

Safeguarding Mutual Correction in Source-Free Domain Adaptation via Cut Statistics

26' NIPS

Seongjun Lee, Changhee Lee[†]

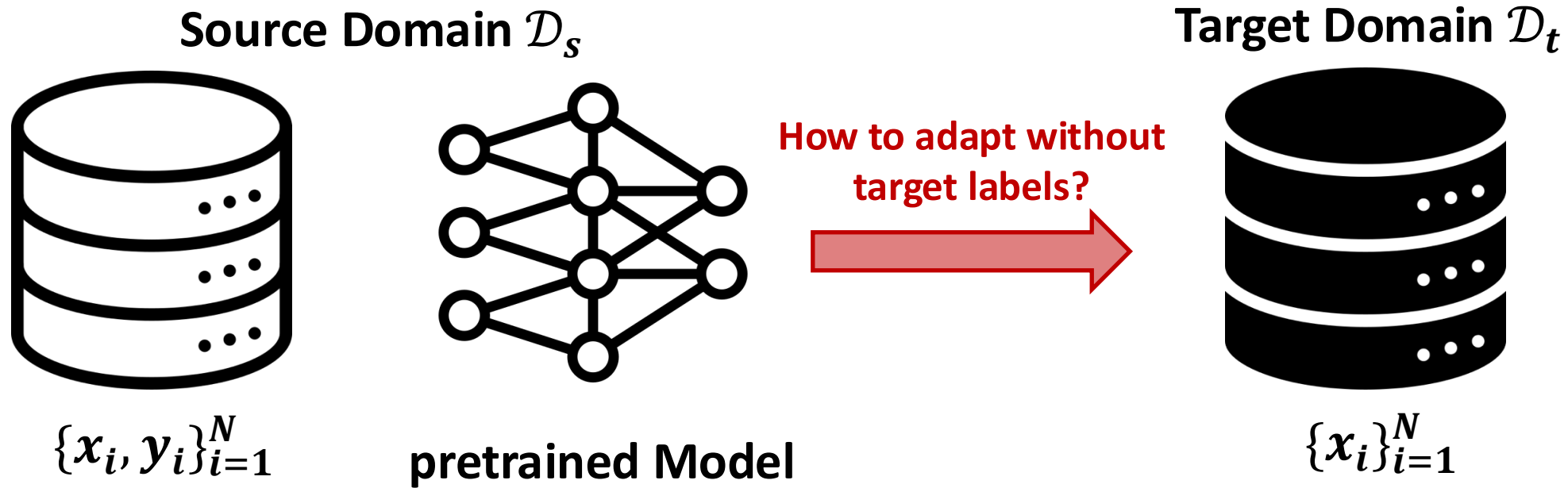
Seongjun Lee

Graduate Student, @Korea University



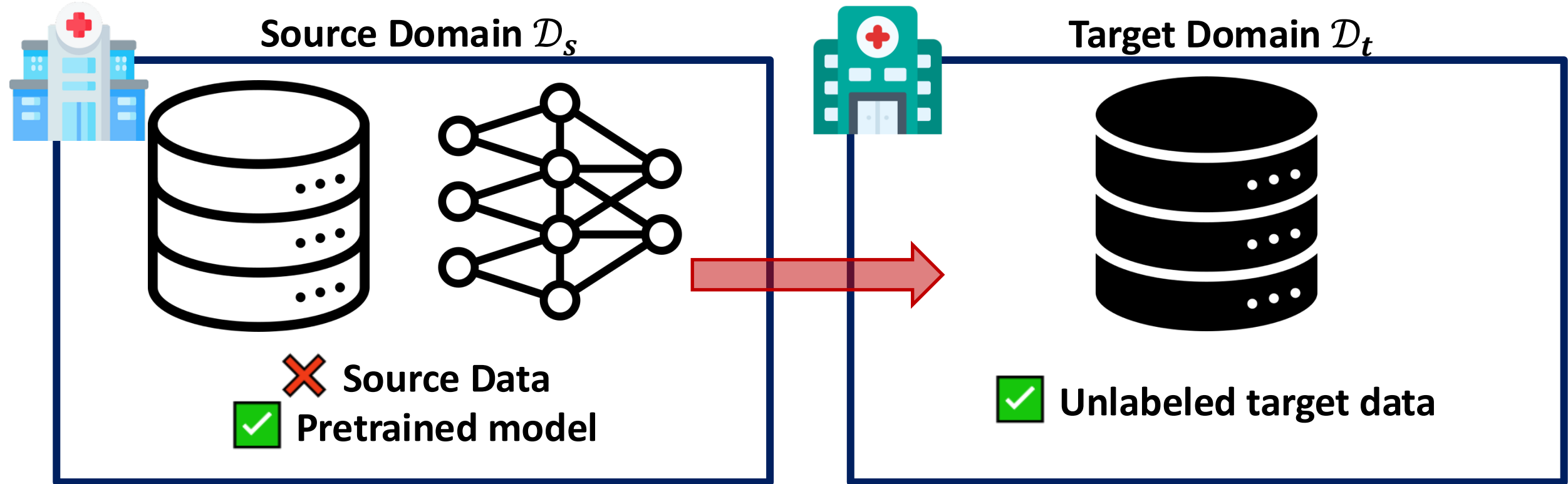
pyoung7307@korea.ac.kr

Unsupervised Domain Adaptation (UDA)



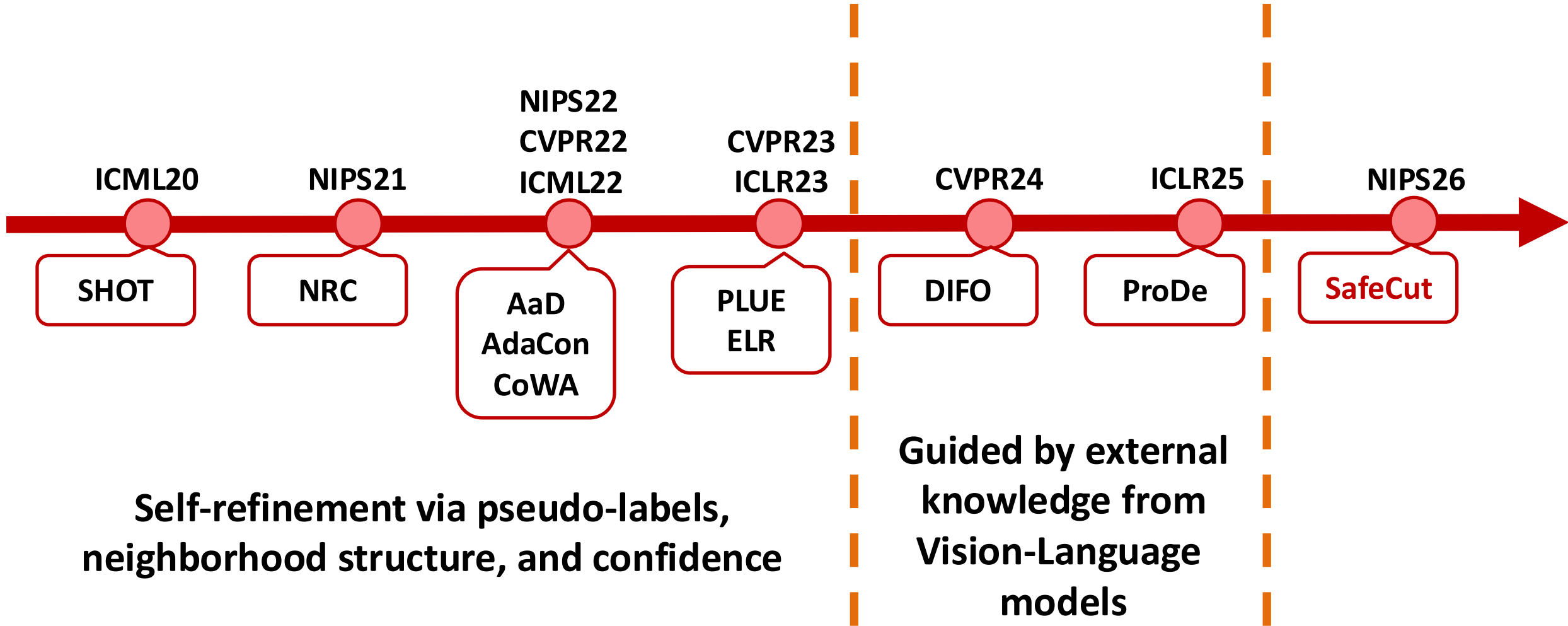
- **UDA:** Adapt a source-pretrained model to an **unlabeled target domain**.
- **Key assumption:** **Source data remain accessible** during adaptation.

Source Free Domain Adaptation (SFDA)



- **Real-world high-stakes scenarios:** Access to the original source data may be restricted due to **privacy and regulatory constraints**.
- **Key question:** How can a source-pretrained model adapt to an **unlabeled target domain without access to source data?**

Related Work



SafeCut

External knowledge sources can help mitigate the systematic errors of self-refinement.

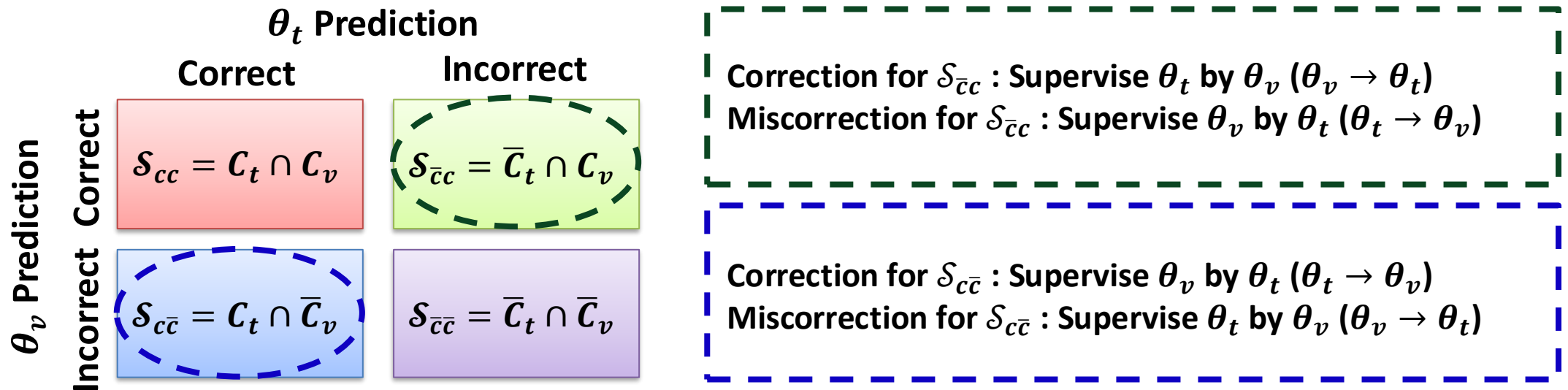
HOWEVER, a key question remains:

How should the source-pretrained model (target model) and the external knowledge source interact to effectively correct each other?

models

Mutual Correction: Correction vs. Miscorrection

- θ_t : target model (t denotes the target model)
- θ_v : peer model (i.e. an external knowledge source) (v denotes the target model)
- $\mathcal{C}_i = \{x \in \mathcal{D}_t : \hat{y}_i(x) = y(x)\}$ where $i \in \{t, v\}$, $\bar{\mathcal{C}}_i = \{x \in \mathcal{D}_t : \hat{y}_i(x) \neq y(x)\}$ where $i \in \{t, v\}$
- $\mathcal{S}_{cc} = \mathcal{C}_t \cap \mathcal{C}_v$, $\mathcal{S}_{c\bar{c}} = \mathcal{C}_t \cap \bar{\mathcal{C}}_v$, $\mathcal{S}_{\bar{c}c} = \bar{\mathcal{C}}_t \cap \mathcal{C}_v$, $\mathcal{S}_{\bar{c}\bar{c}} = \bar{\mathcal{C}}_t \cap \bar{\mathcal{C}}_v$



**However, target labels $y(x)$ are unavailable.
Therefore, these oracle partitions are unobservable.**

Key Concept : Cut statistics

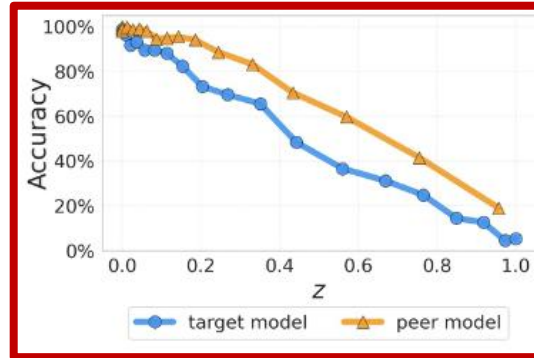
$$z_i(\mathbf{x}) = \sum_{\mathbf{x}' \in \mathcal{N}_K(\mathbf{x})} w(\mathbf{x}, \mathbf{x}') \cdot \mathbf{1}[\hat{y}_i(\mathbf{x}) \neq \hat{y}_i(\mathbf{x}')]]$$

- $i \in [t, v]$
- $\mathbf{x}' \in \mathcal{N}_K(\mathbf{x})$: K -nearest neighbors of $\mathbf{x}$
- $w(\mathbf{x}, \mathbf{x}')$: distance-based weighting function (closer neighbors receive higher weights)

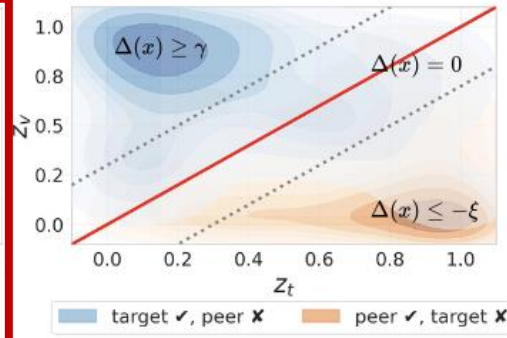
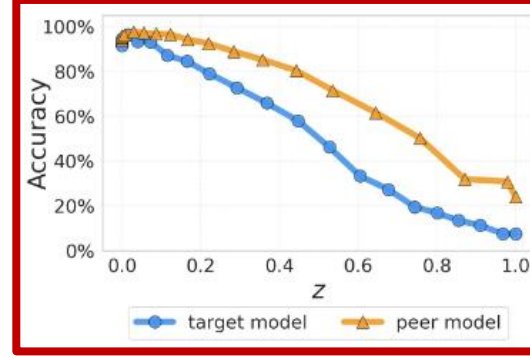
Intuition)

- **Low** cut statistic for $\mathbf{x} \rightarrow$ prediction **agrees with nearby samples (Locally smooth)**
- **High** cut statistic for $\mathbf{x} \rightarrow$ prediction **disagrees with nearby samples (Locally non-smooth)**

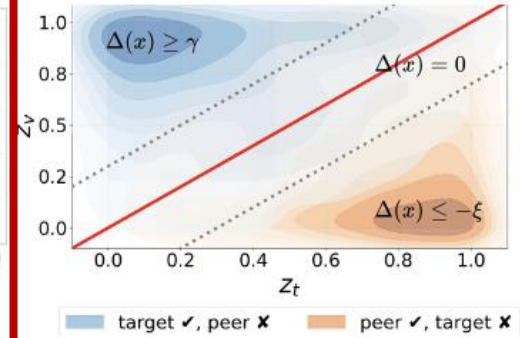
Observation: Lower Cut Tends to Be More Accurate



(a) Cl $\rightarrow$ Pr (Office-Home)



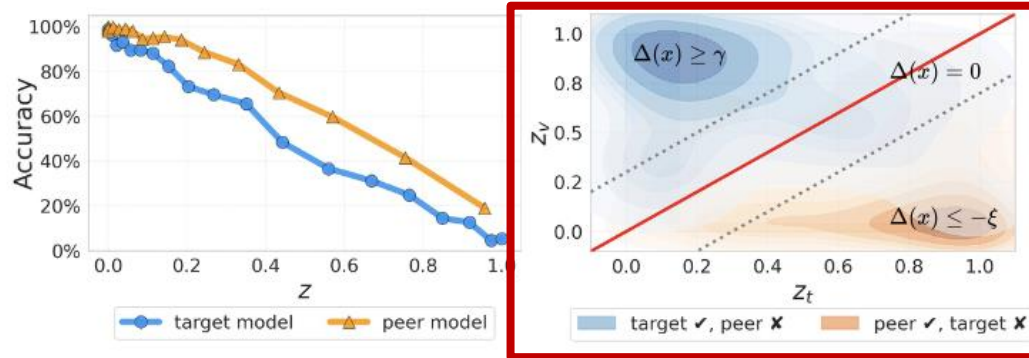
(b) S $\rightarrow$ C (DomainNet-126)



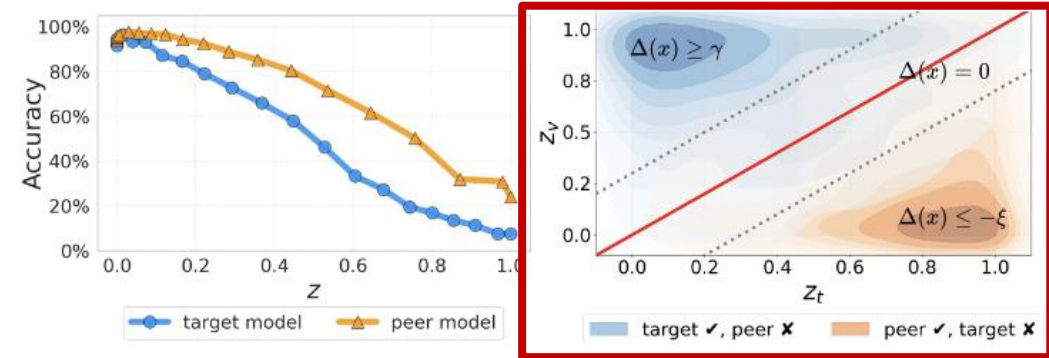
Observation 3.1. (Approximate Monotonicity) For both θ_i with $i \in \{t, v\}$, prediction accuracy is approximately monotonically decreasing with respect to the cut statistic z_i .

- Key implication : Lower cut $\rightarrow$ high prediction accuracy (approximately)

Observation: Which model should we trust?



(a) Cl $\rightarrow$ Pr (Office-Home)



(b) S $\rightarrow$ C (DomainNet-126)

Observation 3.1. (Approximate Monotonicity) For both θ_i with $i \in \{t, v\}$, prediction accuracy is approximately monotonically decreasing with respect to the cut statistic z_i .

- Key implication : Lower cut $\rightarrow$ high prediction accuracy (approximately)
- Cut-margin

$$\Delta(x) = z_v(x) - z_t(x)$$

- $\Delta(x) > 0$: θ_t is more likely to be reliable.
- $\Delta(x) < 0$: θ_v is more likely to be reliable.
- Larger $|\Delta(x)| \rightarrow$ stronger reliability gap

Can a sufficiently large reliability gap guarantee beneficial mutual correction?

$$\hat{\mathcal{S}}_{c\bar{c}}(\gamma) = \{x \in \mathcal{D}_t : \Delta(x) \geq \gamma\}, \quad \hat{\mathcal{S}}_{\bar{c}c}(\xi) = \{x \in \mathcal{D}_t : \Delta(x) \leq -\xi\}. \quad (3)$$

Proposition 3.1 (Reliability-Separated Regions). *There exist baseline margins γ_0, ξ_0 such that:*

$$\mathbb{P}(\hat{y}_t = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)) > \mathbb{P}(\hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)), \quad (4)$$

$$\mathbb{P}(\hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi_0)) > \mathbb{P}(\hat{y}_t = y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi_0)). \quad (5)$$

Please refer to Appendix A.1 for the full proof of Proposition 3.1.

Theorem 3.1 (Correction Dominance). *For the baseline margins γ_0, ξ_0 satisfying Proposition 3.1:*

$$\mathbb{P}(\hat{y}_t = y \wedge \hat{y}_v \neq y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)) > \mathbb{P}(\hat{y}_t \neq y \wedge \hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0)), \quad (6)$$

$$\mathbb{P}(\hat{y}_v = y \wedge \hat{y}_t \neq y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi_0)) > \mathbb{P}(\hat{y}_v \neq y \wedge \hat{y}_t = y \mid x \in \hat{\mathcal{S}}_{\bar{c}c}(\xi_0)). \quad (7)$$

Please refer to Appendix A.1 for the full proof of Theorem 3.1.

Corollary 3.1 (Margin-Driven Strict Dominance). *For sufficiently strict margins $\gamma_{\text{strict}}, \xi_{\text{strict}}$, the ratio of true correction to misclassification increases compared to the baseline margins γ_0, ξ_0 :*

$$\frac{\mathbb{P}(\hat{y}_t = y \wedge \hat{y}_v \neq y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_{\text{strict}}))}{\mathbb{P}(\hat{y}_t \neq y \wedge \hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_{\text{strict}}))} \geq \frac{\mathbb{P}(\hat{y}_t = y \wedge \hat{y}_v \neq y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0))}{\mathbb{P}(\hat{y}_t \neq y \wedge \hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0))} > 1. \quad (8)$$

An analogous result holds for $\hat{\mathcal{S}}_{\bar{c}c}(\cdot)$ with thresholds ξ_0 and ξ_{strict} .

$$\hat{\mathcal{S}}_{c\bar{c}}(\gamma) = \{x \in \mathcal{D}_t : \Delta(x) \geq \gamma\}, \quad \hat{\mathcal{S}}_{\bar{c}c}(\xi) = \{x \in \mathcal{D}_t : \Delta(x) \leq -\xi\}. \quad (3)$$

Optimal Thresholds

HOWEVER selecting optimal thresholds γ and ξ is infeasible in a fully unlabeled setting.

To bridge this gap, we replace hard thresholding with a soft gating mechanism that assigns a supervision weight monotonically increasing in the cut margin $|\Delta(x)|$

$$\frac{\mathbb{P}(y_t = y \wedge y_v \neq y \mid x \in \mathcal{S}_{cc}(\gamma_{strict}))}{\mathbb{P}(\hat{y}_t \neq y \wedge \hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_{strict}))} \geq \frac{\mathbb{P}(y_t = y \wedge y_v \neq y \mid x \in \mathcal{S}_{cc}(\gamma_0))}{\mathbb{P}(\hat{y}_t \neq y \wedge \hat{y}_v = y \mid x \in \hat{\mathcal{S}}_{c\bar{c}}(\gamma_0))} > 1. \quad (8)$$

An analogous result holds for $\hat{\mathcal{S}}_{\bar{c}c}(\cdot)$ with thresholds ξ_0 and ξ_{strict} .

$$\hat{z}_i(x) = \frac{\sum_{x' \in \mathcal{N}_K^{\text{mut}}(x)} \exp(-d(\phi_i(x), \phi_i(x'))) \cdot \mathbf{1}[\hat{y}_i(x) \neq \hat{y}_i(x')]}{\sum_{x' \in \mathcal{N}_K^{\text{mut}}(x)} \exp(-d(\phi_i(x), \phi_i(x')))}$$

- $\phi_i(x)$: penultimate-layer representation of model i
- $d(\cdot)$: cosine distance
- $x' \in \mathcal{N}_K^{\text{mut}}(x)$: mutual K -nearest neighbor of x (i.e. $x' \in \mathcal{N}_K(x)$ and $x \in \mathcal{N}_K(x')$)

- **The reliability score**

$$r_i(x) = \exp(-\hat{z}_i(x)) \in (0,1]$$

Lower cut $\rightarrow$ higher reliability score

- **Note)** Naively updating the cut statistics requires an additional forward pass over the entire target dataset at every epoch. **To avoid this overhead, we reuse latent features cached from the previous epoch.** So, we further stabilize it with EMA:

$$\bar{r}_i^{(e)} = m\bar{r}_i^{(e-1)} + (1 - m)r_i^{(e)}$$

Margin-based Reliability-Gated Supervision

$$T_{v \rightarrow t}(x) = \sigma(\beta(\bar{r}_v(x) - \bar{r}_t(x))), \quad T_{t \rightarrow v}(x) = \sigma(\beta(\bar{r}_t(x) - \bar{r}_v(x)))$$

- σ denotes the sigmoid function and $T_{v \rightarrow t}(x) + T_{t \rightarrow v}(x) = 1$ by the symmetry of the sigmoid.
- β determines the weight sharpness
- **Gating the cross supervision**

$$\mathcal{L}_{cross}^t(x) = T_{v \rightarrow t}(x) \cdot D_{KL}(\text{sg}(\mathbf{p}_v) \parallel \mathbf{p}_t), \quad \mathcal{L}_{cross}^v(x) = T_{t \rightarrow v}(x) \cdot D_{KL}(\text{sg}(\mathbf{p}_t) \parallel \mathbf{p}_v)$$

- **Temporal Anchoring**

$$\mathcal{L}_{anc}^t(x) = T_{t \rightarrow v}(x) \cdot D_{KL}(\tilde{\mathbf{p}}_t \parallel \mathbf{p}_t), \quad \mathcal{L}_{anc}^v(x) = T_{v \rightarrow t}(x) \cdot D_{KL}(\tilde{\mathbf{p}}_v \parallel \mathbf{p}_v)$$

$\tilde{\mathbf{p}}_i$ denotes the predictions from the previous epoch.

- **Overall objective**

$$\mathcal{L}_t = \underbrace{\mathbb{E}_x [T_{v \rightarrow t} \cdot D_{KL}(\text{sg}(\mathbf{p}_v) \parallel \mathbf{p}_t) + T_{t \rightarrow v} \cdot D_{KL}(\tilde{\mathbf{p}}_t \parallel \mathbf{p}_t)]}_{\text{sample-wise supervision}} + \underbrace{\lambda_{\text{IIC}} \mathcal{L}_{\text{IIC}} - \lambda_{\text{div}} H(\bar{\mathbf{p}}_t)}_{\text{batch-wise regularization}} \quad (13)$$

$$\mathcal{L}_v = \underbrace{\mathbb{E}_x [T_{t \rightarrow v} \cdot D_{KL}(\text{sg}(\mathbf{p}_t) \parallel \mathbf{p}_v) + T_{v \rightarrow t} \cdot D_{KL}(\tilde{\mathbf{p}}_v \parallel \mathbf{p}_v)]}_{\text{sample-wise supervision}} + \underbrace{\lambda_{\text{IIC}} \mathcal{L}_{\text{IIC}} - \lambda_{\text{div}} H(\bar{\mathbf{p}}_v)}_{\text{batch-wise regularization}} \quad (14)$$

- **Main Benchmark Evaluation**
 - **Datasets** : Office-31, OfficeHome, DomainNet, VisDA
 - **Target Model** : ResNet-50 (Office-31, OfficeHome, DomainNet) / ResNet-101 (VisDA)
 - **Peer Model** : CLIP ViT-B/32 with **Context Optimization**(CoOp)
 - **Protocol** : Closed-set SFDA | Metric: Accuracy (%) | Last-epoch evaluation
- **Peer Model Generality Experiment**
 - **Target Model** : ResNet-50
 - **Peer Model** : ResNet-50 pretrained on a different source domain
 - **Goal**: Test whether SafeCut works beyond ViL-based peers
- **Ablation & Training Dynamics Analysis**
- **Healthcare Scenarios**
 - **Datasets**:
 - Wilds-Camelyon17 : histopathology binary classification
 - EyePACS→APTOS2019 : diabetic retinopathy severity classification
 - **Goal** : Evaluate SafeCut under weak ViL transfer in high-stakes scenarios.

Experiments: Main Benchmark Evaluation

Table 1: Closed-set SFDA results (Acc, %) on **Office-31**.

Method	Venue	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
Source	–	79.1	76.5	59.9	95.5	61.4	98.8	78.5
SHOT	ICML20	94.4	88.4	73.2	96.7	74.9	99.8	87.9/88.6
NRC	NIPS21	92.4	94.1	75.6	97.2	74.6	98.6	88.7/89.4
AaD	NIPS22	87.6	86.5	58.9	90.6	58.7	90.6	78.8/89.9
AdaCon	CVPR22	87.8	86.3	71.3	96.7	67.2	97.6	84.5/–
CoWA	ICML22	88.0	90.9	74.6	98.1	75.8	99.4	87.8/90.3
PLUE	CVPR23	89.4	89.2	70.6	97.1	67.8	96.6	85.1/–
ELR	ICLR23	92.2	92.2	76.5	97.4	76.1	99.6	89.0/89.6
DIFO-V	CVPR24	96.2	95.0	83.2	96.1	83.0	99.2	92.1/92.5
ProDe-V	ICLR25	96.8	96.6	82.9	97.2	82.0	99.0	92.4/92.6
SafeCut	NIPS26	98.6	95.2	83.9	98.2	83.3	100.0	93.2

Table 2: Comparison on the Office-Home and VisDA datasets (Acc, %).

Method	Venue	Office-Home														VisDA
		Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.	Sy→Re	
Source	–	43.6	67.0	73.9	49.9	60.2	62.6	51.7	40.9	72.6	64.1	46.3	78.1	59.2	49.1	
SHOT	ICML20	55.2	78.5	81.3	68.9	79.0	78.8	68.2	54.3	81.7	73.5	59.1	83.6	71.8/71.8	82.3/82.9	
NRC	NIPS21	54.7	77.1	79.4	64.7	76.2	75.3	63.3	53.6	78.8	70.2	56.5	82.5	69.4/72.2	83.4/85.9	
AaD	NIPS22	57.7	78.5	80.8	68.0	79.6	78.9	66.5	56.6	81.6	70.9	57.3	83.7	71.7/72.7	78.8/88.0	
AdaCon	CVPR22	49.5	76.1	77.0	63.5	73.2	74.4	60.7	47.1	77.9	67.8	49.7	80.3	66.4/–	67.2/78.7	
CoWA	ICML22	56.7	79.0	80.9	68.6	78.7	79.8	68.1	55.6	82.6	72.4	61.7	83.6	72.3/72.5	84.3/86.9	
PLUE	CVPR23	46.3	61.9	71.6	59.2	63.7	65.8	58.0	44.3	72.1	63.0	49.2	76.6	61.0/–	86.6/90.0	
ELR	ICLR23	55.9	77.6	79.9	61.4	76.8	76.5	62.4	53.3	80.9	68.9	58.7	83.6	69.7/72.6	85.8/85.8	
DIFO-V	CVPR24	70.7	90.5	89.2	82.4	90.2	89.2	80.6	71.1	89.4	82.9	71.0	91.0	83.2/83.1	89.8/90.3	
ProDe-V	ICLR25	71.8	92.2	90.5	82.7	91.6	90.6	82.4	72.8	90.8	82.8	72.7	92.0	84.4/84.5	90.6/91.0	
SafeCut	NIPS26	76.9	93.0	91.9	85.9	92.5	91.6	85.0	76.7	92.0	85.7	76.1	93.2	86.7	90.9	

Table 3: Comparison on the DomainNet dataset (Acc, %).

Method	Venue	DomainNet													Avg.
		C→P	C→R	C→S	P→C	P→R	P→S	R→C	R→P	R→S	S→C	S→P	S→R		
Source	–	44.7	59.9	47.5	53.3	75.3	46.2	55.1	62.6	46.4	55.2	50.8	59.6	54.7	
SHOT	ICML20	63.1	78.6	60.1	68.4	80.7	62.0	68.0	68.3	58.7	70.8	64.9	77.7	68.4/–	
NRC	NIPS21	60.0	75.7	55.4	59.6	80.8	58.0	60.0	67.3	54.3	66.9	65.0	77.7	65.1/–	
AaD	NIPS22	62.4	79.1	59.3	67.1	81.6	62.6	69.7	66.7	58.7	71.1	63.6	79.7	68.5/–	
AdaCon	CVPR22	56.9	64.1	51.9	57.5	64.9	52.9	60.7	64.4	51.5	62.7	60.1	61.1	59.0/–	
CoWA	ICML22	64.7	80.6	63.0	70.1	80.6	64.3	70.5	69.5	63.1	71.7	67.3	80.0	70.5/–	
PLUE	CVPR23	55.5	62.0	53.9	57.2	68.4	53.0	58.5	61.8	52.3	60.8	58.5	65.6	58.9/–	
ELR	ICLR23	64.8	76.1	61.3	66.7	81.8	63.7	69.1	69.2	62.0	72.3	66.8	77.5	69.3/79.4	
DIFO-V	CVPR24	76.5	87.8	75.4	78.8	87.7	75.7	80.0	77.3	75.6	79.2	76.4	87.8	79.9/80.0	
ProDe-V	ICLR25	80.9	91.5	77.4	84.1	91.3	76.9	84.2	81.3	76.6	84.0	81.2	91.2	83.4/85.0	
SafeCut	NIPS26	82.8	92.3	81.8	86.2	92.2	81.0	86.9	82.9	81.7	86.6	83.0	92.3	85.8	

Experiments: Peer Model Generality

Table 4: Mutual correction between two ResNet-50 models from different source domains (Office-Home, Acc, %). Bef./Aft.: accuracy before/after adaptation.

Target: Ar						Target: Cl						Target: Pr						Target: Rw					
		θ_1		θ_2				θ_1		θ_2				θ_1		θ_2				θ_1		θ_2	
θ_1	θ_2	Bef.	Aft.	Bef.	Aft.	θ_1	θ_2	Bef.	Aft.	Bef.	Aft.	θ_1	θ_2	Bef.	Aft.	Bef.	Aft.	θ_1	θ_2	Bef.	Aft.	Bef.	Aft.
Cl	Pr	49.9	72.2	51.7	72.2	Ar	Pr	43.6	59.8	40.9	59.9	Ar	Cl	67.0	81.3	60.2	81.2	Ar	Cl	73.9	83.1	62.6	83.1
Cl	Rw	49.9	75.8	64.2	75.8	Ar	Rw	43.6	61.7	46.3	61.6	Cl	Rw	60.2	86.4	78.1	86.4	Ar	Pr	73.9	84.3	72.6	84.2
Pr	Rw	51.7	74.4	64.2	74.4	Pr	Rw	40.9	60.4	46.3	60.4	Ar	Rw	67.0	83.0	78.1	83.0	Cl	Pr	62.6	83.9	72.6	83.9

Table 5: Ablation and sensitivity analysis (Avg. Acc, %).
 $\mathcal{L}_c$: cross-supervision KL, $\mathcal{L}_a$: temporal anchoring.

(a) Component ablation

	O31	OH	DN
$\mathcal{L}_c = 0, \mathcal{L}_a = 0$	91.2	81.8	72.1
$\mathcal{L}_c = 0, \mathcal{L}_a = 1$	91.2	83.3	78.7
$\mathcal{L}_c = 1, \mathcal{L}_a = 0$	92.4	86.0	83.6
$\mathcal{L}_c = 1, \mathcal{L}_a = 1$	92.1	86.2	84.3
SafeCut	93.2	86.7	85.8

(c) CLIP adaptation

	O31	OH	DN
Zero-shot	79.4	79.0	80.2
Adapted	92.8	85.8	84.0

(b) Variant

Variant	O31	OH	DN
Frozen θ_v	89.0	84.8	84.6
Entropy gating	92.5	86.3	85.0
SafeCut	93.2	86.7	85.8
Oracle partition S	96.0	93.3	92.7

(d) β sensitivity (OH)

β	OH
2	86.7
4	86.6
6	86.7

Experiments: Qualitative Analysis

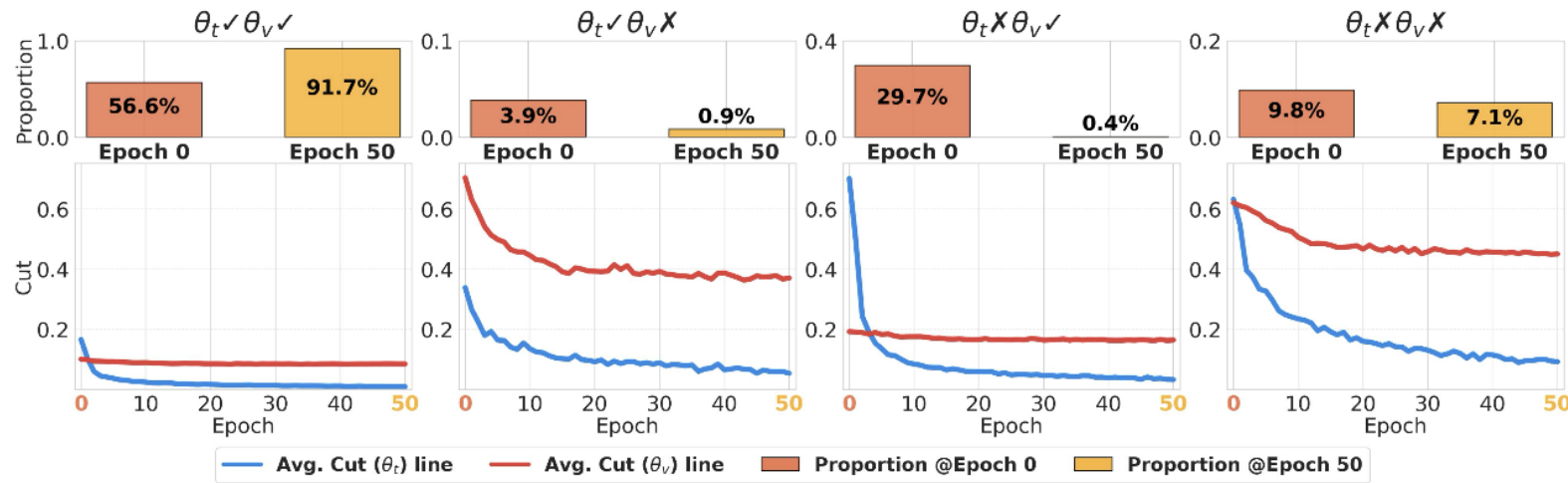


Figure 3: **Correctness pattern (top) and average cut trajectories (bottom) on Office-Home CI→Pr.** Samples are grouped by their initial state and tracked throughout adaptation.

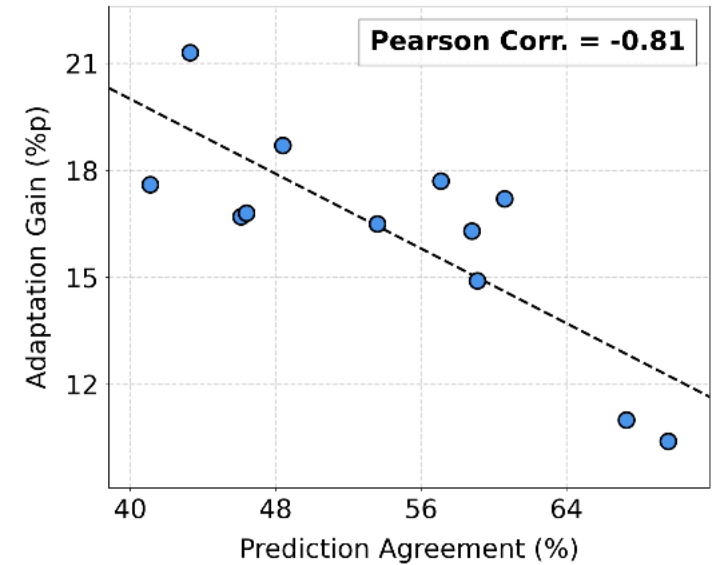


Figure 4: **Peer complementarity.** Prediction agreement vs. adaptation gain in the dual-ResNet setting of Table 4.

Experiments: Healthcare Scenarios

Table 6: SFDA on Camelyon17 (Acc, %).

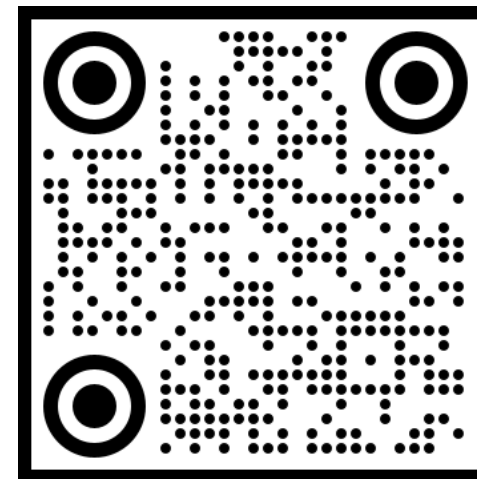
Method	Target	CLIP
Source	64.14	50.01
ProDe	76.27	50.00
$\mathcal{L}_c = 1, \mathcal{L}_a = 0$	77.29	77.20
$\mathcal{L}_c = 1, \mathcal{L}_a = 1$	87.96	87.78
SafeCut	94.96	94.20

Table 7: EyePACS→APTOS2019 (Acc, %).

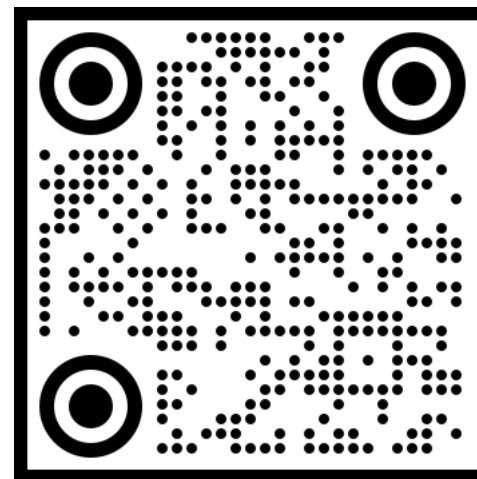
Method	Target	CLIP
Source	39.0	25.4
ProDe	35.2	35.3
SafeCut	46.0	45.9

1. **Extension to a multi-peer setting**
2. **Connection to knowledge distillation in LLMs**
3. **Theoretical analysis of fixed-point convergence in mutual correction**

Thank you



LinkedIn



arXiv_SafeCut