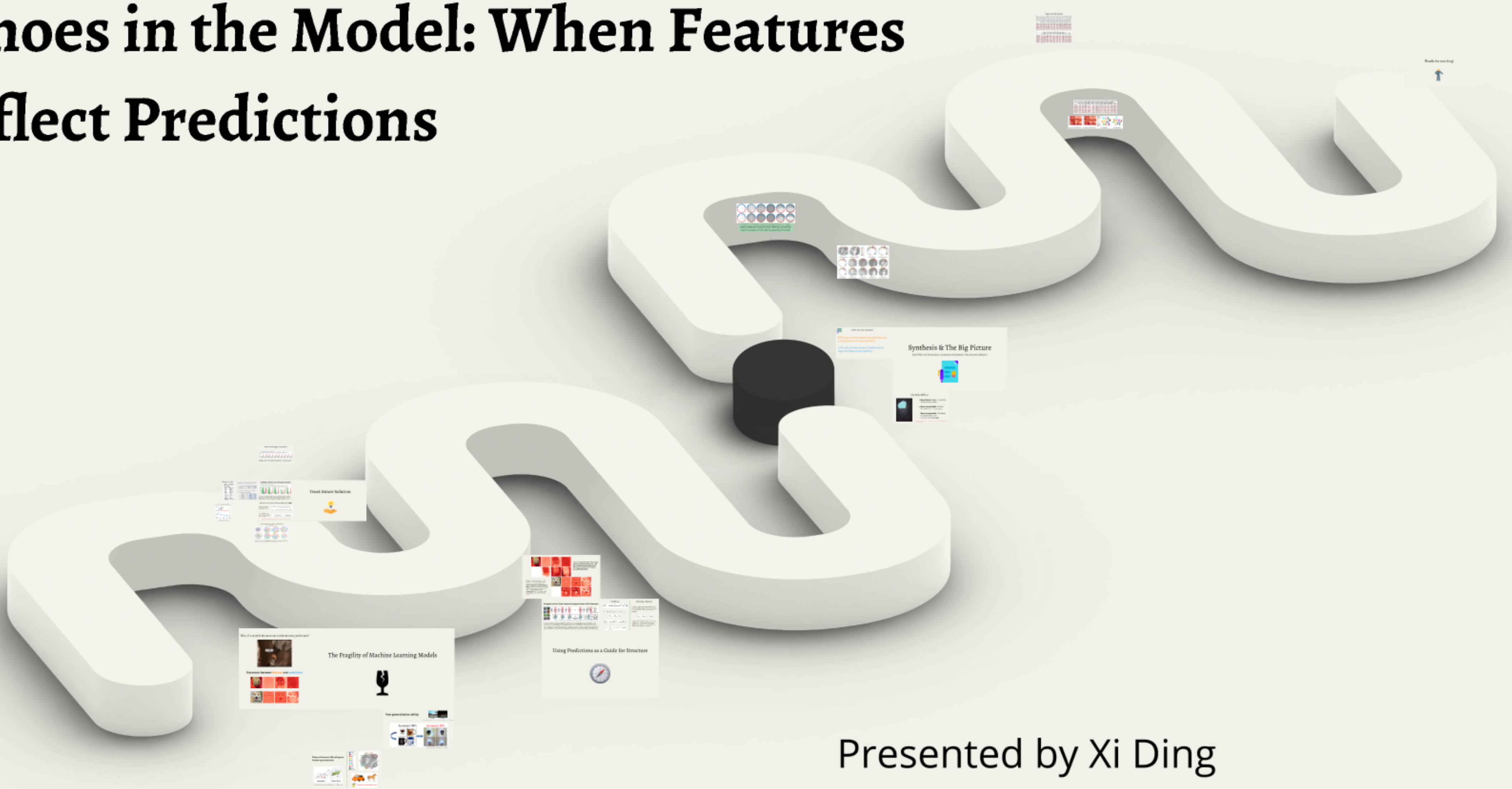
Echoes in the Model: When Features Reflect Predictions

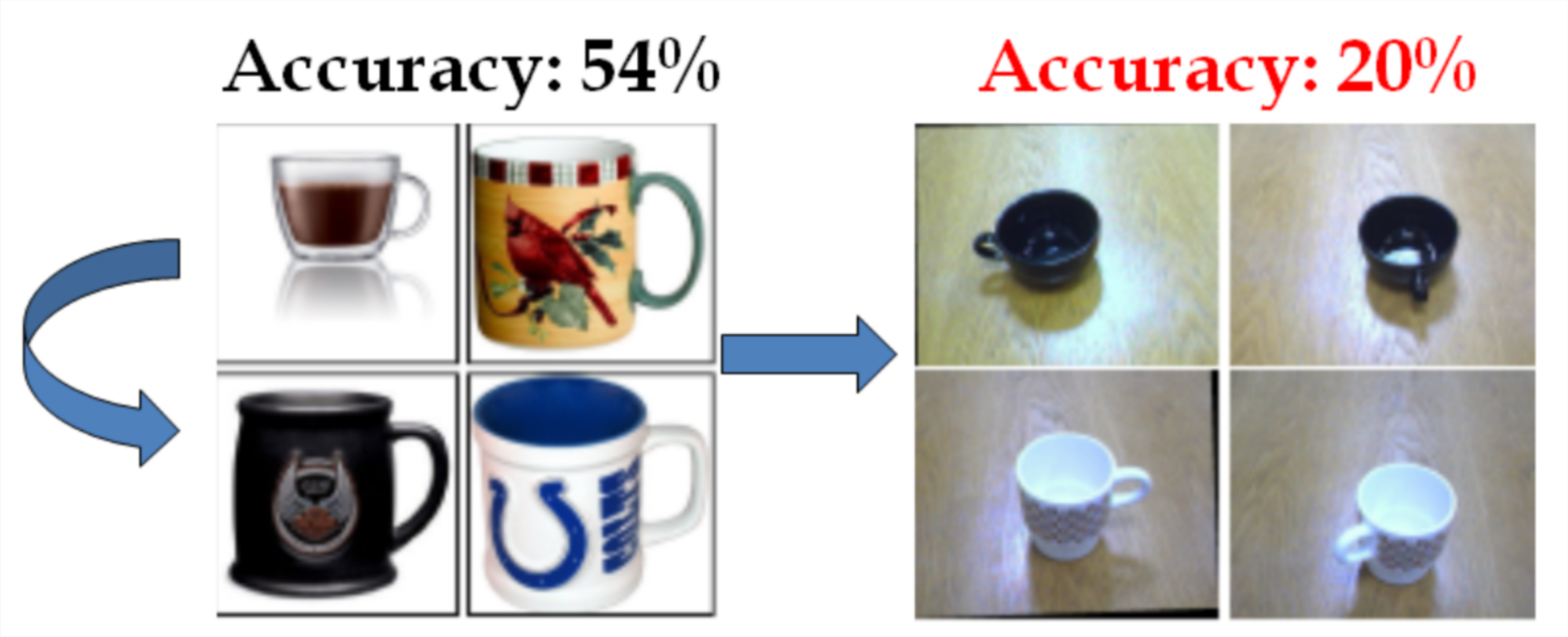


Presented by Xi Ding
TIME Lab @ ARC Hub & Griffith University

The Fragility of Machine Learning Models

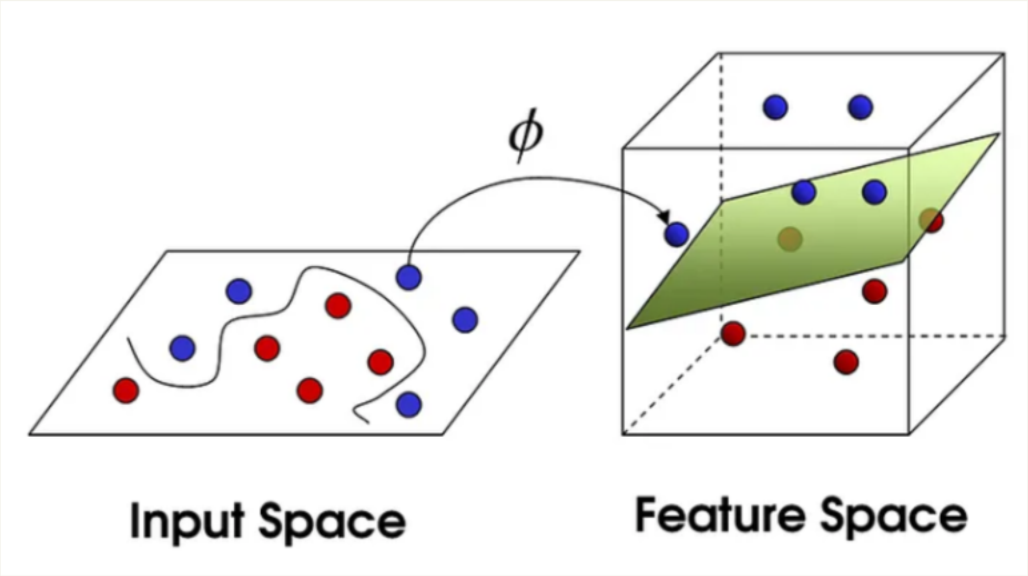
Poor generalization ability

Credit: Tackling Domain Shift in AI: A Deep Dive into Domain Adaptation by Houssem Ben Salem

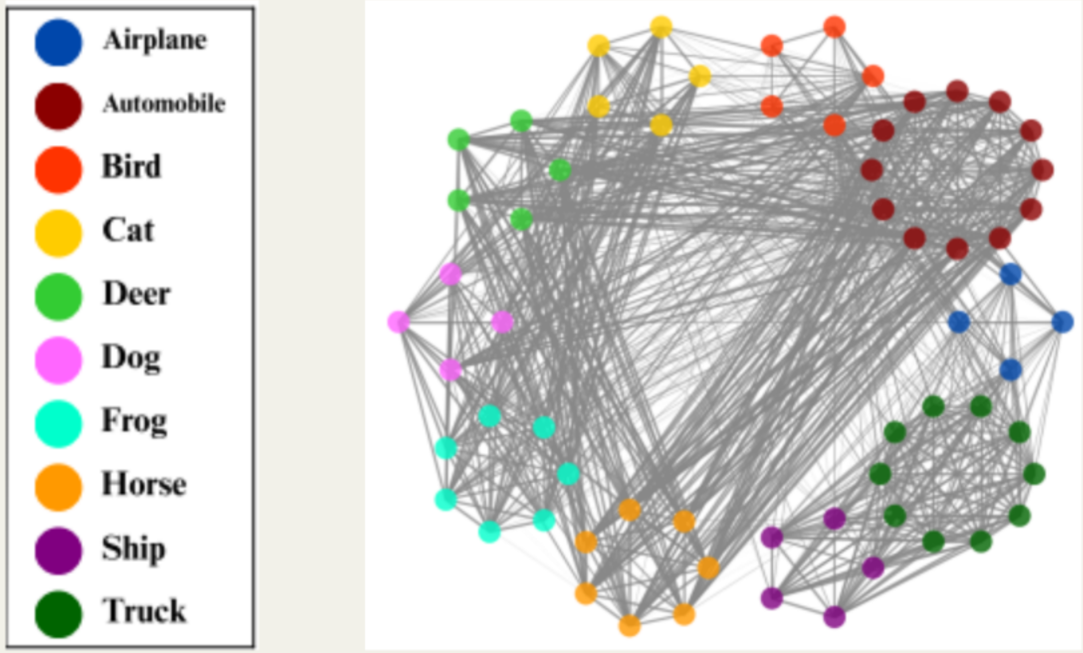


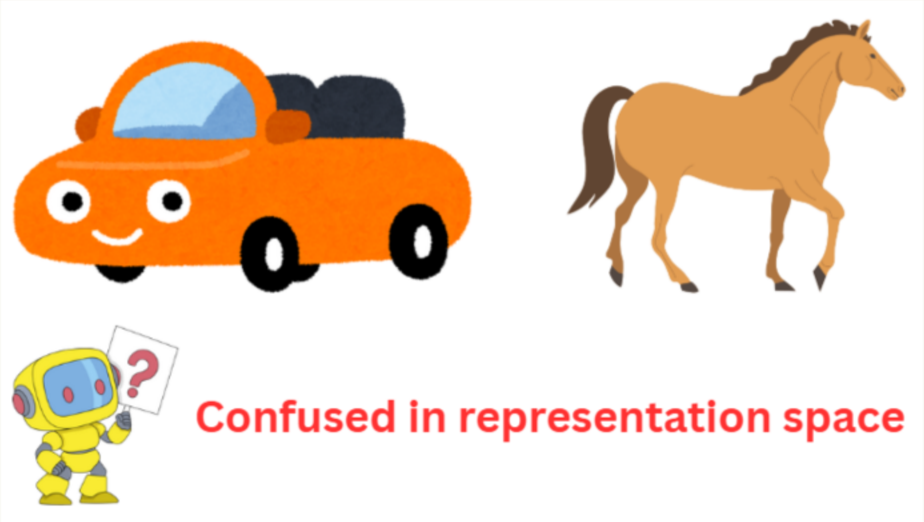
Credit: Domain Adaptation for Object recognition by Kate Saenko

Noisy and semantically ambiguous feature representations



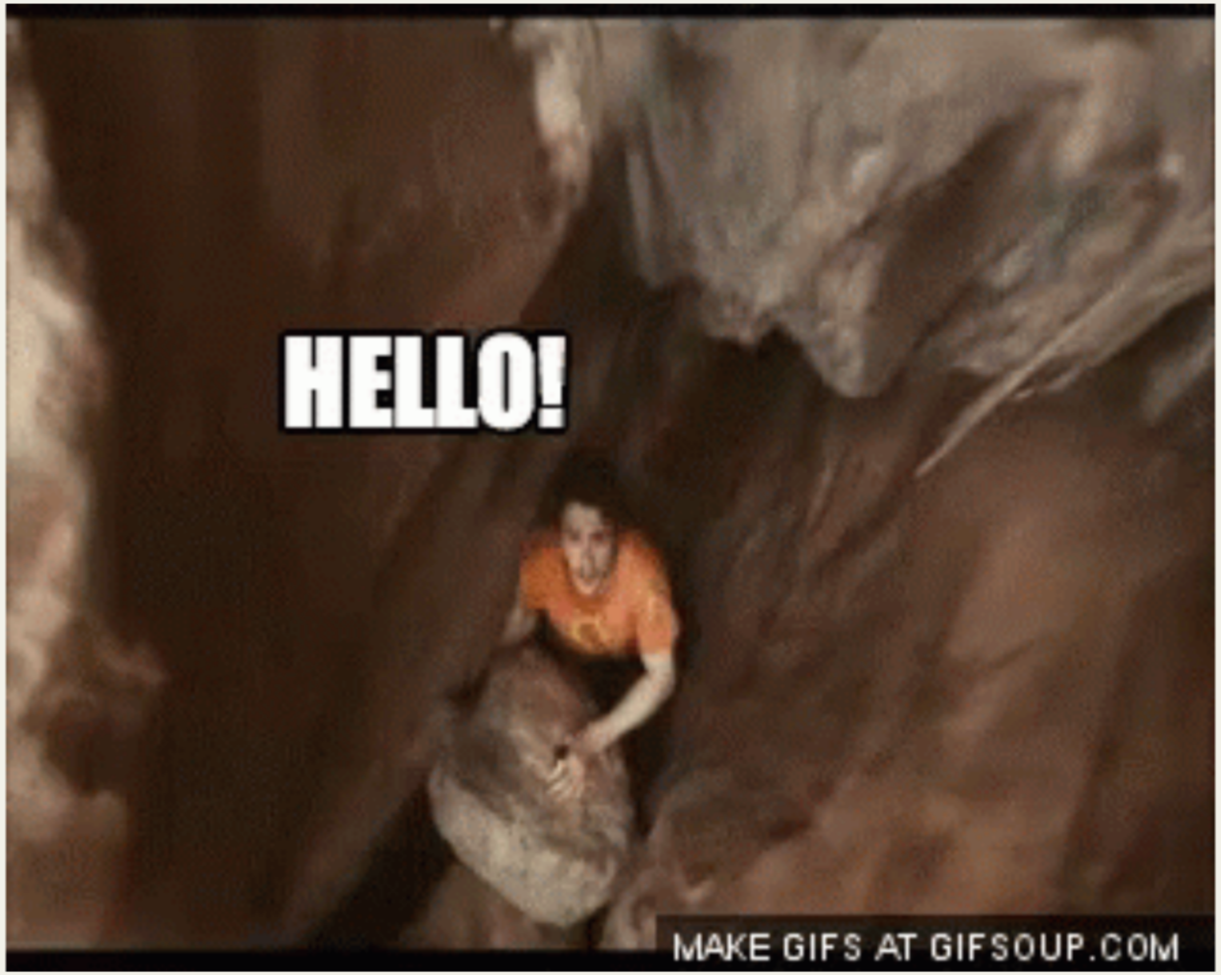
Credit: Representations: The Most Important Thought Framework in Machine Learning





Disconnect between features and predictions

What if a model's features could echo its own predictions?



Trust-Aware Solution

Our Frame work: Joint Feature-Prediction Discrepancy (JFPD)

(i) a trust-weighted discrepancy metric

We combine divergence and trust into the unified JFPD metric:

$$d_{\text{JFPD}} = \alpha \cdot d_{\text{feat}} \cdot \psi + (1 - \alpha) \cdot d_{\text{pred}} \cdot \phi,$$

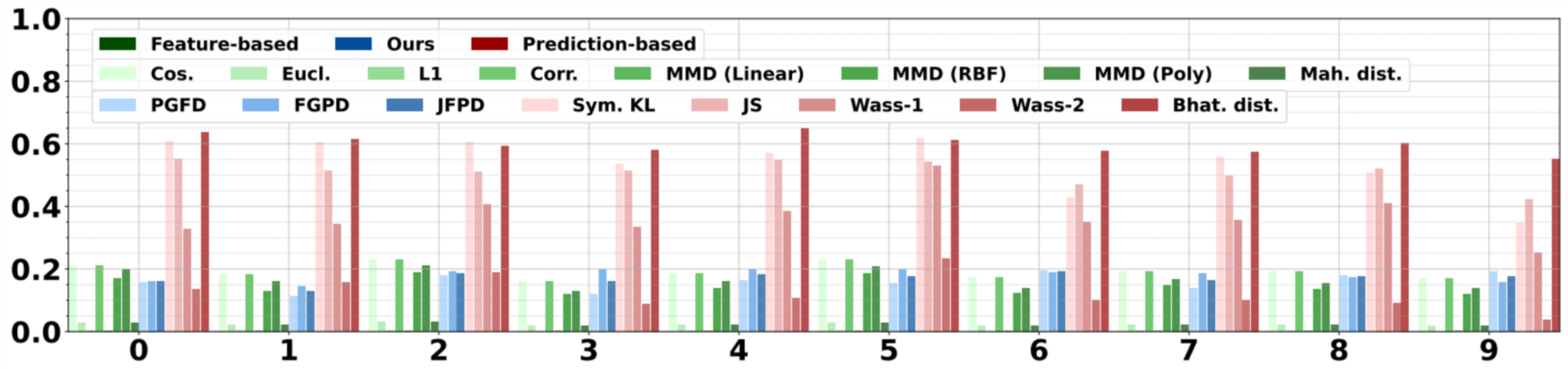
where $\alpha \in [0, 1]$ balances the semantic and categorical components.

(ii) a corresponding fine-tuning loss derived from this metric

$$\mathcal{L}_{\text{JFPD}} = \alpha \cdot \underbrace{\left(d_{\text{feat}}(\mathbf{f}_t, \mathbf{z}^s_{y^t}) \cdot \psi\right)}_{\text{Prediction-Guided Feature Discrepancy}} + (1 - \alpha) \cdot \underbrace{\left(d_{\text{pred}}(\mathbf{p}_t, \mathbf{p}^s_{y^t}) \cdot \phi\right)}_{\text{Feature-Guided Prediction Divergence}},$$

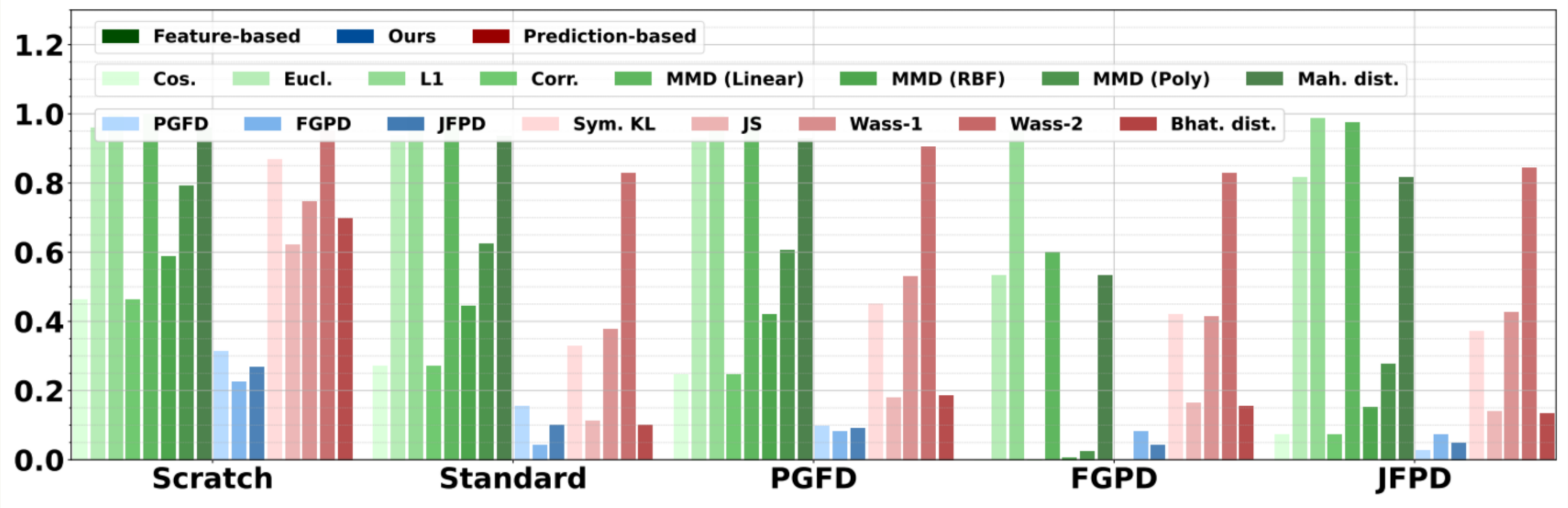
JFPD combines feature and prediction differences, weighted by these trust scores.

Class-wise Δ-divergence comparsion



The vertical axis shows the change in estimated domain gap per digit class and metric, computed as baseline (trained on MNIST only) minus fine-tuned (trained on MNIST, then adapted to SVHN).

Domain gap metrics across fine-tuning methods



Domain gap comparisons for models trained from scratch and fine-tuned using standard cross-entropy, our PGFD, FGPD, and JFPD. Bars represent various feature- and prediction-based discrepancy measures. JFPD consistently reduces domain gaps across both levels, most notably in feature-based metrics, highlighting the need for joint alignment beyond prediction scores.

Experiment results

Table 3: Comparison with SOTA methods on Office-Home.

Method	Venue	Avg
DANN [19]	JMLR 2016	62.7
CDAN+E [57]	NeurIPS 2018	73.9
STA [53]	CVPR 2019	44.7
UAN [98]	CVPR 2019	58.7
ETN [4]	CVPR 2019	64.0
MME [72]	ICCV 2019	73.1
SWD [41]	CVPR 2019	**76.4**
DANCE [73]	NeurIPS 2020	69.1
SHOT [49]	ICML 2020	71.8
APE [36]	ECCV 2020	74.0
HDA+ToAlign [91]	NeurIPS 2021	72.0
FixBi [60]	CVPR 2021	72.7
DCAN+SCDA [46]	ICCV 2021	73.1
CDAC [43]	CVPR 2021	74.8
DECOTA [96]	ICCV 2021	75.7
TransPar-MCC [26]	TIP 2022	73.1
CDTrans-S [94]	ICLR 2022	74.7
ProMM [29]	IJCAI 2023	74.6
MME + SLA [100]	CVPR 2023	75.6
Ours (JFPD)		**76.0**

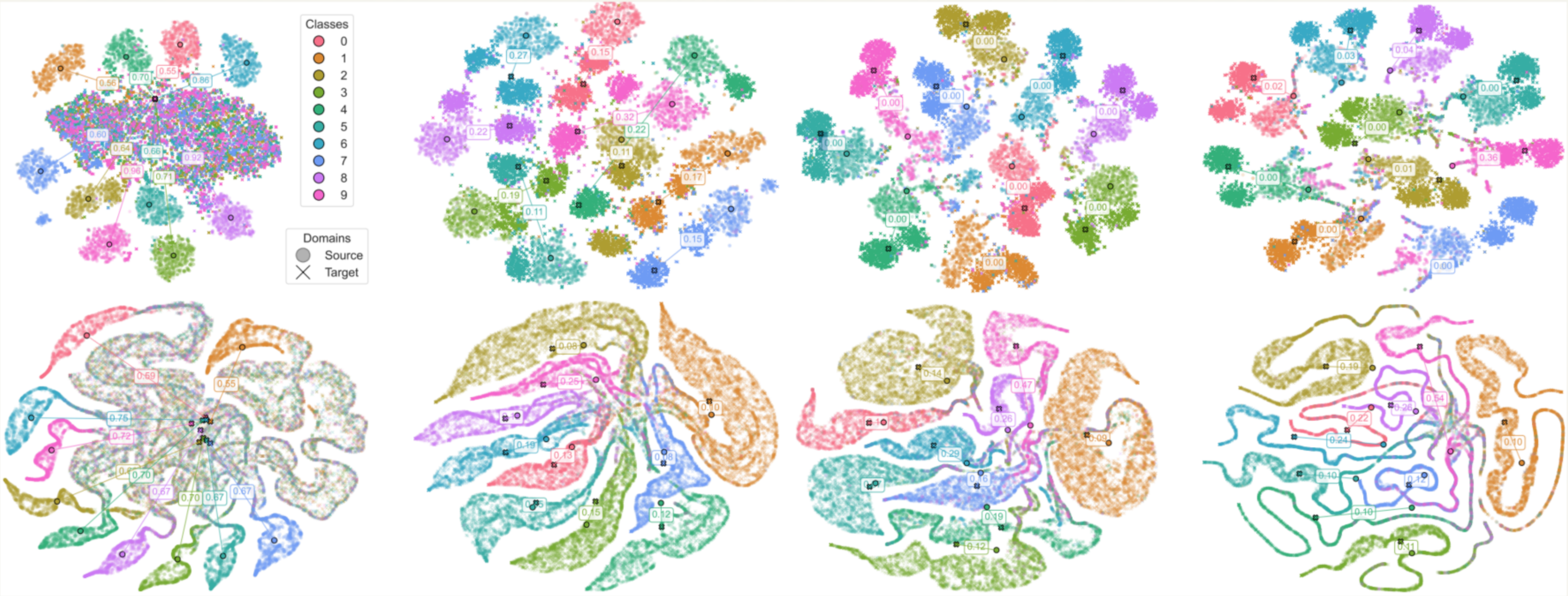
Table 1: Results on Digits. Blue cells indicate improvements over the standard fine-tuning, with darker shades representing larger gains. We evaluate two backbones: a ViT-S-4/9-420 (a small Vision Transformer with 4×4 patch size, 9 layers, and 420-dimensional embeddings) and a lightweight VGG-style CNN (three Conv $\rightarrow$ ReLU $\rightarrow$ Pool blocks). Note that PGFD fine-tuning corresponds to an unsupervised domain adaptation setting.

Model	Source	Target: No fine-tuning				Standard fine-tuning				**FGPD fine-tuning**				**PGFD fine-tuning**				**Our full JFPD fine-tuning**			
		MNIST	SVHN	USPS	Syn	MNIST	SVHN	USPS	Syn	MNIST	SVHN	USPS	Syn	MNIST	SVHN	USPS	Syn	MNIST	SVHN	USPS	Syn
ViT-S	MNIST	99.65	19.88	94.02	18.80	–	94.06	97.46	75.35	–	95.01	98.16	88.50	–	76.64	97.76	60.40	–	95.82	98.06	89.30
	SVHN	70.34	93.79	76.53	39.35	99.49	–	97.01	80.15	99.71	–	97.91	87.95	95.63	–	95.62	52.60	99.63	–	97.86	87.30
	USPS	54.17	16.15	96.66	17.15	99.32	91.58	–	66.15	99.43	92.30	–	77.50	81.70	39.40	–	23.65	99.43	93.38	–	79.90
	Syn	20.33	30.95	62.33	73.30	99.29	91.92	96.66	–	99.47	93.80	97.36	–	69.96	35.60	78.03	–	99.50	93.02	97.56	–
CNN	MNIST	99.61	20.76	85.70	21.40	–	91.57	98.16	74.50	–	92.63	98.46	88.75	–	70.21	97.11	62.55	–	93.89	98.31	93.75
	SVHN	68.65	94.51	75.44	43.35	99.37	–	97.36	81.90	99.50	–	97.91	91.25	93.67	–	90.88	57.00	99.58	–	98.16	94.25
	USPS	94.31	20.62	97.66	19.60	99.46	91.44	–	70.50	99.56	91.78	–	83.65	98.28	47.10	–	39.85	99.60	93.36	–	91.45
	Syn	28.82	38.59	63.08	94.85	99.28	91.13	96.76	–	99.44	91.40	97.76	–	85.99	63.14	84.16	–	99.49	93.67	97.81	–

Table 2: Results on Office-Home with ResNet and ViT backbones. Blue cells indicate improvements over the standard fine-tuning. The darkness of the blue shade represents the degree of improvement.

Backbone	Source	Target: No Fine-tuning				Standard fine-tuning				**FGPD fine-tuning**				**Our full JFPD fine-tuning**			
		Art	Clip	Prod	Real	Art	Clip	Prod	Real	Art	Clip	Prod	Real	Art	Clip	Prod	Real
ResNet-34	Art	66.67	34.27	50.27	58.46	-	51.58	71.23	68.55	-	61.58	79.31	71.58	-	61.95	79.80	71.93
	Clip	33.33	74.80	45.46	45.21	50.52	-	68.33	60.34	56.15	-	78.09	68.63	57.40	-	78.66	69.44
	Prod	34.32	27.10	90.71	56.68	49.32	48.19	-	68.40	57.24	59.94	-	71.44	58.39	60.92	-	73.78
	Real	55.62	36.17	68.85	81.74	61.72	54.41	79.63	-	64.06	62.04	81.36	-	65.21	62.62	81.98	-
ResNet-50	Art	74.56	35.50	56.50	65.59	-	51.30	74.97	73.46	-	60.57	80.79	77.39	-	62.30	81.53	77.71
	Clip	42.21	75.14	51.04	52.45	57.50	-	71.31	63.49	65.16	-	80.48	72.16	67.14	-	81.24	74.44
	Prod	45.96	26.54	93.44	64.59	58.54	47.29	-	73.46	64.32	58.09	-	77.13	67.03	60.89	-	78.78
	Real	64.30	38.86	73.99	82.85	67.14	53.05	80.42	-	69.74	63.19	82.95	-	71.30	64.46	84.93	-
ViT-B/32	Art	70.41	36.17	49.95	64.25	-	58.35	74.57	72.42	-	61.43	76.42	72.16	-	61.00	76.87	72.51
	Clip	43.79	77.60	53.77	55.12	57.45	-	73.81	68.37	57.29	-	76.53	70.08	59.69	-	77.16	71.99
	Prod	42.21	35.05	90.38	63.03	54.64	57.40	-	73.72	56.20	59.25	-	73.37	58.85	61.90	-	75.48
	Real	54.24	40.43	72.35	81.07	64.58	60.77	80.05	-	64.53	63.08	80.56	-	66.30	64.75	81.75	-
ViT-B/16	Art	73.79	37.46	51.92	67.13	-	61.75	76.29	73.47	-	63.84	77.13	73.16	-	64.27	77.85	74.06
	Clip	47.89	78.49	55.61	57.90	60.33	-	74.30	70.08	59.21	-	77.23	71.92	62.82	-	77.91	72.79
	Prod	47.33	38.92	92.14	65.77	58.79	51.67	-	73.72	63.58	60.34	-	74.94	63.01	62.39	-	76.31
	Real	62.77	41.88	74.39	81.62	67.13	56.78	80.88	-	67.56	63.87	81.34	-	69.49	64.72	82.01	-

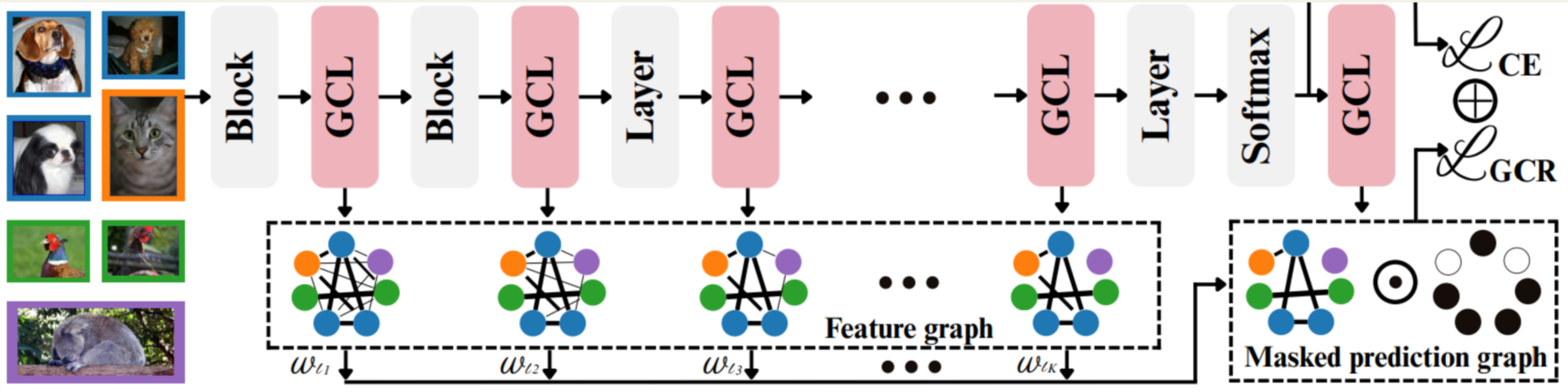
Outcome: The model learns to focus on reliable, trustworthy samples, effectively bridging the domain gap



t-SNE visualizations of MNIST (source, dots) and SVHN (target, crosses). The first row shows learned feature representations; the second row shows softmax predictions. From left to right: (i) pretrained MNIST model (no SVHN fine-tuning), (ii) standard fine-tuning, (iii) our feature-guided prediction divergence, and (iv) our full Joint Feature-Prediction Discrepancy (JFPD) method.

Using Predictions as a Guide for Structure

The pipeline of our Graph Consistency Regularization (GCR) framework



Our parameter-free Graph Consistency Layer (GCL), highlighted in red, can be inserted after any micro-network block (eg., Inception) or specific layer (eg., fully connected). Each GCL constructs a relational graph from batch-level features using a similarity metric (eg., cosine). A reference graph is generated from softmax predictions and masked by intra-class indicators: binary masks identifying semantically consistent pairs. Each GCL enforces alignment between masked prediction graph and the feature-level graphs. The resulting consistency signals are adaptively weighted, forming the Graph Consistency Regularization (GCR) framework, which integrates with the primary loss (eg., cross-entropy), acting as a semantic regularizer to guide learning.

GCR loss

$$\mathbf{F}_{ij}^{(l)} = \mathrm{ReLU}\left(\cos(\mathbf{x}_i^{(l)}, \mathbf{x}_j^{(l)})\right)$$

$$\boldsymbol{S}_{ij} = \mathrm{ReLU}\left(\cos(\mathrm{softmax}(\mathbf{z}_i), \mathrm{softmax}(\mathbf{z}_j))\right)$$

$$\mathbf{P}_{ij} = \mathbf{M}_{ij} \odot \mathbf{S}_{ij}$$

$$\mathcal{L}_{\mathrm{GCR}}^{(l)} = \| \mathrm{triu}(\mathbf{F}^{(l)}) - \mathrm{triu}(\mathbf{P}) \|_F^2$$

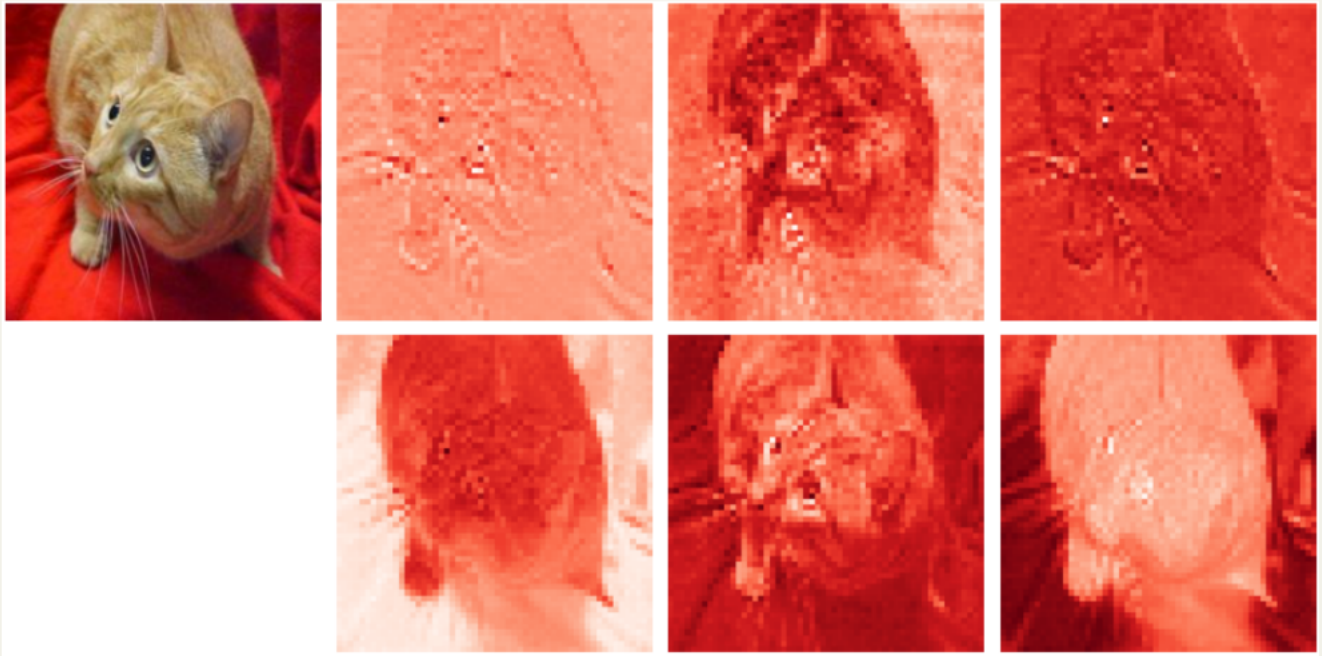
$$\mathcal{L}_{\mathrm{GCR}} = \sum_{l \in \{1,\ldots,K\}} w_l \cdot \| \mathrm{triu}(\mathbf{F}^{(l)}) - \mathrm{triu}(\mathbf{P}) \|_F^2$$

Training objective

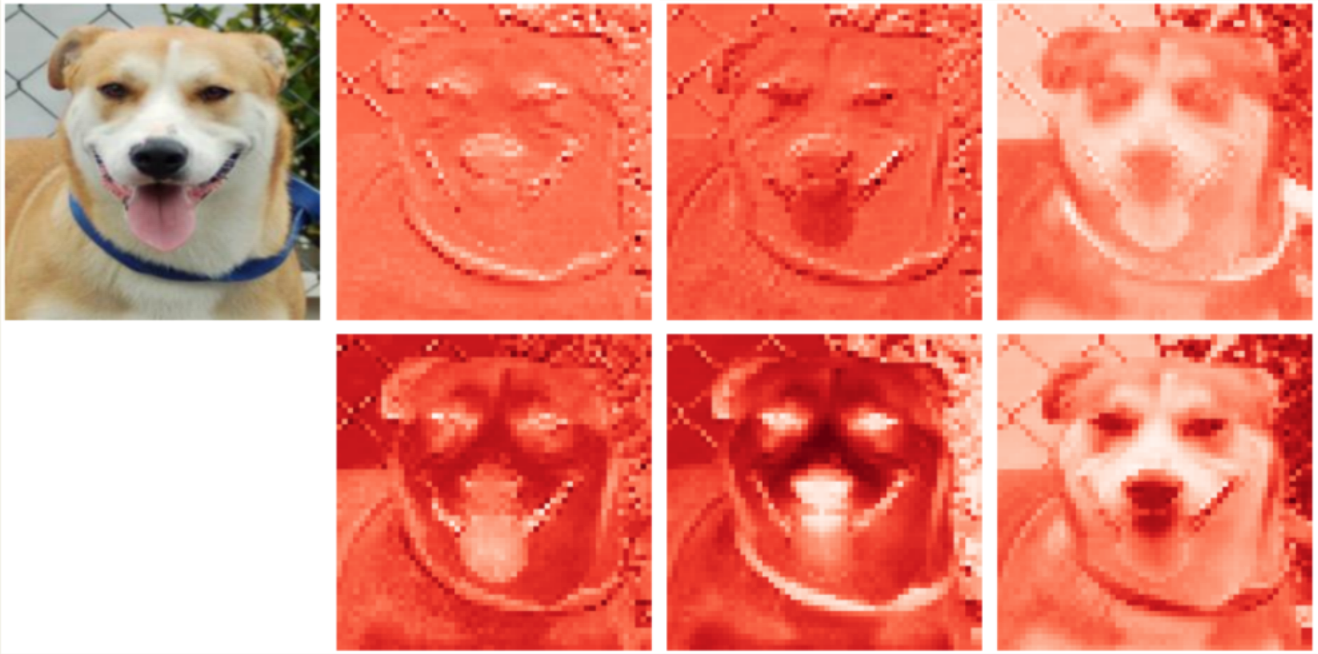
The overall training loss is composed of two components: the standard cross-entropy loss, and the GCR loss. The total loss function is given by:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{CE}} + \lambda \cdot \mathcal{L}_{\mathrm{GCR}}$$

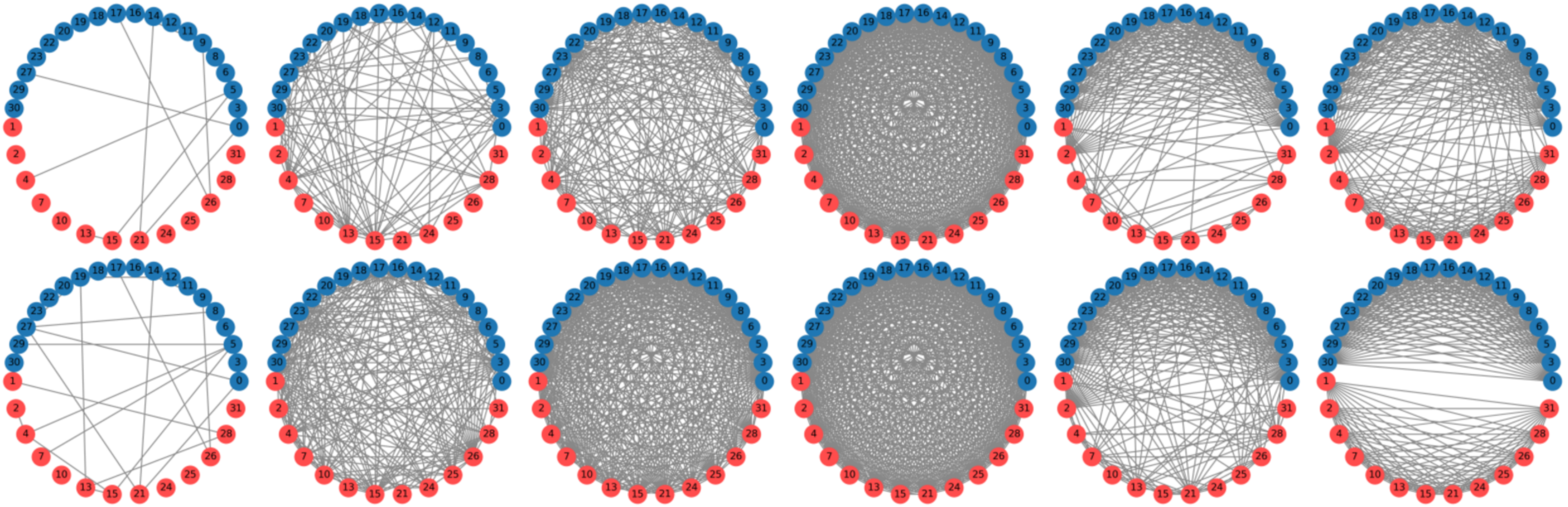
A key advantage of GCR is that it introduces no additional parameters, and its graph alignment loss relies on matrix operations well-suited to modern hardware.



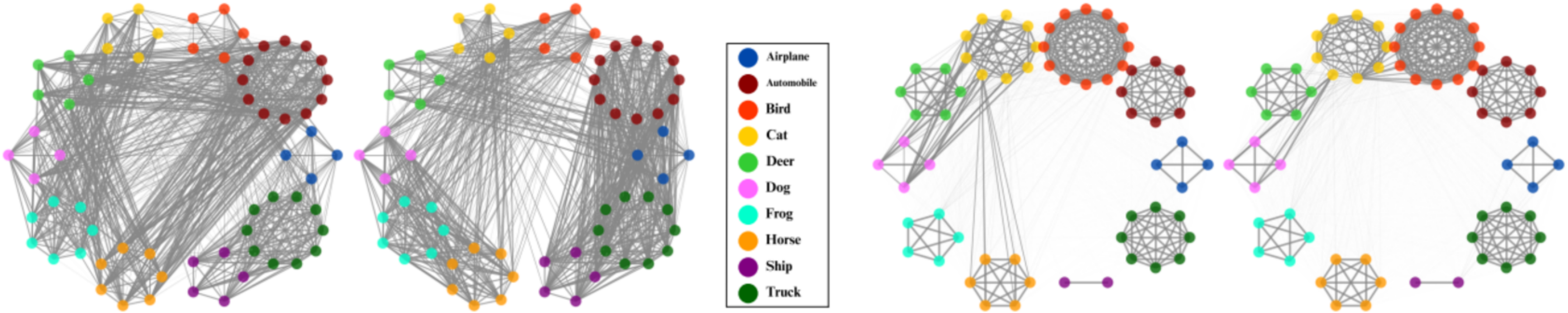
Feature map visualizations from models trained on identical data batches: (the first and the third row) baseline and (the second and the fourth row) our GCL-augmented model.



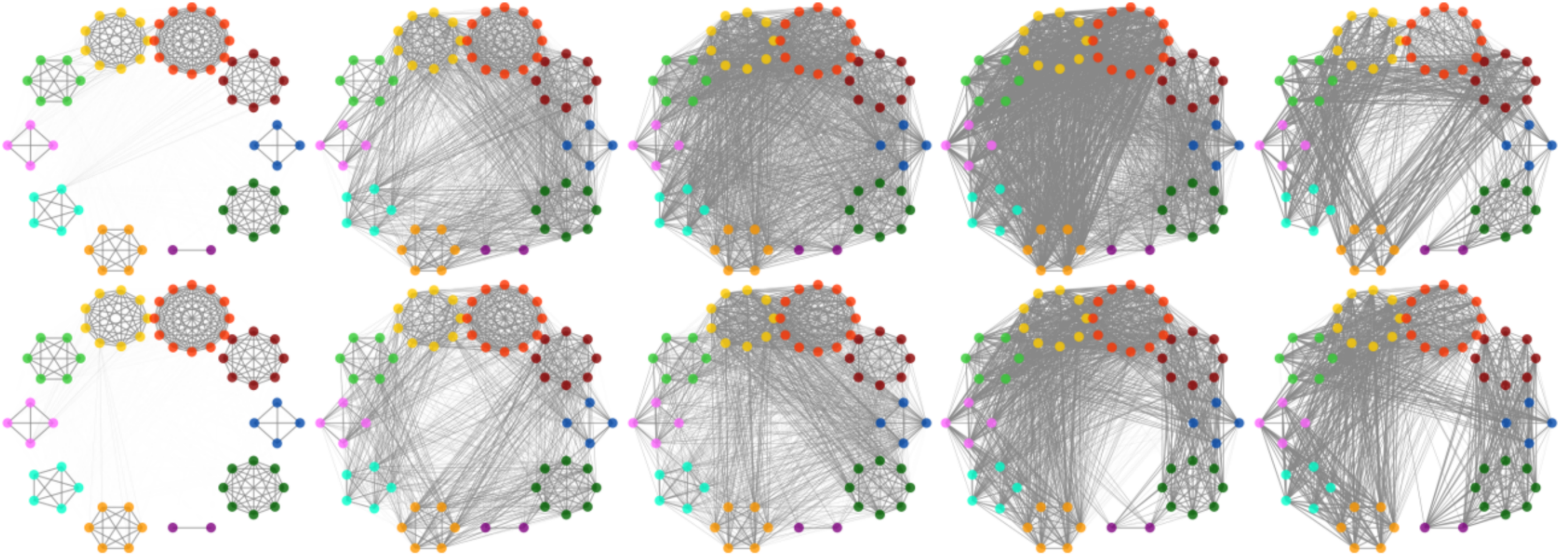
Brighter red regions indicate stronger feature activations. Compared to the baseline, GCL-enhanced maps more **clearly emphasize class-discriminative cues**, eg., cat faces, ears, and eyes, and for dogs, tongues, noses, and facial contours, reflecting **improved focus and interpretability**. GCL also yields higher classification accuracy (from 98.1% to 99.8%)



Relational graph visualization on Kaggle cats vs dogs. We compare the best baseline model and our GCL-augmented model using the same batch of 32 samples (red = cat, blue = dog). The baseline consists of four convolutional blocks and two fully connected layers; our method inserts a Graph Consistency Layer (GCL) after each, totaling six GCLs. The top row shows the baseline (without GCLs); the bottom row shows our GCL-enhanced model. Each column visualizes the relational graph at a specific layer, from early features (left) to final predictions (right). Early layers exhibit weak connectivity, as low-level features poorly capture class semantics. For clarity, edges with similarity < 0.4 are omitted.



(a) DenseNet-121 (b) With our GCLs (c) MobileNet (d) With our GCLs



(a) ShuffleNet (b) GoogLeNet (c) ResNet-50 (d) ResNet-101 (e) DenseNet-121

Experiment results

Table 1: Accuracy (%) on CIFAR-10 across models. Results are shown for MobileNet (MNet), ShuffleNet (SN), SqueezeNet (SQNet), GoogLeNet (GLNet), ResNeXt-50/101 (Rx-50/101), ResNet-34/50/101 (R34/R50/R101), DenseNet-121 (D121), and MAE under various GCL configurations (Early, Mid, Late, combinations, Full). Bold indicates the best improvements over baselines; underlines mark the second-best. The final column shows the average accuracy for each configuration.

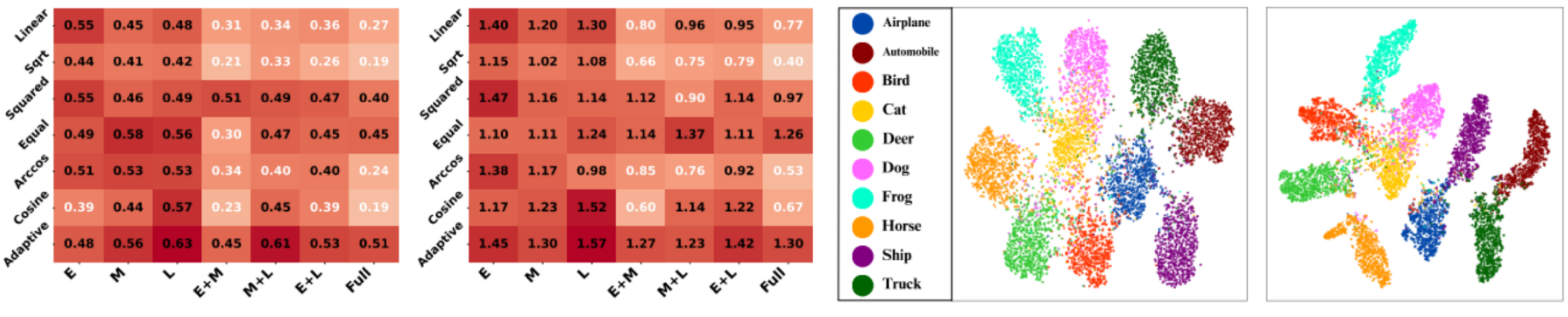
	MAE	MNet	SN	SQNet	GLNet	Rx-50	Rx-101	R34	R50	R101	D121	Mean
Baseline	89.01	90.17	91.29	92.25	94.13	94.68	95.11	94.75	95.07	95.15	95.07	93.3 $\pm$ 2.2
Early GCL	89.38	91.18	92.36	92.57	**94.84**	**95.56**	95.54	**95.63**	95.50	95.45	**95.77**	94.0 $\pm$ 2.1
Mid GCL	**89.76**	91.15	92.55	92.46	94.74	95.39	95.46	95.61	95.55	**95.69**	95.58	94.0 $\pm$ 2.0
Late GCL	89.74	**91.38**	92.38	**92.82**	94.81	95.42	**95.64**	**95.63**	**95.58**	95.58	95.64	**94.1** $\pm$ 2.0
Early+Mid	89.52	90.83	92.53	92.30	94.72	95.41	95.47	95.53	95.45	95.43	95.57	93.9 $\pm$ 2.1
Mid+Late	89.51	91.17	**92.75**	**92.82**	94.68	95.43	95.45	95.38	95.40	95.45	95.64	94.0 $\pm$ 2.0
Early+Late	89.65	91.05	92.34	92.66	94.76	95.33	95.42	95.57	95.39	95.41	95.59	93.9 $\pm$ 2.0
Full GCL	89.58	90.98	92.47	92.62	94.63	95.43	95.41	95.40	95.55	95.44	95.43	93.9 $\pm$ 2.0

Table 2: Accuracy (%) on CIFAR-100 across models.

	MAE	MNet	SN	SQNet	Rx-50	Rx-101	R34	R50	D121	Mean
Baseline	64.25	65.98	70.06	69.41	77.77	77.78	76.76	77.39	77.01	72.93 $\pm$ 5.19
Early GCL	64.99	67.47	**71.94**	70.87	79.24	79.61	77.97	79.31	79.46	74.54 $\pm$ 5.46
Mid GCL	65.02	67.84	71.83	70.29	78.99	79.34	77.77	78.86	79.33	74.36 $\pm$ 5.33
Late GCL	**65.52**	**68.34**	71.39	70.58	**79.48**	**79.77**	**78.16**	**79.34**	**79.77**	**74.71** $\pm$ 5.38
Early+Mid	65.25	67.55	71.56	70.44	78.96	79.18	77.34	78.64	79.19	74.23 $\pm$ 5.24
Mid+Late	65.24	68.32	71.60	70.33	78.97	79.49	77.36	78.79	79.49	74.40 $\pm$ 5.23
Early+Late	65.20	67.28	71.61	**71.00**	78.98	79.49	78.11	78.64	79.39	74.41 $\pm$ 5.36
Full GCL	65.36	68.19	71.28	70.80	79.08	79.23	77.72	78.77	79.20	74.40 $\pm$ 5.18

Table 3: Accuracy (%) on Tiny ImageNet across models. All results are obtained by training models from scratch. We also evaluate Stochastic ResNet-18 (R18SD) and SE-ResNet-18 (SER18).

	ViT$_{/32}$	**ViT$_{/16}$**	**CeiT**	**MViT$_{XXS}$**	**MViT$_{XS}$**	**MViT**	**Swin**	**MNet**	**R18SD**	**SER18**	**R34**	**Mean**
Baseline	37.86	40.01	49.89	49.35	51.54	52.65	54.20	57.84	63.42	65.71	67.45	53.6±9.2
Early GCL	39.00	41.00	**51.18**	50.19	51.35	53.86	54.93	**57.92**	63.88	**66.46**	**67.85**	**54.3**±9.0
Mid GCL	38.56	40.98	50.36	49.87	51.42	53.94	55.16	57.63	63.96	65.72	67.57	54.1±8.9
Late GCL	38.05	40.29	50.77	49.81	**51.95**	**54.08**	55.53	57.88	63.85	65.78	67.68	54.2±9.1
Early+Mid	**39.02**	**41.25**	50.27	49.73	51.63	53.95	54.94	57.53	64.11	65.92	67.68	54.2±8.9
Mid+Late	38.38	40.50	50.08	**50.56**	51.53	53.87	**55.57**	57.62	**64.23**	65.89	67.64	54.2±9.1
Early+Late	38.33	40.65	50.67	50.22	51.39	53.63	54.95	57.87	63.95	65.77	67.68	54.1±9.0
Full GCL	38.38	40.84	49.91	50.16	51.85	54.04	54.94	57.70	64.04	65.94	67.72	54.1±9.0



(a) Δgain on CIFAR-10 (b) Δgain on CIFAR-100 (c) ShuffleNet (d) With GCLs

Synthesis & The Big Picture

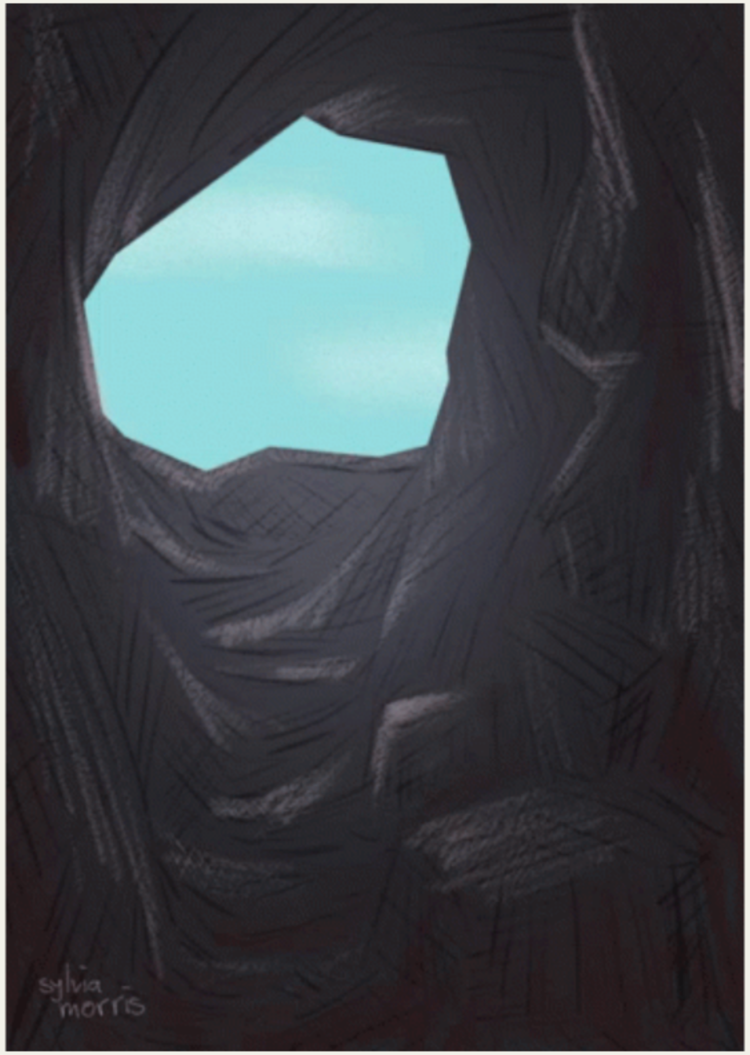
Both JFPD and GCR are about creating harmony between features and predictions

From Trust to Structure

JFPD uses trust to selectively align features and predictions in new domains.

GCR uses the structure of predictions to organize features everywhere.

The Echo Effects

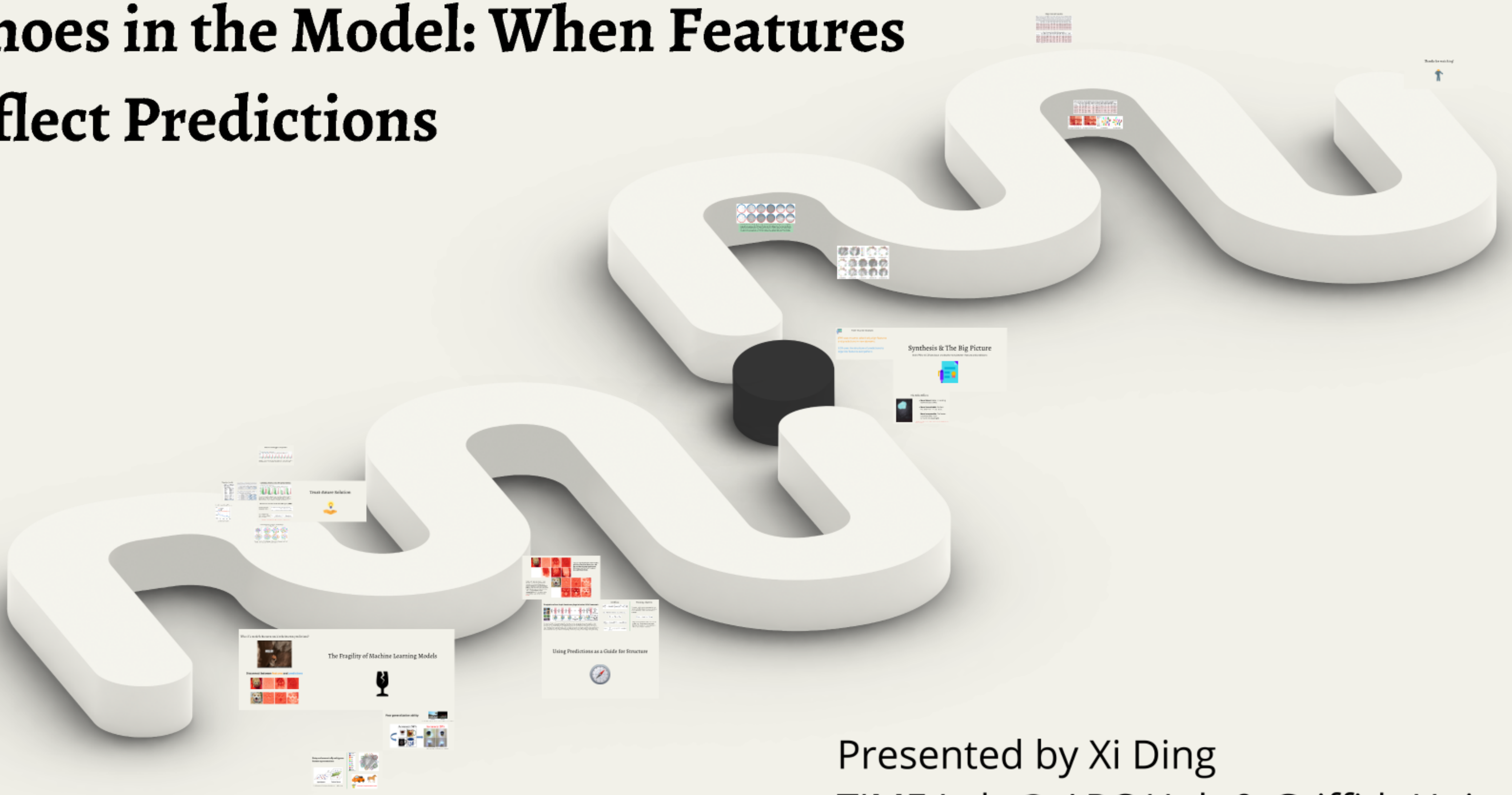


- **More Robust**: Better at handling real-world data shifts.
- **More Generalizable**: Perform well even with limited labels.
- **More Interpretable**: The feature space becomes more semantically meaningful.

Aligning what a model sees with what it believes is a fundamental step toward more human-like learning

Thanks for watching!

Echoes in the Model: When Features Reflect Predictions



Presented by Xi Ding
TIME Lab @ ARC Hub & Griffith University