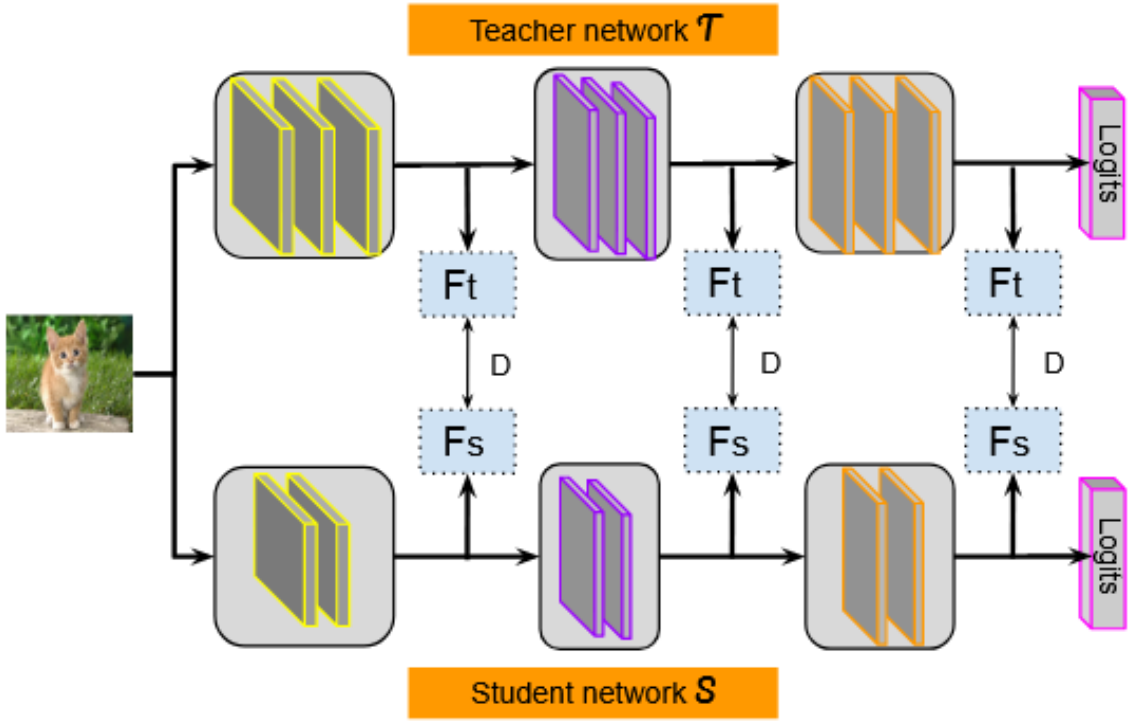
NEURAL INFORMATION PROCESSING SYSTEMS

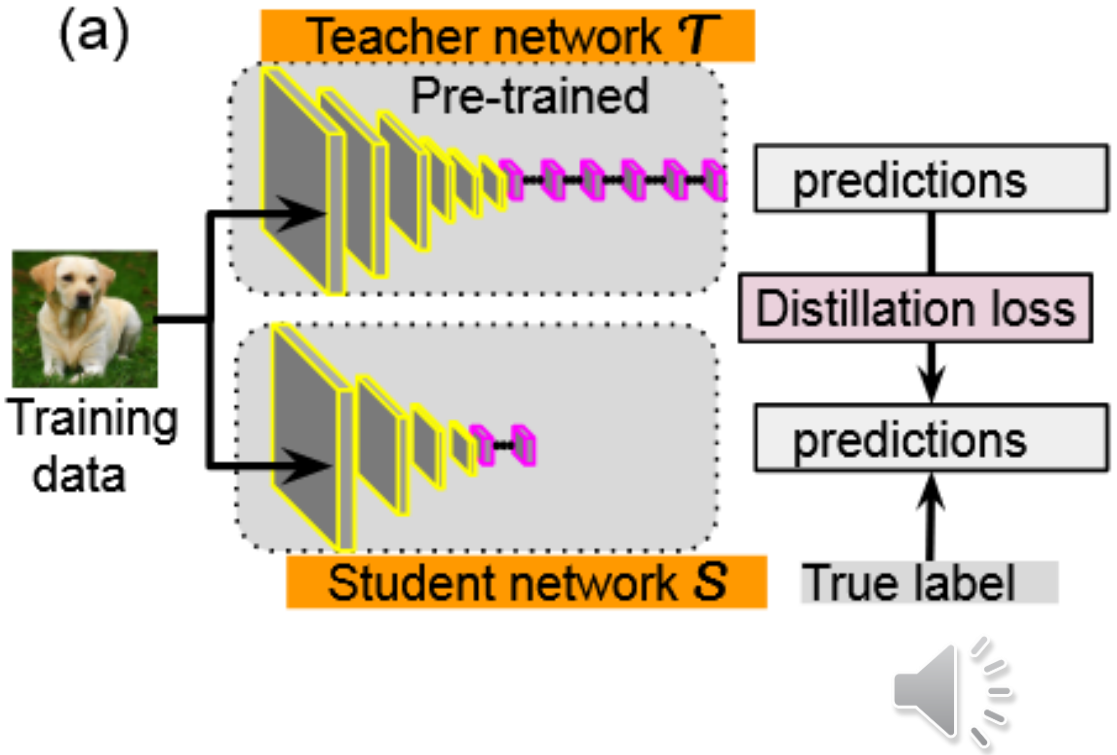
KD-Zero: Evolving Knowledge Distiller for Any Teacher-Student Pairs

Lujun Li Peijie Dong Anggeng Li Zimian Wei Ya Yang

Hello everyone, it's a great honour to introduce our NeurIPS 2023 work

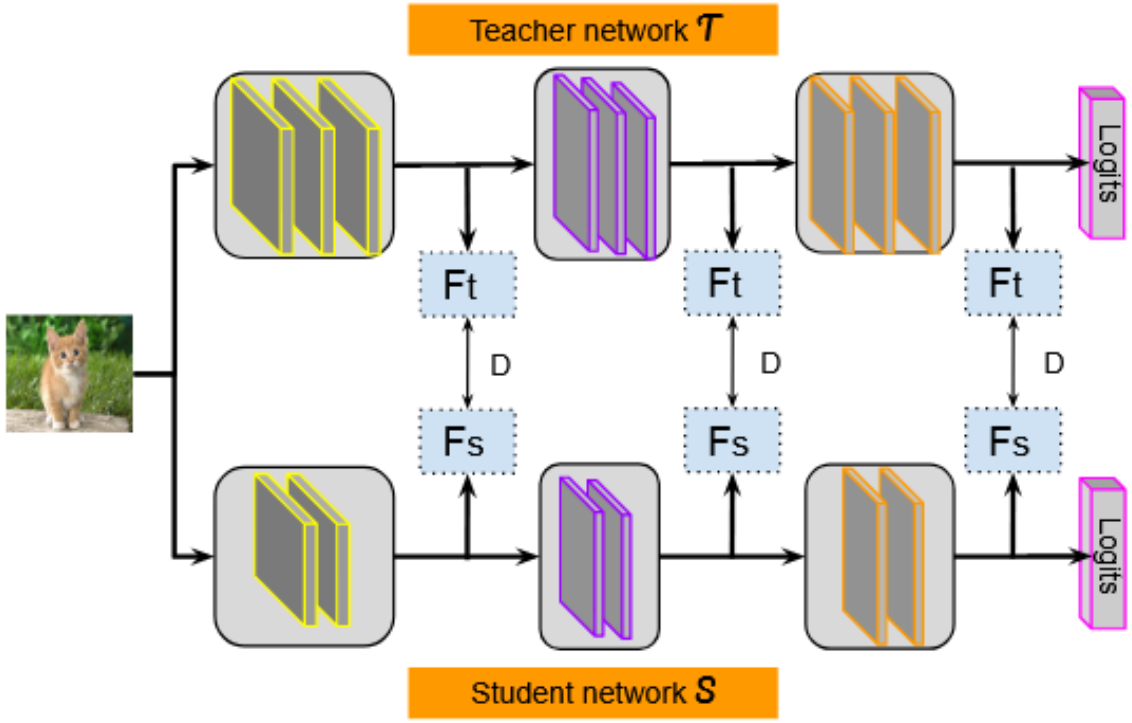
Background: Knowledge distillation

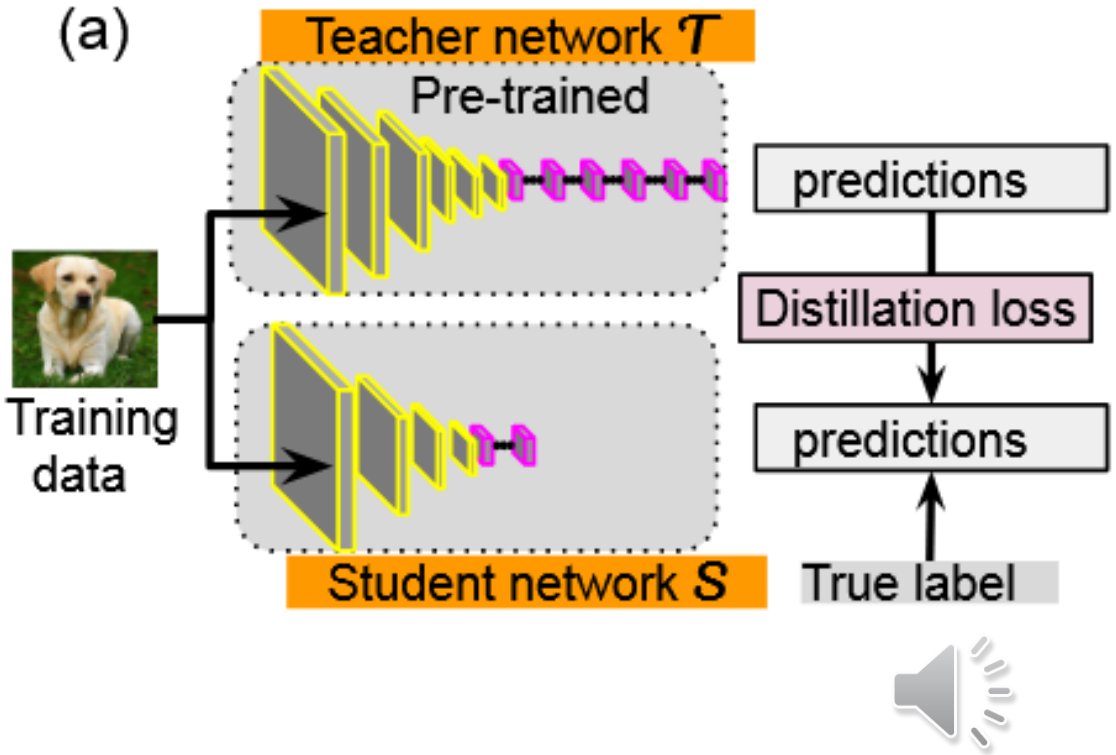




Knowledge distillation uses features or logits of teacher models to improve student models

Background: Knowledge distillation





Knowledge distillation can effectively improve performance on different tasks

$$\mathcal{L}(F_T, F_S) = D(TF_t(F_T), TF_s(F_S))$$

Method	Teacher's TF_t	Student's TF_s	Distillation position	Distance metric	Lost knowledge
FitNet [52]	None	1×1 Conv	Middle layer	L_1	None
AT [36]	Attention map	Attention map	End of layer group	L_2	Channel dims
KP [56]	Projection matrix	Projection matrix	Middle layers	L_1 + KP loss	Spatial dims
FSP [57]	FSP matrix	FSP matrix	End of layer group	L_2	Spatial dims
FT [54]	Encoder-decoder	Encoder-decoder	End of layer group	L_1	Channel + Spatial dims
AT [36]	Attention map	Attention map	End of layer group	L_2	Channel dimensions
MINILM [58]	Self-ttention	Self-attention	End of layer group	KL	Channel dimensions
Jacobian [59]	Gradient penalty	Gradient penalty	End of layer group	L_2	Channel dims
SVD [57]	Truncated SVD	Truncated SVD	End of layer group	L_2	Spatial dims
VID [8]	None	1×1 Conv	Middle layers	KL	None
IRG [18]	Instance graph	Instance graph	Middle layers	L_2	Spatial dims
RCO [60]	None	None	Teacher's train route	L_2	None
SP [61]	Similarity matrix	Similarity matrix	Middle layer	Frobenius norm	None
MEAL [62]	Adaptive pooling	Adaptive pooling	End of layer group	$L_{1/2}$/KL/L_{GAN}	None
Heo [62]	Margin ReLU	1×1 Conv	Pre-ReLU	Partial L_2	Negative features
AB [63]	Binarization	1×1 Conv	Pre-ReLU	Margin L_2	feature values
Chung [64]	None	None	End of layer	L_{GAN}	None
Wang [65]	None	Adaptation layer	Middle layer	Margin L_1	Channel + Spatial dims
KSANC [66]	Average pooling	Average pooling	Middle layers	L_2 + L_{GAN}	Spatial dims
Kulkarni [67]	None	None	End of layer group	L_2	None
IR [68]	Attention matrix	Attention matrix	Middle layers	KL+ Cosine	None
Liu [18]	Transform matrix	Transform matrix	Middle layers	KL	Spatial dims
NST [55]	None	None	Intermediate layers	MMD	None
Gao [69]	None	None	Intermediate layers	L_2	None

Different transformations, distance functions make up different KD designs

$$\mathcal{L}(F_T, F_S) = D(TF_t(F_T), TF_s(F_S))$$

Method	Teacher's TF_t	Student's TF_s	Distillation position	Distance metric	Lost knowledge
FitNet [52]	None	1×1 Conv	Middle layer	L_1	None
AT [36]	Attention map	Attention map	End of layer group	L_2	Channel dims
KP [56]	Projection matrix	Projection matrix	Middle layers	L_1 + KP loss	Spatial dims
FSP [57]	FSP matrix	FSP matrix	End of layer group	L_2	Spatial dims
FT [54]	Encoder-decoder	Encoder-decoder	End of layer group	L_1	Channel + Spatial dims
AT [36]	Attention map	Attention map	End of layer group	L_2	Channel dimensions
MINILM [58]	Self-ttention	Self-attention	End of layer group	KL	Channel dimensions
Jacobian [59]	Gradient penalty	Gradient penalty	End of layer group	L_2	Channel dims
SVD [57]	Truncated SVD	Truncated SVD	End of layer group	L_2	Spatial dims
VID [8]	None	1×1 Conv	Middle layers	KL	None
IRG [18]	Instance graph	Instance graph	Middle layers	L_2	Spatial dims
RCO [60]	None	None	Teacher's train route	L_2	None
SP [61]	Similarity matrix	Similarity matrix	Middle layer	Frobenius norm	None
MEAL [62]	Adaptive pooling	Adaptive pooling	End of layer group	$L_{1/2}$/KL/L_{GAN}	None
Heo [62]	Margin ReLU	1×1 Conv	Pre-ReLU	Partial L_2	Negative features
AB [63]	Binarization	1×1 Conv	Pre-ReLU	Margin L_2	feature values
Chung [64]	None	None	End of layer	L_{GAN}	None
Wang [65]	None	Adaptation layer	Middle layer	Margin L_1	Channel + Spatial dims
KSANC [66]	Average pooling	Average pooling	Middle layers	L_2 + L_{GAN}	Spatial dims
Kulkarni [67]	None	None	End of layer group	L_2	None
IR [68]	Attention matrix	Attention matrix	Middle layers	KL+ Cosine	None
Liu [18]	Transform matrix	Transform matrix	Middle layers	KL	Spatial dims
NST [55]	None	None	Intermediate layers	MMD	None
Gao [69]	None	None	Intermediate layers	L_2	None

These KD methods also lost knowledge in the knowledge transfer

KD-Zero: Motivation

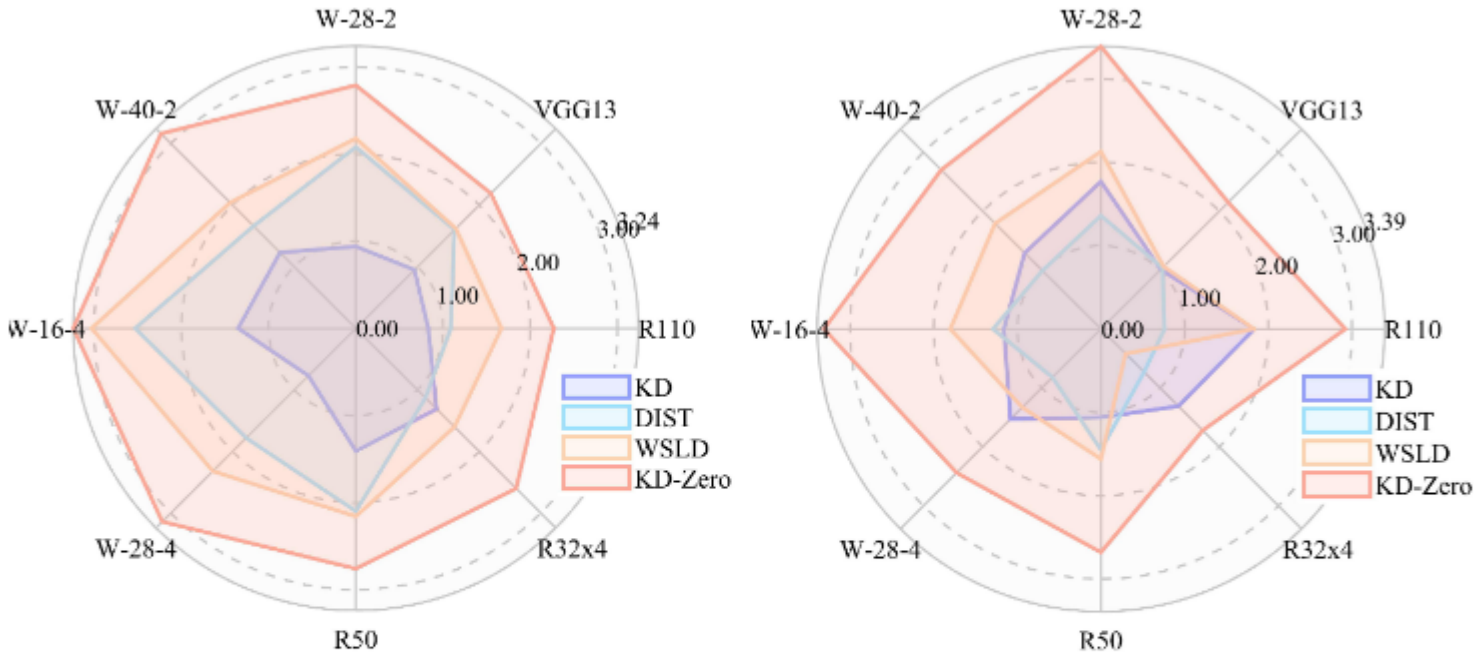
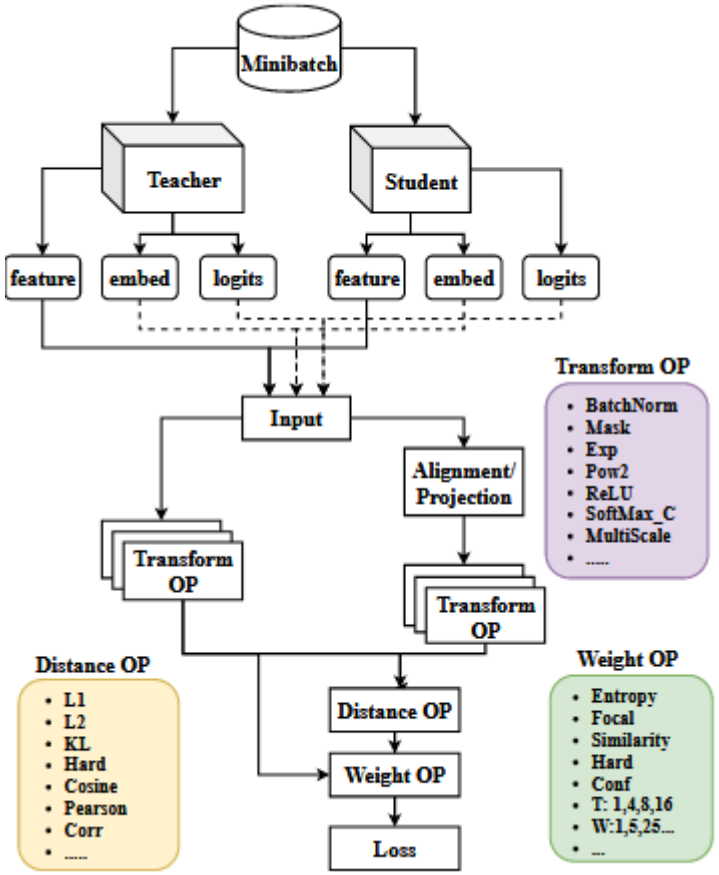


Figure 2: Top-1 gain (%). of WRN-16-2 (*left*) & ResNet-20 (*right*) distilled via various capacities teacher models on CIFAR-100.

However, we observed that KD sensitive with teacher-student architecture

KD-Zero: Our New framework



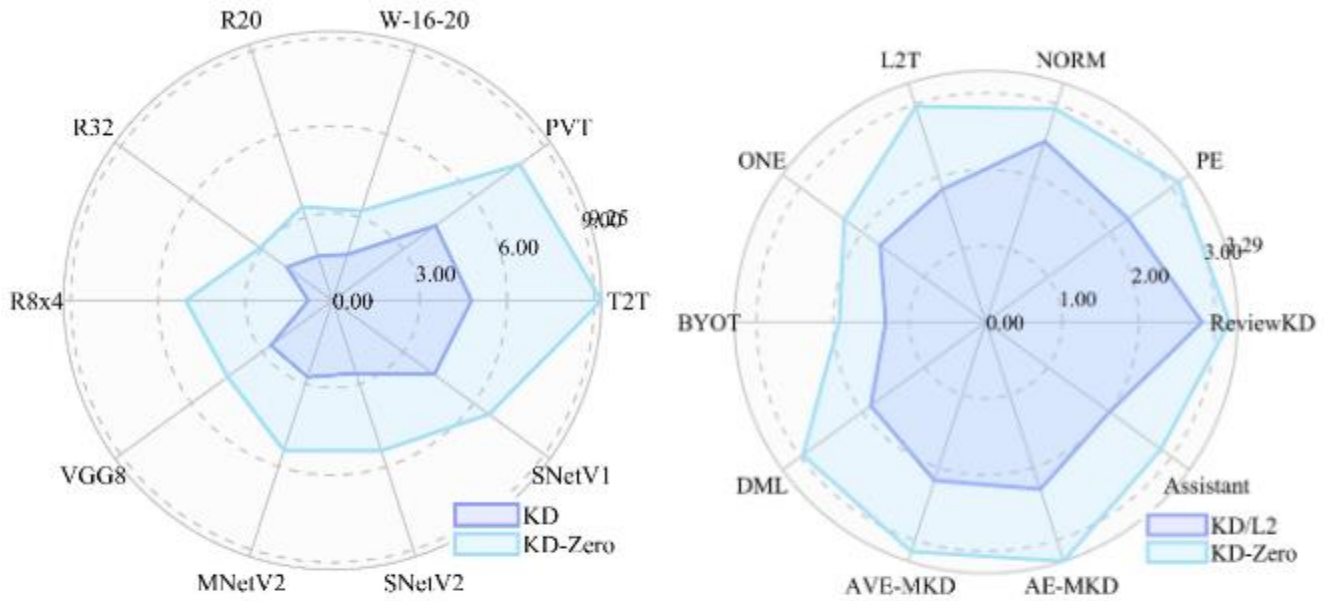
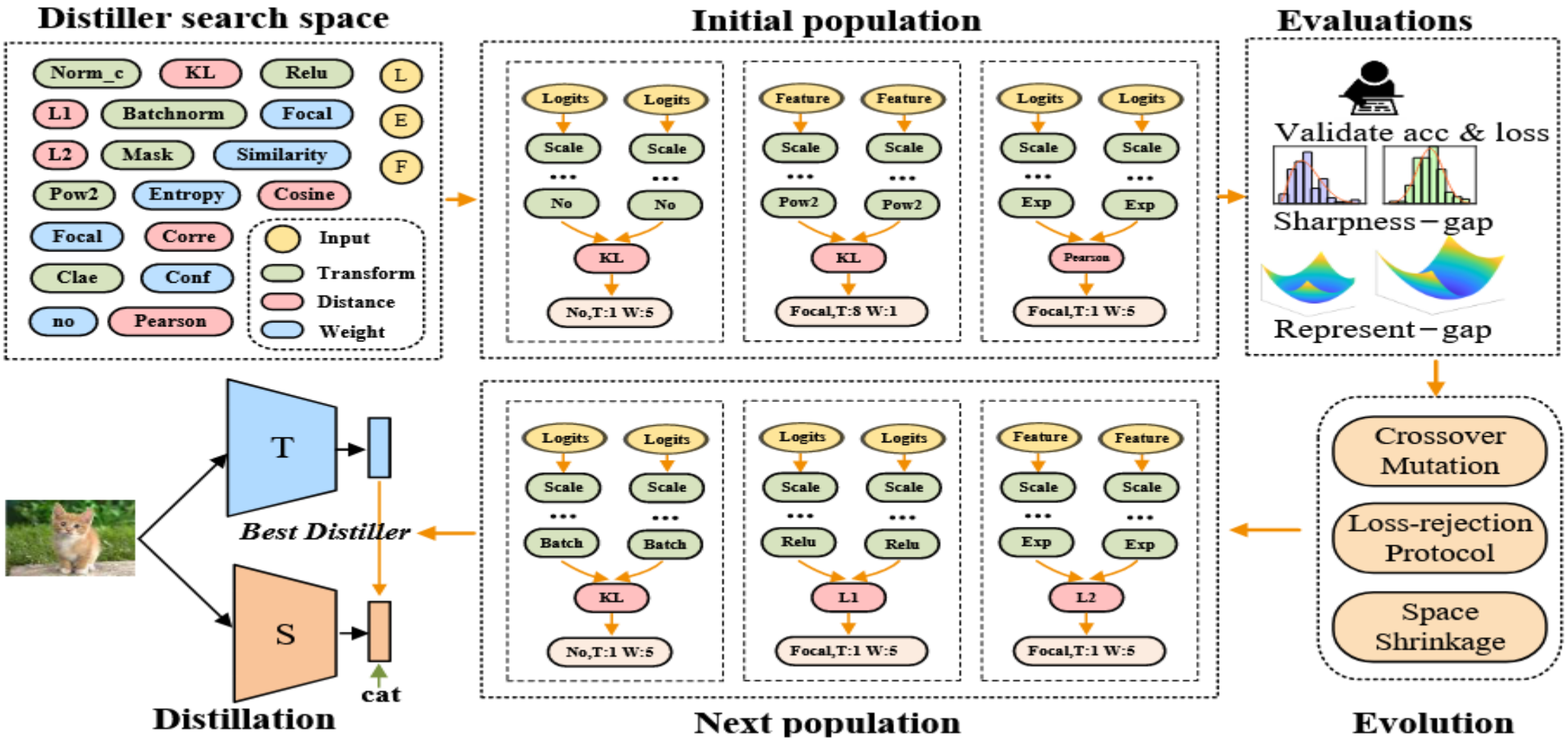


Figure 3: Top-1 gain (%). of various student with KD-Zero (*left*). Various KD methods combined with KD-Zero (*right*) for ResNet-20 on C-100.

We search for transforms, distances and weights for different models and methods

KD-Zero: Our New framework



We use evolutionary algorithms to crossover & mutation for best-performing distiller.

KD-Zero: Our New framework

Table 1: Specific operations in KD-Zero. More details of their formulas are available in the Appendix.

Transform	norm-based: $batchnorm, min-max, norm_{HW,C,N}, softmax_{HW,C,N}, logsoftmax_{HW,C,N}$ activation-based: $exp, mish, leaky, relu, tanh, sigmoid, pow2, pow4, log, sqrt$ scale-based: $scale, multi-scale, scale_{r1,r2}, local_{r1,r2,r4}, batch, channel$ attention&mask-based: $drop, satt, natt, catt, mask$, other: no, bmm, mm
Distance	no-norm loss: smooth ℓ_1, ℓ_1, ℓ_2, ℓ_{KL},, ℓ_{hard}; norm loss: $\ell_{Cosine}, \ell_{Pearson}, \ell_{Correlation}$
Weight	calibration: $entropy, focal, sim, conf, no$; weight values: 0.01,..., 100; τ values:1,4, 8

$$\alpha^* = \arg\min_{\alpha \in \mathcal{A}} (\mathcal{L}_{CE}(f_S, Y) + \overbrace{(log(\exp(f_S)) - log(\exp(f_T)))}^{Sharpness-gap} - \overbrace{\frac{\mathrm{HSIC}(f_S, f_T)}{\sqrt{\mathrm{HSIC}(f_S, f_S)\mathrm{HSIC}(f_T, f_T)}}}^{CKA-gap}), \quad (2)$$

We valuate their performance and sharpness & represent the gap between teacher-student.

KD-Zero: Results

Model	Same architectural style					Different architectural style				
Teacher	W-40-2	R110	R110	R32×4	VGG13	VGG13	R32×4	W-40-2	R56	R56
Student	W-16-2	R20	R32	R8×4	VGG8	MNetV2	SNetV2	SNetV1	DeiT	T2T
Teacher	75.61	74.31	74.31	79.42	74.64	74.64	79.42	75.61	72.34	72.34
Student	73.26	69.06	71.14	72.50	70.36	64.60	71.82	70.50	65.08	69.37
FitNets [58]	73.58	68.99	71.06	73.50	71.02	64.14	73.54	73.73	70.82	71.96
AT [74]	74.08	70.22	72.31	73.44	71.43	59.40	72.73	73.32	73.51	74.01
SP [63]	73.83	70.04	72.69	72.94	72.68	66.30	74.56	74.52	67.36	72.26
RKD [49]	73.35	69.25	71.82	71.90	71.48	64.52	73.21	72.21	70.39	71.88
CRD [61]	75.48	71.46	73.48	75.51	73.94	69.73	75.65	76.05	NA	NA
SRRL [30]	75.46	71.51	73.80	75.92	73.23	69.34	75.66	76.61	NA	NA
KD	74.92	70.67	73.08	73.33	72.98	67.37	74.45	74.83	73.25	74.15
DIST [28]	75.35	71.68	73.86	75.79	73.86	69.17	76.08	75.85	72.56	73.86
WSLD [80]	75.30	71.53	73.36	74.79	74.36	68.79	75.93	75.09	74.56	75.28
IPWD [48]	NA	71.32	73.91	76.03	NA	NA	NA	76.03	NA	NA
KD-Zero	**76.42**	**72.05**	**74.19**	**77.85**	**75.26**	**70.42**	**77.45**	**77.52**	**78.25**	**78.32**
Gain$_{\pm STD}$	3.16$_{\pm 0.16}$	2.99$_{\pm 0.21}$	3.05$_{\pm 0.12}$	5.35$_{\pm 0.22}$	4.90$_{\pm 0.11}$	5.82$_{\pm 0.18}$	5.63$_{\pm 0.25}$	7.02 $_{\pm 0.14}$	13.17$_{\pm 0.36}$	8.95$_{\pm 0.29}$

On Cifar, our method achieves SOTA results on both features and logits distillation

KD-Zero: Results

Teacher	Student	Acc.	Teacher	Student	KD [24]	AT [74]	OFD [23]	SRRL [30]	CRD [61]	Review [51]	MGD [72]	KD-Zero
ResNet-34	ResNet-18	Top-1	73.40	69.75	70.66	70.69	70.81	71.73	71.17	71.61	71.58	**72.17**$\pm$0.15
		Top-5	91.42	89.07	89.88	90.01	89.98	90.60	90.13	90.51	90.35	**90.46**$\pm$0.25
ResNet-50	MobileNet	Top-1	76.16	70.13	70.68	70.72	71.25	72.49	71.37	72.56	72.35	**73.02**$\pm$0.22
		Top-5	92.86	89.49	90.30	90.03	90.34	90.92	90.41	91.00	90.71	**91.05**$\pm$0.26

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage detectors						
T: Cascade Mask RCNN-X101	45.6	64.1	49.7	26.2	49.6	60.0
S: Faster RCNN-R50	38.4	59.0	42.0	21.5	42.1	50.3
KD [25]	39.7	61.2	43.0	23.2	43.3	51.7
FKD [75]	41.5	62.2	45.1	23.5	45.0	55.3
CWD [60]	41.7	62.0	45.5	23.3	45.5	55.5
DIST [28]	40.4	61.7	43.8	23.9	44.6	52.6
KD-Zero	41.9	62.7	45.5	23.6	45.6	55.6
One-stage detectors						
T: RetinaNet-X101	41.0	60.9	44.0	23.9	45.2	54.0
S: RetinaNet-R50	37.4	56.7	39.6	20.0	40.7	49.7
KD [25]*	37.2	56.5	39.3	20.4	40.4	49.5
FKD [75]	39.6	58.8	42.1	22.7	43.3	52.5
FGD [71]	40.4	59.9	43.3	23.4	44.7	54.1
DIST [28]	39.8	59.5	42.5	22.0	43.7	53.0
KD-Zero	40.9	60.4	43.5	23.2	45.2	54.8

Method	mIoU (%)
T: DeepLabV3-R101	78.07
S: DeepLabV3-R18	74.21
SKD [45]	75.42
IFVD [66]	75.59
CWD [60]	75.55
CIRKD [69]	76.38
DIST [28]	77.10
KD-Zero	77.38
S: PSPNet-R18	72.55
SKD [45]	73.29
IFVD [66]	73.71
CWD [60]	74.36
CIRKD [69]	74.73
KD-Zero	76.25

Moreover, our method can scale to large-scale datasets and downstream tasks like detection, segmentation

KD-Zero: Results

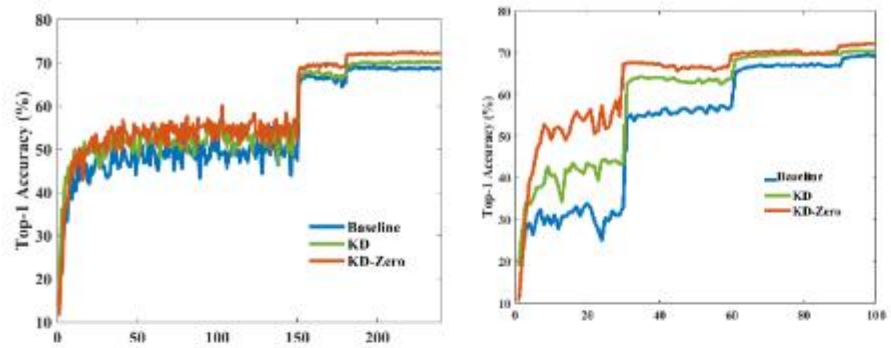


Figure 6: Comparison of training curves (*left*) of ResNet-20 on CIFAR-100 and ResNet-18 on ImageNet (*right*).

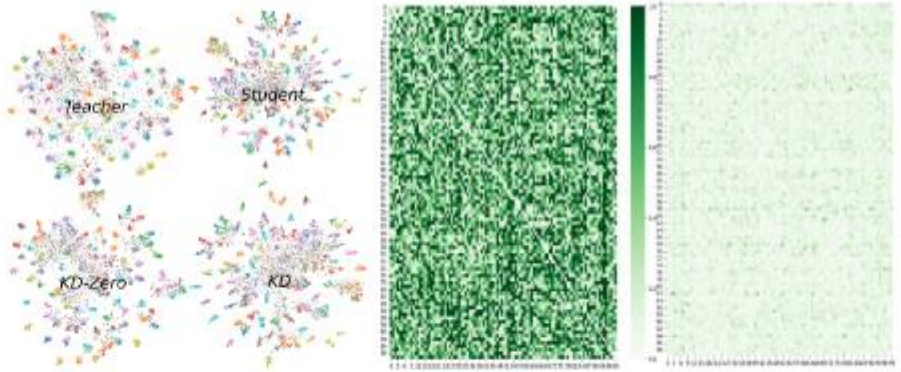


Figure 7: Penultimate-layer visualization (*left*), logits-correlation map of teacher-student (ResNet-110/20) via KD (*middle*) & KD-Zero (*right*).

Figure 8: Comparison on the Grad-CAM++ [4] visualization results between the features of the baseline student model, the student model trained with KD and our KD-Zero, and the teacher model. Results are obtained on ImageNet with ResNet18 (left) and MobileNet (right).

Thanks for listening!

That's all. Thanks for listening.