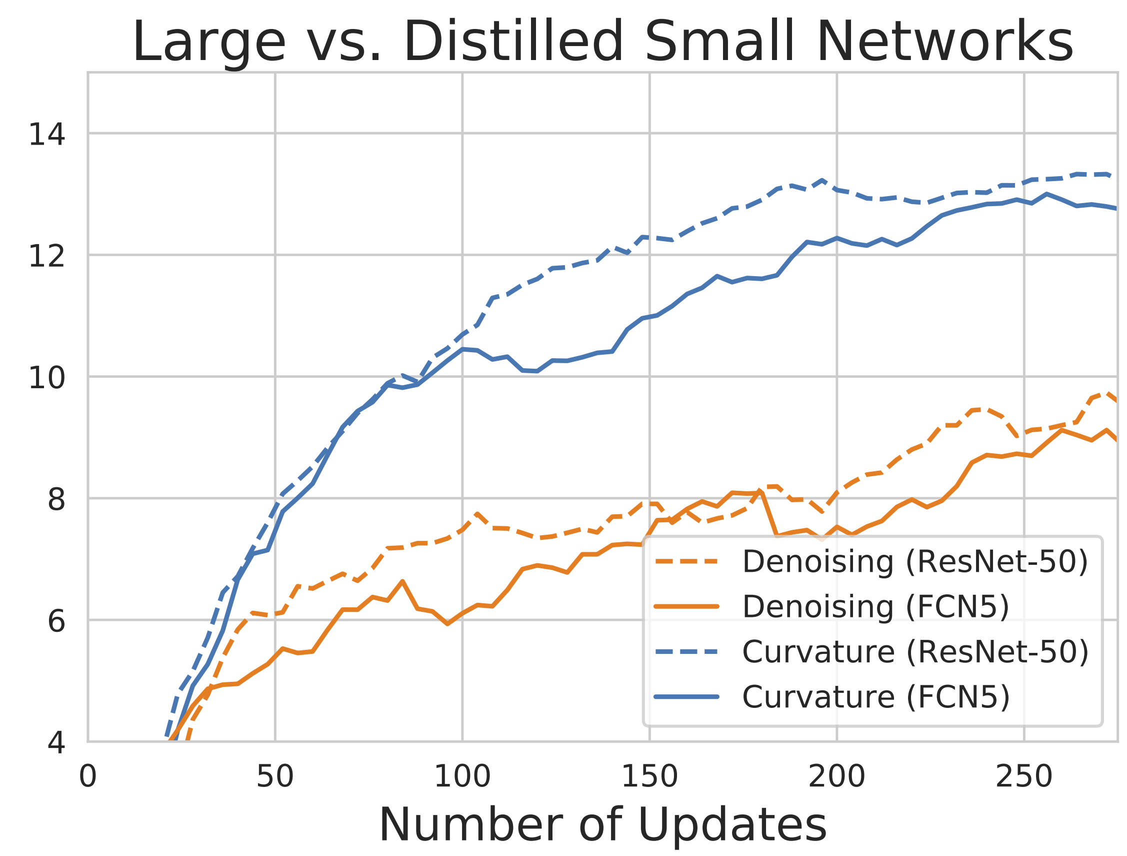


- Base model needn't be a network
- Decision tree, or oracle for some other task
- Can actual curvature label
- Works really well.

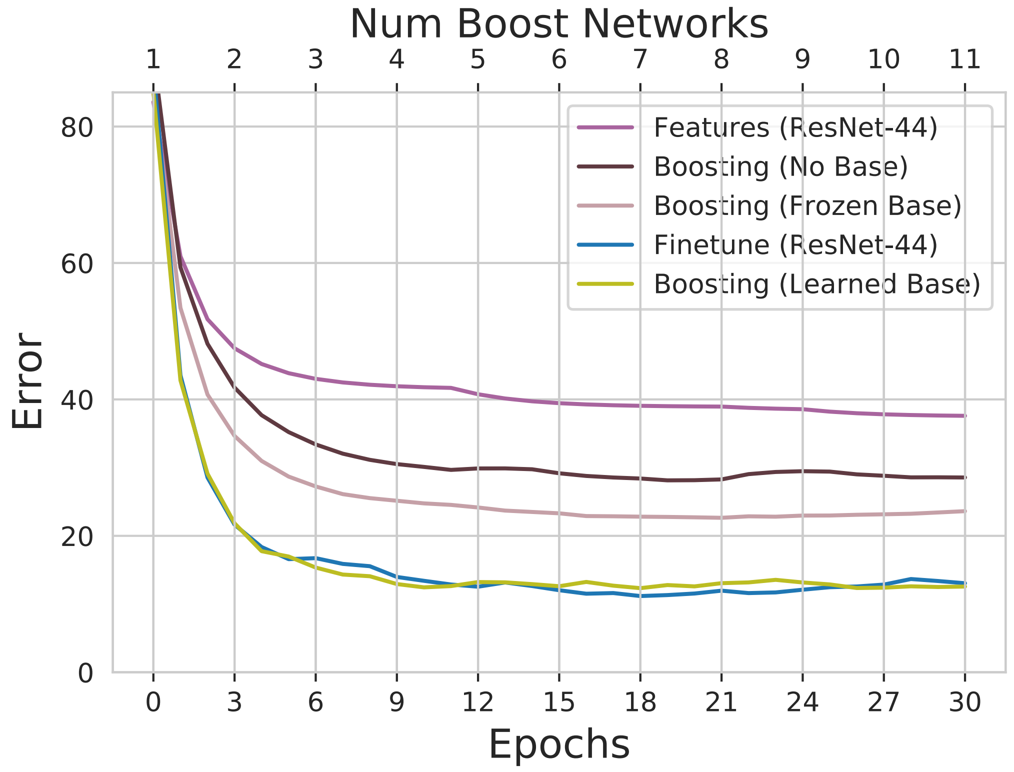
More in our paper + on the website:

sidetuning.berkeley.edu

Analysis of network size:



Comparison to boosting

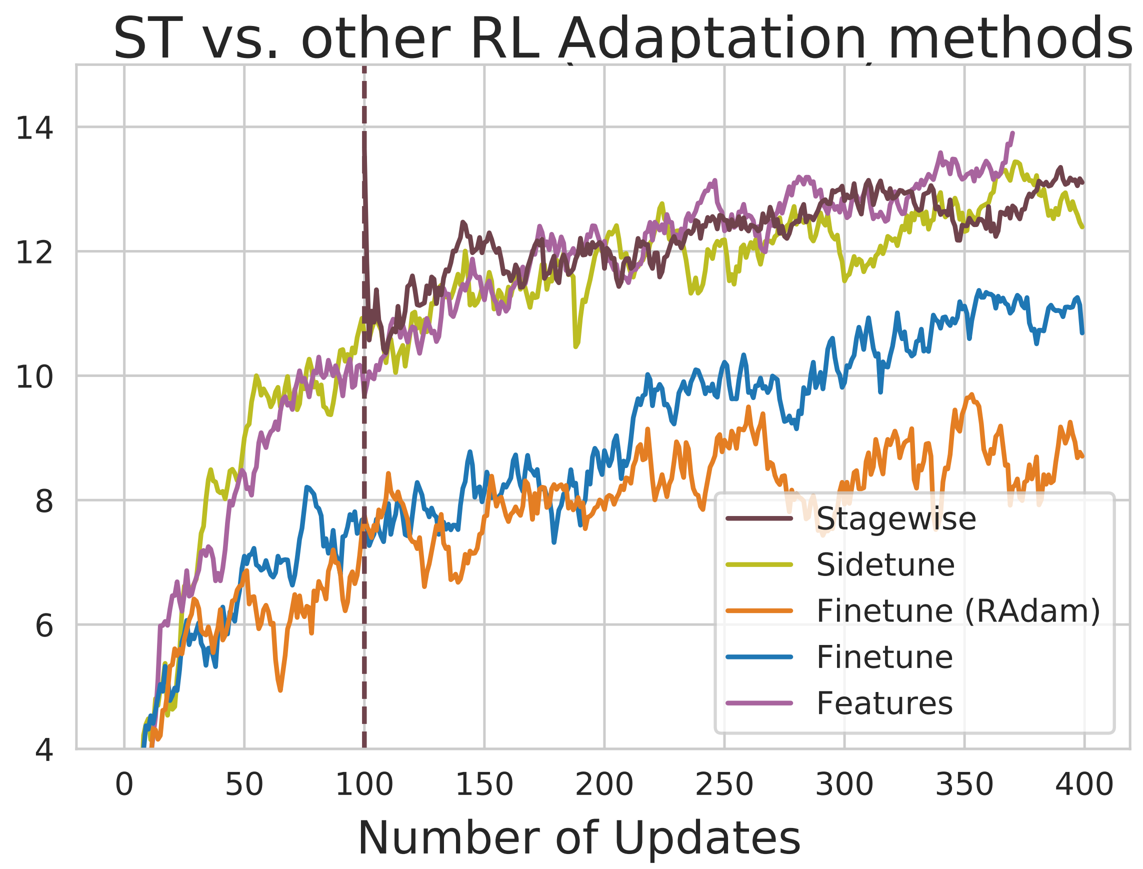


Ablation studies with different design choices

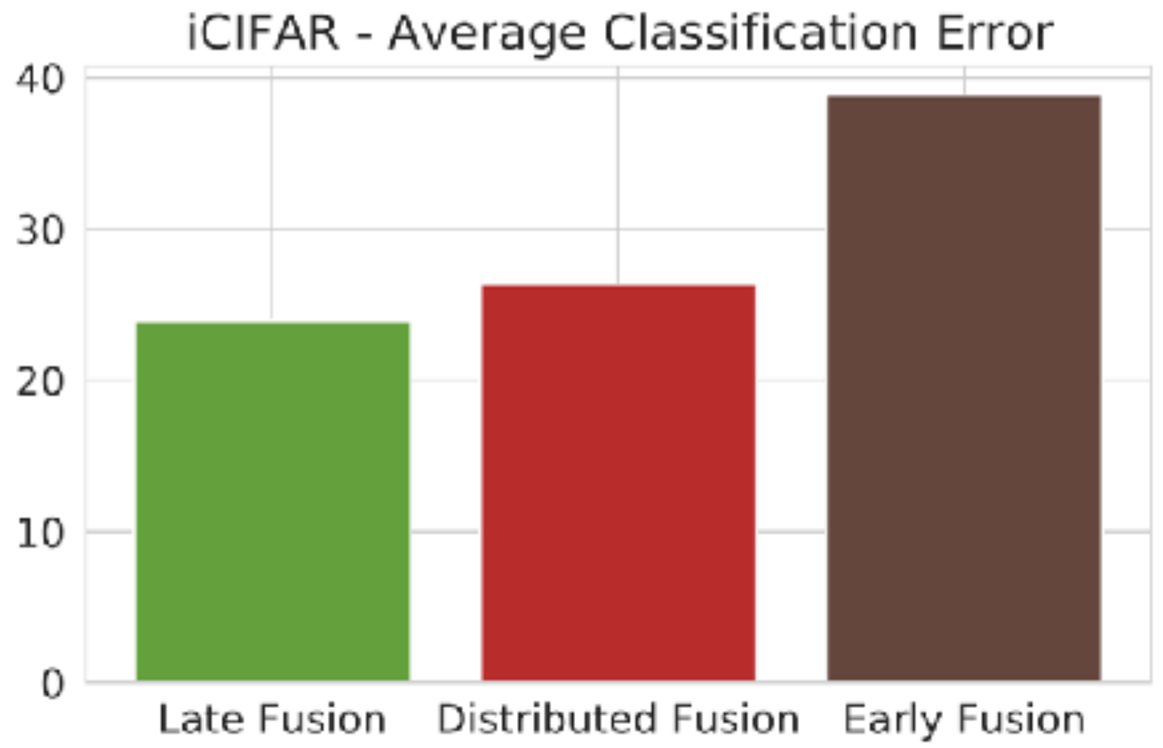
Method	Avg. Rank (↓) iTaskonomy
Product (Element-wise)	3.64
Summation (α-blending)	2.27
MLP ([29])	2.18
FiLM [21]	1.91

Method	Avg. Rank (↓) iTaskonomy
Base-Only	2.55
Side-Only	2.10
Side-tuning	1.36

Comparison to other adaptation methods in RL



Fusion: Early vs. Mid vs. Late



And more:

- Experiments with non-neural network base
- Analysis of α w.r.t. task relatedness
- Code (github repo)
- Environments to reproduce experiments (via docker)
- Full qualitative and quantitative results for all methods

Side-Tuning: A Baseline for Network Adaptation via Additive Side Networks

http://sidetuning.berkeley.edu

Jeffrey O. Zhang

Alexander Sax

Amir Zamir

Leonidas Guibas

Jitendra Malik