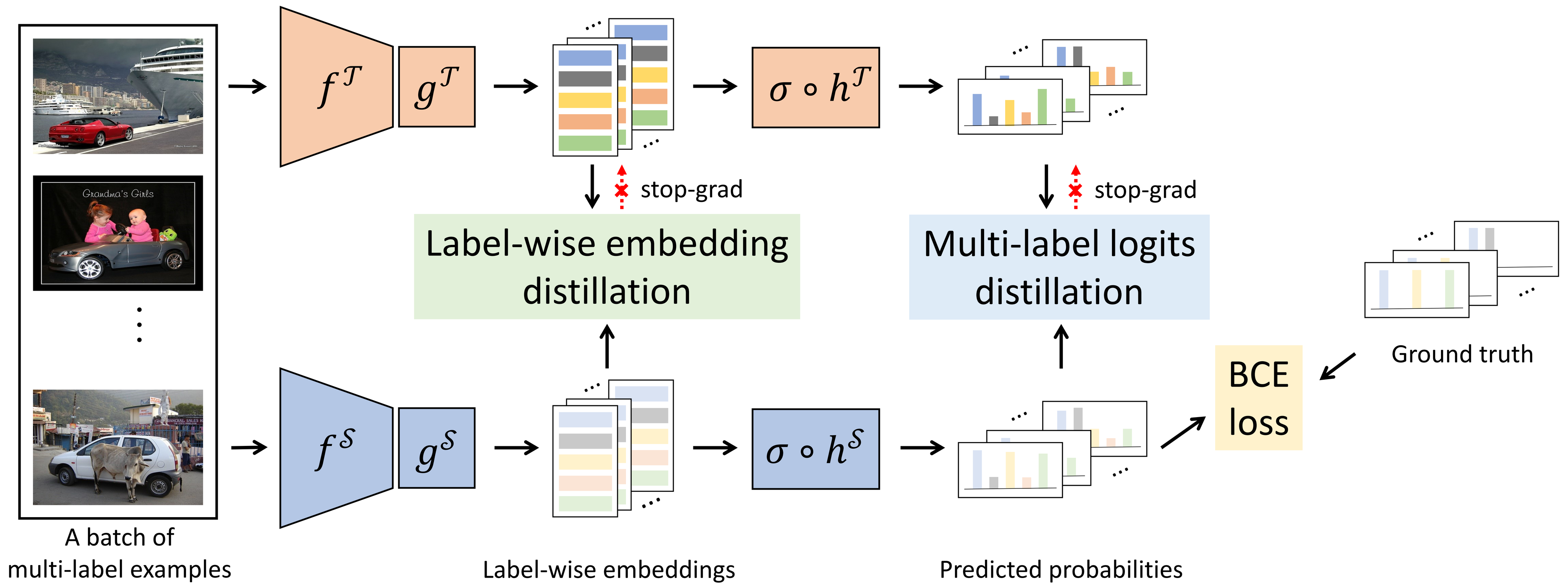
Multi-Label Knowledge Distillation

Penghui Yang, Ming-Kun Xie, Chen-Chen Zong, Lei Feng, Gang Niu, Masashi Sugiyama, Sheng-Jun Huang

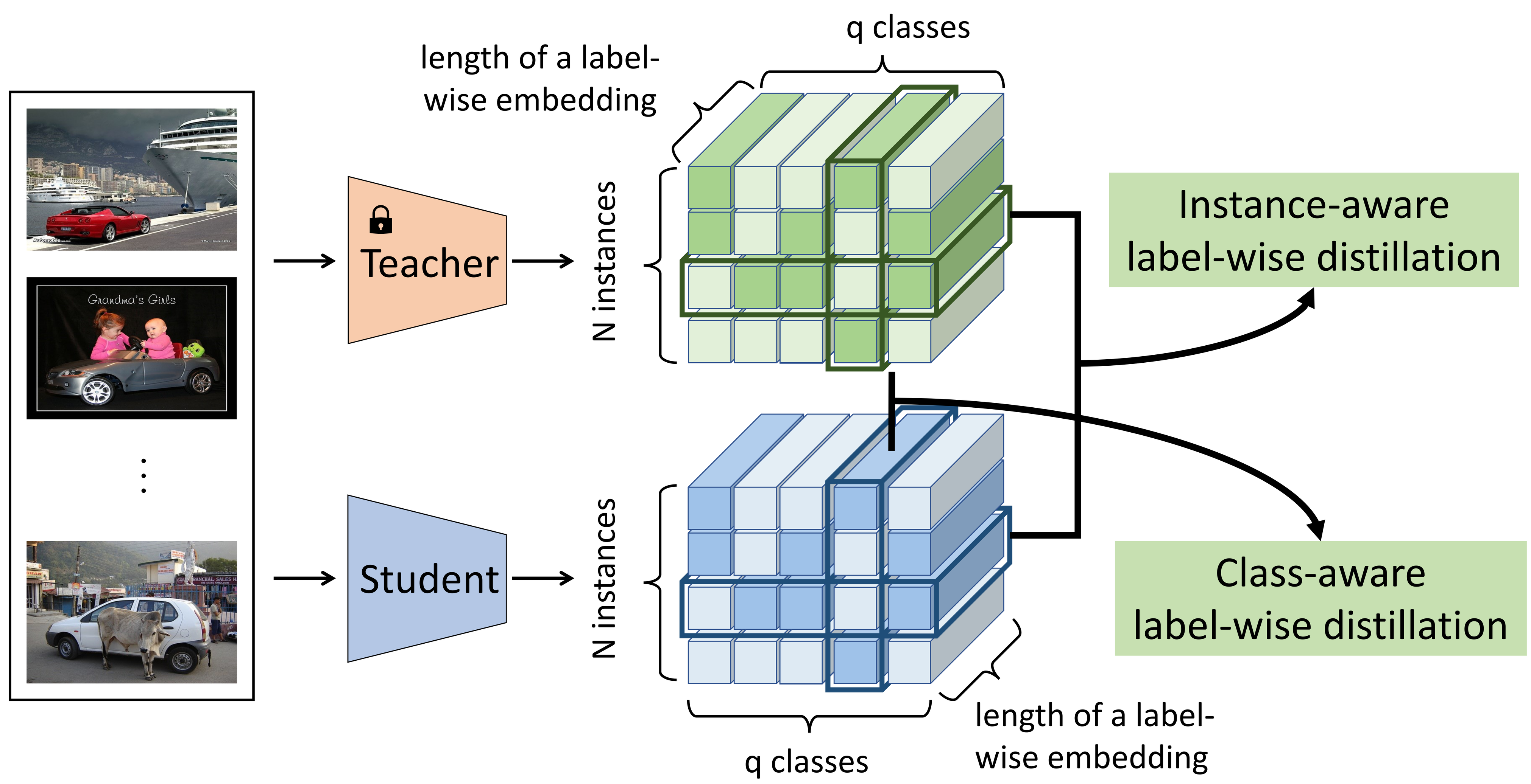
Why can't we extend existing KD methods to MLKD?

	Traditional KD	MLKD
Logits-based methods	The sum of predicted probabilities = 1, so we can use KL divergence.	The sum of predicted probabilities ≠ 1, so we cannot use KL divergence.
Feature-based methods	Only one object in a image, which makes the semantic information in the feature map very clear	Multiple semantics in one image, which makes the semantic information in the feature map unclear

Our proposed L2D framework

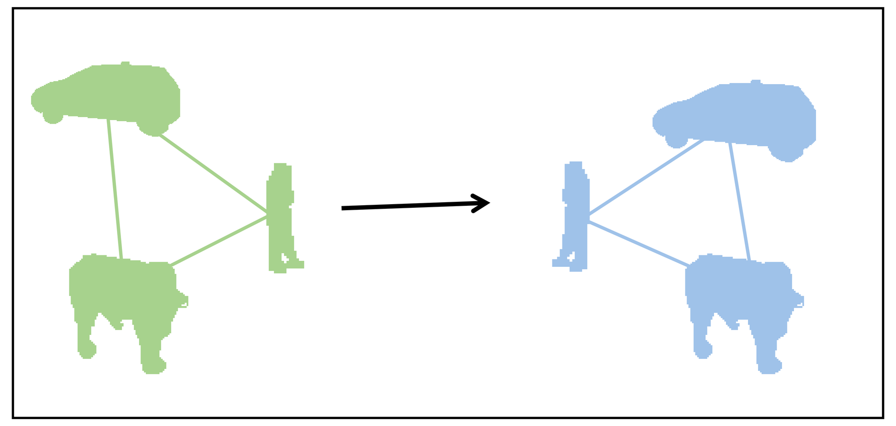


Class/Instance-aware Label-wise Embedding Distillation



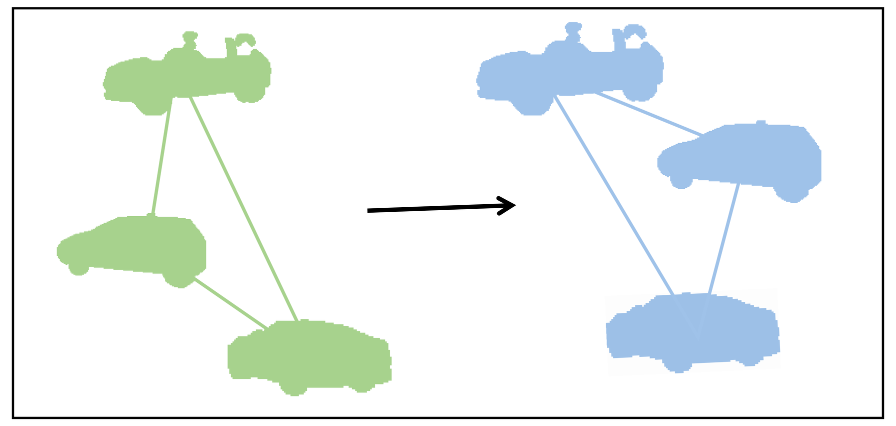
Instance-aware label-wise distillation

Take  as an example.



Class-aware label-wise distillation

Take class **car** as an example.



Performance on MS-COCO

Table 1. Results on MS-COCO where teacher and student models are in the **same** architectures.

Teacher	RepVGG-A2			ResNet-101			WRN-101			Swin-S		
Student	RepVGG-A0			ResNet-34			WRN-50			Swin-T		
Metrics	mAP	OF1	CF1	mAP	OF1	CF1	mAP	OF1	CF1	mAP	OF1	CF1
Teacher	72.71	74.11	68.63	73.62	73.89	68.61	74.70	75.56	70.73	81.70	80.48	77.12
Student	70.02	72.49	66.77	70.31	72.49	66.82	74.45	75.43	70.61	79.59	79.18	75.42
RKD	70.08	72.39	66.73	70.13	72.44	66.78	74.70	75.71	70.84	79.63	79.19	75.57
PKT	70.11	72.47	66.80	70.43	72.64	66.68	74.54	75.47	70.58	79.64	79.09	75.39
ReviewKD	70.00	72.35	66.82	70.39	72.62	66.76	74.03	75.29	70.36	79.81	79.18	75.55
MSE	70.26	72.54	66.99	70.54	72.75	66.85	74.53	75.60	70.71	79.67	79.20	75.52
PS	70.65	72.89	67.60	70.86	72.66	67.12	75.12	76.05	71.63	79.96	79.64	76.20
MLD	70.74	72.81	67.46	70.68	72.69	67.19	74.92	75.75	71.21	80.11	79.68	76.44
L2D	**72.81**	**74.59**	**69.49**	**72.87**	**74.45**	**69.43**	**76.61**	**77.08**	**72.79**	**81.59**	**81.03**	**77.86**

Table 2. Results on MS-COCO where teacher and student models are in the **different** architectures.

Teacher	ResNet-101			Swin-T			ResNet-101			Swin-T		
Student	RepVGG-A0			ResNet-34			MobileNet v2			MobileNet v2		
Metrics	mAP	OF1	CF1	mAP	OF1	CF1	mAP	OF1	CF1	mAP	OF1	CF1
Teacher	73.62	73.89	68.61	79.43	78.77	75.07	73.62	73.89	68.61	79.43	78.77	75.07
Student	70.02	72.49	66.77	70.31	72.49	66.82	71.85	73.59	68.26	71.85	73.59	68.26
RKD	70.08	72.35	66.72	70.00	72.34	66.64	71.76	73.68	68.40	71.74	73.68	68.37
PKT	69.99	72.35	66.56	70.26	72.39	66.82	71.88	73.60	68.35	71.84	73.76	68.37
ReviewKD	70.00	72.33	66.62	70.29	72.39	66.58	71.92	73.73	68.48	71.73	73.71	68.36
MSE	70.07	72.50	66.85	70.33	72.57	66.72	71.91	73.68	68.28	71.80	73.74	68.38
PS	70.30	72.61	67.10	70.94	72.93	67.57	72.11	73.89	68.42	72.42	74.14	68.94
MLD	70.48	72.77	67.10	71.14	72.99	67.63	72.17	73.84	68.52	72.35	74.10	68.91
L2D	**72.14**	**74.08**	**68.78**	**73.42**	**74.97**	**70.20**	**73.24**	**74.85**	**69.72**	**74.21**	**75.72**	**70.87**

Performance on VOC

Table 3. Results on Pascal VOC 2007 validation teacher and student models are in the **same** architectures.

Teacher	RepVGG-A2			ResNet-50			WRN-101			Swin-S		
Student	RepVGG-A0			ResNet-18			WRN-50			Swin-T		
Metrics	mAP	OF1	CF1	mAP	OF1	CF1	mAP	OF1	CF1	mAP	OF1	CF1
Teacher	86.20	85.63	82.62	86.73	84.92	81.21	88.00	87.03	83.72	92.75	91.05	88.82
Student	83.79	83.36	79.83	84.01	83.60	79.42	88.52	87.21	84.08	91.31	89.98	88.00
RKD	83.75	83.41	79.85	84.48	83.54	79.83	88.21	87.33	84.55	91.52	90.44	88.51
PKT	83.63	83.53	80.04	84.12	83.10	79.31	87.69	87.07	84.14	91.28	90.17	88.03
ReviewKD	83.87	83.98	80.54	83.71	83.01	79.25	88.23	87.13	84.20	91.45	90.17	88.06
MSE	84.02	83.67	79.94	84.23	83.16	79.29	88.04	86.49	83.57	91.06	89.99	87.66
PS	83.77	83.74	80.28	84.44	83.78	79.95	88.30	86.92	83.91	91.21	90.25	88.12
MLD	83.65	83.66	80.02	84.48	84.07	80.29	88.29	87.16	84.25	91.43	90.72	88.81
L2D	**84.56**	**84.37**	**80.82**	**85.71**	**85.70**	**82.11**	**89.52**	**88.25**	**85.69**	**91.92**	**91.34**	**89.58**

Table 4. Results on Pascal VOC 2007 validation where teacher and student models are in the **different** architectures.

Teacher	ResNet-50			Swin-T			ResNet-50			Swin-T		
Student	RepVGG-A0			ResNet-18			MobileNet v2			MobileNet v2		
Metrics	mAP	OF1	CF1	mAP	OF1	CF1	mAP	OF1	CF1	mAP	OF1	CF1
Teacher	86.73	84.92	81.21	91.43	89.81	87.63	86.73	84.92	81.21	91.43	89.81	87.63
Student	83.79	83.36	79.83	84.01	83.60	79.42	86.12	85.01	81.76	86.12	85.01	81.76
RKD	84.26	84.29	80.70	83.27	83.05	79.55	86.22	84.97	81.76	85.68	85.31	81.57
PKT	83.93	83.79	80.03	83.45	83.25	79.64	86.10	84.84	81.66	85.67	85.22	81.68
ReviewKD	84.07	83.62	80.34	83.37	83.08	78.93	85.87	85.04	81.73	85.69	85.10	81.56
MSE	84.01	84.05	80.52	83.60	83.06	79.46	86.20	84.94	81.84	85.80	85.51	81.98
PS	84.80	84.46	81.13	83.97	83.75	79.86	86.26	85.47	82.06	86.07	85.73	82.39
MLD	85.07	84.91	81.55	84.61	84.26	80.78	86.38	85.67	82.43	86.11	85.98	82.55
L2D	**86.26**	**85.85**	**82.55**	**85.87**	**85.67**	**82.17**	**87.32**	**86.48**	**83.26**	**87.37**	**86.88**	**83.68**

Performance on NUS-WIDE

Teacher	ResNet-101			Swin-T		
Student	ResNet-34			MobileNet v2		
Metrics	mAP	OF1	CF1	mAP	OF1	CF1
Teacher	55.32	75.56	61.31	59.73	77.30	65.44
Student	53.41	75.10	60.08	54.49	75.72	61.74
RKD	53.62	75.20	59.91	54.76	75.69	61.74
PKT	53.55	75.08	60.35	54.59	75.69	61.74
ReviewKD	53.52	75.23	60.44	54.85	75.84	61.75
MSE	53.52	75.13	59.94	54.86	75.80	61.69
PS	54.14	75.43	60.79	55.18	75.91	62.35
MLD	54.44	75.36	60.73	55.36	76.00	62.52
L2D	**55.31**	**76.17**	**62.79**	**56.91**	**76.92**	**63.89**

Performance on Reversed KD

ResNet34 → ResNet101

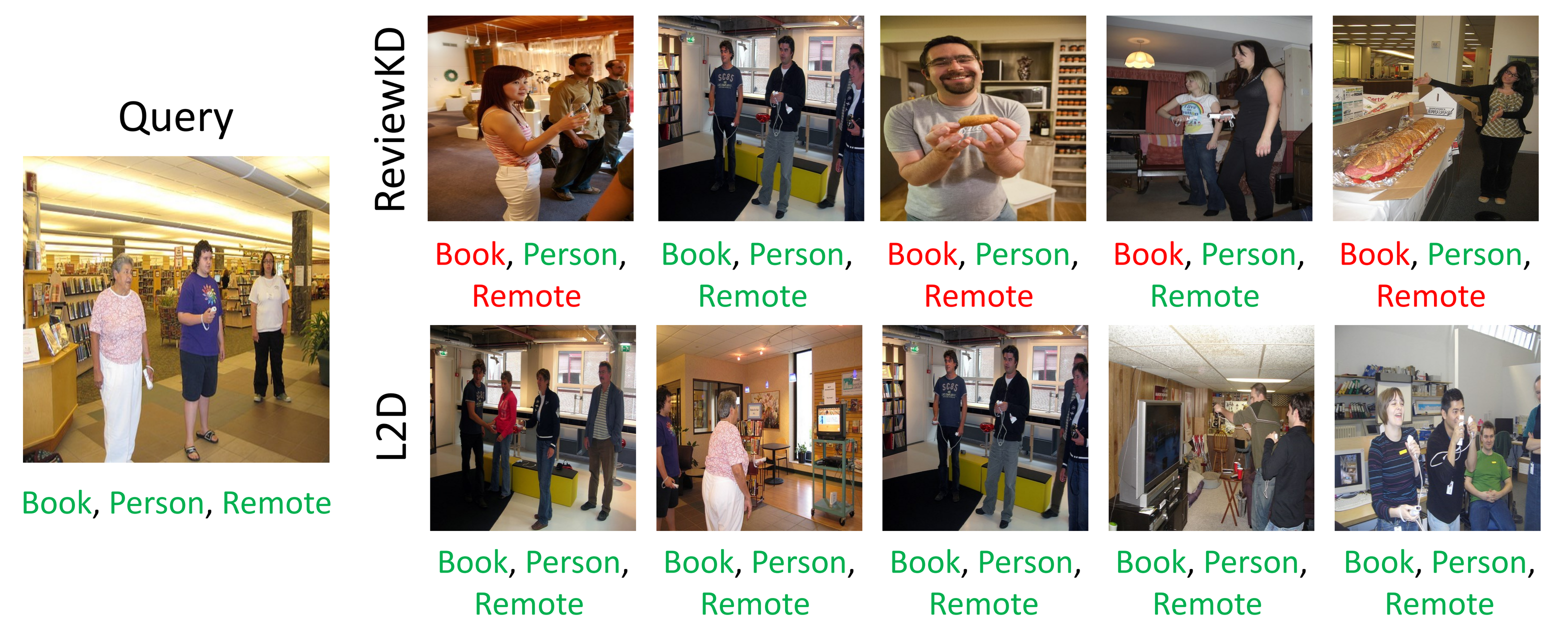
Metrics	mAP	OF1	CF1
Teacher	70.19	72.30	66.50
Student	73.98 (+3.79)	75.01 (+2.71)	70.12 (+3.62)
RKD	74.03 (+3.84)	74.96 (+2.66)	70.01 (+3.51)
PKT	73.95 (+3.76)	74.94 (+2.64)	69.98 (+3.48)
ReviewKD	74.02 (+3.83)	74.96 (+2.66)	70.07 (+3.57)
MSE	74.21 (+4.02)	75.12 (+2.82)	70.18 (+3.68)
PS	74.70 (+4.51)	75.78 (+3.48)	71.08 (+4.58)
MLD	74.64 (+4.45)	75.78 (+3.48)	71.10 (+4.60)
L2D	**75.51(+5.32)**	**76.25(+3.95)**	**71.75(+5.25)**

Per-Class Performance on VOC

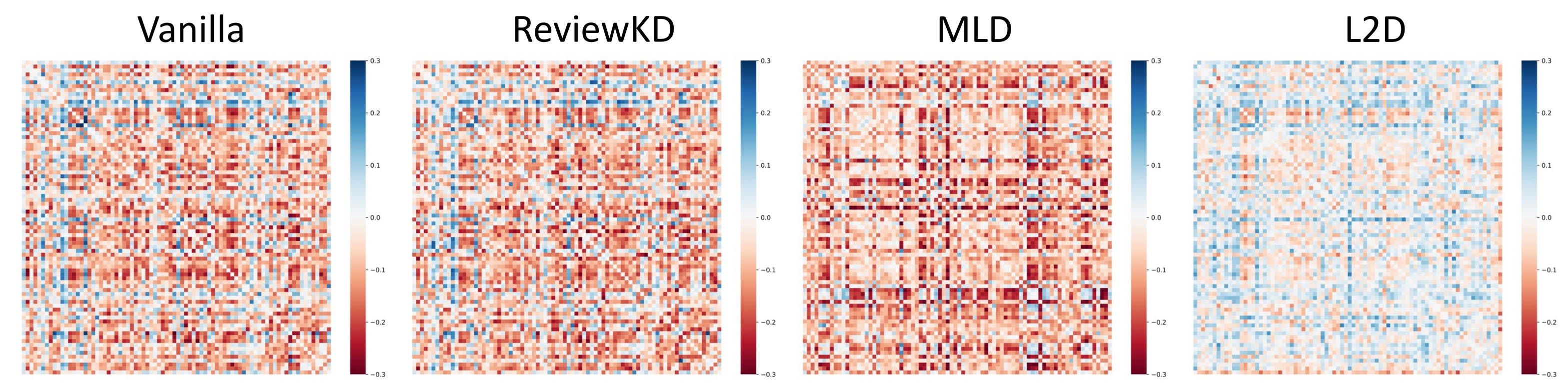
Comparison results of the comparing methods on VOC in terms of AP and mAP (%), where the backbones of teacher and student model are respectively ResNet-50 and ResNet-18. The best performance is highlighted in red, and second best performance is highlighted in blue.

Methods	bottle	plant	chair	sofa	table	cow	tv	bus	sheep	mbike	dog	bird	bike	cat	boat	car	horse	person	train	aero	mAP
Vanilla	57.18	67.43	70.35	73.17	76.14	82.65	82.30	85.53	84.16	88.38	90.18	90.65	91.40	90.67	92.54	92.72	94.22	95.73	96.09	97.27	84.01
RKD	59.01	67.01	71.31	72.86	76.90	81.11	81.70	84.79	84.77	88.86	90.03	91.65	91.53	91.86	92.53	92.06	93.97	95.88	97.08	97.00	84.18
PKT	57.29	66.16	71.16	73.22	76.89	81.75	81.95	85.06	83.66	88.73	90.10	91.31	91.77	91.83	92.35	92.07	93.12	95.80	96.99	97.09	83.86
MSE	58.26	68.02	70.68	72.06	77.93	81.06	82.36	86.03	83.72	88.30	90.46	90.60	91.58	91.07	91.64	92.01	94.05	95.69	97.13	97.10	84.23
MLD	58.32	68.88	71.12	73.77	78.65	84.60	82.42	86.41	83.77	88.58	90.78	91.02	91.92	91.63	92.25	92.34	94.53	96.04	96.97	97.30	84.48
L2D	59.71	70.52	74.77	75.01	78.93	83.87	84.13	85.45	85.67	89.83	90.60	91.48	91.90	92.14	92.57	93.40	94.67	96.46	96.81	97.39	85.71

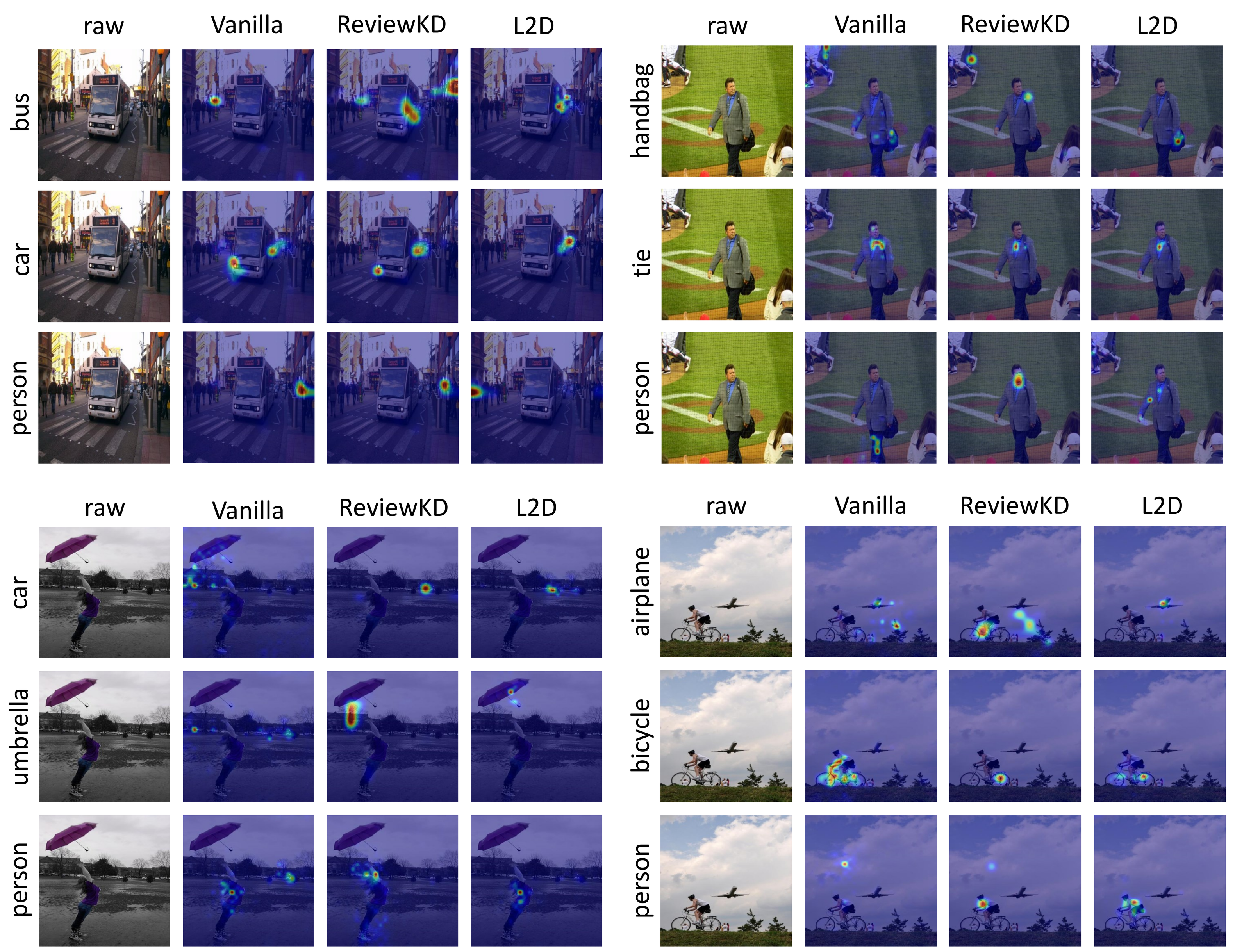
Performance on Image Retrieval



Distilling Inter-class Correlations



Visualization of Attention Maps



Acknowledgement: Sheng-Jun Huang was supported by the National Key R&D Program of China(2020AAA0107000), the Natural Science Foundation of Jiangsu Province of China(BK20222012, BK20211517), and NSFC(62222605). Lei Feng was supported by the Joint NTU-WeBank Research Centre on Fintech(Award No: NWJ-2021-005), Nanyang Technological University, Singapore, Chongqing Overseas Chinese Entrepreneurship and Innovation Support Program, Chongqing Artificial Intelligence Innovation Center, and CAAI-Huawei MindSpore Open Fund. Masashi Sugiyama was supported by JST CREST Grant Number JPMJCR18A2.