Wasserstein Distance Rivals Kullback-Leibler Divergence for Knowledge Distillation

Jiaming Lv*, Haoyuan Yang*, Peihua Li †

Dalian University of Technology

Jiaming Lv

Haoyuan Yang

Peihua Li

*Equal contribution.

†Corresponding author.

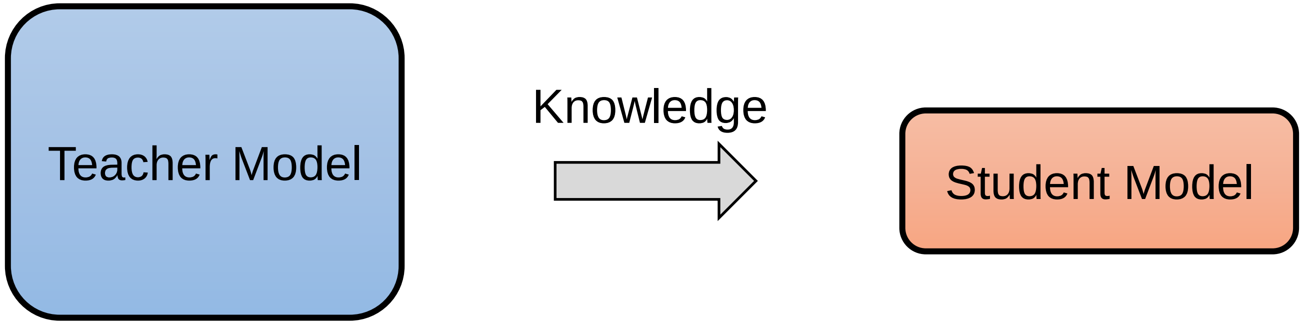
Content

- Introduction
  - What is knowledge distillation?
  - Two downsides of KL-Div for knowledge distillation
- WD for Knowledge Transfer
  - Discrete WD for Logit Distillation
  - Continuous WD for Feature Distillation
- Experiments
  - Ablation
  - Comparison with SOTAs
- Conclusion

Introduction

What is knowledge distillation?

Transfer knowledge from a high-performance teacher model with large capacity to a lightweight student model.



Teacher Model:
- High-performance
- Large capacity

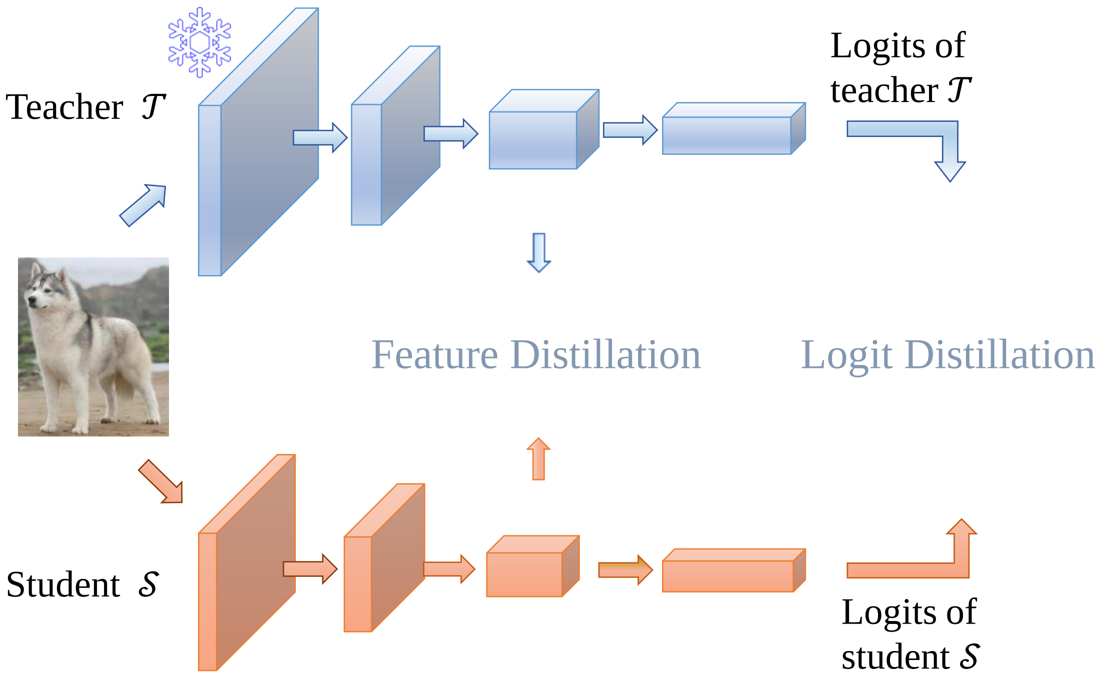
Student Model:
- Lightweight

- ➢ **Align** the student with the teacher to reduce deployment costs.
- ➢ **Improve** student model performance via knowledge transfer.

Introduction

Two strategies in knowledge distillation:

- Logit Distillation
- Feature Distillation



Strategy	Knowledge
Logit	The prediction of category probabilities
Feature	Intermediate layer features

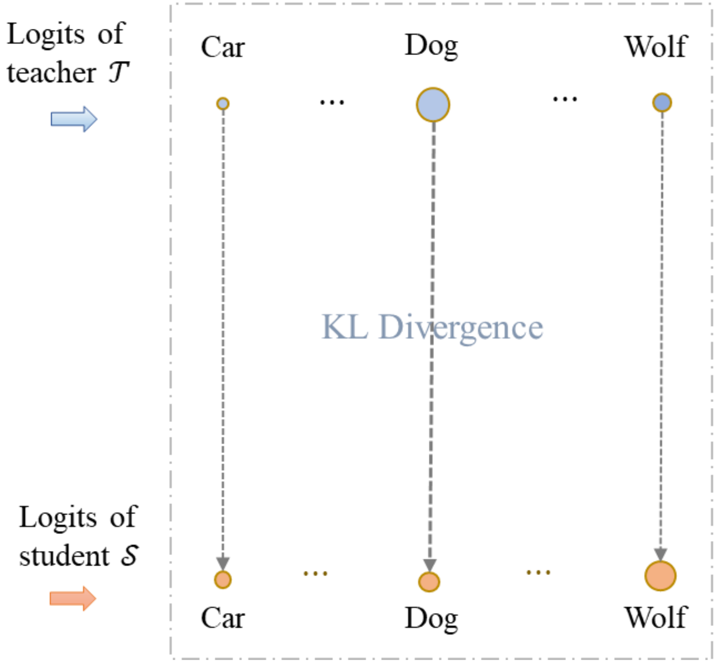
Introduction

KL Divergence (KL-Div) for knowledge distillation

- KL-Div has been predominant in logit distillation.

Method	Dis-similarity
KD [1]	KL-Div
DKD [2]	
NKD [3]	
WTTM [4]	

- KL-Div is complementary to many methods that transfer knowledge from intermediate layers. [5, 6, 7]



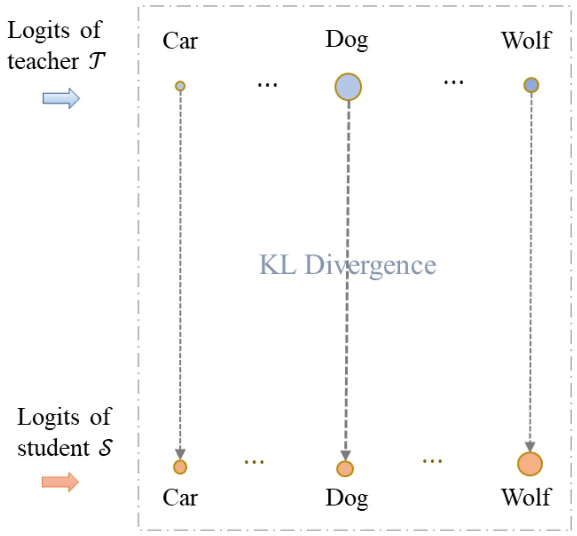
[1] G. Hinton, O. Vinyals, J. Dean. Distilling the knowledge in a neural network. arXiv, 2015.
[2] B. Zhao, Q. Cui, R. Song, Y. Qiu, J. Liang. Decoupled knowledge distillation. In CVPR, 2022.
[3] Z. Yang, A. Zeng, C. Yuan, Y. Li. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In ICCV, 2023.
[4] K. Zheng, E.-H. Yang. Knowledge distillation based on transformed teacher matching. In ICLR, 2024.
[5] L. Liu, Q. Huang, S. Lin, H. Xie, B. Wang, X. Chang, X. Liang. Exploring inter-channel correlation for diversity-preserved knowledge distillation. In ICCV, 2021.
[6] M. Zong, Z. Qiu, X. Ma, K. Yang, C. Liu, J. Hou, S. Yi, W. Ouyang. Better teacher better student: Dynamic prior knowledge for knowledge distillation. In ICLR, 2023.
[7] D. Liu, M. Kan, S. Shan, X. Chen. Function-consistent feature distillation. In ICLR, 2023.

Introduction

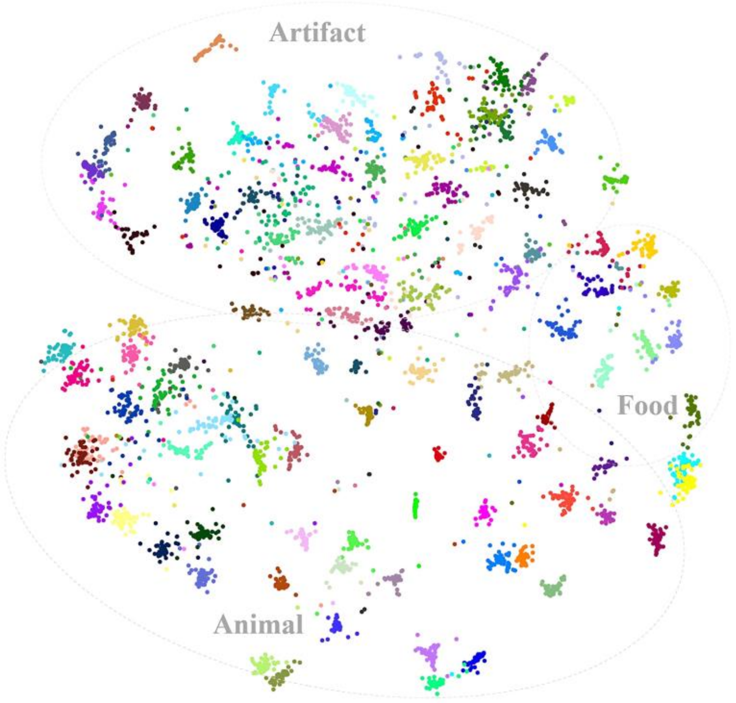
Two downsides of KL-Div for knowledge distillation

- KL-Div lacks a mechanism to perform **cross-category** comparison.
- KL-Div is problematic for distilling knowledge from intermediate layers.

KL-Div is a **category-to-category** measure.



Real-world categories exhibit rich **interrelations (IRs)** in feature space.



It is essential to **explicitly** leverage the relationships among categories.

Introduction

Two downsides of KL-Div for knowledge distillation

- KL-Div lacks a mechanism to perform cross-category comparison.
- KL-Div is problematic for distilling knowledge from intermediate layers.

- **KL-Div struggles with high-dimensional features**: deep features are sparsely in feature space [8]
  - Make **non-parametric density estimation (e.g., histogram)** that KL-Div requires infeasible due to the curse of dimensionality.
  - Lead to **non-overlapping discrete distributions** that KL-Div fails to deal with. [9]

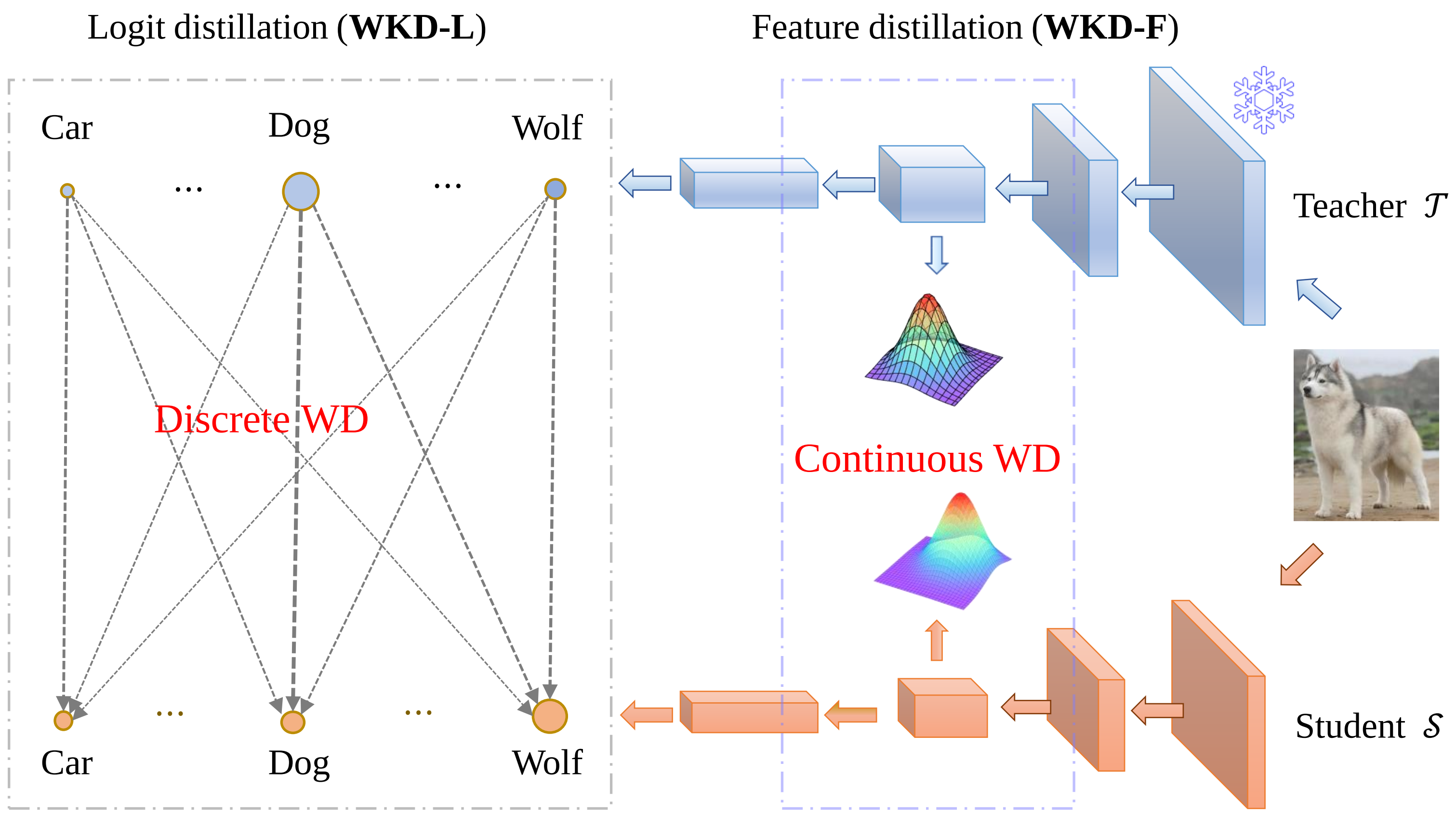
- **KL-Div has limited ability for continuous distributions**: KL-Div is not a metric [10] and is unaware of **geometric structure** of the underlying manifold [11].

[8] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning, 2009.
[9] M. Arjovsky, S. Chintala, L. Bottou. Wasserstein generative adversarial networks. In ICML, 2017.
[10] K. T. Abou-Moustafa, F. P. Ferrie. A note on metric properties for some divergence measures: The Gaussian case. In ACML, 2012.
[11] S. Ozair, C. Lynch, Y. Bengio, A. van den Oord, S. Levine, P. Sermanet. Wasserstein dependency measure for representation learning. In NeurIPS, 2019.

WD for Knowledge Transfer

WD (Wasserstein Distance) for Knowledge Transfer

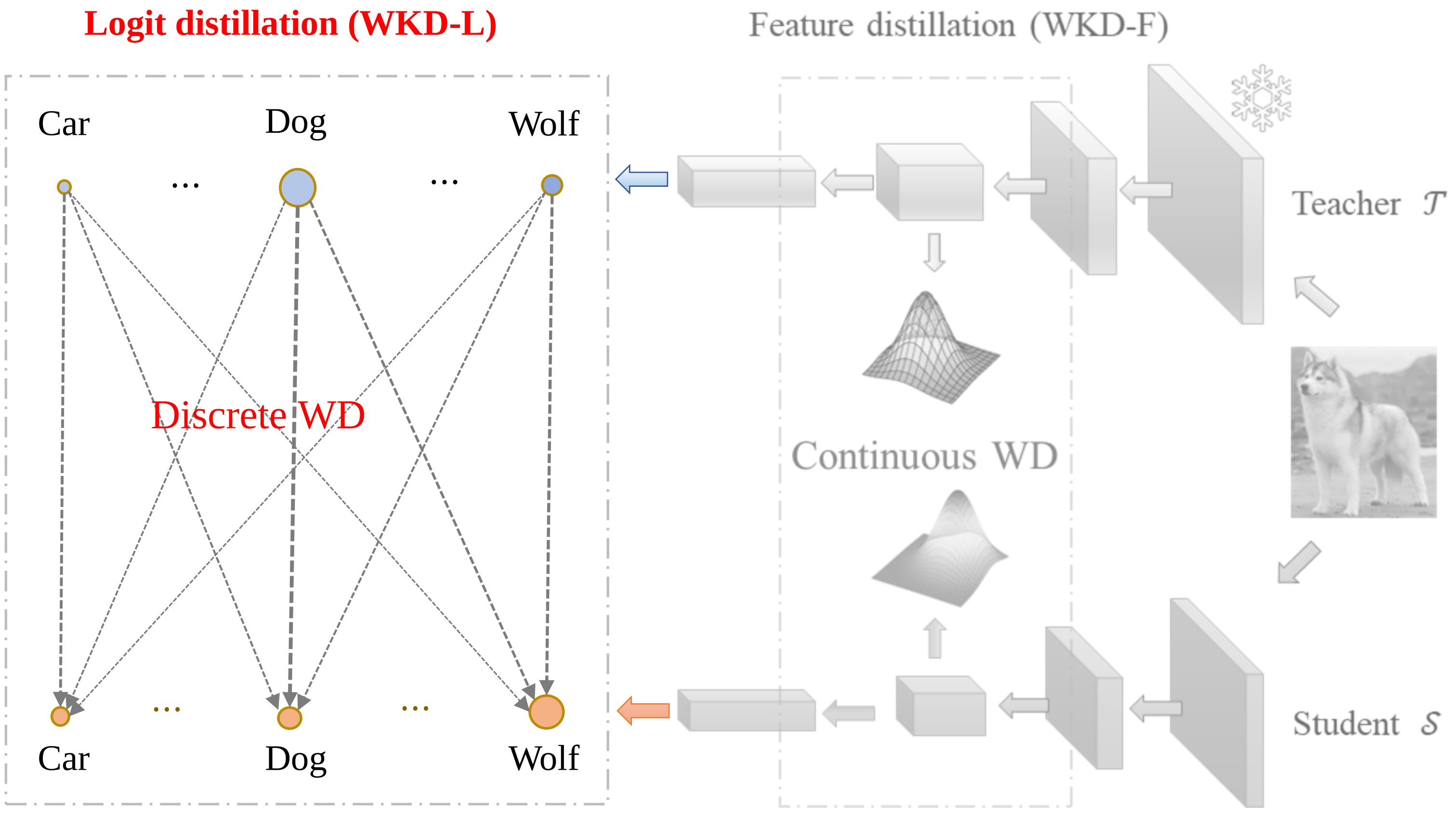
- Discrete WD for Logit distillation: WKD-L
  - Explicitly leverage interrelations among categories via cross-category comparison.
- Continuous WD for Feature distillation: WKD-F
  - Leverages geometric structure of the Riemannian space of Gaussians



WD for Knowledge Transfer

WD (Wasserstein Distance) for Knowledge Transfer

- Discrete WD for Logit distillation: WKD-L
  - Explicitly leverage interrelations among categories via cross-category comparison.
- Continuous WD for Feature distillation: WKD-F
  - Leverages geometric structure of the Riemannian space of Gaussians



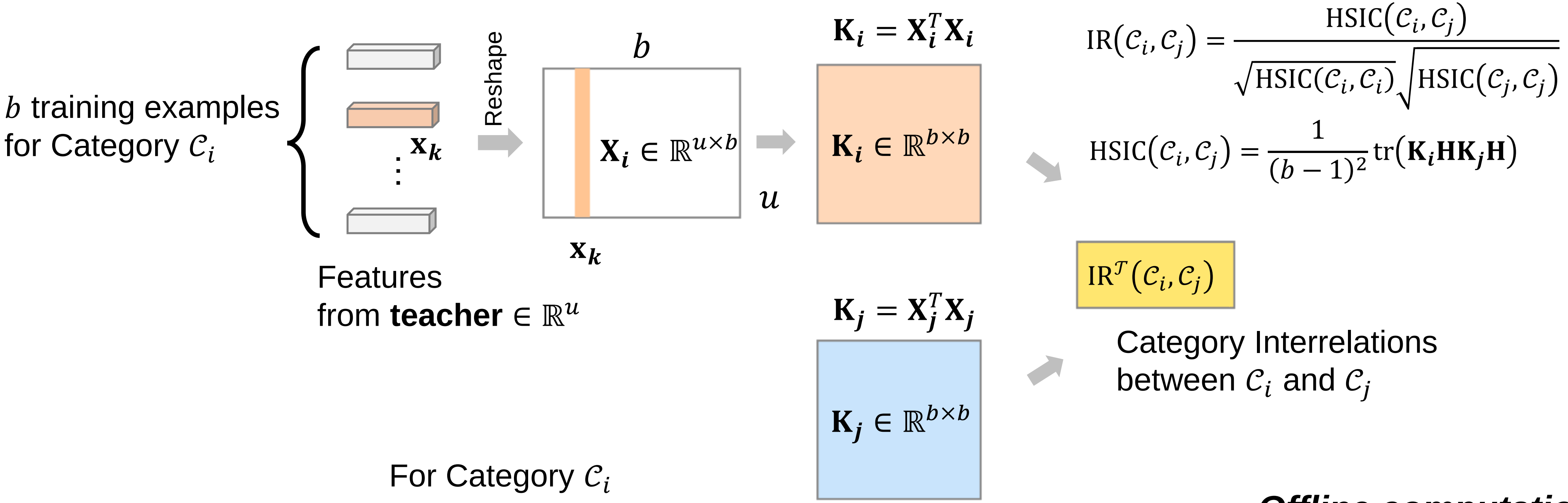
- Quantify category Interrelations (IRs)
- Discrete WD Loss

WD for Knowledge Transfer – Discrete WD

Discrete WD for Logit Distillation

- ➢ Interrelations (IRs) among categories
- ➢ Discrete WD Loss

Quantify category Interrelations (IRs) explicitly based on CKA [12]



Offline computation

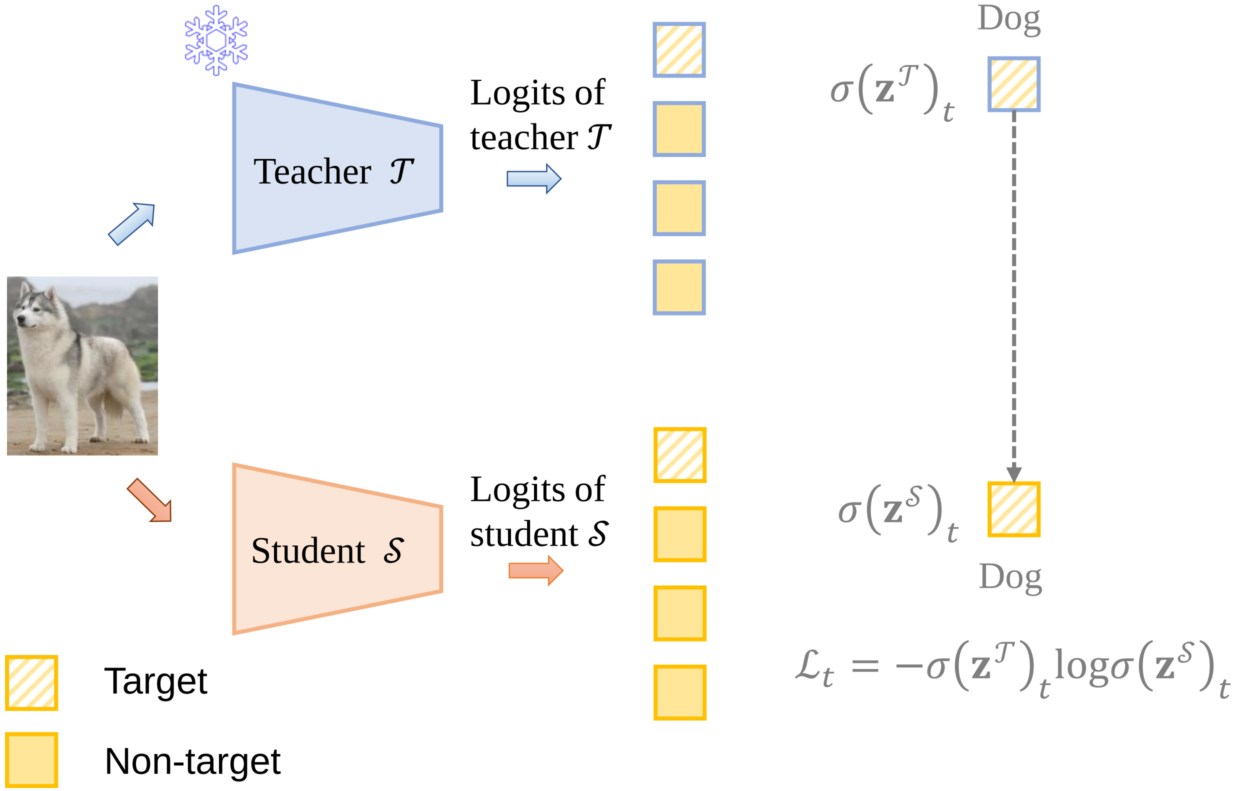
[12] C. Cortes, M. Mohri, A. Rostamizadeh. Algorithms for learning kernels based on centered alignment. JMLR, 2012.

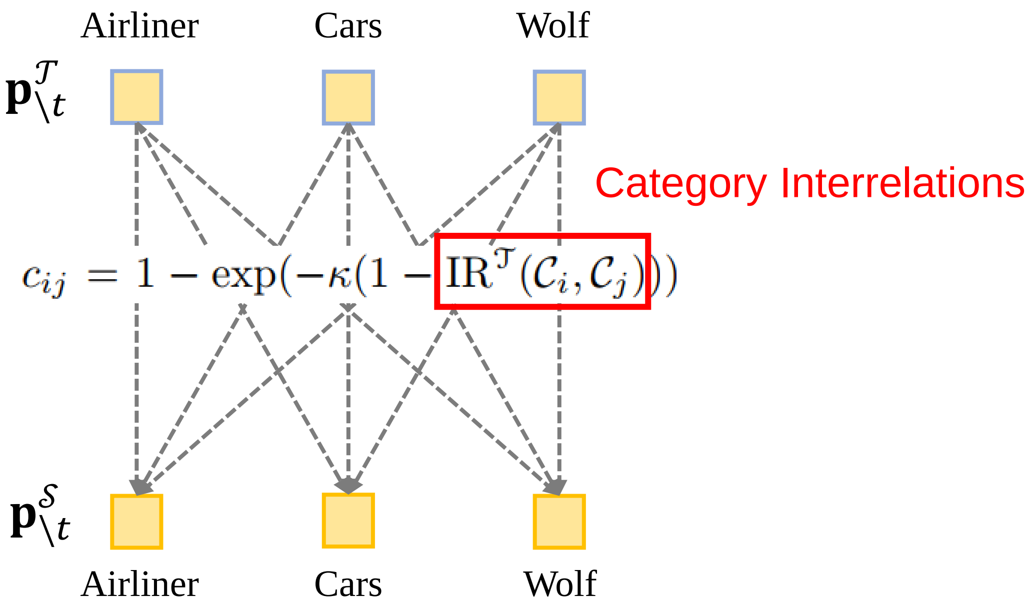
WD for Knowledge Transfer – Discrete WD

Discrete WD for Logit Distillation

- Interrelations (IRs) among categories
- Discrete WD Loss

Loss: $\mathcal{L}_{\text{WKD-L}} = \lambda \mathrm{D}_{\text{WD}}(\mathbf{p}^{\mathcal{T}}_{\backslash t}, \mathbf{p}^{\mathcal{S}}_{\backslash t}) + \mathcal{L}_{\text{t}}$





Entropy regularized linear programming:

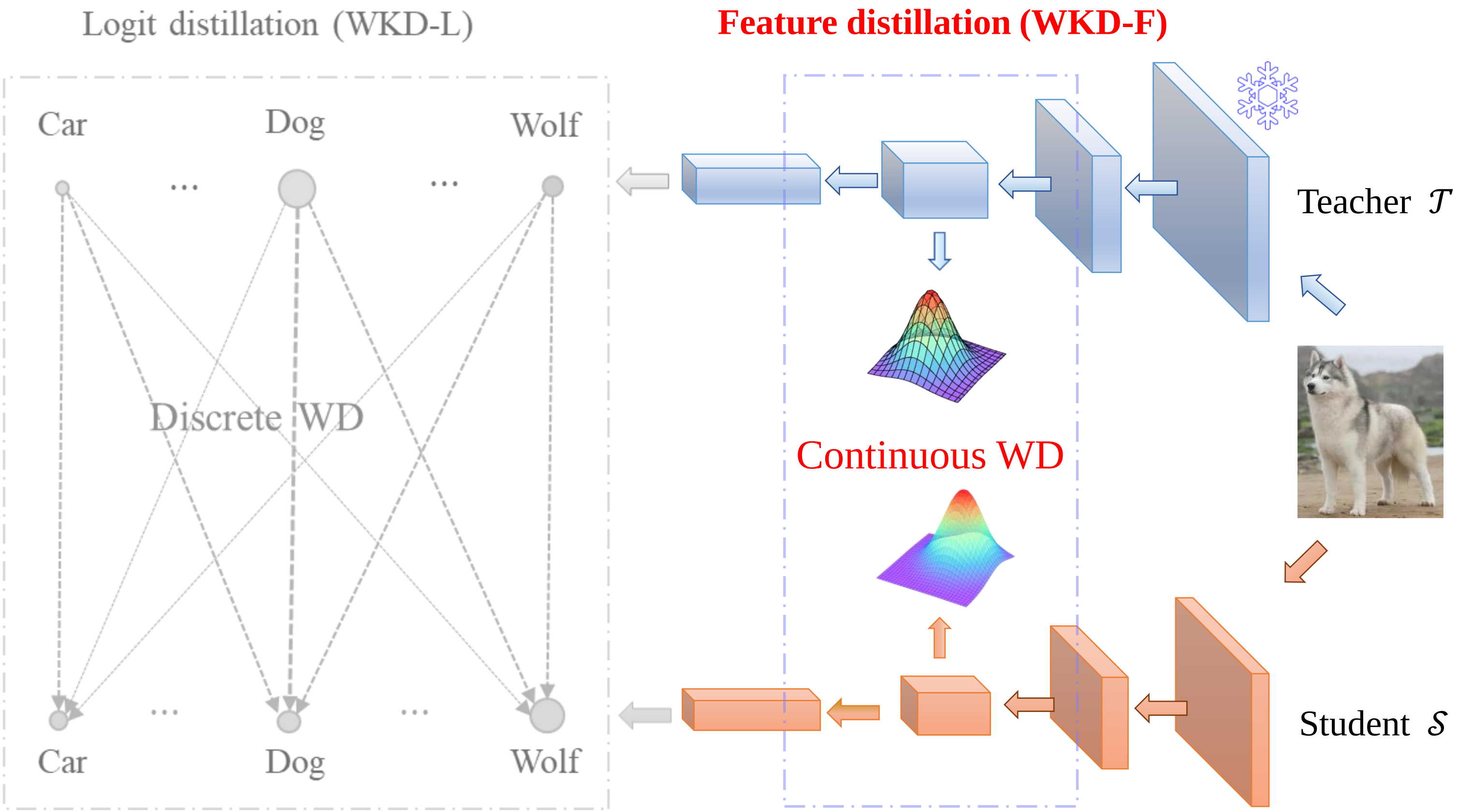
$$\mathrm{D}_{\text{WD}}(\mathbf{p}^{\mathcal{T}}_{\backslash t}, \mathbf{p}^{\mathcal{S}}_{\backslash t}) = \min_{q_{i,j}} \sum_{i,j} c_{ij} q_{ij} + \eta q_{ij} \log q_{ij}$$

$$\text{s.t. } q_{ij} \geq 0, \sum_j q_{ij} = p_i^{\mathcal{T}}, \sum_i q_{ij} = p_j^{\mathcal{S}}, i, j \in S_n$$

WD for Knowledge Transfer

WD (Wasserstein Distance) for Knowledge Transfer

- ➢ Discrete WD for Logit distillation: WKD-L
  - Explicitly leverage interrelations among categories via cross-category comparison.
- ➢ Continuous WD for Feature distillation: WKD-F
  - Leverages geometric structure of the Riemannian space of Gaussians

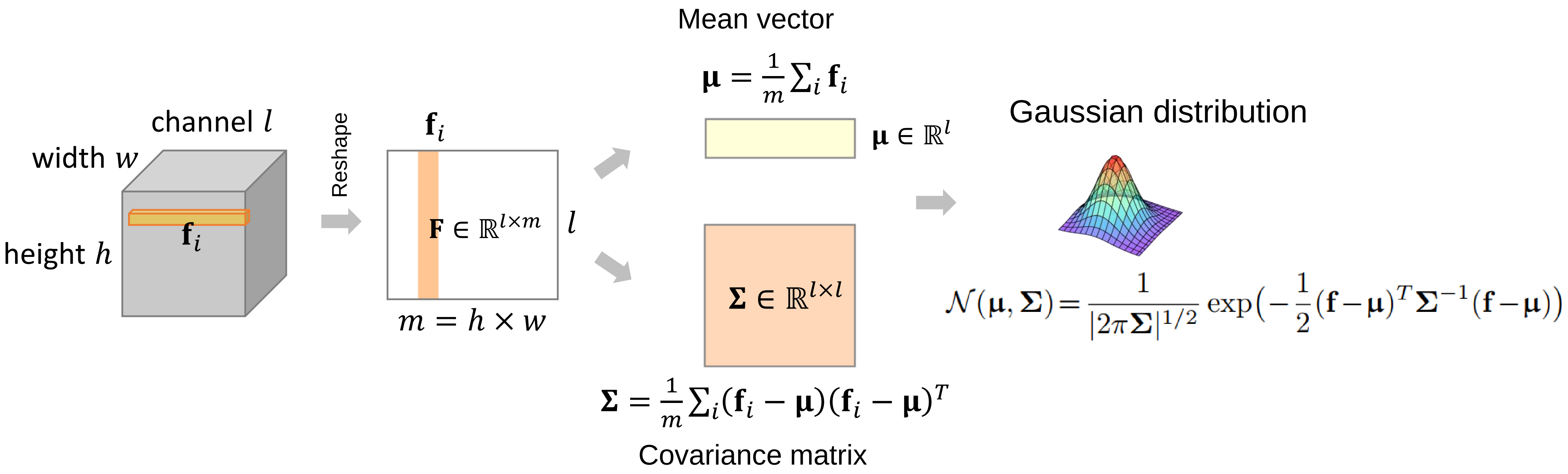


- ➢ Feature distribution modeling
- ➢ Continuous WD Loss

WD for Knowledge Transfer – Continuous WD

Continuous WD for Feature Distillation

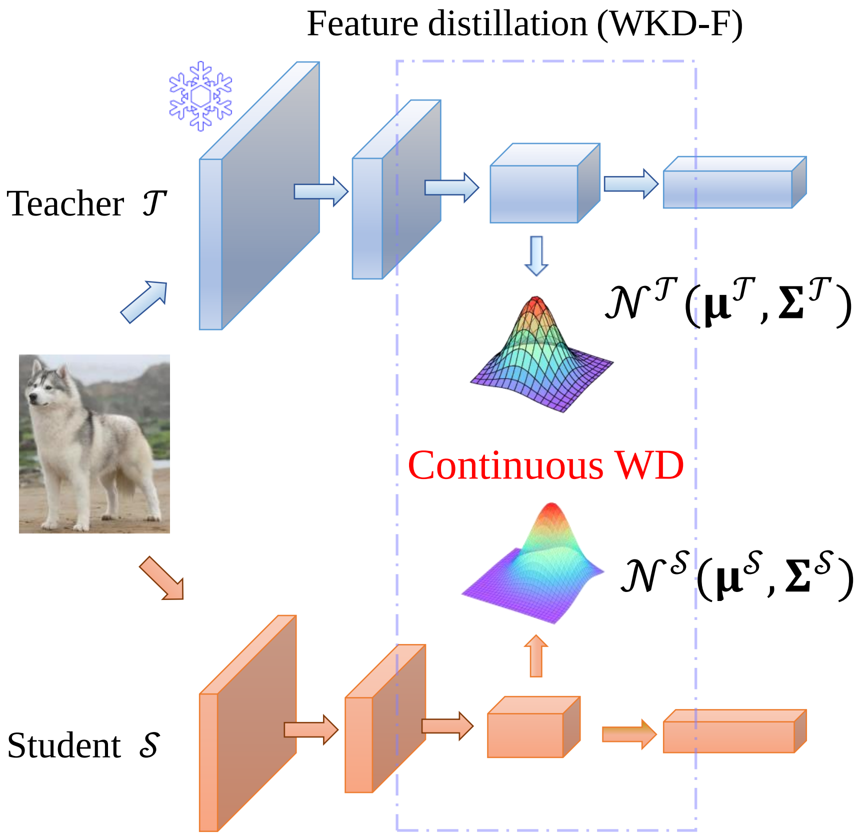
- Feature distribution modeling
- Continuous WD Loss



WD for Knowledge Transfer – Continuous WD

Continuous WD for Feature Distillation

- ➢ Feature distribution modeling
- ➢ Continuous WD Loss



Loss: The continuous WD between the two Gaussians

$$\mathrm{D_{WD}}(\mathcal{N}^{\mathcal{T}}, \mathcal{N}^{\mathcal{S}}) = \inf_q \int_{\mathbb{R}^l}\int_{\mathbb{R}^l} \|\mathbf{f}^{\mathcal{T}} - \mathbf{f}^{\mathcal{S}}\|^2 q(\mathbf{f}^{\mathcal{T}}, \mathbf{f}^{\mathcal{S}}) d\mathbf{f}^{\mathcal{T}} d\mathbf{f}^{\mathcal{S}},$$

Closed-form distance

$$\mathrm{D_{WD}}(\mathcal{N}^{\mathcal{T}}, \mathcal{N}^{\mathcal{S}}) = \mathrm{D_{mean}}(\boldsymbol{\mu}^{\mathcal{T}}, \boldsymbol{\mu}^{\mathcal{S}}) + \mathrm{D_{cov}}(\boldsymbol{\Sigma}^{\mathcal{T}}, \boldsymbol{\Sigma}^{\mathcal{S}})$$

The diagonals of Σ

$$\mathrm{D_{mean}}(\boldsymbol{\mu}^{\mathcal{T}}, \boldsymbol{\mu}^{\mathcal{S}}) = \left\|\boldsymbol{\mu}^{\mathcal{T}} - \boldsymbol{\mu}^{\mathcal{S}}\right\|^2 \quad \mathrm{D_{cov}}(\boldsymbol{\Sigma}^{\mathcal{T}}, \boldsymbol{\Sigma}^{\mathcal{S}}) = \left\|\boldsymbol{\delta}^{\mathcal{T}} - \boldsymbol{\delta}^{\mathcal{S}}\right\|^2$$

Experiments — Comparison with SOTA

➢ Image Classification

Image classification on ImageNet

- *Teacher and the student are CNNs.*

Setting		$\mathcal{T}$	$\mathcal{S}$	Logit						Feature					Logit + Feature				
				KD [2]	DKD [3]	NKD [4]	CTKD [54]	WTTM [5]	WKD-L (ours)	FitNet [24]	CRD [25]	Review-KD [29]	CAT [55]	WKD-F (ours)	CRD+KD [25]	DPK [7]	FCFD [8]	KD-Zero [56]	WKD-L+WKD-F (ours)
RN34 →RN18	Top-1	73.31	69.75	71.03	71.70	71.96	71.51	72.19	**72.49**	70.53	71.17	71.61	71.26	**72.50**	71.38	72.51	72.25	72.17	**72.76**
	Top-5	91.42	89.07	90.05	90.41	–	90.47	–	**90.75**	89.87	90.13	90.51	90.45	**91.00**	90.49	90.77	90.71	90.46	**91.08**
RN50 →MNV1	Top-1	76.16	68.87	70.50	72.05	72.58	–	73.09	**73.17**	70.26	71.37	72.56	72.24	**73.12**	–	73.26	73.26	73.02	**73.69**
	Top-5	92.86	88.76	89.80	91.05	–	–	–	**91.32**	90.14	90.41	91.00	91.13	**91.39**	–	91.17	91.24	91.05	**91.63**

Image classification on CIFAR-100

- *Teacher is a CNN and the student is a Transformer or vice versa.*

Teacher (Acc)	Student (Acc)	Logit					Feature				
		KD [2]	DKD [3]	DIST [63]	OFA [46]	WKD-L (ours)	FitNet [24]	CC [64]	RKD [65]	CRD [25]	WKD-F (ours)
Transformer→CNN											
Swin-T (89.26)	RN18 (74.01)	78.74	80.26	77.75	80.54	**81.42**±0.22	78.87	74.19	74.11	77.63	**81.57**±0.12
ViT-S (92.04)	RN18 (74.01)	77.26	78.10	76.49	80.15	**80.81**±0.21	77.71	74.26	73.72	76.60	**81.12**±0.24
ViT-S (92.04)	MNV2 (73.68)	72.77	69.80	72.54	78.45	**79.04**±0.05	73.54	70.67	68.46	78.14	**79.11**±0.07
CNN→Transformer											
ConvNeXt-T (88.41)	DeiT-T (68.00)	72.99	74.60	73.55	75.76	**76.11**±0.18	60.78	68.01	69.79	65.94	**73.27**±0.22
ConvNeXt-T (88.41)	Swin-P (72.63)	76.44	76.80	76.41	78.32	**78.94**±0.17	24.06	72.63	71.73	67.09	**74.80**±0.13

RN: ResNet
MN: MobileNet

Experiments – Comparison with SOTA

➢ Self-Knowledge Distillation

- Implement WKD in the framework of **Born-Again Network** (BAN [13]). Teacher and Student are *same networks*.
- Conduct experiments with ResNet18 on ImageNet.

Method	Self-KD	Dis-similarity	Top-1
Standard train	×	NA	69.75
Tf-KD [67]	✓	KL-Div	70.14
FRSKD [68]	✓	KL-Div	70.17
Zipf's KD [69]	✓	KL-Div	70.30
USKD [4]	✓	KL-Div	70.75
BAN [66]	✓	KL-Div	70.50
WKD-L	✓	WD	**71.35**

[13] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, A. Anandkumar. Born again neural networks. In ICML, 2018.

Experiments – Comparison with SOTA

➢ Object detection

Object detection on MS-COCO

- Extend WKD to object detection in the framework of Faster-RCNN.

Faster RCNN-FPN		RN101→RN18			RN50→MNV2		
		mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
Strategy	Teacher	42.04	62.48	45.88	40.22	61.02	43.81
	Student	33.26	53.61	35.26	29.47	48.87	30.90
Logit	KD [2]	33.97	54.66	36.62	30.13	50.28	31.35
	DKD [3]	35.05	56.60	37.54	32.34	53.77	34.01
	WKD-L (Ours)	**35.24**	**56.73**	**37.91**	**32.48**	**53.85**	**34.21**
Feature	FitNet [24]	34.13	54.16	36.71	30.20	49.80	31.69
	FGFI [50]	35.44	55.51	38.17	31.16	50.68	32.92
	ICD [51]	35.90	56.02	38.75	32.88	52.56	34.93
	ReviewKD [29]	36.75	56.72	34.00	33.71	53.15	36.13
	WKD-F (Ours)	**37.21**	**57.32**	**40.15**	**34.47**	**54.67**	**36.85**
Logit + Feature	DKD+ ReviewKD [3]	37.01	57.53	39.85	34.35	54.89	36.61
	WKD-L+ WKD-F (ours)	**37.49**	**57.76**	**40.39**	**34.80**	**55.27**	**37.28**
	FCFD† [8]	37.37	57.60	40.34	34.97	55.04	37.51
	WKD-L+ WKD-F† (ours)	**37.79**	**57.95**	**41.08**	**35.48**	**55.21**	**38.45**

RN: ResNet

Experiments — Comparison with SOTA

Training latency on ImageNet.

Strategy	Method	Top-1 (%)	Params (M)	Latency (ms)
Logit	KD [2]	71.03	0	215
	NKD [4]	71.96	0	214
	WKD-L (Ours)	72.49	0	280
Feature	ReviewKD [29]	71.61	7.2	349
	EMD+IPOT [16]	70.46	0.25	258
	WKD-F (Ours)	72.50	0.81	207
Logit + Feature	FCFD [8]	72.25	5.98	303
	ICKD-C [6]	72.19	0.33	222
	WKD-L+ WKD-F (ours)	72.76	0.81	292

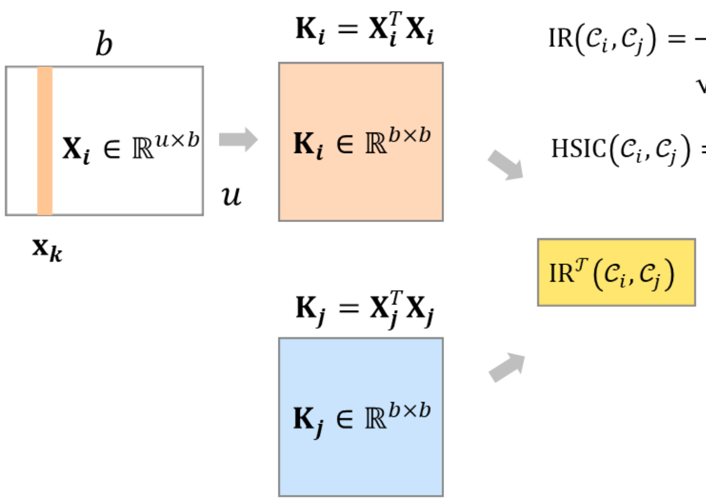
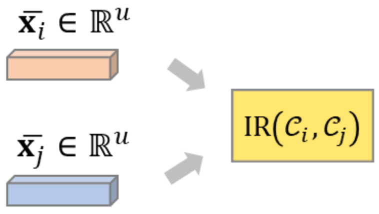
Experiments — Ablation

- Ablation of WKD-L
  - How to model category interrelations (IRs)?
- Ablation of WKD-F
  - How to model distributions?

Experiments — Ablation

Ablation of WKD-L

- How to model category interrelations (IRs)?

CKA	 $\mathrm{IR}(\mathcal{C}_i,\mathcal{C}_j) = \frac{\mathrm{HSIC}(\mathcal{C}_i,\mathcal{C}_j)}{\sqrt{\mathrm{HSIC}(\mathcal{C}_i,\mathcal{C}_i)}\sqrt{\mathrm{HSIC}(\mathcal{C}_j,\mathcal{C}_j)}}$ $\mathrm{HSIC}(\mathcal{C}_i,\mathcal{C}_j) = \frac{1}{(b-1)^2}\mathrm{tr}(\mathbf{K}_i\mathbf{H}\mathbf{K}_j\mathbf{H})$	**Linear kernel** $\mathbf{K}_i^{\mathrm{lin}} = \mathbf{X}_i^T\mathbf{X}_i$	**72.49**
		Polynomial kernel $\mathbf{K}_i^{\mathrm{poly}} = \left(\mathbf{X}_i^T\mathbf{X}_i + 1\right)^k$	72.10
		RBF kernel $\mathbf{K}_i^{\mathrm{rbf}} = \exp\left(-\frac{\mathbf{D}_i}{2\alpha^2\mathrm{Med}(\mathbf{D}_i)}\right)$ $\mathbf{D}_i = 2\left(\mathrm{diag}(\mathbf{X}_i^T\mathbf{X}_i)\mathbf{1}^T\right)_{\mathrm{sym}} - 2\mathbf{X}_i^T\mathbf{X}_i$	72.30
Cosine	 $\mathrm{IR}(\mathcal{C}_i,\mathcal{C}_j) = \frac{\bar{\mathbf{x}}_i{}^T\bar{\mathbf{x}}_j}{\sqrt{\bar{\mathbf{x}}_i{}^T\bar{\mathbf{x}}_i}\sqrt{\bar{\mathbf{x}}_j{}^T\bar{\mathbf{x}}_j}}$	**Classifier weight** $\bar{\mathbf{x}}_i = \mathbf{w}_i$,the i-th column of weight $\mathbf{W} \in \mathbb{R}^{u\times n}$ in the last FC layer	72.19
		Class centroid $\bar{\mathbf{x}}_i = \frac{1}{b}\mathbf{X}_i\mathbf{1}$	72.02

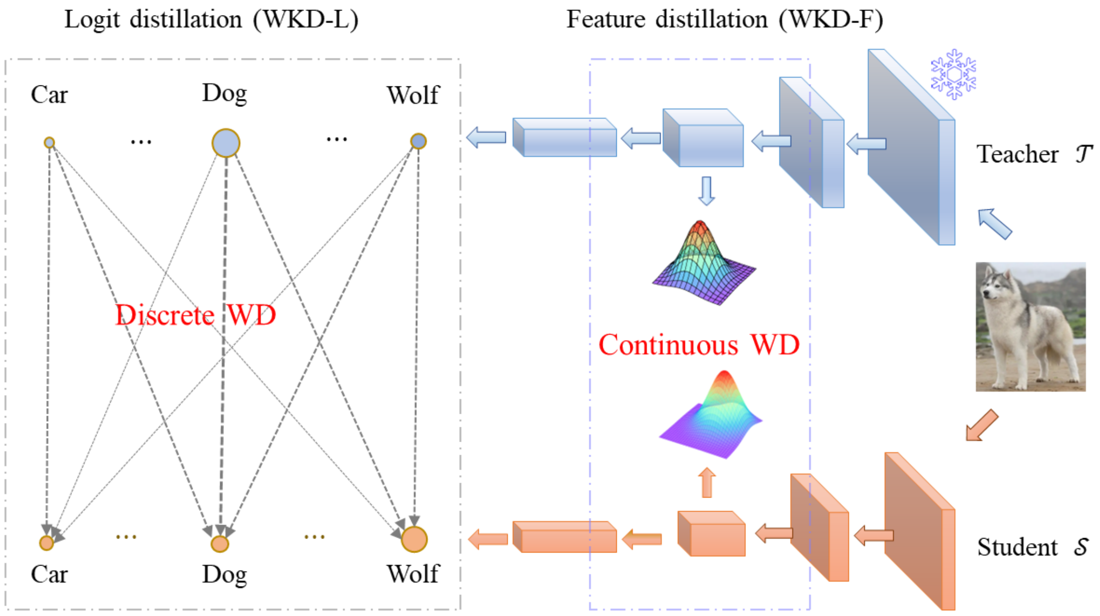
Experiments — Ablation

Ablation of WKD-F

- How to model distributions?

Distribution		Dis-similarity	Top-1
FitNet	--	Frobenius	70.53
Parametric	Gaussian (Diag)	WD	**72.50**
		KL-Div	71.75
		Sym KL-Div	71.93
	Laplace	KL-Div	71.38
	Exponential	KL-Div	70.14
Non-Parametric	Probability Mass Function	WD	71.57

Conclusion



- WKD-L leverages rich interrelations among classes via **cross-category comparisons** for logit distillation.
- WKD-F leverages **geometric structure** of the Riemannian space of Gaussians for feature distillation.
- WKD-L and WKD-F outperform KL-Div counterparts on **both** classification and detection tasks. Their **combination** further improves the performance.

Thanks for your attention

Our Lab

Code