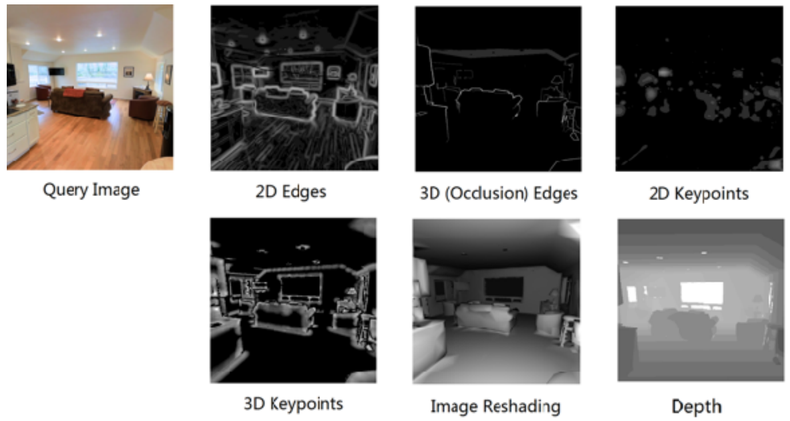
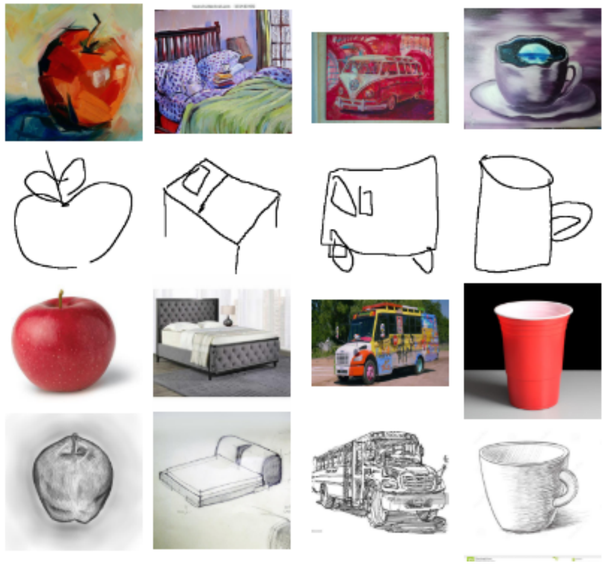
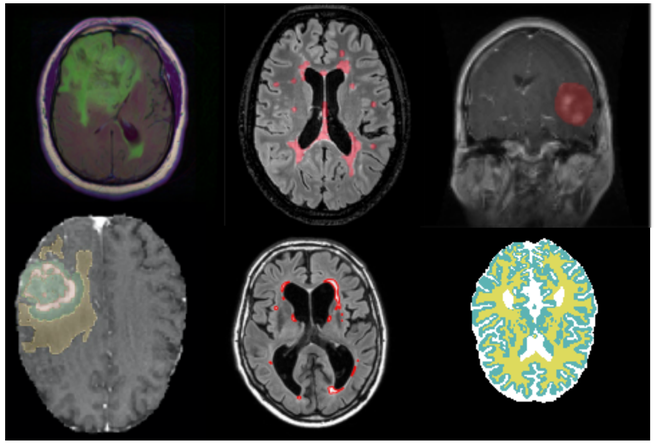
TBSI 清华-伯克利深圳学院 Tsinghua-Berkeley Shenzhen Institute







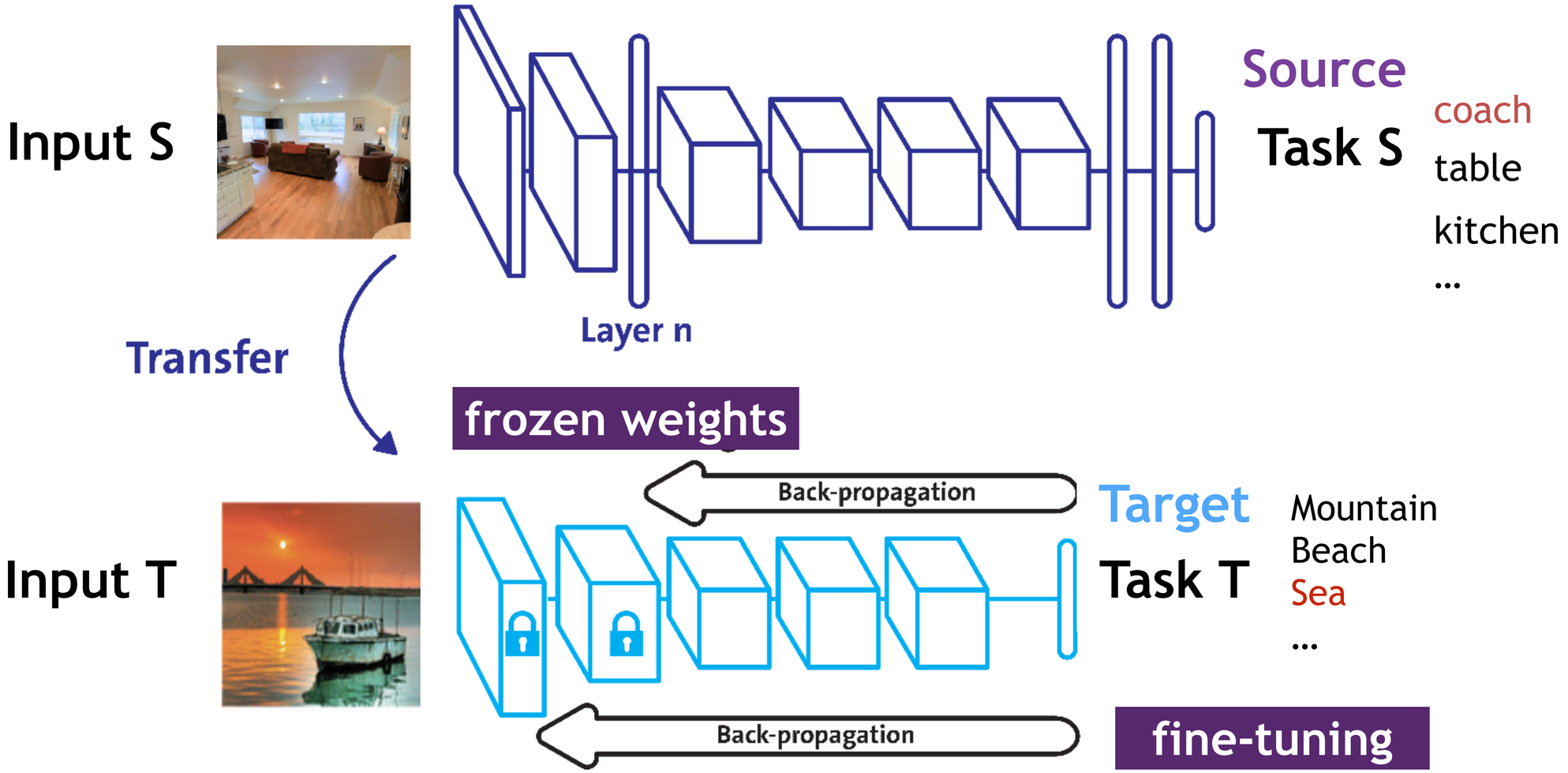
Measuring Transferability in Transfer Learning
迁移学习中的可迁移性度量

Yang Li

Dec 13,2021

中科院深圳先进技术研究院

Transfer Learning using a Pre-Trained Model

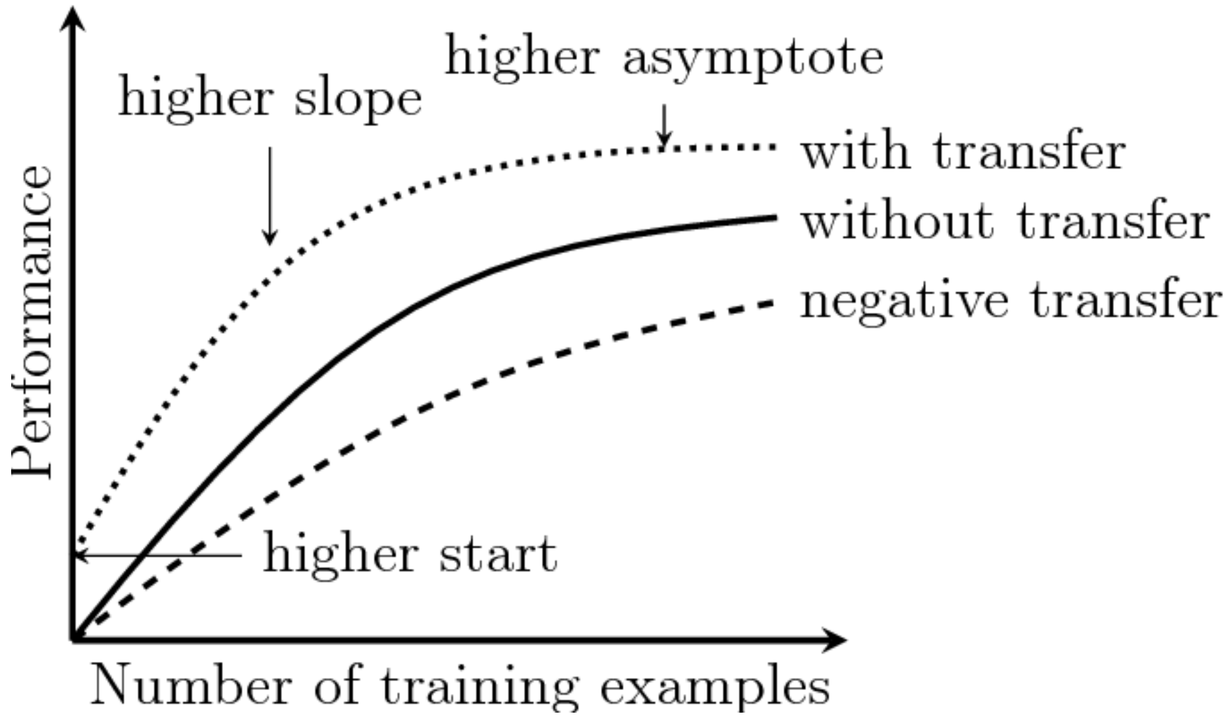


- Commonly used in computer vision,NLP etc

Transfer Learning with Pre-Trained Model

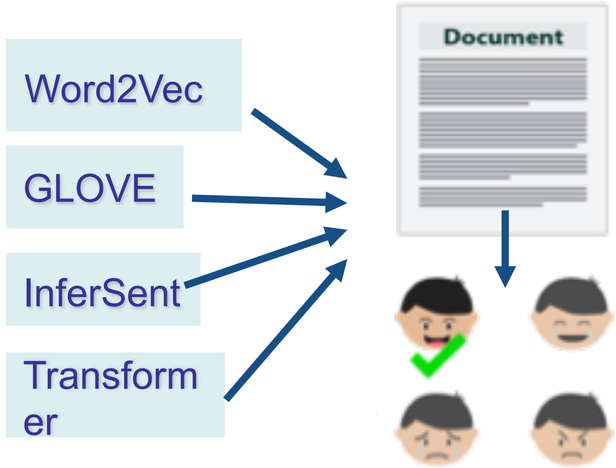
- Improve target training efficiency, reduce number of target labeled data needed

Assumes represenation of S is *transferable* to T



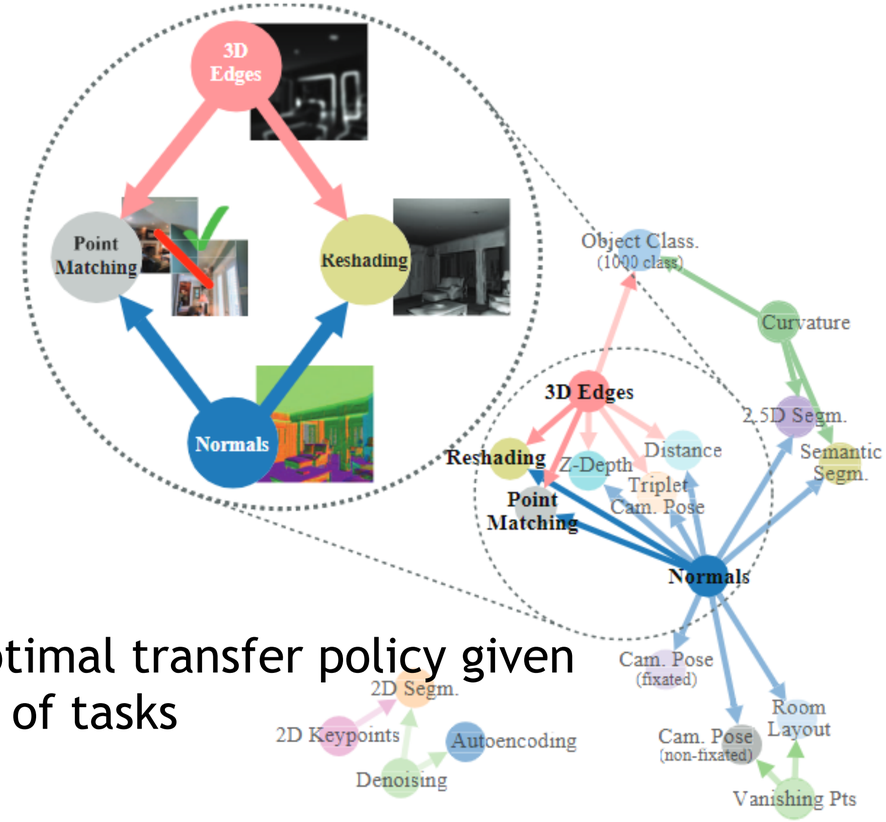
Why is Knowing Task Transferability Important?

- Source-task selection



e.g. Select the best word/sentence encoder for NLP tasks

- Task transfer policy learning

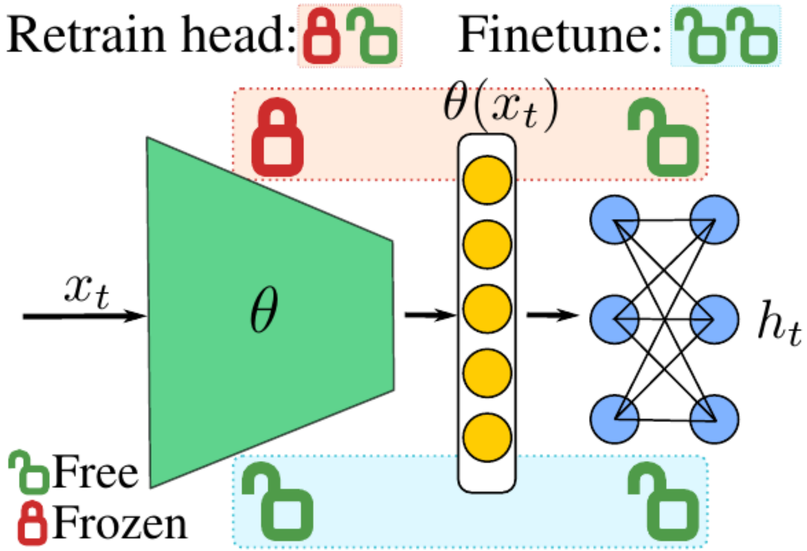


e.g. Find optimal transfer policy given a collection of tasks

- Learn more transferable features

(Empirical) Transferability

Given source data (X_s, Y_s) and source model (θ_s, h_s), compute **expected log loss (or accuracy)** of the transferred network on target data



$$\mathrm{Trf}(S \to T) = \begin{cases} \mathbb{E}\left[\log\ P(y_t|x_t;\theta_s,h_t)\right] & \text{(Retrain head)} \\ \mathbb{E}\left[\log\ P(y_t|x_t;\theta_t,h_t)\right] & \text{(Finetune)} \end{cases}$$

Related Works — Empirical Transferability

- Feature transferability in Neural Network (Yosinski 2014)
- Taskonomy (Zamir et. al 2018) for 2D/3D scene understanding tasks.
  Shape Inductive Biases (Feinman & Lake 2018) for 3D shapes

Limitation:

- need to train the transfer network using gradient descend, subject to local minimum
- inefficient to compute

Can we estimate the transfer performance without any training of the target network?

A. R. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik and S. Savarese, "Taskonomy: Disentangling Task Transfer Learning," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018

Analytical Transferability Metric

- An analytical way to *estimate* empirical transferability

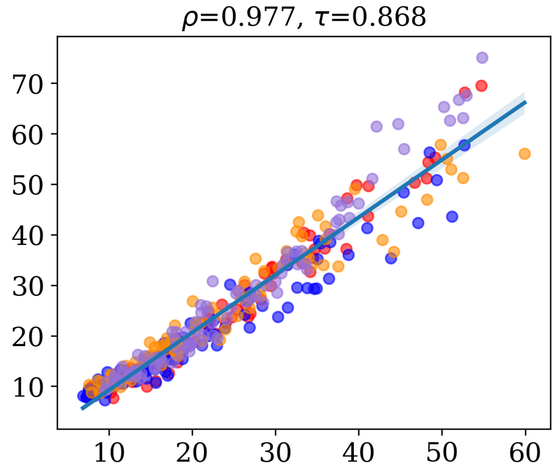
Algorithm	Cross-Task $P(Y_S \mid X_S) \neq P(Y_T \mid X_T)$	Cross-Instance $X_S \neq X_T$	Cross-Domain $P(X_S) \neq P(X_T)$
NCE (Tran et al. 2019) *	✓	x	x
H-Score (Bao et al. 2019)	✓	✓	x
LEEP (Nguyen et al. 2020)**	✓	✓	x
OTCE (Tan et al. 2021)	✓	✓	✓
LogME (K. You 2021) ***	✓	✓	✓

first transferability work for task transfer learning

* Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transfer- ability and hardness of supervised classification tasks. ICCV, 2019.
** Cuong V Nguyen, Tal Hassner, Cedric Archambeau, and 952 Matthias Seeger. Leep: A new measure to evaluate transferability of learned representations.ICML, 2020.

*** K. You, Y. Liu, J. Wang, and M. Long, “Logme: Practical assessment of pre-trained models for transfer learning,” in *ICML*. PMLR, 2021, pp. 12 133–12 143.

Talk Outline



- Analytical Transferability Metrics
  - H-score (efficient cross-task transferability metric)
  - OT-CE (cross-domain, cross-task transferability metric)
- Transferability-guided fine-tuning
- Measuring Transferability for Medical Segmentation

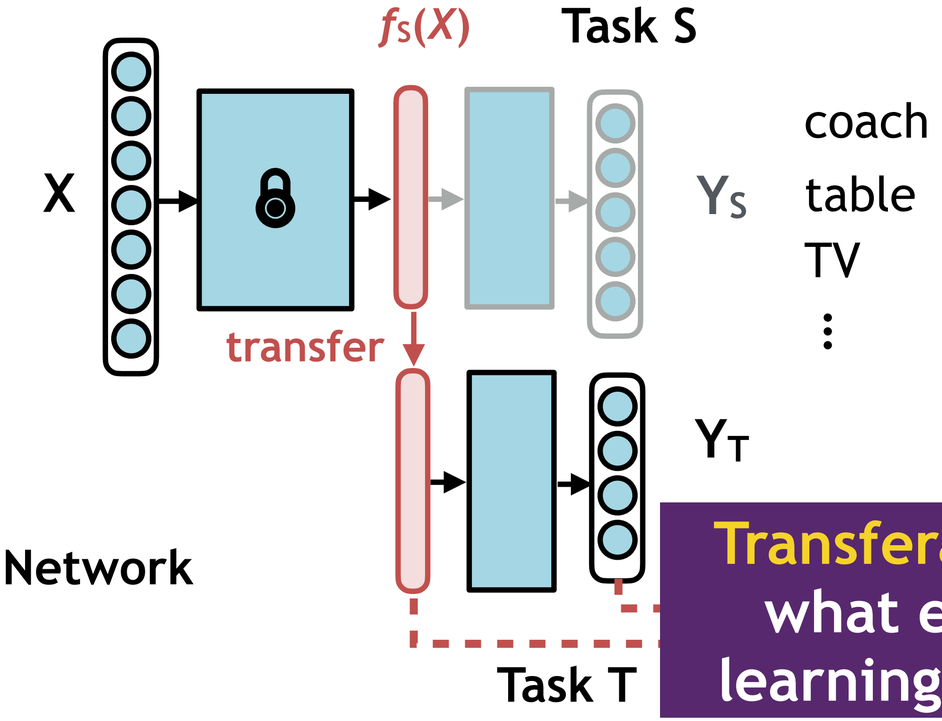
The Task Transferability Problem

Given:

- Input X, source task label Y_S, target task label Y_T
- Trained source model with optimal feature $f_S(X)$

Assume same domain
$P(X_s)=P(X_t)=P(X)$

Retrain-head setting



The Transfer Network

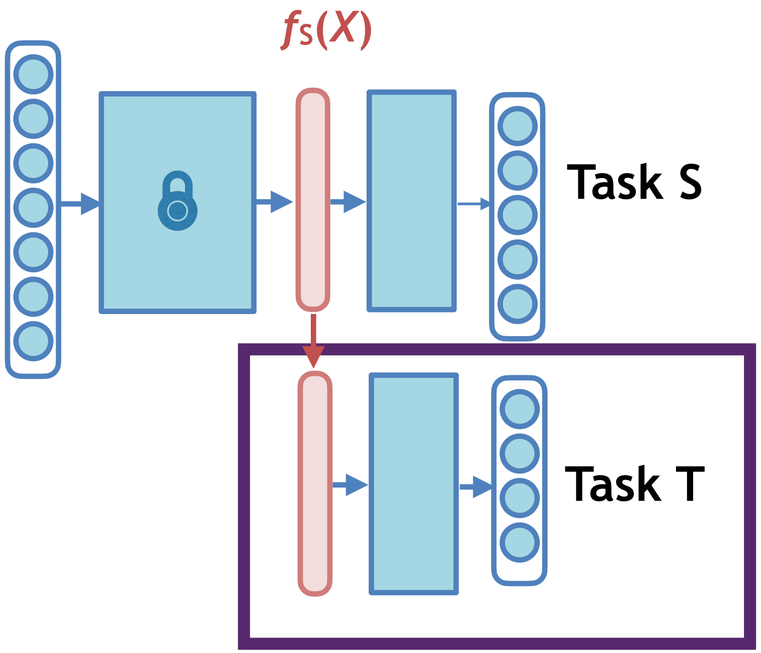
Transferability of S→T : to what extent can f_S help learning target task (X, Y_T)?

Task Transferability

Bao, Yajie, Yang Li et al. "An information-theoretic approach to transferability in task transfer learning." *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019.

Transferability from Task S to Task T

$$\mathfrak{T}(S,T) \triangleq \frac{\textbf{Target Performance of } f_S}{\textbf{Optimal Target Performance}}$$



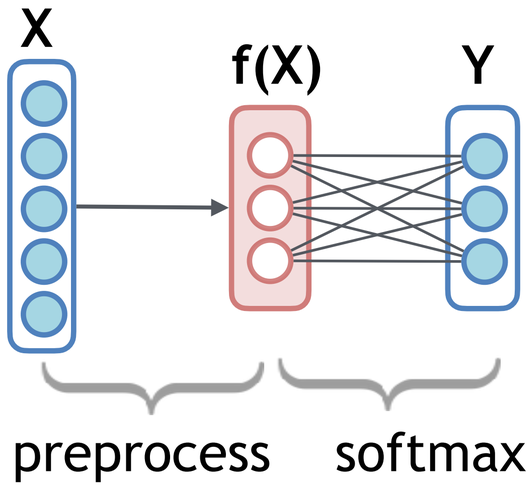
$$\begin{cases} \mathfrak{T}(S,T) = 1 & \text{☺} \\ 0 \le \mathfrak{T}(S,T) \le 1 & \\ \mathfrak{T}(S,T) = 0 & \text{☹} \end{cases}$$

How to measure the performance of $f_S(X)$ on target task (X, Y_T) ?

joint work with Yaojie Bao

Measuring Feature Effectiveness - Neural Network Perspective

- Classification using log-loss:
  - X, Y random variables; f(X) a zero-mean feature
  - Expected log loss: $L(f;\theta) = \mathbb{E}_{X,Y}[L(f(X), Y; \theta)]$



- By Local information geometry [Huang 2018], given feature f(X), the optimal loss is

$$L(f, \theta^\star) = Const(X, Y) - \underbrace{H(f)}_{\text{H-score of f(X)}} + o(\epsilon^2)$$

Higher H-score => Better Performance

$$\mathcal{H}(f) = \mathrm{tr}(\mathrm{cov}(f(X))^{-1}\mathrm{cov}(\mathbb{E}_{P_{X|Y}}[f(X)|Y]))$$

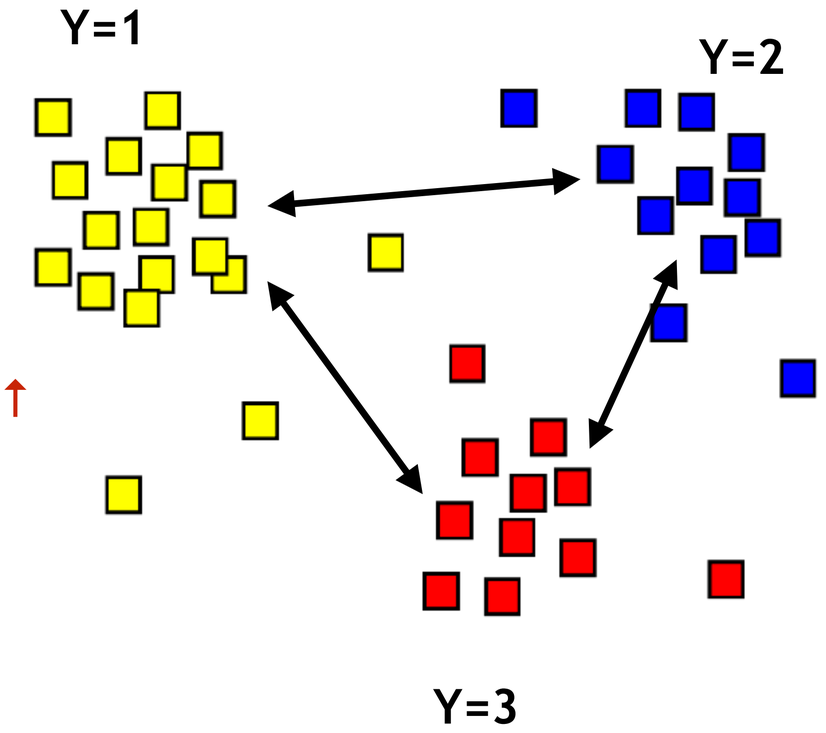
Interpretation of $\mathscr{H}(f)$

Intuition in latent space

$$\mathscr{H}(f) = \mathrm{tr}(\mathrm{cov}(f(X))^{-1}\mathrm{cov}(\mathbb{E}_{X|Y}[f(X)\,|\,Y]))$$

H-score ↑ **feature redundancy ↓** **average intra-class distance ↑**

$$\mathbb{E}[\|\mathbb{E}[f(X)\,|\,Y]\|^2]$$



Statistical operational meaning

- H-score characterizes the asymptotic error probability of the likelihood ratio predictor based on $f(X)$
- Higher H-score ↔ Faster error decay rate

Computing Transferability

$$\mathfrak{T}(S,T) = \frac{\mathscr{H}_T(f_S)}{\mathscr{H}_T(f_T)}$$

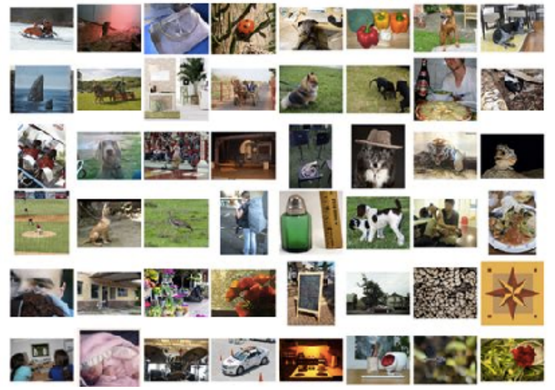
- H-score of source feature $\mathscr{H}_T(f_S)$
  - Easy to compute
  - O(mk^2) time complexity
- Maximal H-score: $\mathscr{H}_T(f_T)$
  - Discrete X: Alternating Conditional Expectation (ACE) algorithm (Huang et. al. 2015)
  - Continuous X: Neural network formulation (Wang et al. 2017)

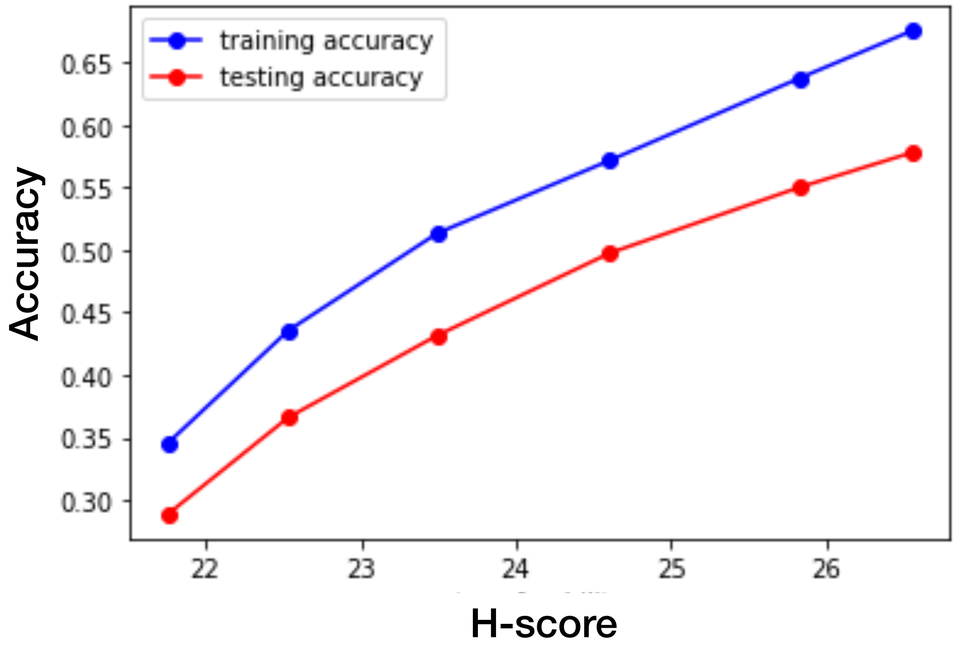
Python Code for H-Score

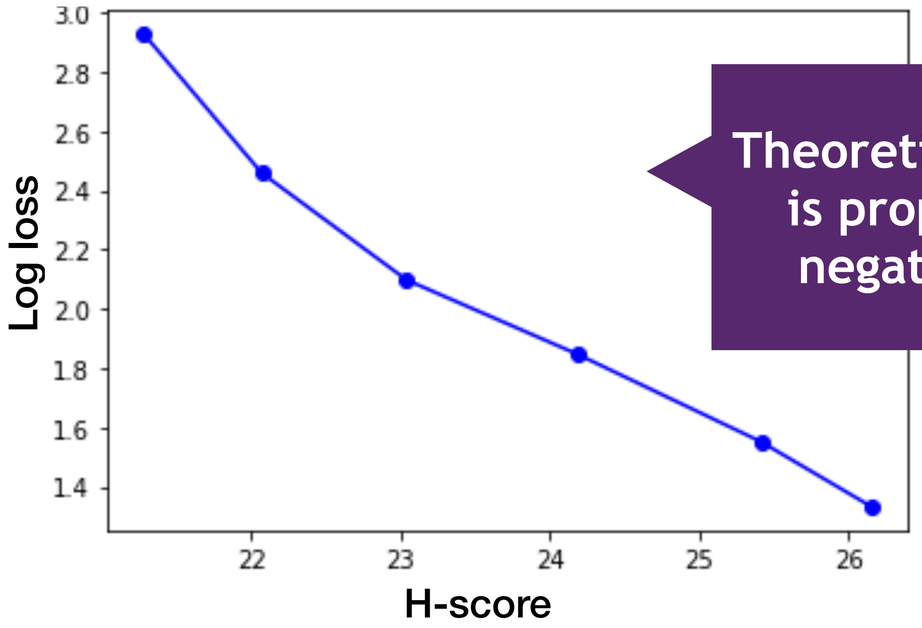
```
def Hscore(f,Y):
    Covf=np.cov(f)
    alphabetY=list(set(Y))
    g=np.zeros_like(f)
    for z in alphabetY:
        g[Y==y]=np.mean(f[Y==y,:], axis=0)
    Covg=np.cov(g)
    score=np.trace(np.dot(np.linalg.pinv(Covf,
                rcond=1e-15), Covg))
    return score
```

Results: Image Classification Feature Selection

- Source task: ImageNet 1000 classification (ResNet50 features from 6 layers 4a-5f)
- Target task: Cifar 100-class classification on 20,000 images



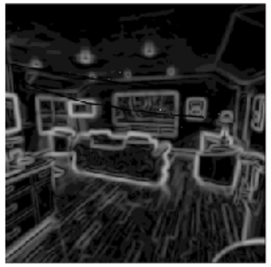




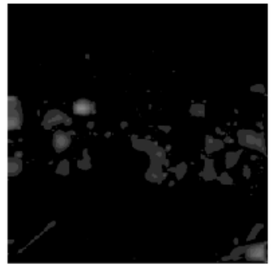
Theoretically, H-score is proportional to negative log loss

Results: Source Task Selection for 3D Scene Understanding


Query Image


2D Edges


3D (Occlusion) Edges


2D Keypoints

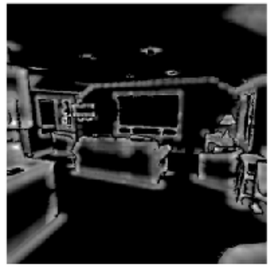
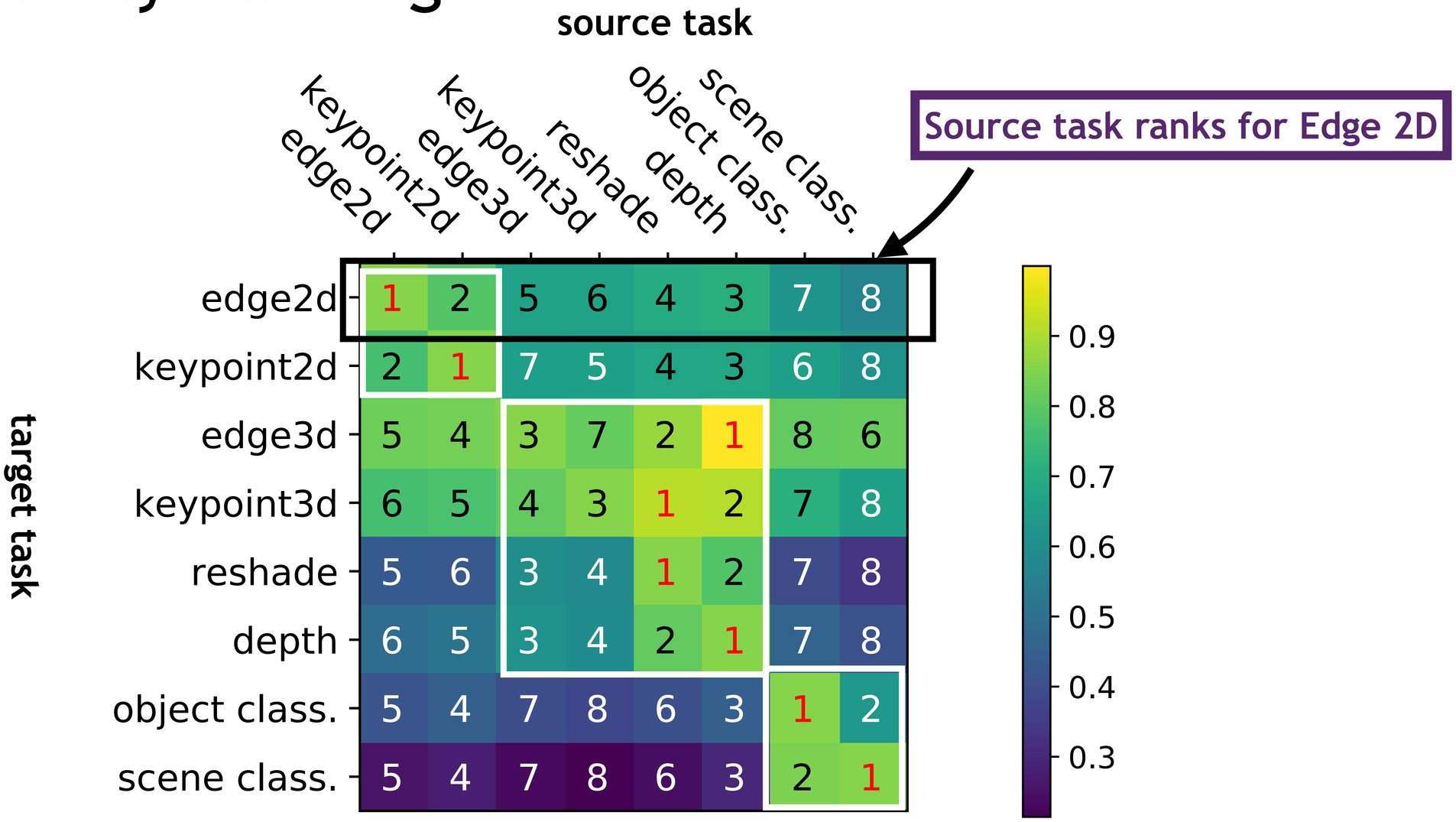

3D Keypoints


Image Reshading


Depth

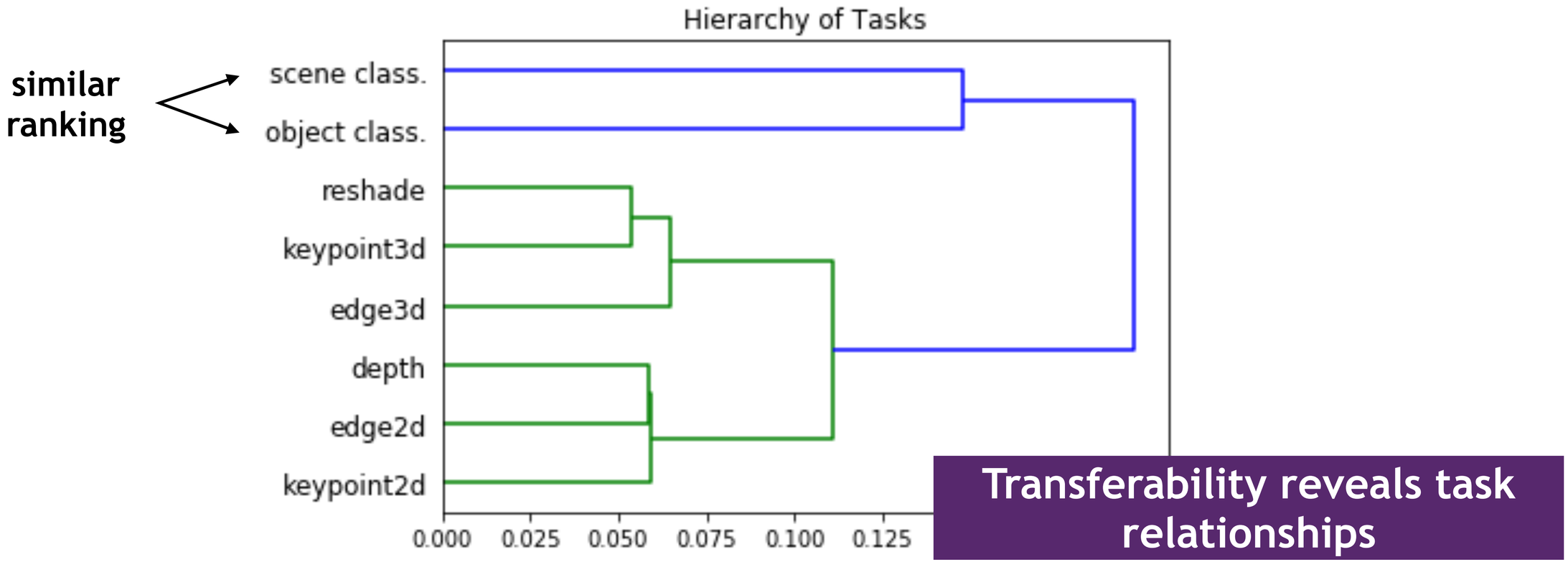
- 8 image-based tasks from Taskonomy dataset (Zamir et al. 2018)
  - **2 classification tasks**: object-class, scene-class
  - **6 2D/3D image-to-image tasks**: average H-score over all superpixels
- **Source models**: pre-trained task-specific models (4,000,000 training samples);
- **Target model**: linear feature transfer using 20,000 images (64 x 64)

Transferability Ranking



Task Relationships

Cluster the source task transferability scores for each target task.

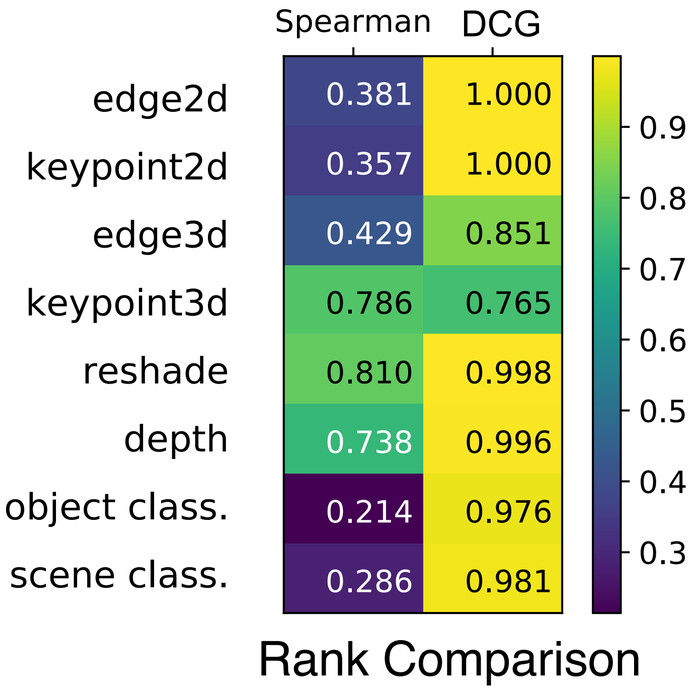
- 

Transferability reveals task relationships

Comparison with Task Affinity

Reference metric: task affinity, an empirical transferability score (Amir et al. 2018)

- Ranking comparison metric: Spearman's correlation and Discounted cumulative gain (DCG)



Advantage of our approach:

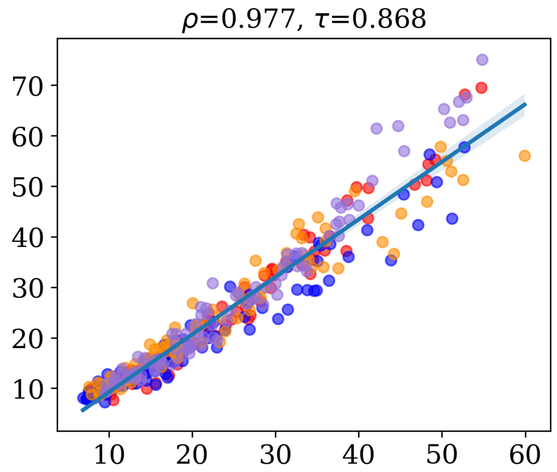
- Efficiency: five times more efficient than Affinity
- Clear operational meaning based on statistics & information theory

Discussion on H-Score

- Proposed an efficient, easy-to-compute task transferability metric with clear operational meaning.
  - Theoretically proven for classification tasks.
  - Validated on image-processing, vision recognition applications
- Shrinkage based H-score (Ibrahim et. al. 2021) : improved the stability of covariance estimation in H-score computation

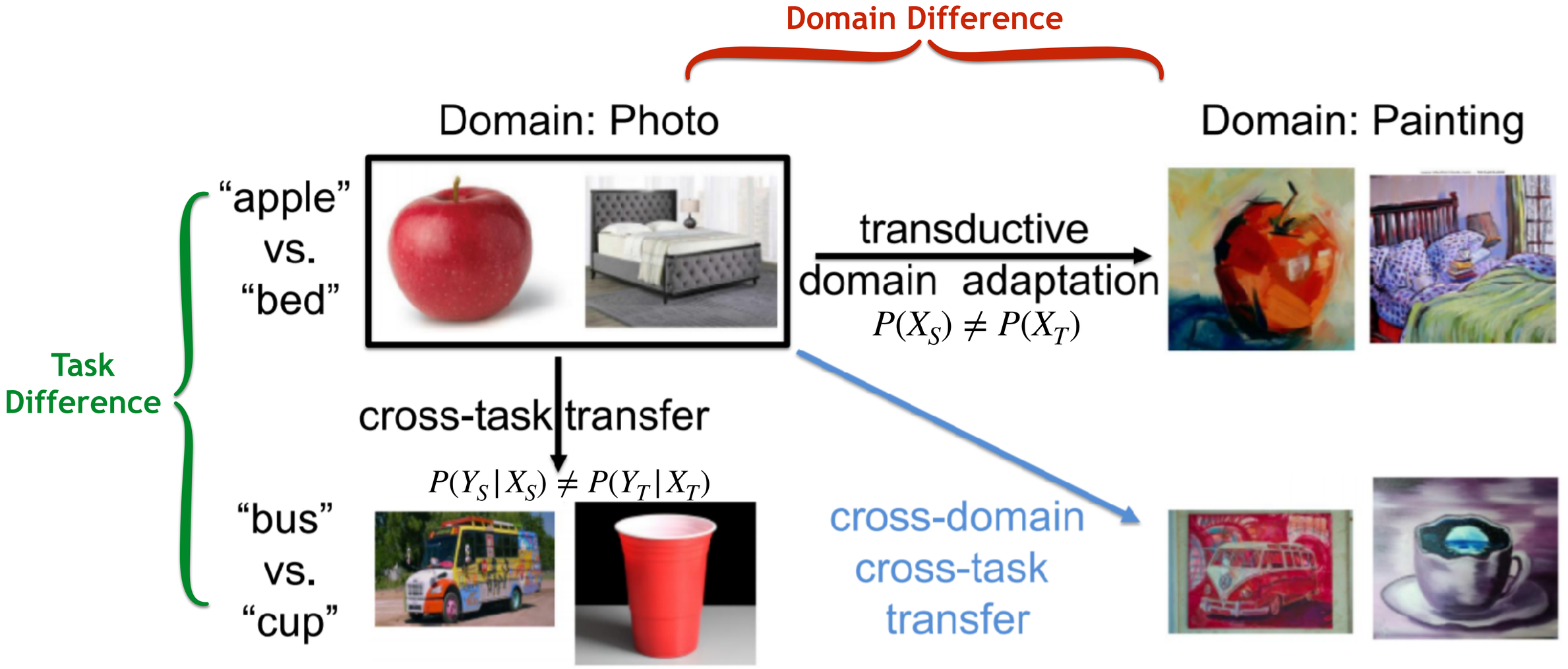
Ibrahim, S., Ponomareva, N., & Mazumder, R. (2021). Newer is not always better: Rethinking transferability metrics, their peculiarities, stability and performance. *arXiv preprint arXiv:2110.06893*.

Talk Outline



- Analytical Transferability Metrics
  - H-score (efficient cross-task transferability metric)
  - OT-CE (cross-domain, cross-task transferability metric)
- Transferability-guided fine-tuning
- Measuring Transferability for Medical Segmentation

Transferability across tasks and domains



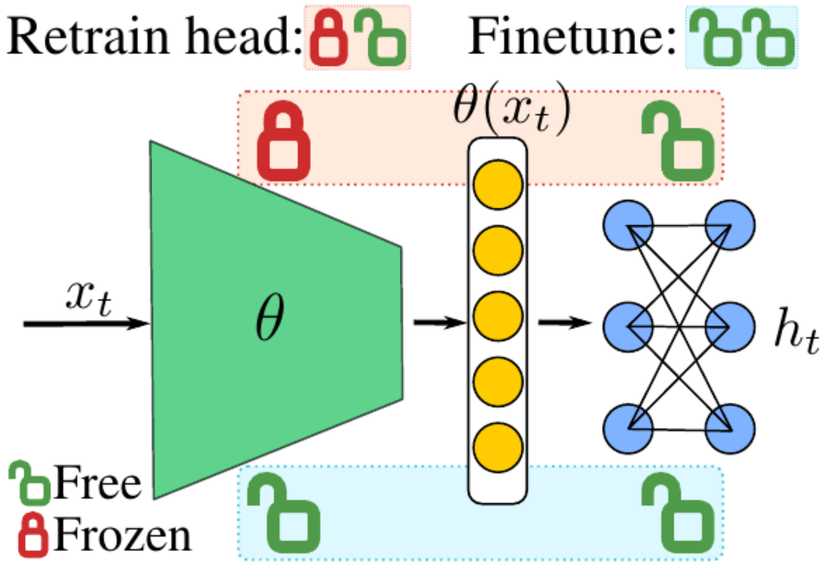
Terminologies

- Source dataset: $D_s = \{(x_s^i, y_s^i)\}_{i=1}^m \sim P_s(x, y)$
- Target dataset: $D_t = \{(x_t^i, y_t^i)\}_{i=1}^n \sim P_t(x, y)$

where $x_s^i, x_t^i \in \mathcal{X}$, $y_s^i \in \mathcal{Y}_s, y_t^i \in \mathcal{Y}_t$

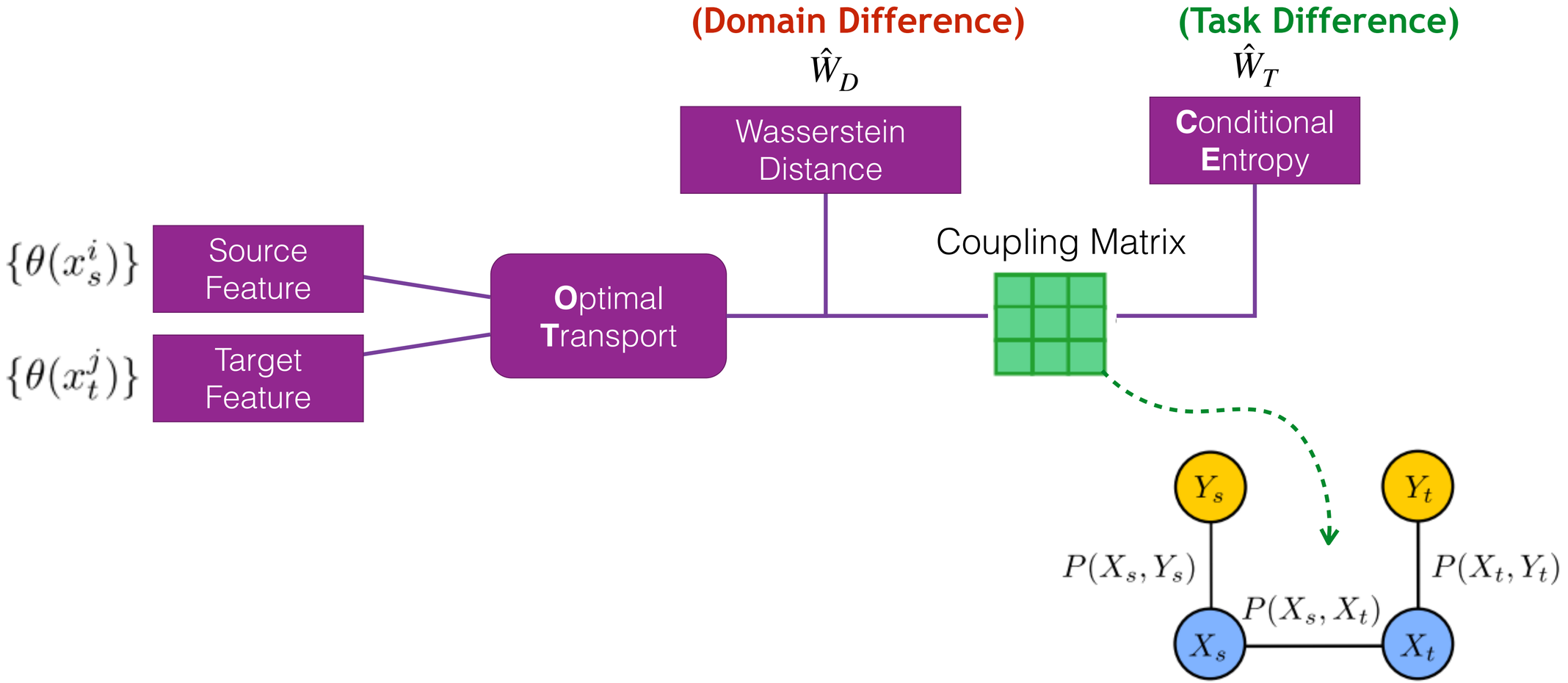
- Source model: (θ, h_s) where $\theta : \mathcal{X} \to \mathbb{R}^d$ $h_s : \mathbb{R}^d \to \mathcal{P}(\mathcal{Y}_s)$

Source data is needed to compute domain difference



OTCE Transferability Score

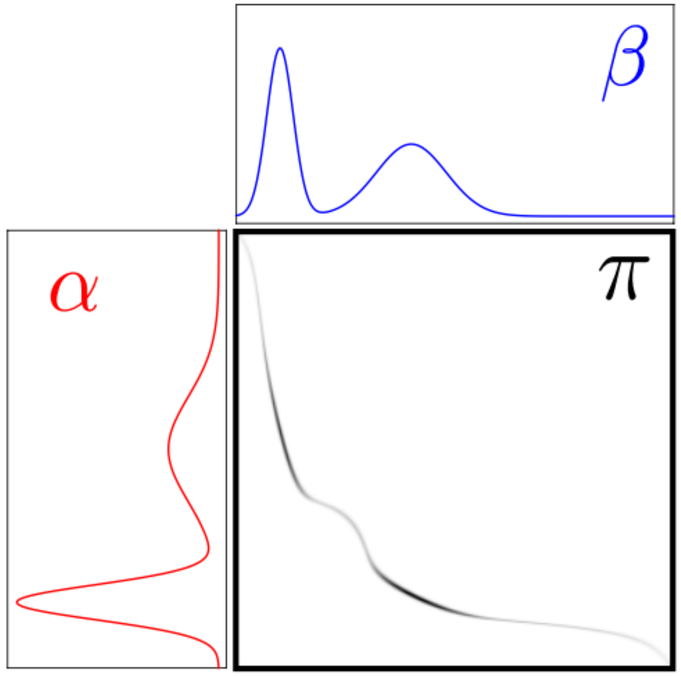
Yang Tan, Yang Li*, and Shao-Lun Huang. "OTCE: A Transferability Metric for Cross-Domain Cross-Task Representations." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15779-15788. 2021.Code and Supplementary Material:



joint work with Yang Tan

Compute Domain Difference

- Optimal transport problem



$$\alpha \in \mathcal{P}(\mathcal{X}), \beta \in \mathcal{P}(\mathcal{X})$$

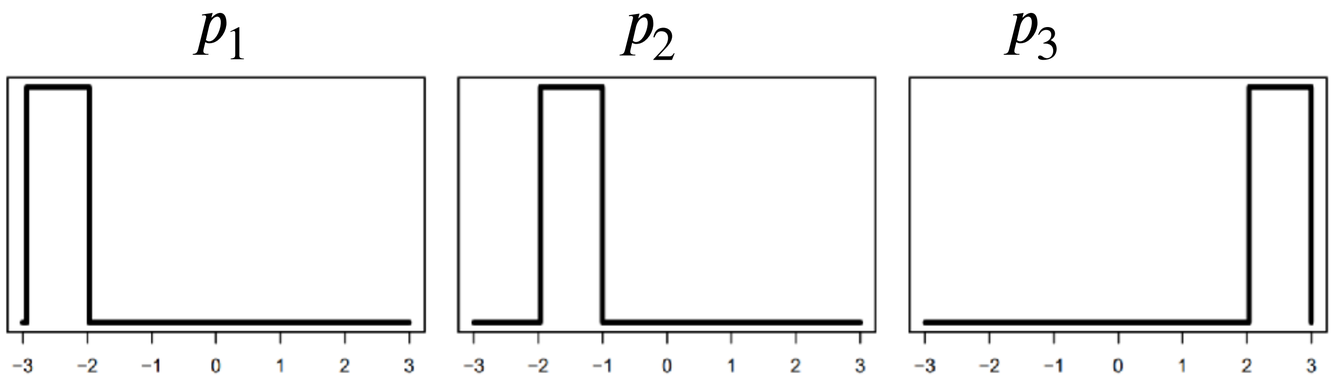
$$\Pi(\alpha, \beta) \triangleq \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) | P_{1\#}\pi = \alpha, P_{2\#}\pi = \beta\}$$

$$OT(\alpha, \beta) \triangleq \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} c(x, y) d\pi(x, y)$$

When c is $c(x, y)^d$, OT is the pth-Wasserstain Distance W^p

Advantage of Wasserstein Distance

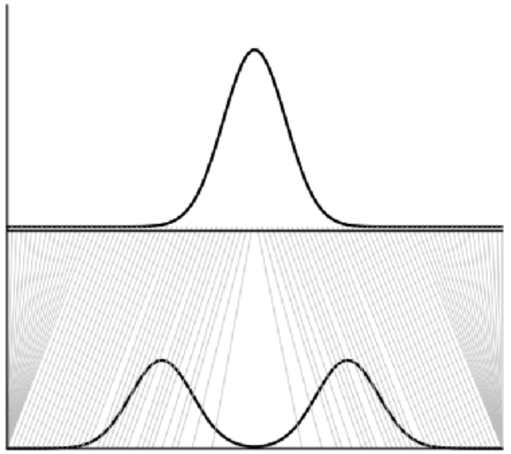
- Observe the underlying geometry



$$KL(p_1; p_2) = KL(p_1; p_3)$$
$$MMD(p_1, p_2) = MMD(p_1, p_3)$$

$$W(p_1, p_2) < W(p_1, p_3)$$

- Provide a map about moving mass



Compute Domain Difference

- Given two datasets

$$D_A = (x_A^i, y_A^i)_{i=1}^m \sim P_A(x, y) \qquad D_B = (x_B^j, y_B^j)_{j=1}^n \sim P_B(x, y)$$

- Regularized optimal transport (Sinkhorn Algorithm)

$$OT(D_A, D_B) \triangleq \min_{\pi} \int_{\mathcal{X}\times\mathcal{X}} c(x_A, x_B) d\pi(x_A, x_B) + \epsilon H(\pi)$$

$$c(x_A^i, x_B^j) = \|x_A^i - x_B^j\|_2^2$$

[1]

[1] Cuturi, Marco. "Sinkhorn distances: Lightspeed computation of optimal transport." *Advances in neural information processing systems*. 2013.

- Domain difference as wasserstein distance

$$W(D_A, D_B) = \sum_{i,j=1}^{m,n} \pi^*(x_A^i, x_B^j) c(x_A^i, x_B^j)$$

Task Difference: Theoretical Motivation

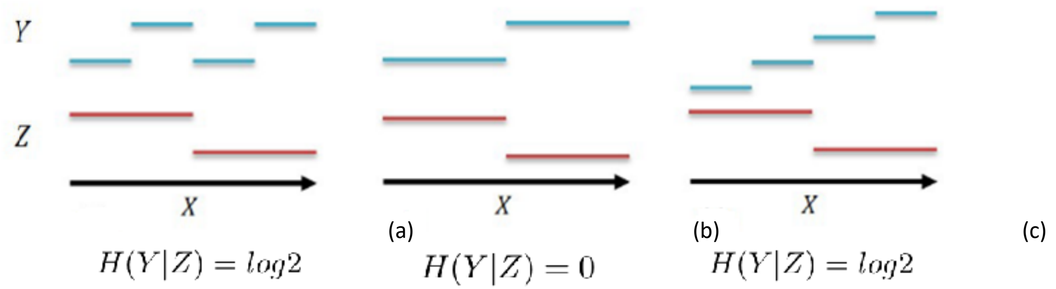
- Given X and multi-task labels Y and Z

$$k_Y = \underset{k \in K}{\operatorname{argmin}} \mathcal{L}_Y(w_Z, k)$$

- Approximat transferability by conditional entropy

$$\widetilde{\mathrm{Trf}}(T^Z \to T^Y) = l_Y(w_Z, k_Y)$$

Tran, Anh T., Cuong V. Nguyen, and Tal Hassner. "Transferability and hardness of supervised classification tasks." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.



Task Difference: (a) = (c) > (b)

$$\widetilde{\mathrm{Trf}}(T^Z \to T^Y) \geq l_Z(w_Z, h_Z) - \boxed{H(Y|Z)}$$

Challenge: how to compute H(Y|Z) when source and target inputs are not the same?

Computing Task Difference

- Build sample correspondence using the Coupling Matrix (joint distribution) π^* from the OT step

$$\pi^* = \begin{bmatrix} P(x_A^1, x_B^1) & P(x_A^1, x_B^2) & \dots & P(x_A^1, x_B^n) \\ P(x_A^2, x_B^1) & P(x_A^2, x_B^2) & \dots & P(x_A^2, x_B^n) \\ \vdots & \vdots & \ddots & \vdots \\ P(x_A^n, x_B^1) & P(x_A^n, x_B^2) & \dots & P(x_A^n, x_B^n) \end{bmatrix}$$

- Compute empirical distributions of labels

$$\hat{P}(y_s, y_t) = \sum_{i,j: y_s^i = y_s, y_t^j = y_t} \pi_{ij}^*, \quad \hat{P}(y_s) = \sum_{y_t \in \mathcal{Y}_t} \hat{P}(y_s, y_t).$$

$$W_T = H(Y_t|Y_s) = -\sum_{y_t \in \mathcal{Y}_t} \sum_{y_s \in \mathcal{Y}_s} \hat{P}(y_s, y_t) \log \frac{\hat{P}(y_s, y_t)}{\hat{P}(y_s)}$$

Conditional Entropy

OTCE Score

- Linear combination between Domain Difference and Task Difference

$$OTCE = \lambda_1 \hat{W}_D + \lambda_2 \hat{W}_T + b$$

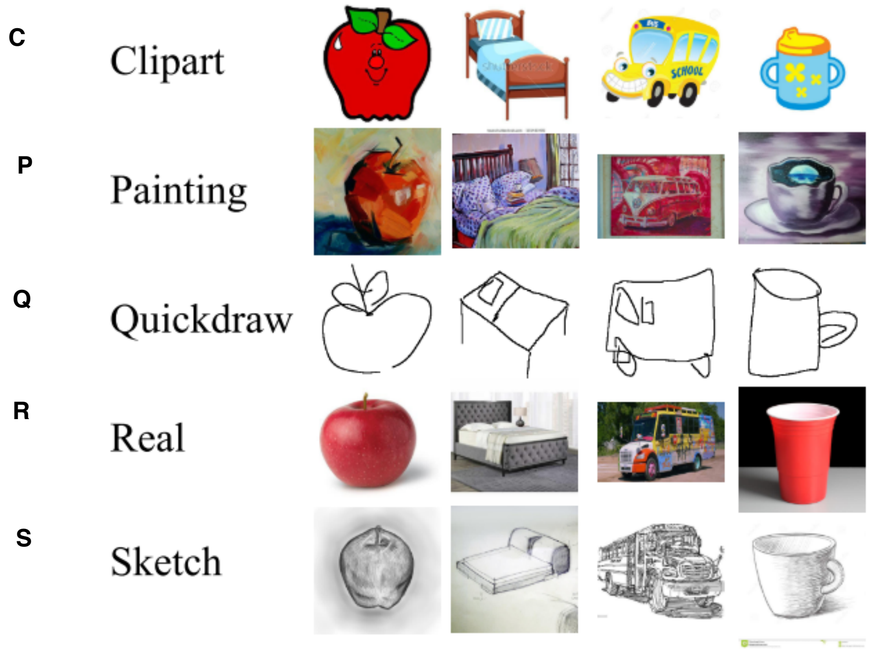
- Learn coefficients from auxiliary tasks with known empirical transferability

Practical consideration: to avoid using auxiliary task, we can use task difference WT alone to characterize transferability

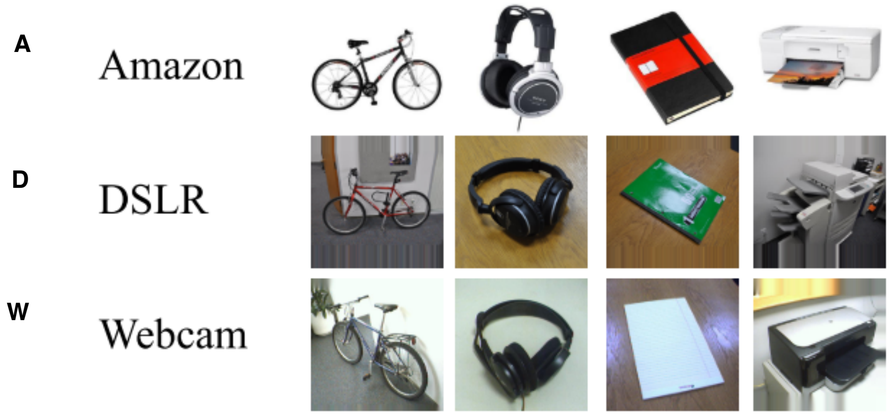
$$OTCE_{sim} = W_T$$

Experiment Datasets (Classification)

- **DomainNet**: 345 categories in 5 domains



- **Office31**: 32 categories in 3 domains.



For a source domain D_{src}, randomly sample 100 classification tasks in each target domain $D_{tgt} \in D \backslash \{D_{src}\}$

Experiment Setup

- **Transfer settings**

Setting Names	target sample size	label size
standard	all target task samples	1-50
few-shot	10 samples	1-50
fixed category size	all target task samples	50

tasks with diverse hardness

tasks with similar hardness

- **Evaluation metric**: compare the *ranking* of target tasks ordered by OT-CE score and empirical transferability
  - Spearsman's rank correlation ρ
  - Kendall's rank τ

Experiment Result

- **Standard & Few-shot settings**

Setting	Dataset	Source domain	Target domain	Spearman / Kendall				
				OTCE	OTCE$_{sim}$	LEEP [6]	NCE [4]	H-score [5]	LogME [15]
Standard	DomainNet	C	P,Q,R,S	**0.976 / 0.861**	0.966 / 0.839	0.932 / 0.779	0.825 / 0.670	0.920 / 0.748	0.867 / 0.667
		P	C,Q,R,S	**0.977 / 0.868**	0.960 / 0.822	0.906 / 0.743	0.849 / 0.686	0.937 / 0.777	0.929 / 0.761
		Q	C,P,R,S	0.961 / 0.826	**0.963 / 0.832**	0.953 / 0.810	0.943 / 0.793	0.942 / 0.784	0.912 / 0.744
		R	C,P,Q,S	**0.975 / 0.863**	0.951 / 0.808	0.910 / 0.747	0.872 / 0.707	0.942 / 0.786	0.855 / 0.670
		S	C,P,Q,R	**0.969 / 0.842**	0.967 / 0.839	0.965 / 0.834	0.962 / 0.830	0.950 / 0.802	0.908 / 0.733
	Office31	A	D,W	**0.848 / 0.643**	0.831 / 0.619	0.817 / 0.606	0.817 / 0.604	0.601 / 0.417	0.660 / 0.459
		D	A,W	**0.885 / 0.702**	0.839 / 0.651	0.862 / 0.676	0.851 / 0.664	0.464 / 0.335	0.172 / 0.119
		W	A,D	**0.859 / 0.676**	0.801 / 0.614	0.812 / 0.626	0.805 / 0.616	0.524 / 0.371	0.470 / 0.318
			Average	**0.931 / 0.785**	0.910 / 0.753	0.895 / 0.728	0.866 / 0.696	0.785 / 0.628	0.722 / 0.559
Few-shot	DomainNet	C	P,Q,R,S	**0.926 / 0.756**	0.909 / 0.729	0.836 / 0.640	0.745 / 0.576	0.762 / 0.567	0.731 / 0.524
		P	C,Q,R,S	**0.931 / 0.772**	0.886 / 0.701	0.803 / 0.618	0.746 / 0.575	0.811 / 0.608	0.849 / 0.649
		Q	C,P,R,S	0.821 / 0.631	0.829 / 0.636	0.798 / 0.602	0.782 / 0.584	0.813 / 0.614	**0.866 / 0.682**
		R	C,P,Q,S	**0.929 / 0.769**	0.853 / 0.666	0.770 / 0.589	0.728 / 0.559	0.845 / 0.652	0.774 / 0.574
		S	C,P,Q,R	**0.914 / 0.742**	0.895 / 0.710	0.872 / 0.680	0.872 / 0.679	0.838 / 0.645	0.867 / 0.684
	Office31	A	D,W	**0.859 / 0.662**	0.845 / 0.640	0.818 / 0.609	0.811 / 0.602	0.651 / 0.456	0.659 / 0.460
		D	A,W	**0.929 / 0.773**	0.925 / 0.766	0.925 / 0.764	0.924 / 0.765	0.429 / 0.308	0.002 / 0.021
		W	A,D	0.927 / 0.765	**0.929 / 0.767**	0.916 / 0.749	0.919 / 0.752	0.316 / 0.235	0.250 / 0.181
			Average	**0.905 / 0.734**	0.884 / 0.702	0.842 / 0.656	0.816 / 0.637	0.683 / 0.511	0.625 / 0.472

Experiment Result

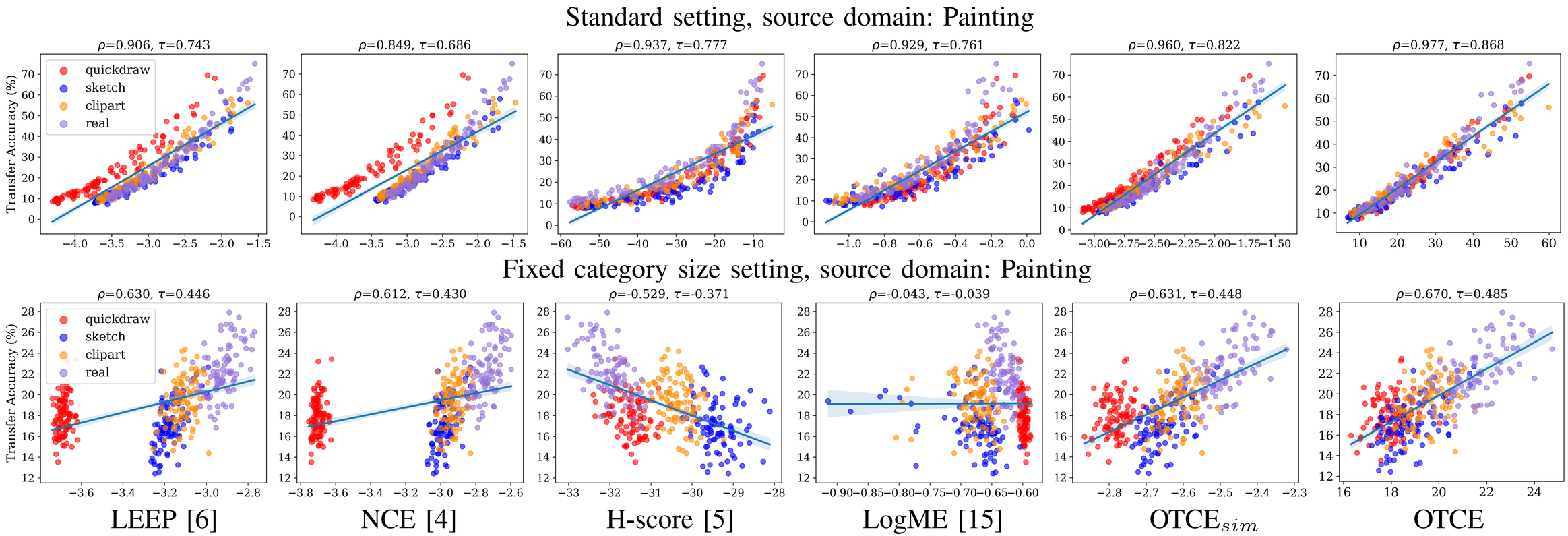
- **Same category set size setting: similar target task complexity**

Source domain	Target domain	Spearman / Kendall					
		OTCE	OTCE$_{sim}$	LEEP [6]	NCE [4]	H-score [5]	LogME [15]
C	P,Q,R,S	**0.701 / 0.500**	0.687 / 0.487	0.685 / 0.486	0.666 / 0.472	-0.438 / -0.290	-0.222 / -0.151
P	C,Q,R,S	**0.670 / 0.485**	0.631 / 0.448	0.630 / 0.446	0.612 / 0.430	-0.529 / -0.371	-0.043 / -0.039
Q	C,P,R,S	**0.341 / 0.225**	0.316 / 0.211	0.210 / 0.136	0.291 / 0.191	-0.256 / -0.172	0.066 / 0.037
R	C,P,Q,S	**0.637 / 0.455**	0.598 / 0.415	0.587 / 0.407	0.586 / 0.406	-0.094 / -0.063	-0.382 / -0.252
S	C,P,Q,R	0.428 / 0.292	**0.436 / 0.299**	0.404 / 0.277	0.432 / 0.298	-0.247 / -0.164	0.027 / 0.006
	Average	**0.555 / 0.391**	0.534 / 0.372	0.503 / 0.350	0.517 / 0.359	-0.313 / -0.212	-0.111 / -0.080

Transferability between different task pairs less distinguishable => More challenging scenario

Experiment Results

- Visualization of transferability experiment results on DomainNet



Experiments: Transferability Applications

- Source Model Selection

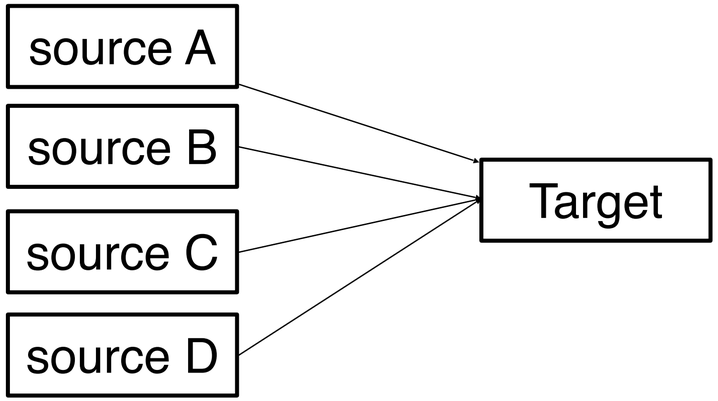
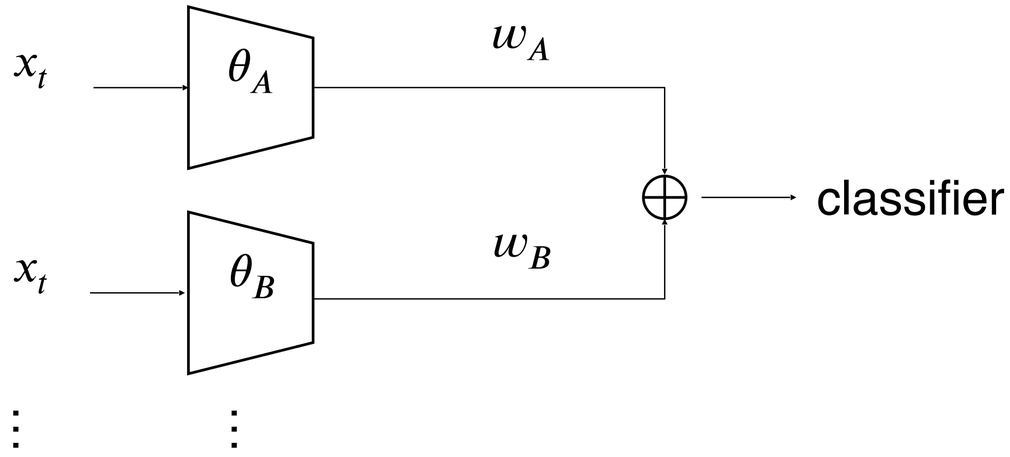
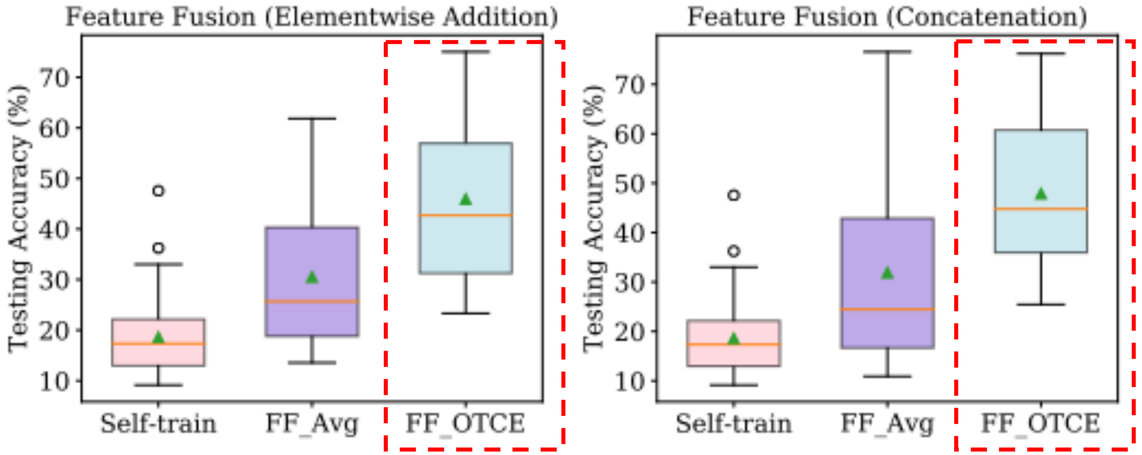


Table 2. Quantitative comparisons of source model selection accuracy (%) among transferability metrics on DomainNet.

Method	Target domain					average
	C	P	Q	R	S	
LEEP[24]	31.1	26.7	5.6	97.8	**100.0**	52.2
NCE[38]	**41.1**	**94.4**	2.2	**100.0**	**100.0**	67.5
OTCE	**41.1**	93.3	**97.8**	**100.0**	**100.0**	**86.4**

- Multi-Source Feature Fusion



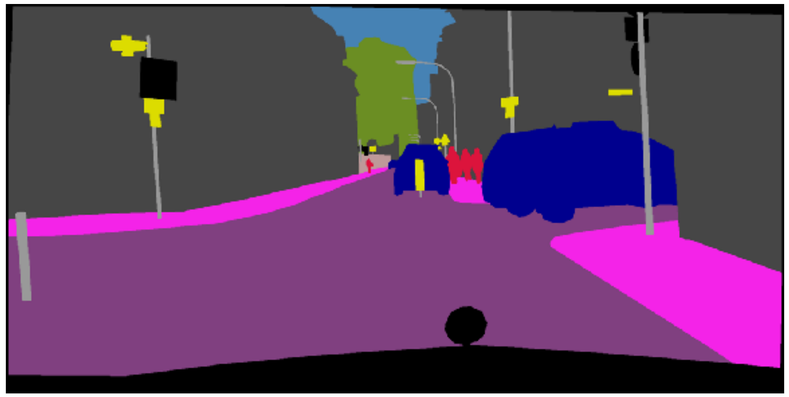


Semantic Segmentation Task

- A fundamental task in Autonomous driving, medical image analysis etc

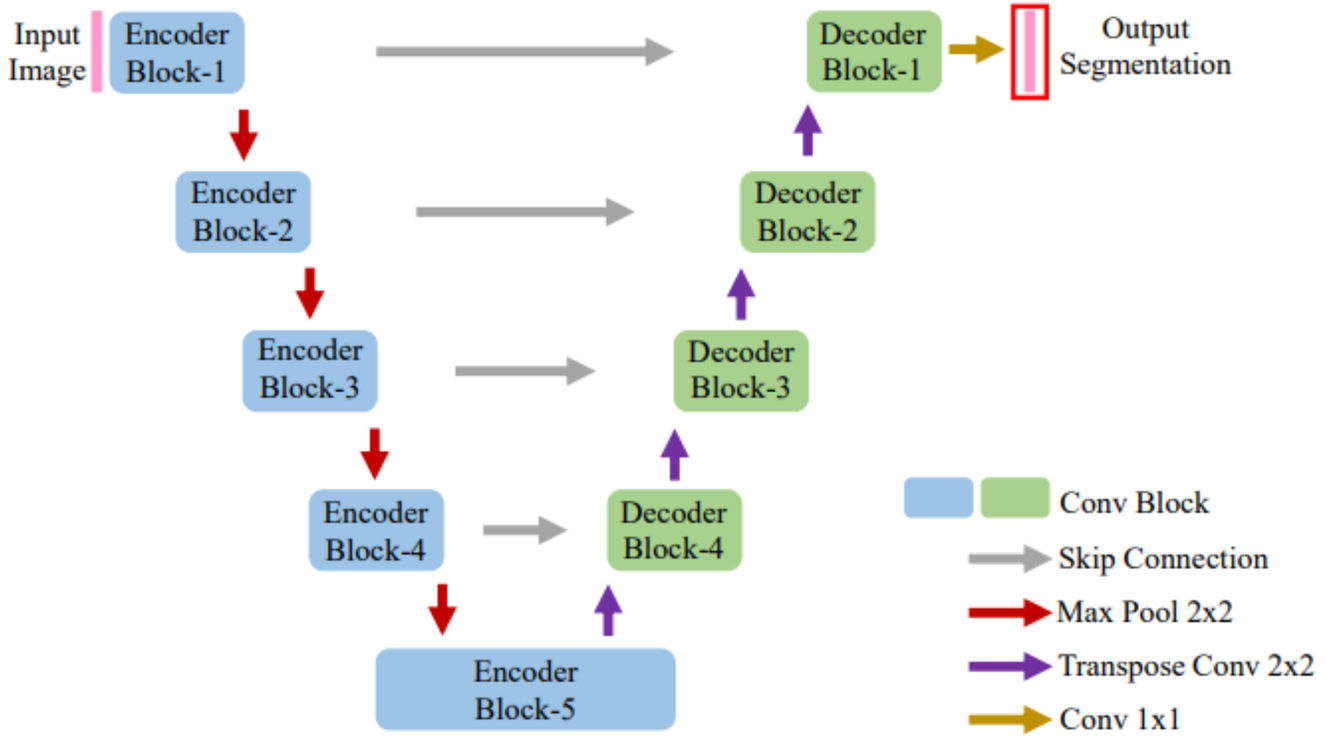
Input image

Semantic label



Which feature to transfer?

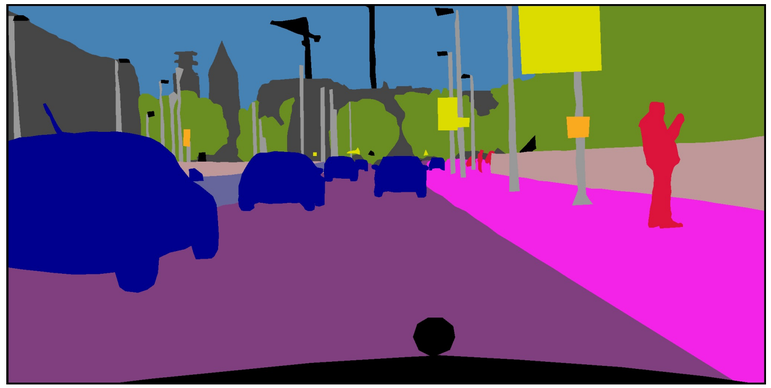
- U-Net architecture commonly used for semantic segmentation



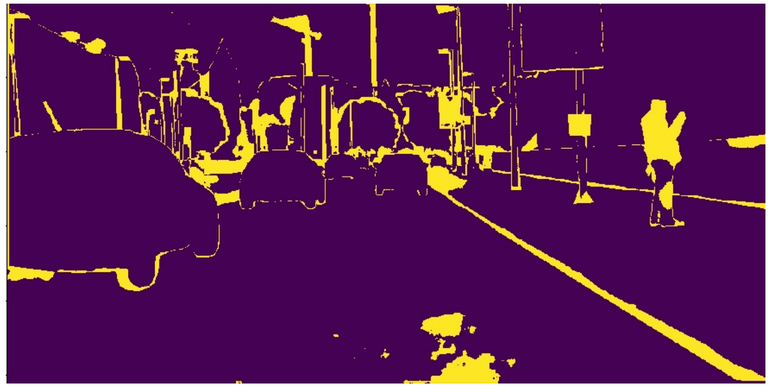
Challenge in computing transferability: Output Y is high-dimensional

Observation on prediction error map

Input image



Semantic label

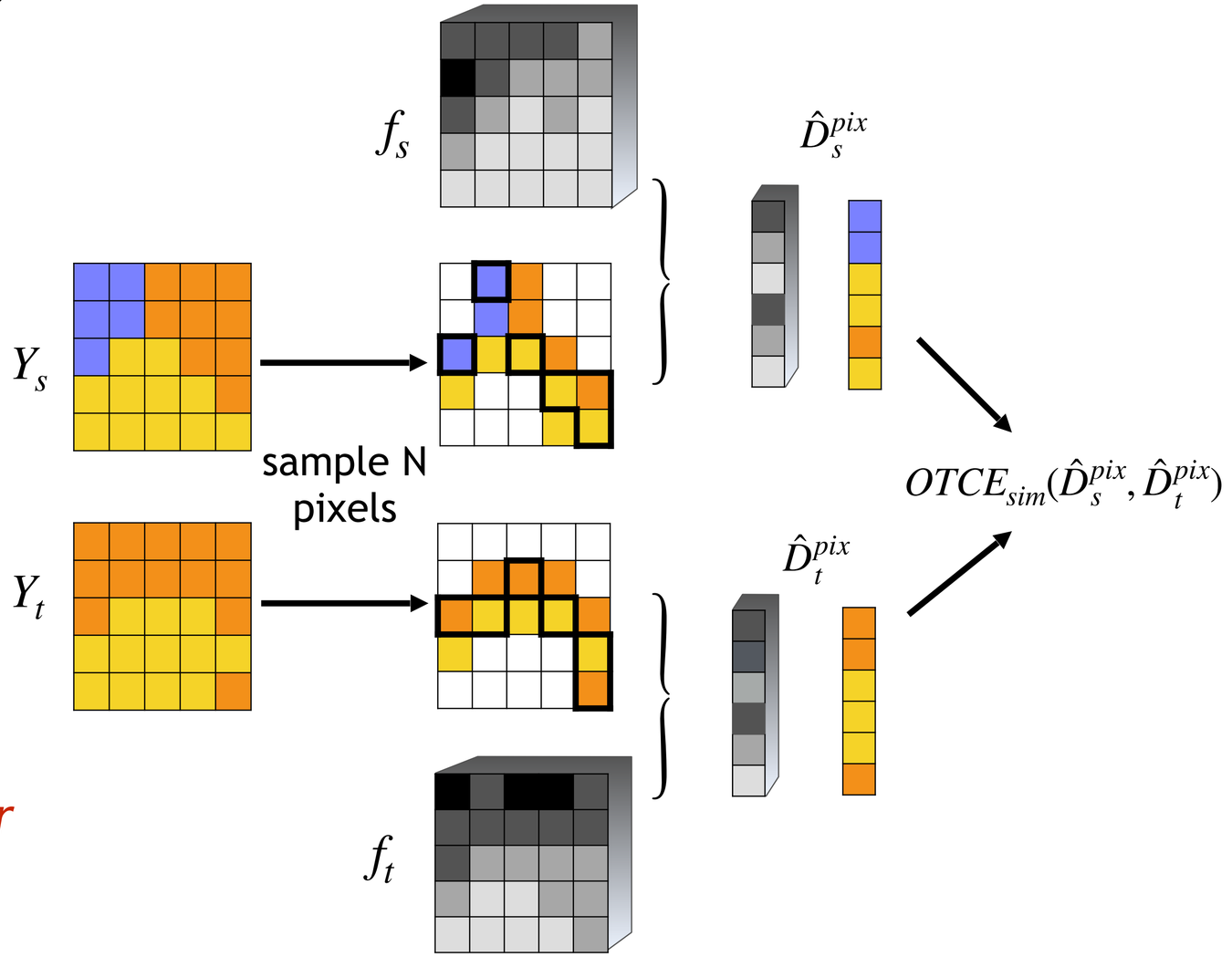


Predicted error map

Extract edge from semantic label

OTCE for Semantic Segmenation

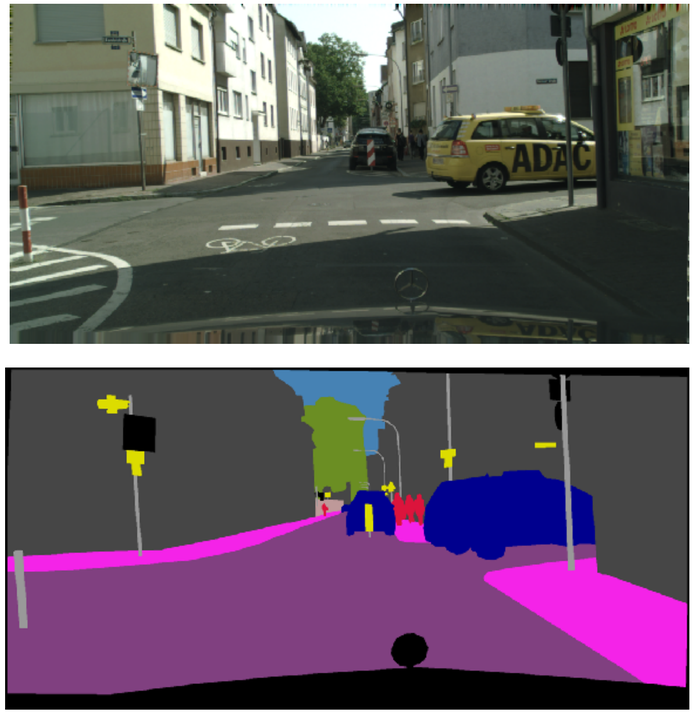
- Solution:
  - Sample N pixels from all images
  - Compute $OTCE_{sim}$ over the feature set of the sampled labels



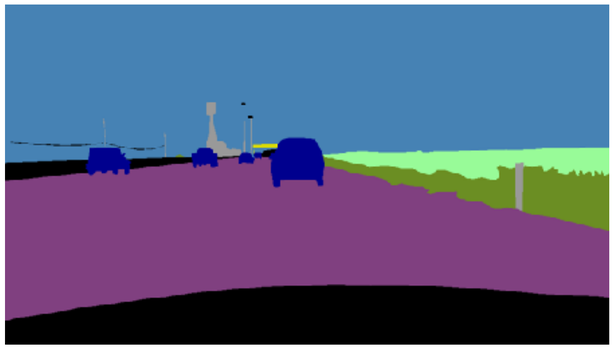
Sample Heuristic: sample pixels near segmentation boundaries

Experiment Datasets (Segmentation)

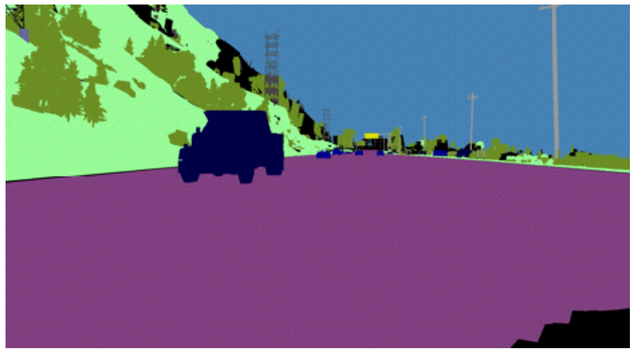
Cityscapes
(real captured)



BDD100K
(real captured)

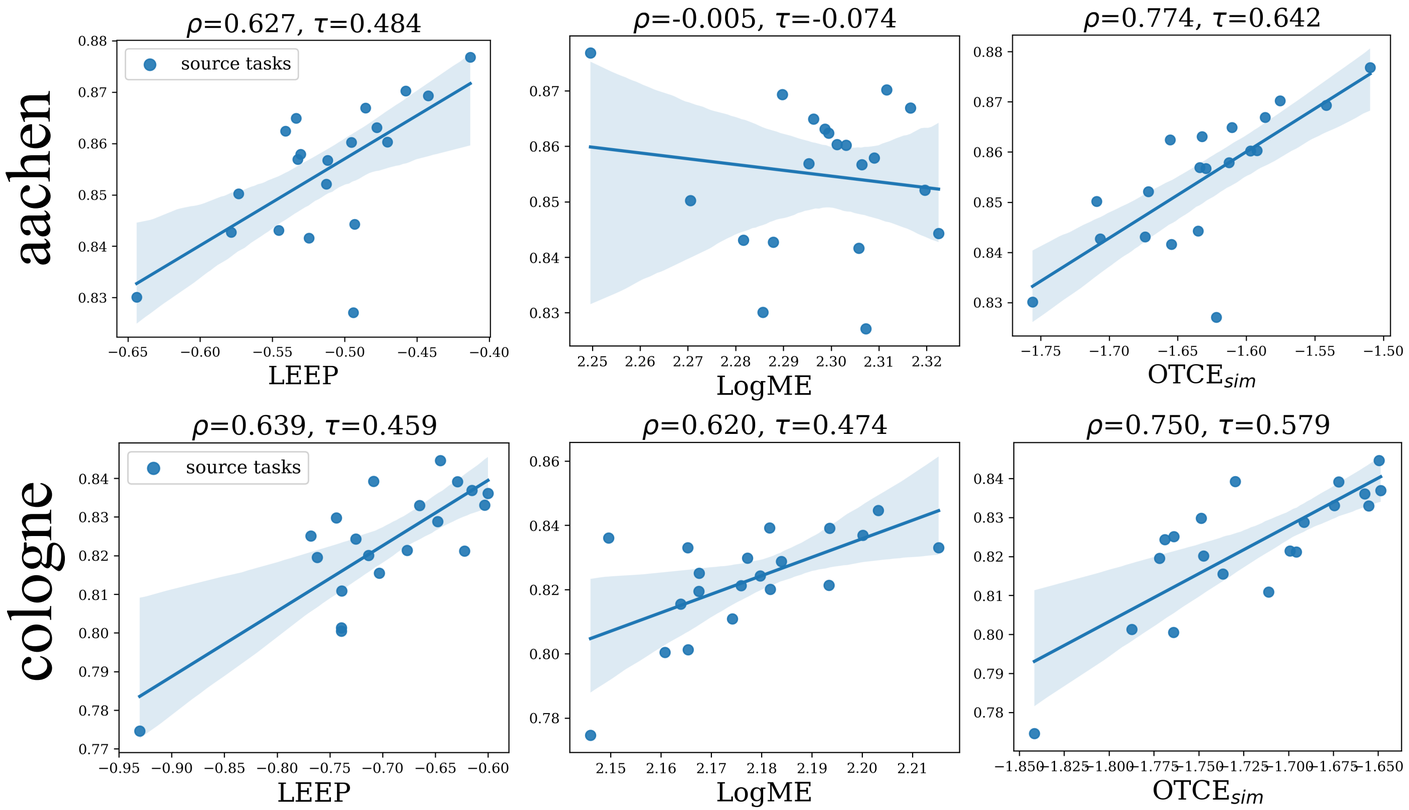


GTA5
(computer game)



Results: Intra-dataset transfer

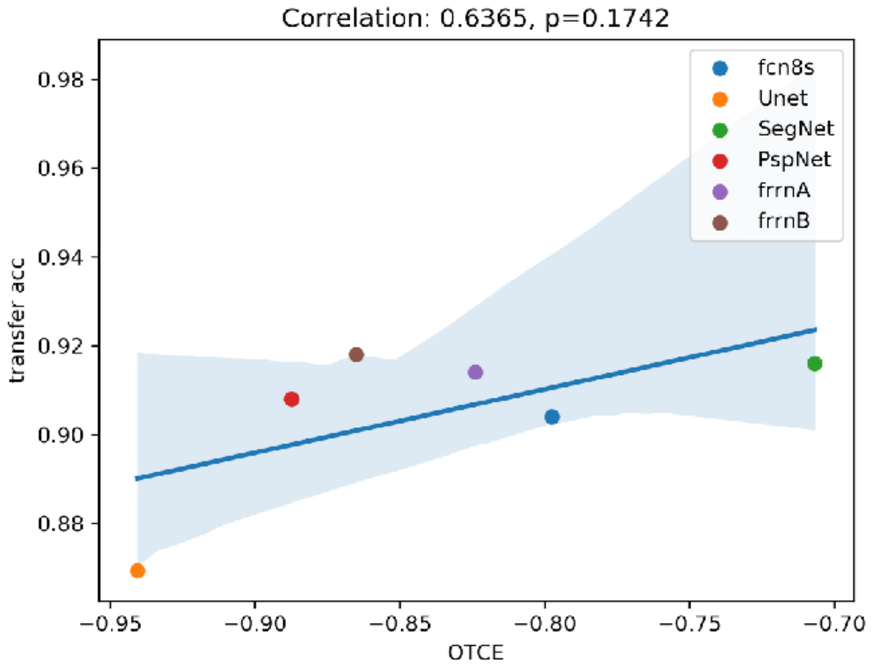
- **Two cities in Cityscapes**
- Comparing OT-CE_{sim} with LEEP and LogME



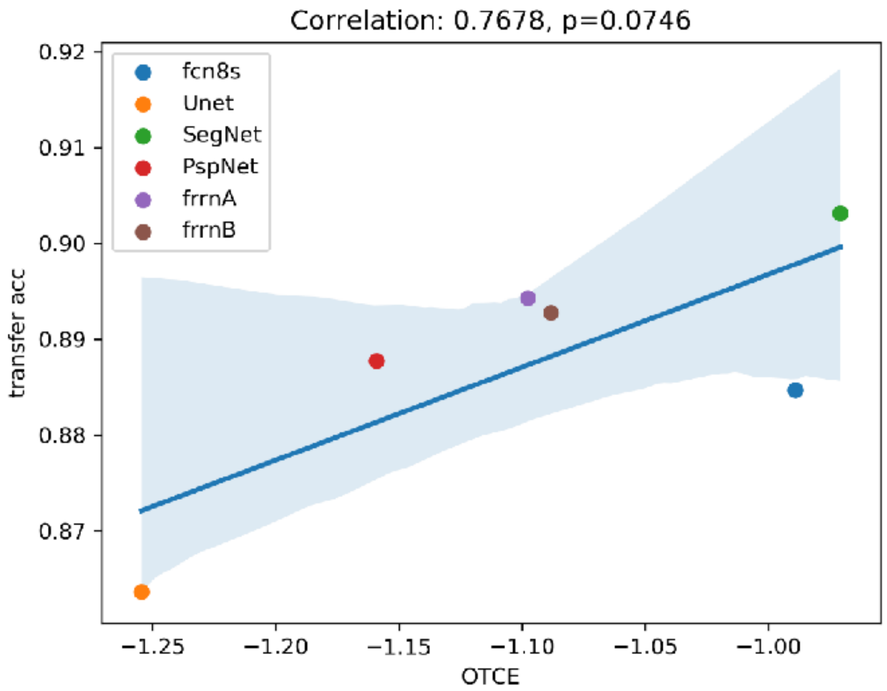
Results: cross-dataset transfer

- 6 model architectures: Fcn8s, UNet, SegNet, PspNet, FrrnA, FrrnB
- Comparing OT-CE_{sim} with transfer accuracy

source: BDD100k
target: Cityscapes



source: GTA5
target: Cityscapes



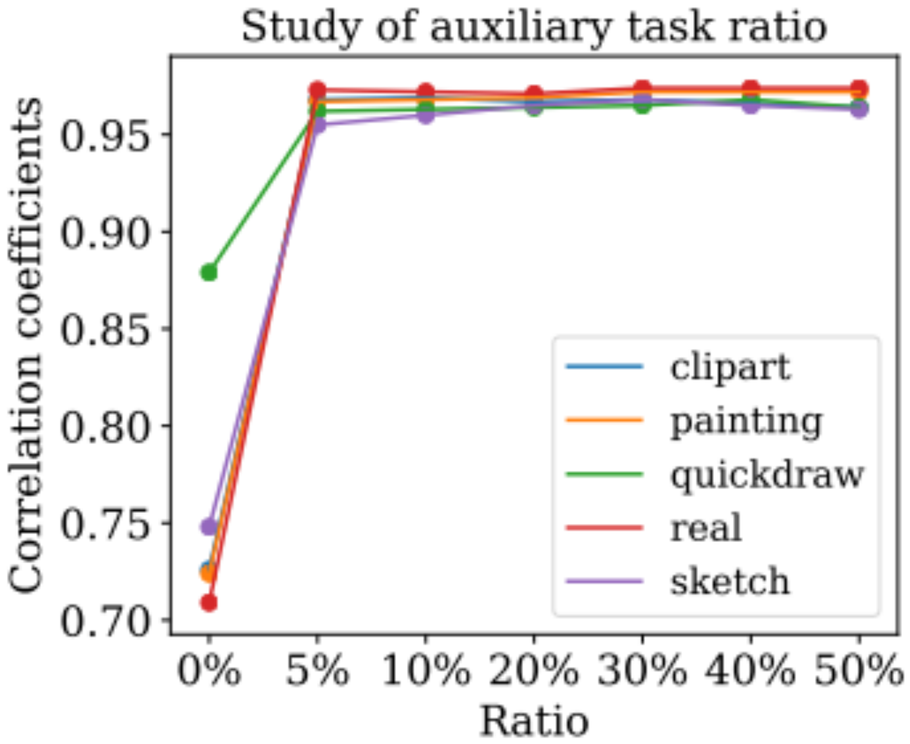
Comparison with Other Works

Transfer setting	Target task	Spearman			Kendall		
		OTCE$_{sim}$	LEEP [6]	LogME [15]	OTCE$_{sim}$	LEEP [6]	LogME [15]
Intra-dataset Transfer	aachen	**0.774**	0.627	-0.005	**0.642**	0.484	-0.074
	cologne	**0.750**	0.639	0.620	**0.579**	0.459	0.474
	erfurt	**0.750**	0.585	0.565	**0.575**	0.432	0.400
	jena	0.735	**0.868**	0.561	0.579	**0.695**	0.400
	strasbourg	0.791	0.680	**0.838**	0.632	0.505	**0.684**
Inter-dataset Transfer (source: BDD100K)	aachen	**0.771**	0.371	0.371	**0.600**	0.333	0.200
	cologne	**0.657**	0.371	0.600	**0.467**	0.333	0.333
	erfurt	0.086	**0.714**	0.257	0.200	**0.600**	0.067
	jena	0.600	0.314	**0.657**	**0.467**	0.200	**0.467**
	strasbourg	**0.657**	0.429	**0.657**	**0.467**	0.333	**0.467**
Inter-dataset Transfer (source: GTA5)	aachen	0.200	0.314	**0.429**	0.067	0.200	**0.333**
	cologne	**0.829**	0.429	0.771	**0.733**	0.200	0.600
	erfurt	0.600	0.543	**0.943**	0.467	0.333	**0.867**
	jena	0.714	0.257	**0.886**	0.467	0.200	**0.733**
	strasbourg	**0.886**	-0.029	0.429	**0.733**	0.067	0.200
	Average	**0.653**	0.474	0.572	**0.512**	0.358	0.410

bold denotes the best result and underline denotes the 2nd best result.

OT-CE Discussion

- Shown to be effective for both classification and semantic segmentation tasks
- OT-CE and OT-CEsim can out-perform other state-of-the-art transferability metrics in cross-domain cross-task settings
- Only a small number of auxiliary tasks are needed for OT-CE

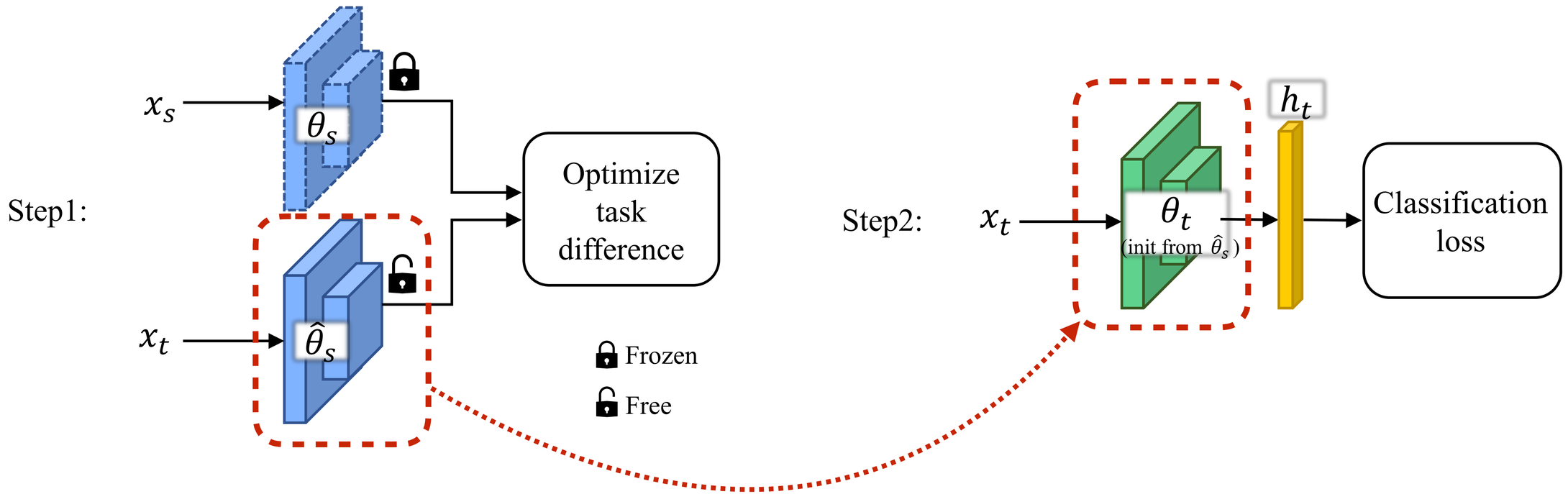


Talk Outline

- Analytical Transferability Metrics
  - H-score (efficient cross-task transferability metric)
  - OT-CE (cross-domain, cross-task transferability metric)
- Transferability-guided fine-tuning for few-shot transfer learning
- Measuring Transferability for Medical Segmentation

OT-CE Score based Fine Tuning

- Step 1: Minimize task difference
- Step 2: Fine-tune on target data



Learning transferable feature

Source data

$$D_s = \{(x_s^i, y_s^i)\}_{i=1}^m \sim P_s(x, y)$$

Target data

$$D_t = \{(x_t^i, y_t^i)\}_{i=1}^n \sim P_t(x, y)$$

Source Model

$$(\theta, h_s)$$

1) Solve the Optimal Transport problem via Sinkhorn Algorithm

$$OT(D_s, D_t) \triangleq \min_{\pi \in \Pi(D_s, D_t)} \sum_{i,j=1}^{m,n} c(\theta(x_s^i), \theta(x_t^j))\pi_{ij} + \epsilon H(\pi)$$

$$\hat{P}(y_s, y_t) = \sum_{i,j: y_s^i = y_s, y_t^j = y_t} \pi_{ij}^*, \ \hat{P}(y_s) = \sum_{y_t \in \mathcal{Y}_t} \hat{P}(y_s, y_t).$$

2) Minimize the conditional entropy (task difference)

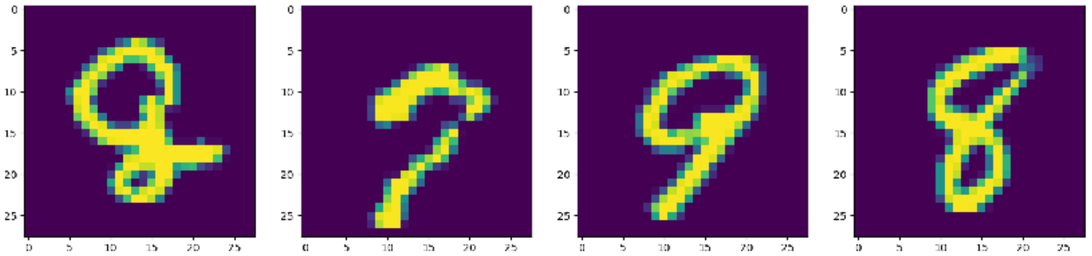
$$\min_\theta \ H(Y_t|Y_s) = -\min_\theta \sum_{y_t \in \mathcal{Y}_t} \sum_{y_s \in \mathcal{Y}_s} \hat{P}(y_s, y_t) \log \frac{\hat{P}(y_s, y_t)}{\hat{P}(y_s)}$$

Then we can finetune (theta, h_t) on the target training data and evaluate the transfer performance

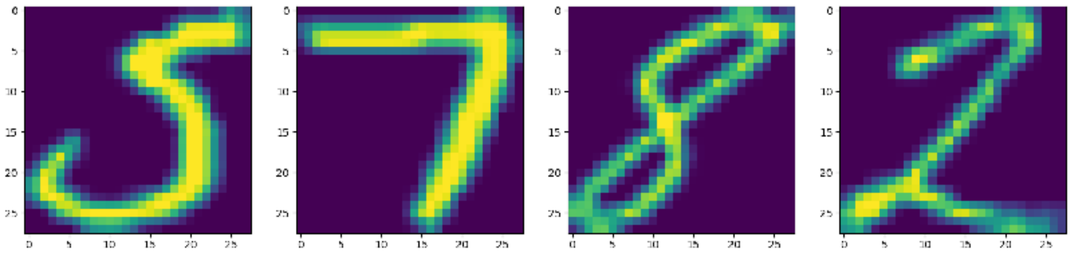
Few-Shot Transfer Learning Dataset

- Dataset visualization

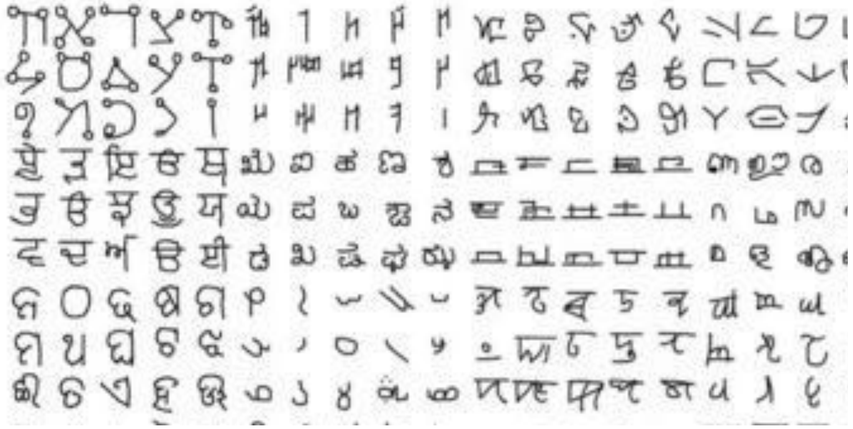
MNIST (number 0-9)



USPS (number 0-9)



Omniglot (characters of 1,623 classes)



Few-Shot Transfer Learning Results

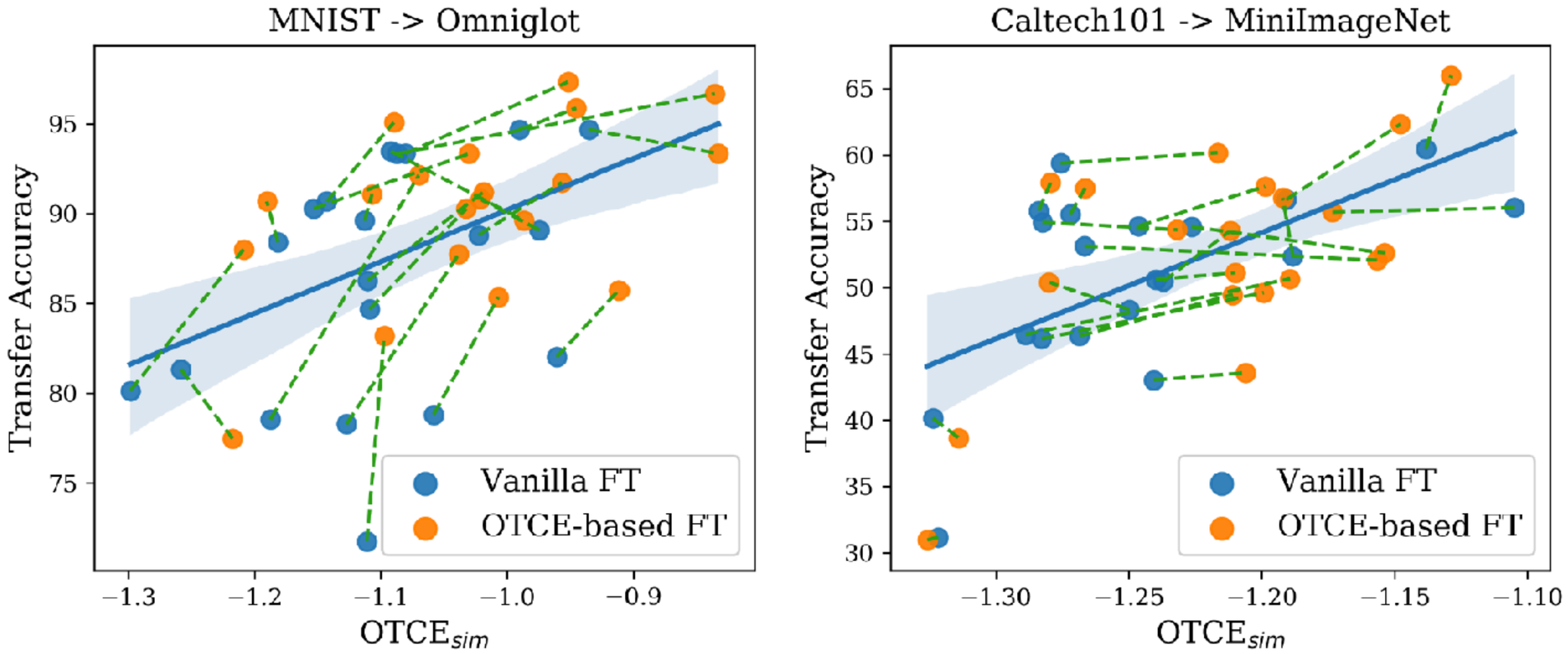
source: **MNIST** (number **0-6**)
target: **USPS** (number **7-9**) (3-way-5-shot task)

Method	Target testing acc
Train from scratch	81.16%
Vanilla finetune	86.43%
OTCE-based finetune	**91.29% (+4.86%)**

source: **MNIST** (number **0-9**)
target: **Omniglot** (randomly sample 100 5-way-5-shot tasks)

Method	Target testing acc
Vanilla finetune	86.11%
OTCE-based finetune	**90.52% (+4.41%)**

Few-Shot Transfer Learning Results



- On average. OTCE-based fine tuning has higher transferability score and higher accuracy than vanilla fine-tuning.

Comparison with other few-shot learning methods

Model	Method	MNIST → Omniglot	Caltech101 → MiniImageNet
FewshotNet	MAML [29]	$88.60 \pm 1.14\%$	$28.23 \pm 0.44\%$
	MatchingNet [27]	$87.92 \pm 1.10\%$	$44.75 \pm 1.30\%$
	ProtoNet [30]	$83.11 \pm 1.34\%$	$50.40 \pm 1.35\%$
	RelationNet [28]	$69.35 \pm 1.62\%$	$29.55 \pm 0.61\%$
	Vanilla finetune	$91.30 \pm 0.95\%$	$49.49 \pm 1.27\%$
	OTCE-based finetune	$\mathbf{92.32 \pm 0.87\%}$	$\mathbf{51.36 \pm 1.33\%}$
LeNet	Vanilla finetune	$86.11 \pm 1.10\%$	-
	OTCE-based finetune	$\mathbf{90.52 \pm 0.94\%}$	-
ResNet-18	Vanilla finetune	-	$48.48 \pm 1.39\%$
	OTCE-based finetune	-	$\mathbf{50.02 \pm 1.34\%}$

meta learning based methods

Fine-Tuning Discussion

- Meta learning doesn’t work as well as Vanilla fine-tuning in cross-domain setting
- Minimizing the task difference characterized by OTCE is an effective way to obtain a better initial source weights so that we can achieve higher transfer performance on target task.

Talk Outline

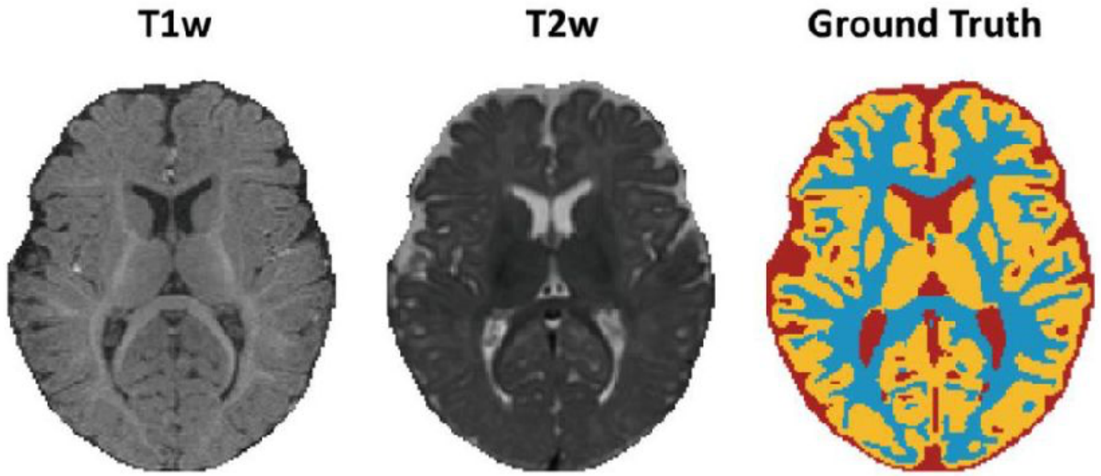
- Analytical Transferability Metrics
  - H-score (efficient cross-task transferability metric)
  - OT-CE (cross-domain, cross-task transferability metric)
- Transferability-guided fine-tuning for few-shot transfer learning
- Measuring Transferability for Medical Segmentation

Transfer Learning in MRI Segmentation Tasks

- An important problem in federated learning and distributed meidcal AI systems

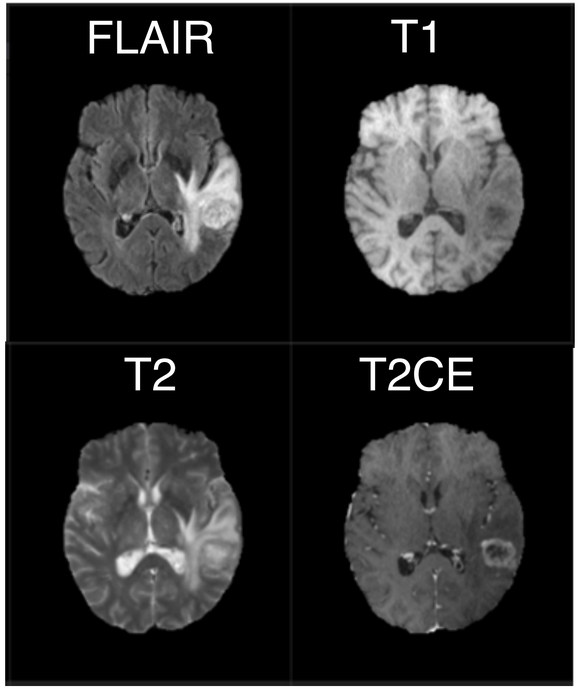
iSeg2019: brain matter segmentation

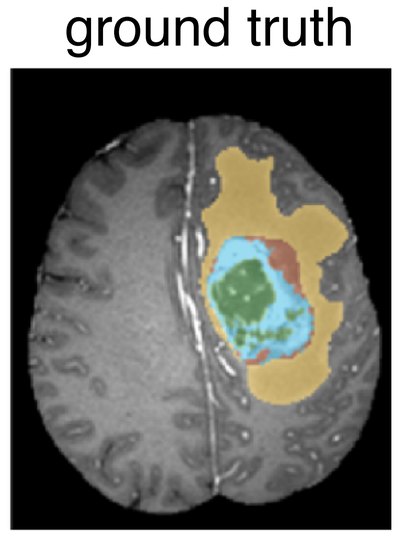
3 tasks, 2 modalities



FeTS Challenge: brain tumor segmentation

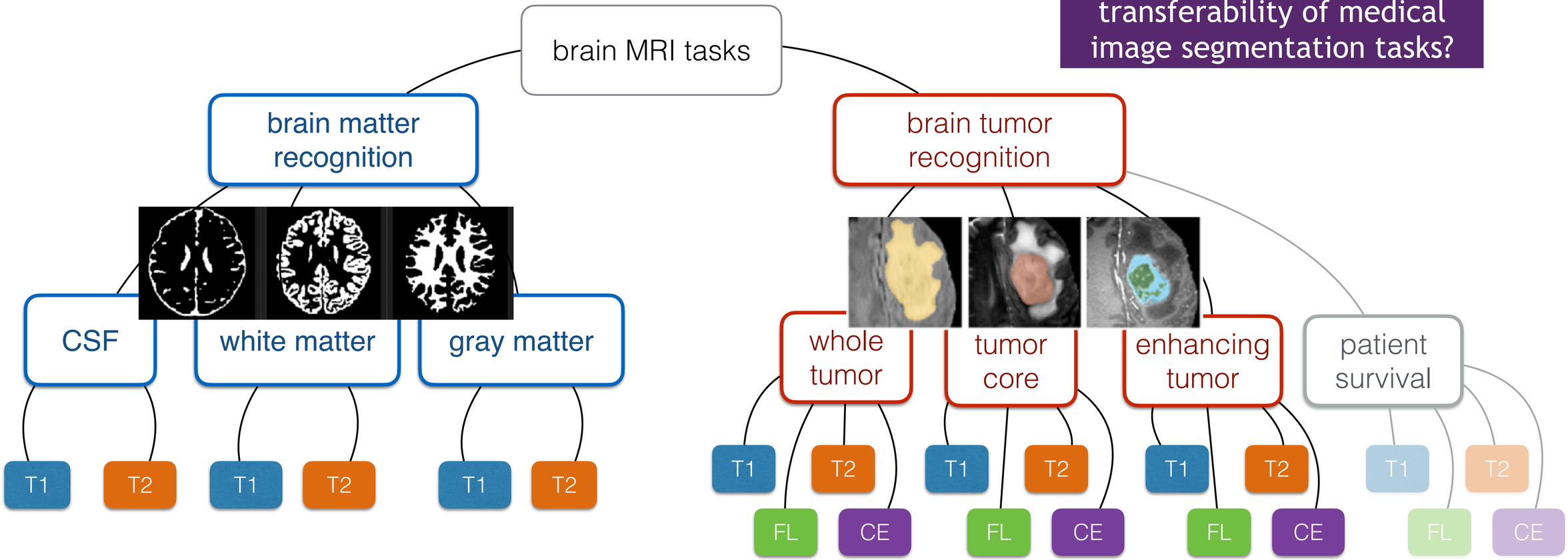
4 tasks, 4 modalities





Empirical Study on Transferability for Brain MRI Segmentation

What affects the transferability of medical image segmentation tasks?



6 modality-task combinations

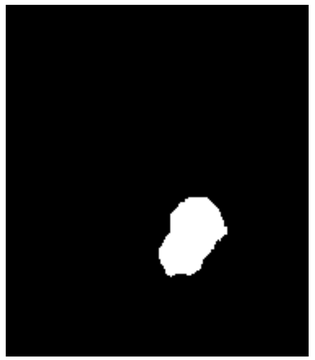
12 modality-task combinations

Finding #1: Transferring from related tasks is easier then unrelated tasks

brain tumor recognition

brain matter recognition

ground truth label



tumor core

white matter

transfer from related source

whole tumor -> tumor core

gray matter -> white matter

transfer from unrelated source

gray matter -> tumor core

whole tumor -> white matter

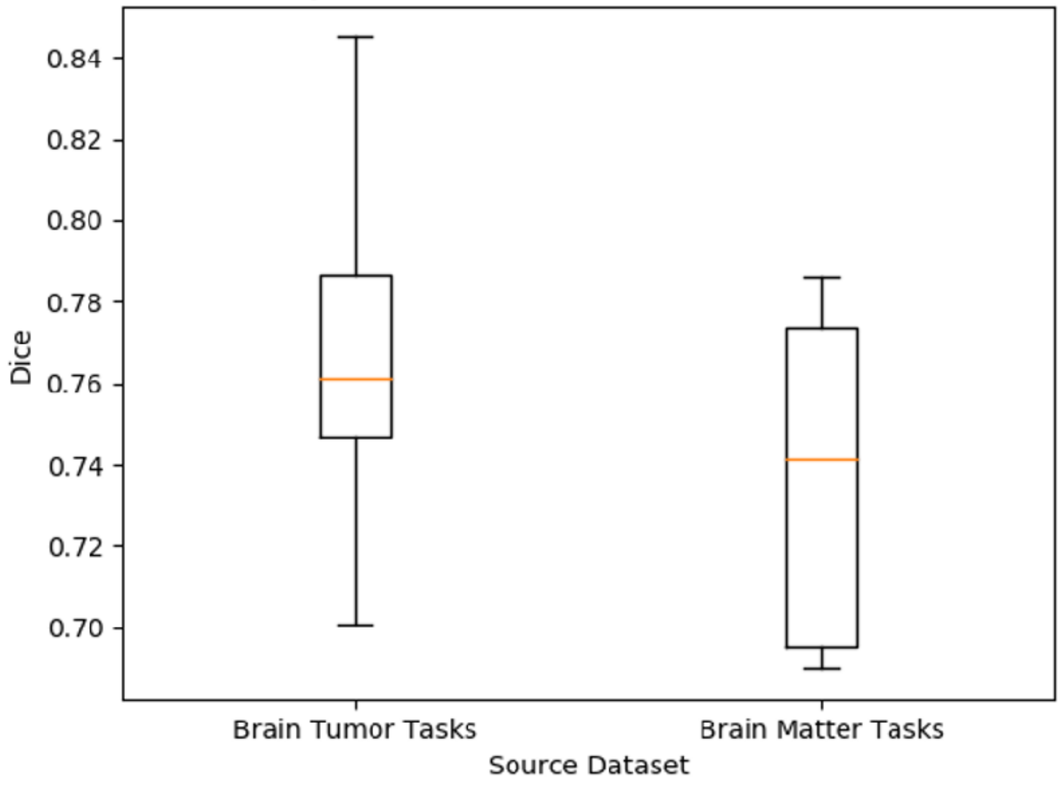
Finding #1: Transferring from related tasks is easier then unrelated tasks

- many-to-one transfer experiments

target: enhancing tumor

modality: T1CE

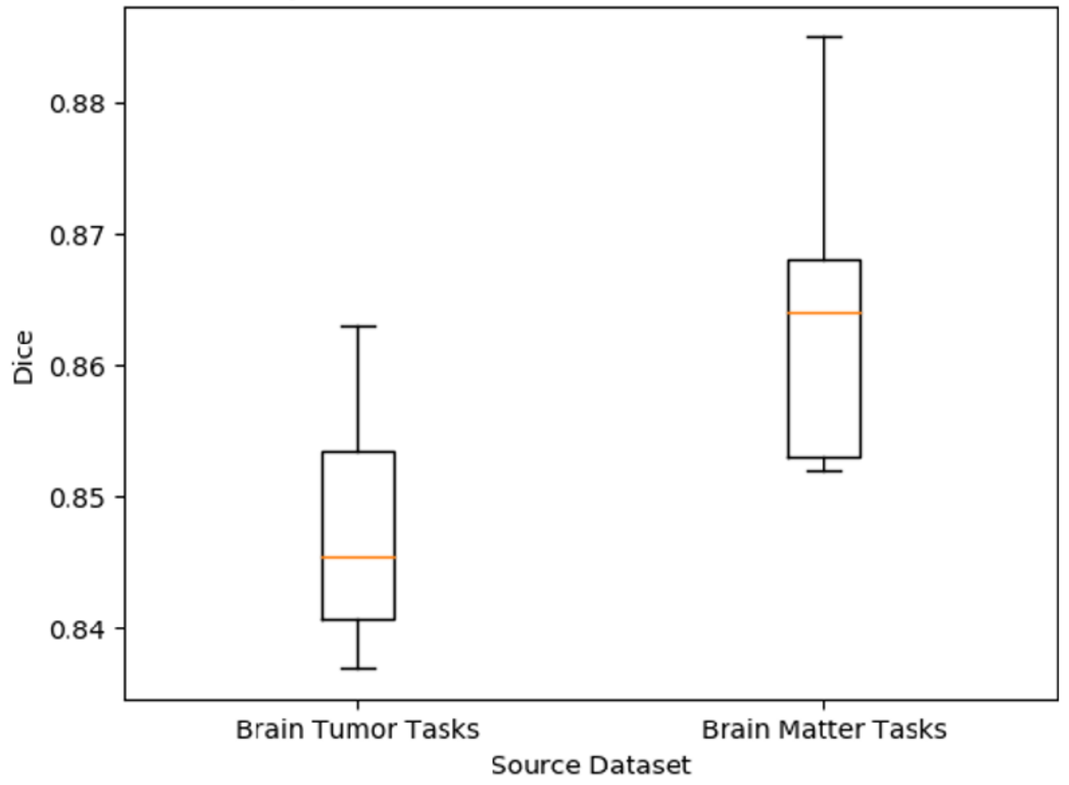
brain tumor recognition



target: white matter

modality: T1

brain matter recognition

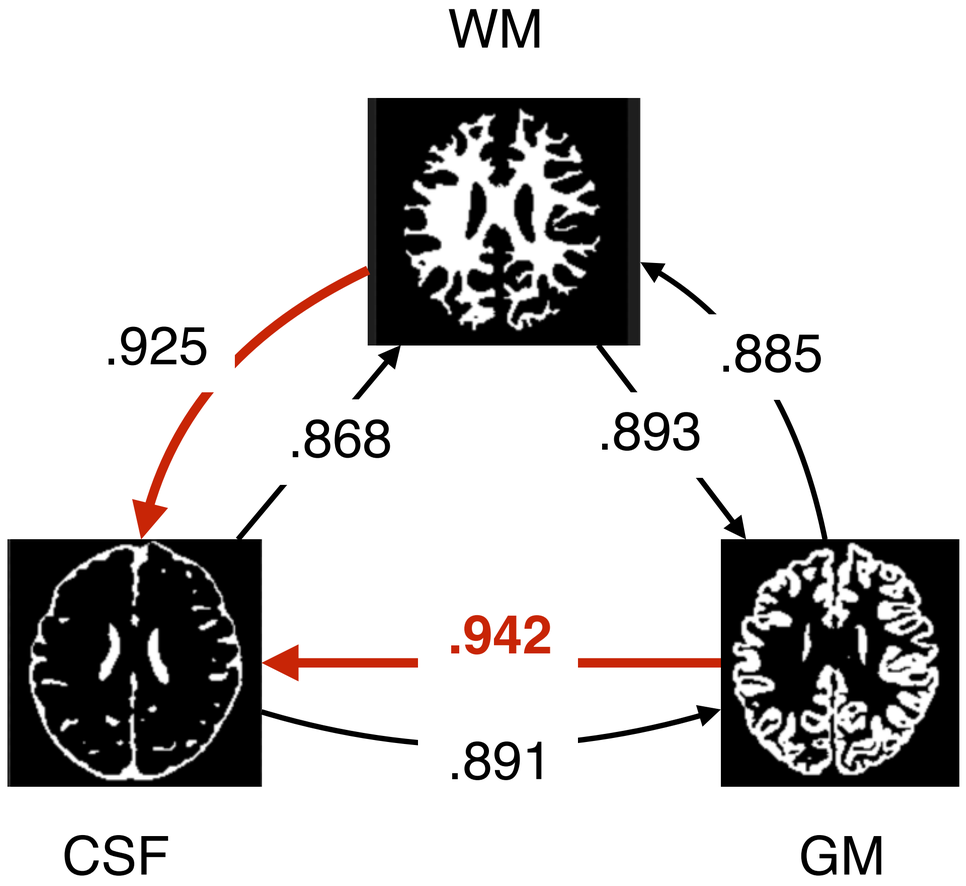


Finding #2: Cross-task transfer is easier than cross-modality transfer

Mean Transfer Accuracy (dice loss)

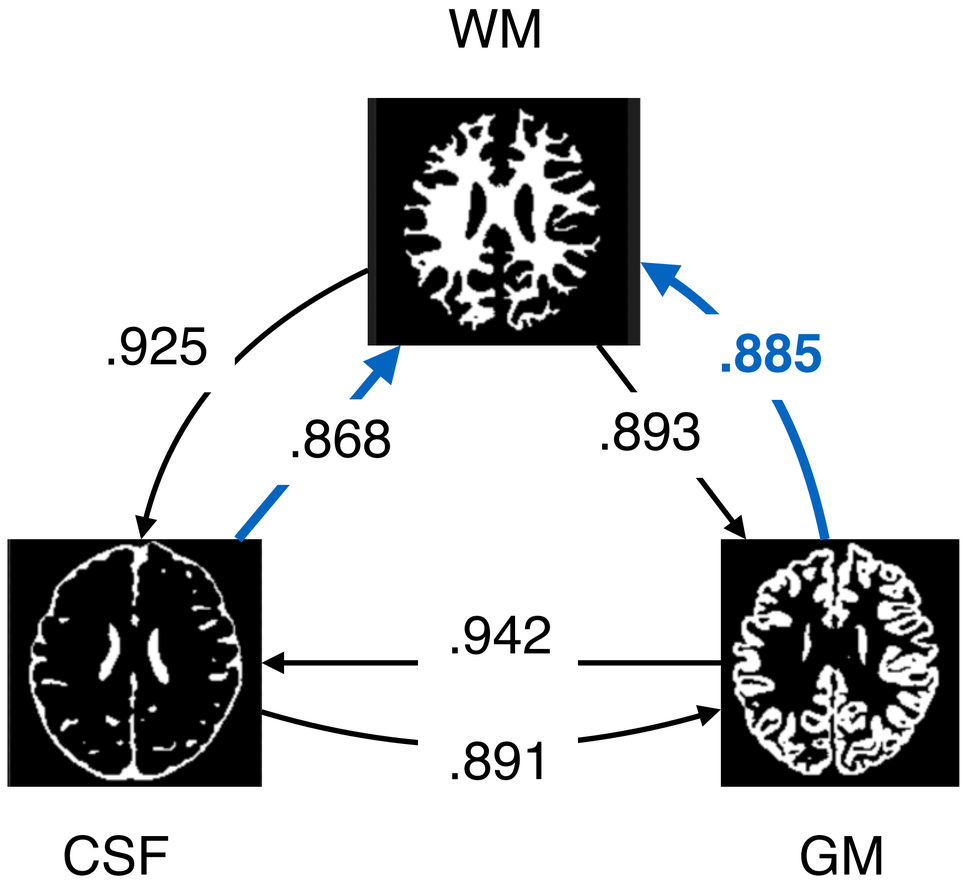
		Target Task					
		ET,T1CE	ED,T1CE	NCR,T1CE	WM,T1	GM,T1	CSF,T1
Source Tasks	**same task different modality**	0.755	0.731	0.726	0.864	0.881	0.935
	different task same modality	0.821	0.786	0.782	0.877	0.892	0.934

Finding #3: geometric task relationships affect transferability



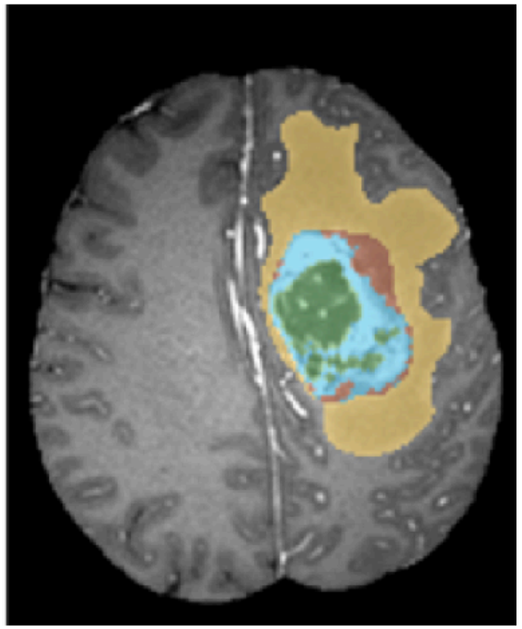
Mutually exclusive case: “complementary tasks” are more transferable than “non-complementary tasks”

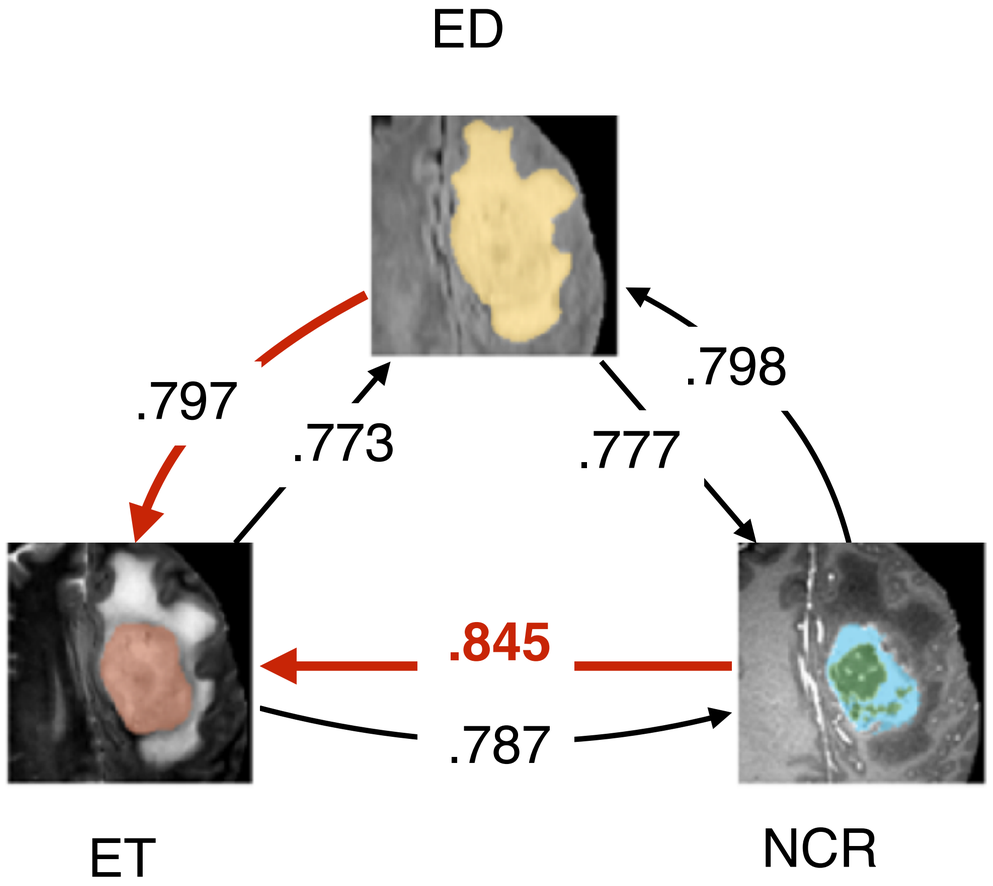
Finding #3: geometric task relationships affect transferability



Mutually exclusive case: “complementary tasks” are more transferable than “non-complementary tasks”

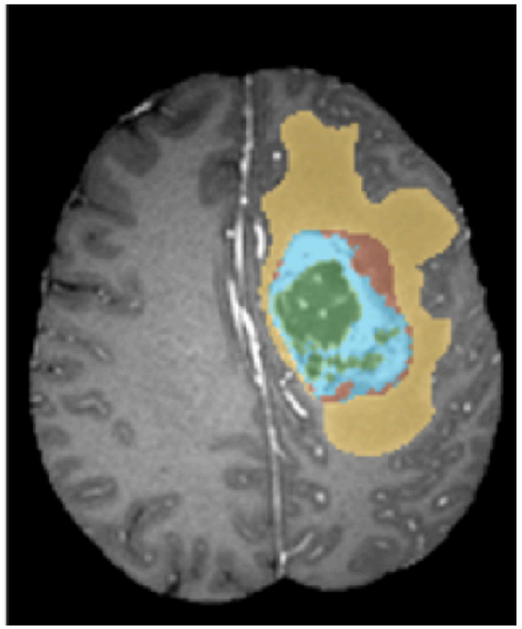
Finding #3: geometric task relationships affect transferability

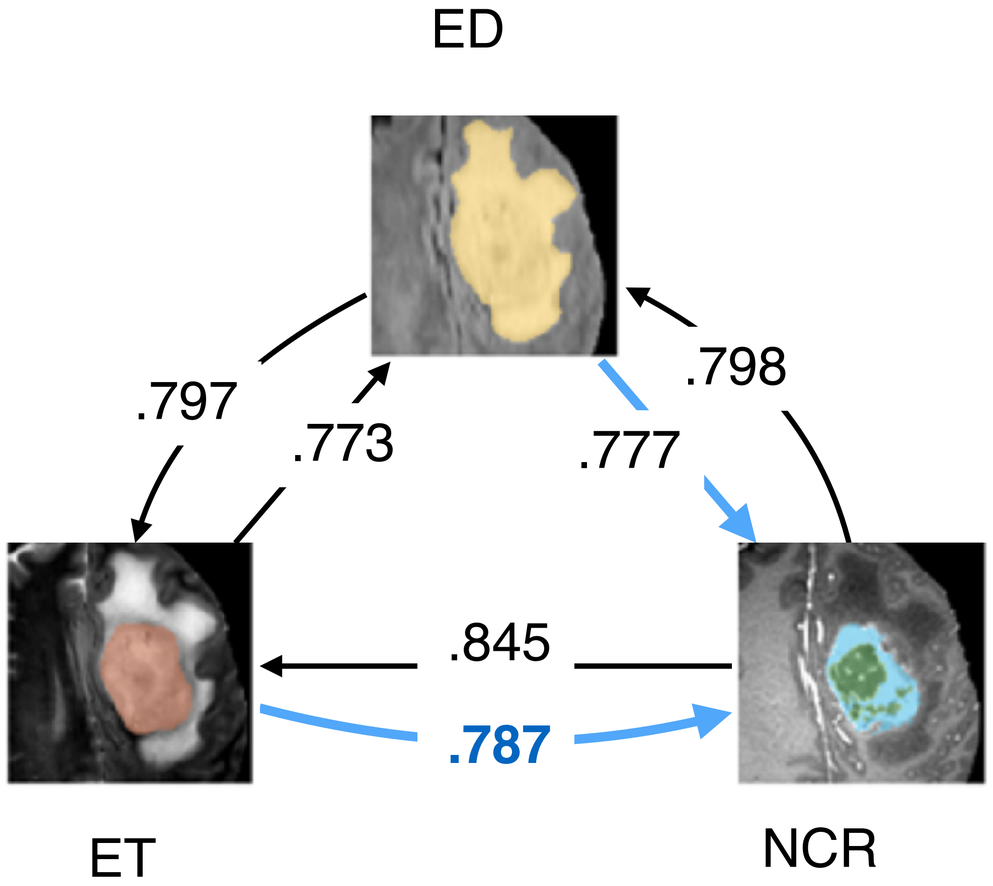




Subset case: “smaller task” is more transferable than “bigger task”

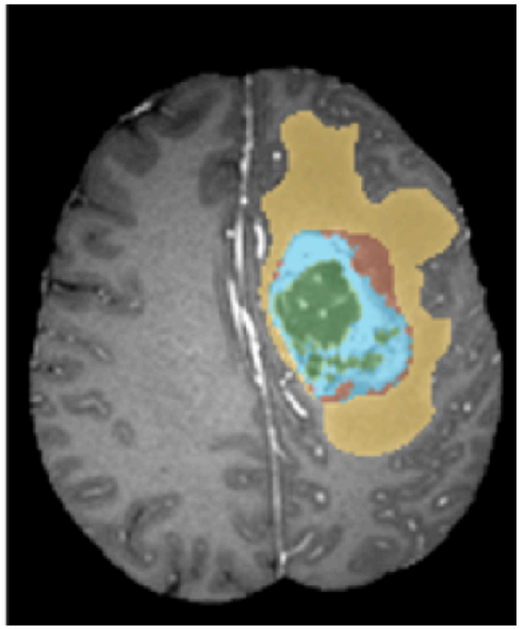
Finding #3: geometric task relationships affect transferability

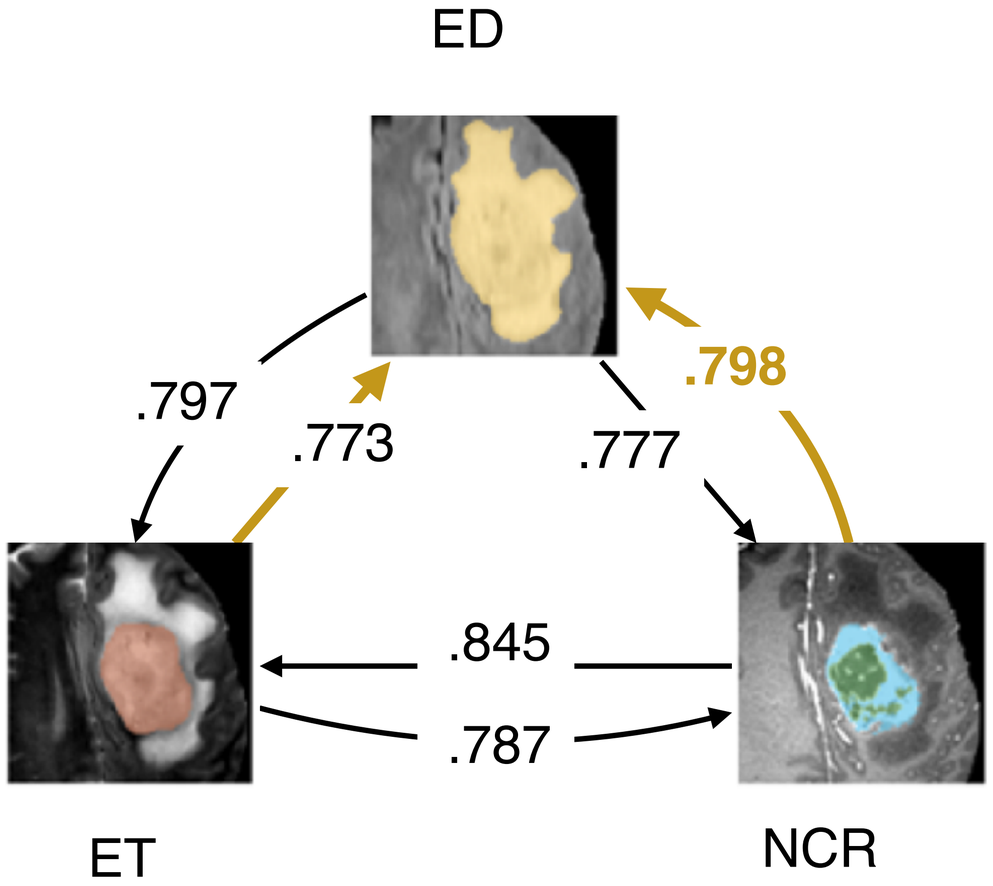




Subset case: “smaller task” is more transferable than “bigger task”

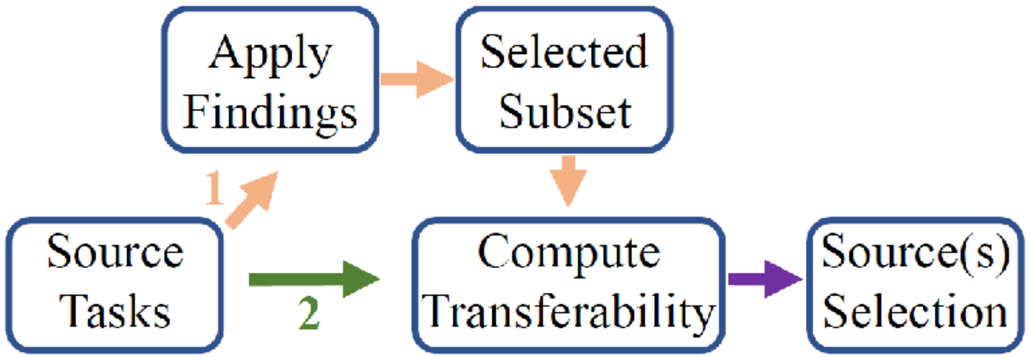
Finding #3: gemetric task relationships affect transferability



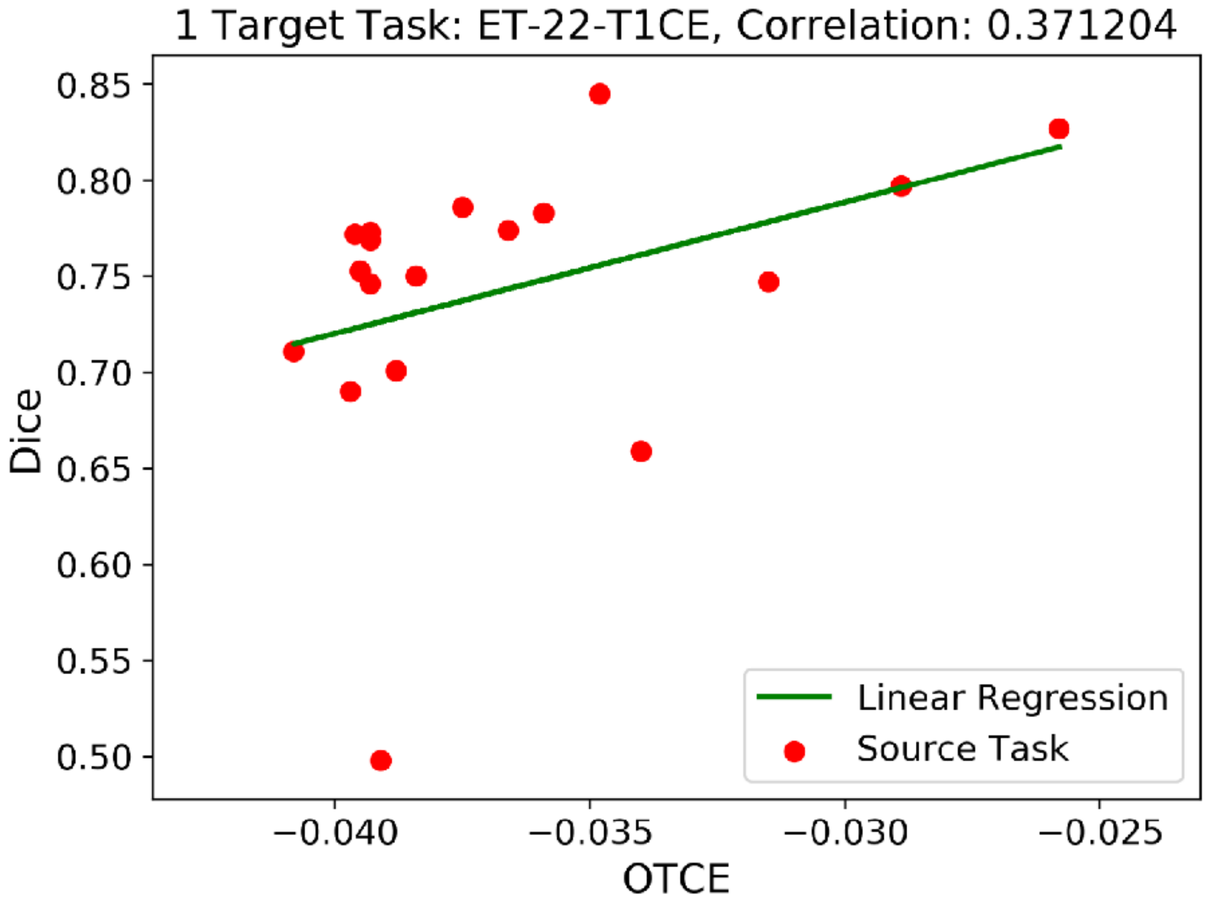


Subset case: “smaller task” is more transferable than “bigger task”

Source Task Selection



Source(s) Selection Pipeline



Source Task Selection

- Reduced the size of candidate source task pool to 16
- Analytical transferability metrics are more accurate on the filtered candidate set (smaller rank displacement)

Spearsman's Footrule (Total Rank Displacement)

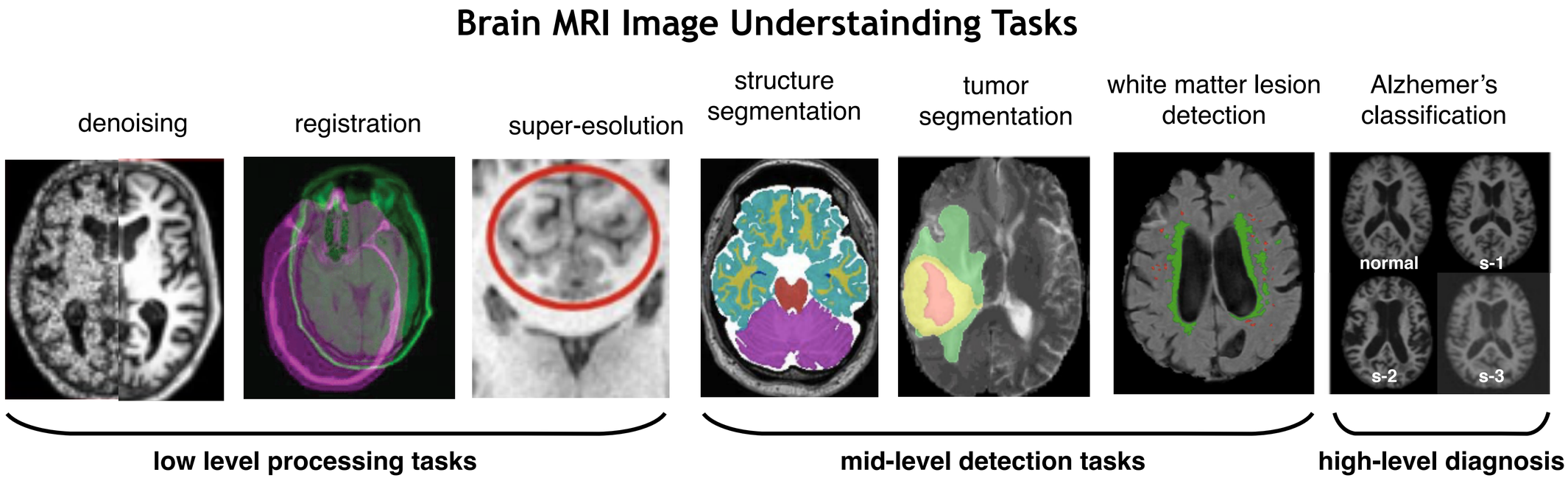
Target	Method	Top 1	Top 2	Top 3	Top 4
ET-22-T2	H-score w/o F	5	10	22	27
	H-score w/ F	4	5	6	7
	OTCE w/o F	2	2	4	12
	OTCE w/ F	2	2	4	7
ET-20-T1	H-score w/o F	14	24	30	40
	H-score w/ F	0	9	9	13
	OTCE w/o F	2	14	17	23
	OTCE w/ F	2	11	13	17

still work in progress

Future Works on Medical Image Transfer Learning

Inspired by how neuroradiologist uses expert knowledge in diagnosis

- Use transferability to automatically decide which tasks to be learned together in multi-task learning
- Design a multi-task transfer curriculum to perform hierachical transfer



Summary

Estimating transferability is important in practical transfer learning

- For same-domain transferability, H-score is very efficient and theoretically proven in information theory (Shrinkage-based H-Score resolves numerical issues)
- For cross-domain case, OT-CE and OT-CE_{sim} are robust to many challenging scenarios, with less data assumption
- Transferability metrics can lead to cross task/domain feature learning algorithms (HGR-regression and OT-CE based fine tuning)

Future work: efficient transferability for regression problems, transfer strategy design based on transferability

Acknowledgement

黄绍伦
(TBSI)

Leonidas Guibas
(Stanford)

awesome collaborators

鲍亚捷
(TBSI,MS 2016)
University of Georgia

谭杨
(TBSI,PhD 2019)

李一悰
(TBSI,MS 2019)

Thank You!
Q&A

Related Papers and Code

- Yajie Bao, **Yang Li**, Shao-Lun Huang*, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. "An information-theoretic approach to transferability in task transfer learning." In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 2309-2313. IEEE, 2019.

- Yang Tan, **Yang Li***, and Shao-Lun Huang. "OTCE: A Transferability Metric for Cross-Domain Cross-Task Representations." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15779-15788. 2021.

- Code and Supplementary Material:

  - http://yangli-feasibility.com/home/ttl.html

  - https://github.com/tanyang1231/OTCE_Transferability_CVPR21