Knowledge Distillation Layer that Lets the Student Decide

Ada Görgün, Yeti Z. Gürbüz and A. Aydın Alatan

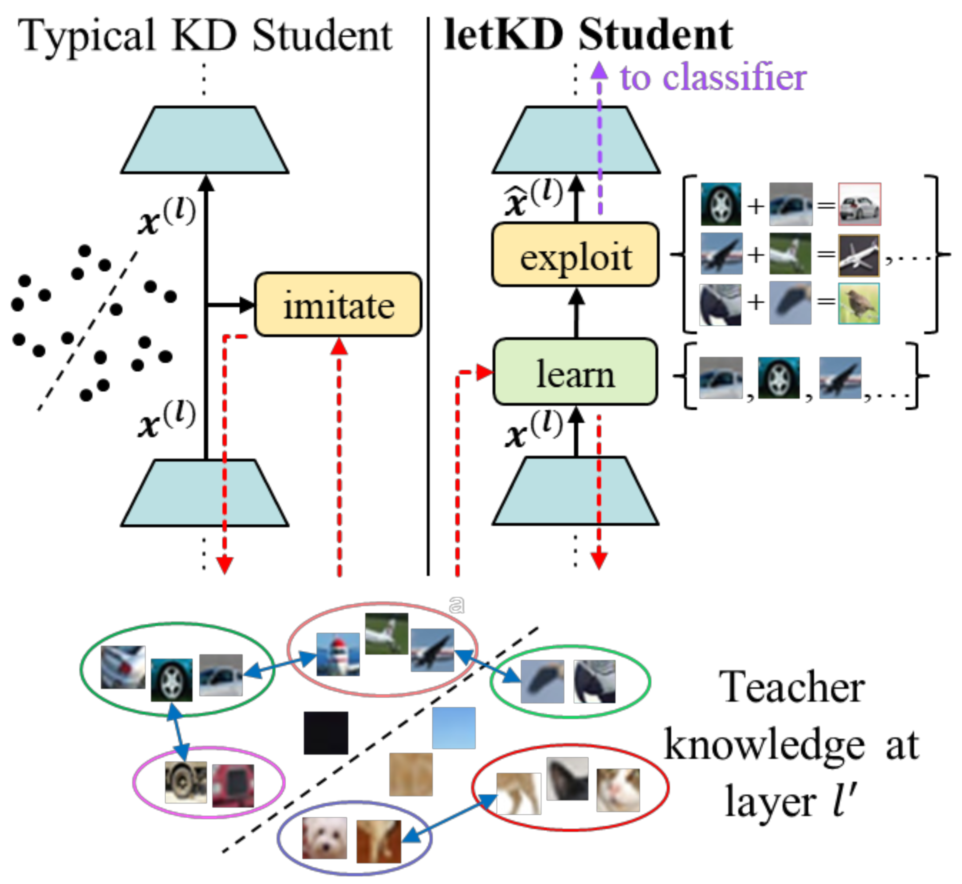
{ada.gorgun, yeti, alatan}@metu.edu.tr

BMVC 2023

20th - 24th November 2023, Aberdeen, UK

Problem

Typical KD methods use regularization while pushing the student to **imitate** the feature geometry of the teacher.



: teacher knowledge propagation

: inter-class relationships captured by the teacher

Considering the *architectural* differences in between, **forcing** the student to imitate the teacher's responses would be demanding, especially for the intermediate layers.

Method: Formulation

Key Idea: Learn the semantic entities that the teacher finds useful and **exploit** them in feature transform, enabling us to feed forward the knowledge during inference as well.

Given a set of matching kernels ω_k and features x_i at spatial location i, **we define** feature embedding by template matching as:

$$p_{|i} = \underset{\substack{p,q \geqslant 0 \\ q+\Sigma_k p_k = 1}}{\arg\max}\, q\,\mu + \sum_k p_k\, \omega_k^{\intercal} x_i$$

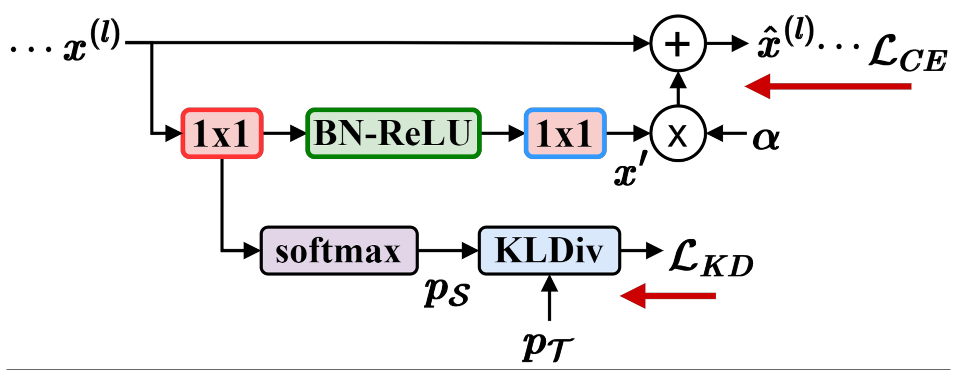
learned through the teacher

Feature Embedding

$$x_i' = \sum_k p_{k|i} \nu_k$$

Solver (see paper for details): 1x1-BN-ReLU-1x1 is equivalent to feature embedding by template matching.

Proposed KD Layer:



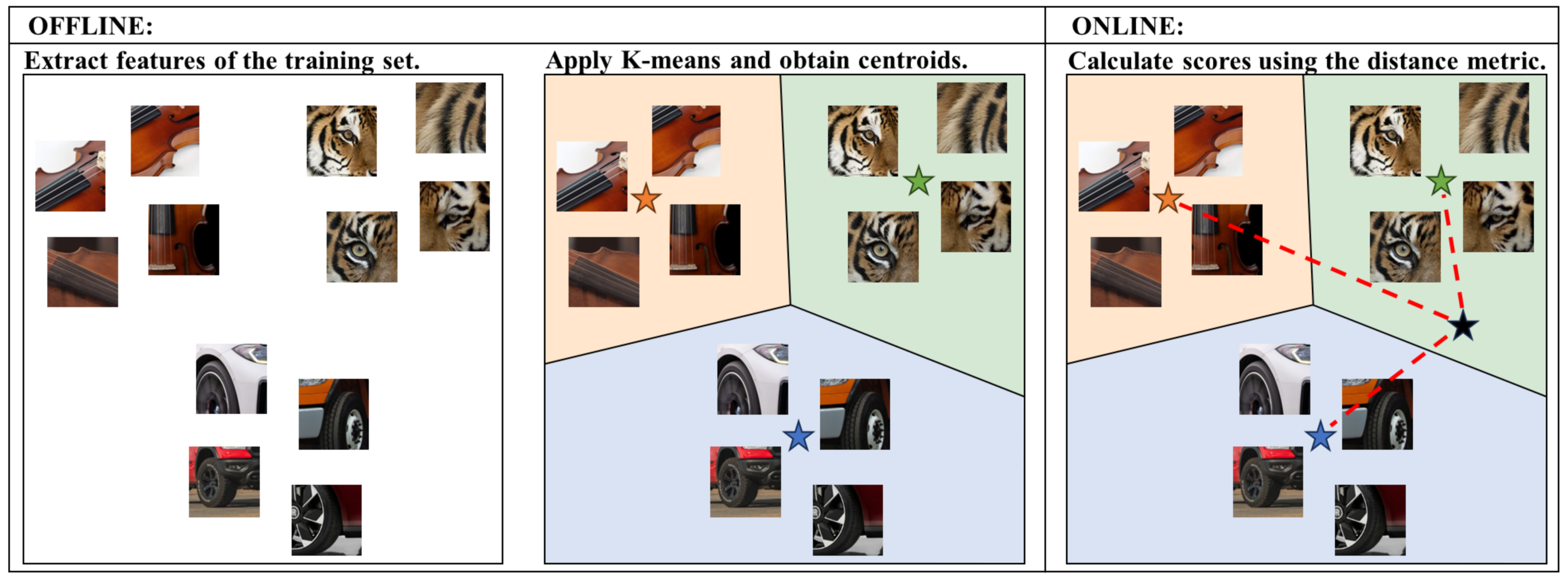
conv. : learning feedback

p_S / p_T : student's/teacher's soft predictions

Penultimate Layer Teacher Supervision

Viewing each pixel in a feature map as a *semantic entity*, **we propose** to employ *K-means* to the teacher's features to obtain fine-grained labels for these entities.

$$p_{\mathcal{T}}(\star) = \text{softmax}\left(-\frac{[\|c-\star\|_2^2]_{c=[\star,\star,\star]}}{\tau}\right)$$

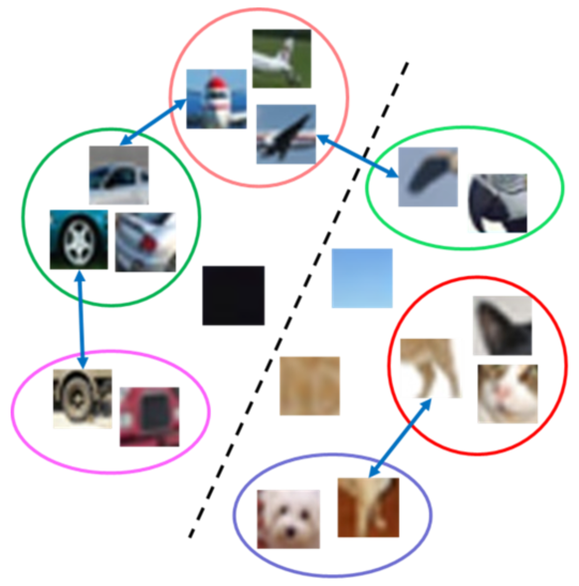


Intermediate Layer Teacher Supervision

Data-driven mechanisms (*e.g.* *K*-means) lead to imitation of the teacher's geometry, hurting the performance of the student due to capacity differences.

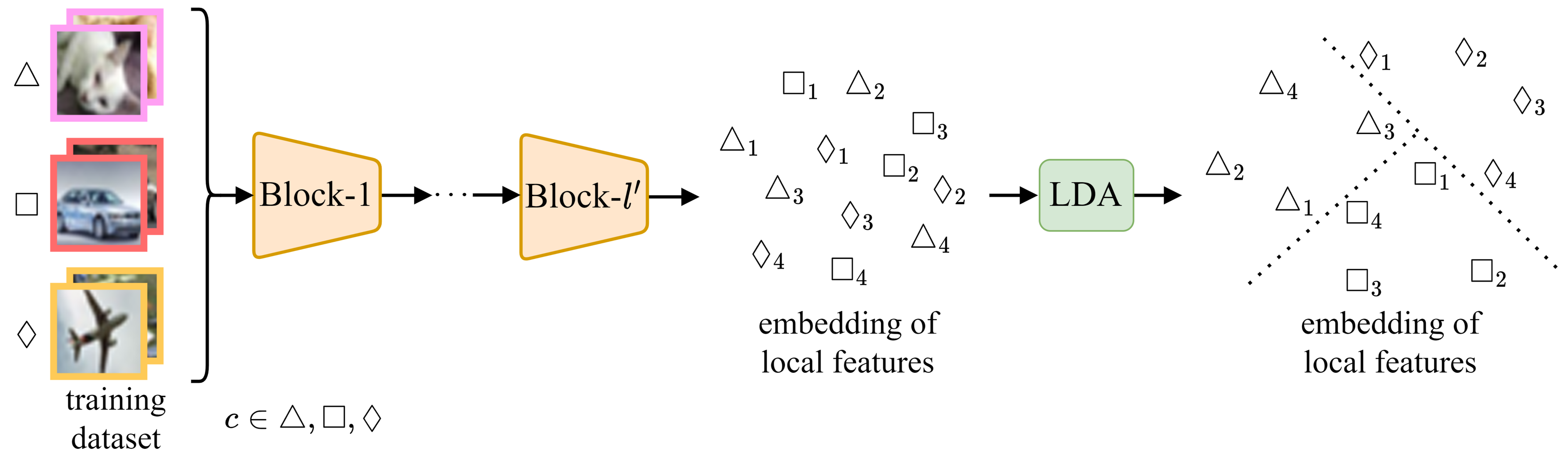
Key Idea: Exploit the relationship between different classes (*e.g.* *leg* for *horse* and *deer*) that are learned by the teacher.

$p_{\mathcal{T}}(i)$: probability of classifying feature i as a particular prototype given i is assigned to its corresponding prototype.

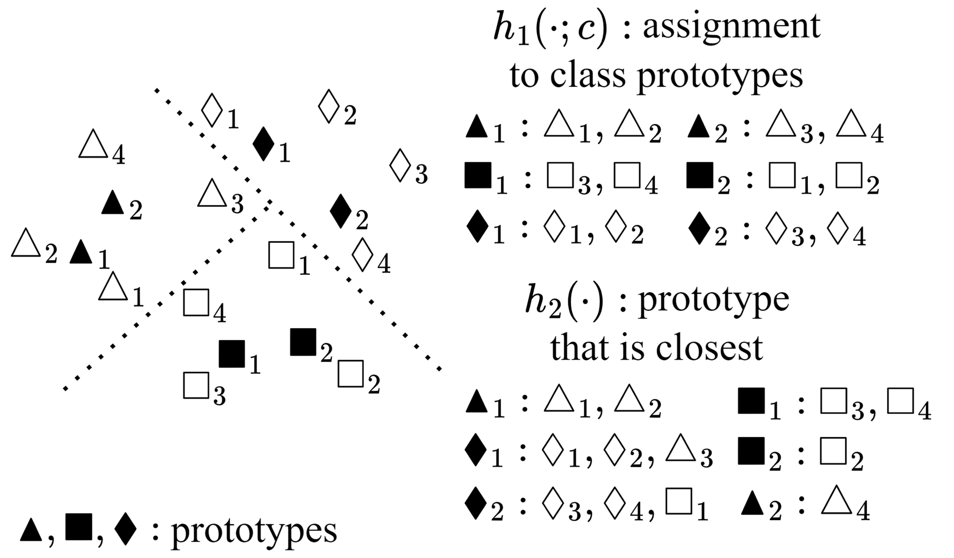


Offline Teacher Supervision:

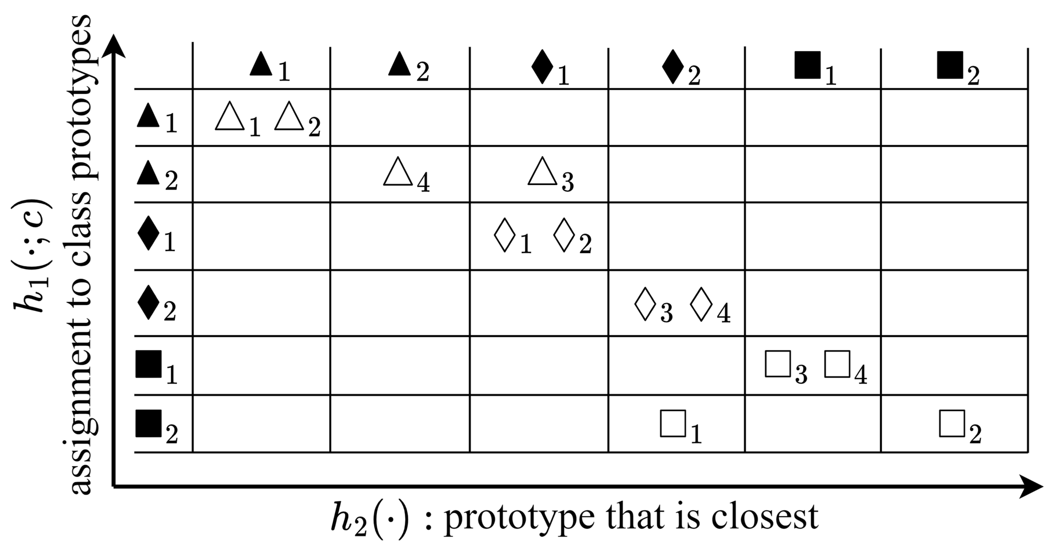
Step-1: Extract features of the training set and apply LDA.



Step-2: Apply *K*-means for each class separately. Obtain prototypes as the cluster centers (sub-classes).

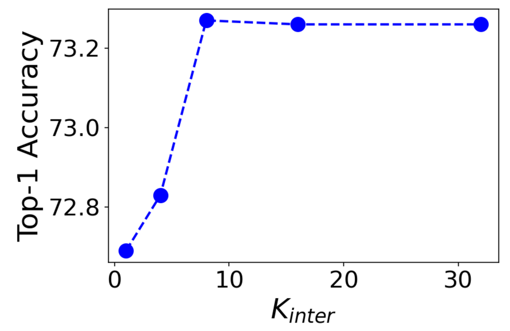


Step-3: Assign the neareast-neighbor prototype to each sample and normalize each row to calculate $p_{\mathcal{T}}(\cdot) = p(h_2(\cdot) \mid h_1(\cdot; c))$, where $c \in \triangle, \square, \lozenge$.



Ablation Studies

Clusters → / Archs. ↓	3x3 Kernels QUEST	3x3 Kernels letKD-1	K-Means QUEST	K-Means letKD-1	Teacher	Student
RN56-RN20	71.92	72.11	71.84	72.44	72.34	69.06
RN110-RN32	74.31	74.44	74.08	74.40	74.31	71.14
RN83-RN29	72.41	72.61	72.48	73.33	73.84	70.53



Method	Top-1 Acc.↑
QUEST	71.84
QUEST (2)	71.79
letKD-1	72.44
letKD-1 (2)	72.73

Without K**-means.** Kernels of 3x3 of the residual blocks in architectures such as ResNet correspond to learnable templates (*i.e.*, cluster centers) of some semantic entities.

Employing sub-classes (K_{inter}) for the intermediate layer distillation enables better knowledge transfer.

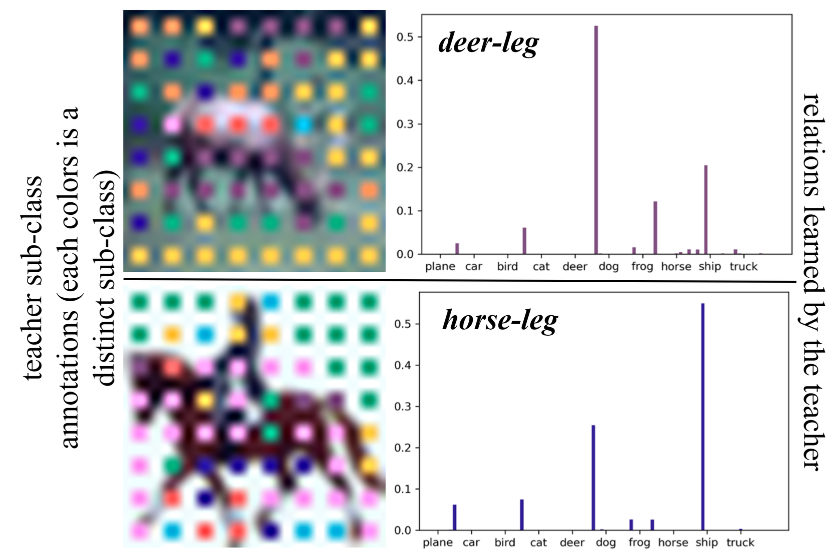
Naive multi-layer distillation (denoted as **(2)**) without our KD layer hurts the performance, suggesting that a better way should be found.

Effect of our KD layer:

Classification capacity in the intermediate layers is improved with our KD layer.

Marginal increase in the computation is not the main source of improvement.

The major contribution to the performance occurs upon combining the KD layer with our supervision.



	letKD-2 ($\alpha_{inter} = 0$) x	letKD-2 ($\alpha_{inter} = 1$) $\hat{x}$	
Intermediate Layer Top-1 Acc.↑	52.71	51.71	56.36

Methods	Top-1 Acc.↑
FitNet	71.59
FitNet+KD layer without supervision	71.80
FitNet+KD layer with supervision	73.36

Inter.	α_{inter}	Penult.	α_{penult}	Top-1 Acc.↑
✓	0	-	0	70.64
✓	1	-	0	70.80
-	0	✓	0	71.84
-	0	✓	1	72.44
✓	0	✓	0	71.70
✓	0	✓	1	72.13
✓	1	✓	0	72.78
✓	1	✓	1	73.27

Image Recognition Experiments: Top-1 Acc.↑

CIFAR100	Homogeneous						Heterogeneous			
Teacher	WRN-40-2	WRN-40-2	RN56	RN110	RN110	RN32x4	WRN-40-2	RN32x4	RN32x4	RN50
Student	WRN-16-2	WRN-40-1	RN20	RN20	RN32	RN8x4	SNV1	SNV1	SNV2	MNV2
Methods ↓	75.61 73.26	75.61 71.98	72.34 69.06	74.31 69.06	74.31 71.14	79.42 72.50	75.61 70.50	79.42 70.50	79.42 71.82	79.34 64.60
SimKD	76.06	74.92	68.95	69.35	72.15	**78.08**	76.95	77.18	77.78	68.91
TDD	75.01	74.04	71.53	-	-	-	75.60	-	-	68.37
QUEST	76.10	74.58	71.84	71.89	74.08	75.88	76.75	76.28	77.09	69.81
letKD-1	76.29 $\mp$0.15	75.01 $\mp$0.09	72.44 $\mp$0.24	72.68 $\mp$0.31	74.40 $\mp$0.14	76.70 $\mp$0.06	76.93 $\mp$0.16	76.65 $\mp$0.24	77.75 $\mp$0.17	69.97 $\mp$0.18
letKD-2	**76.56** **$\mp$0.22**	**75.19** **$\mp$0.13**	**73.27** **$\mp$0.16**	**73.38** **$\mp$0.14**	**74.62** **$\mp$0.20**	**77.09** **$\mp$0.18**	**77.08** **$\mp$0.12**	**77.30** **$\mp$0.12**	**77.95** **$\mp$0.06**	**70.39** **$\mp$0.23**

ImageNet		Teacher	Student	KD	DKD	QUEST	**letKD-1**	**letKD-2**
RN34-RN18	Top-1	73.31	69.75	70.66	71.70	71.67	72.33	**72.38**
	Top-5	91.42	89.07	89.88	90.41	90.67	91.06	**91.15**
RN50-MNV2	Top-1	76.13	68.87	68.58	72.05	72.54	73.78	**73.98**
	Top-5	92.86	88.76	88.98	91.05	91.13	91.81	**92.00**