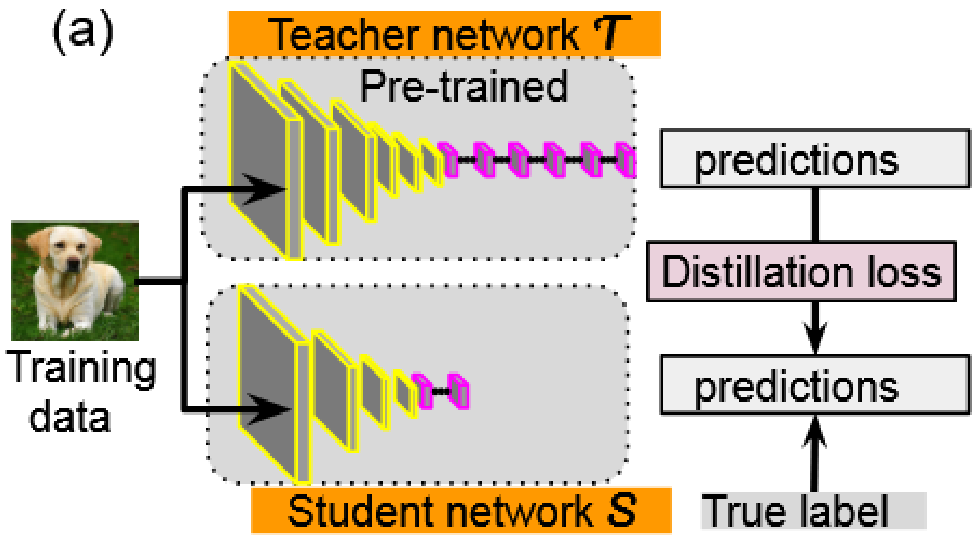
DisWOT: Student Architecture Search for Distillation WithOut Training

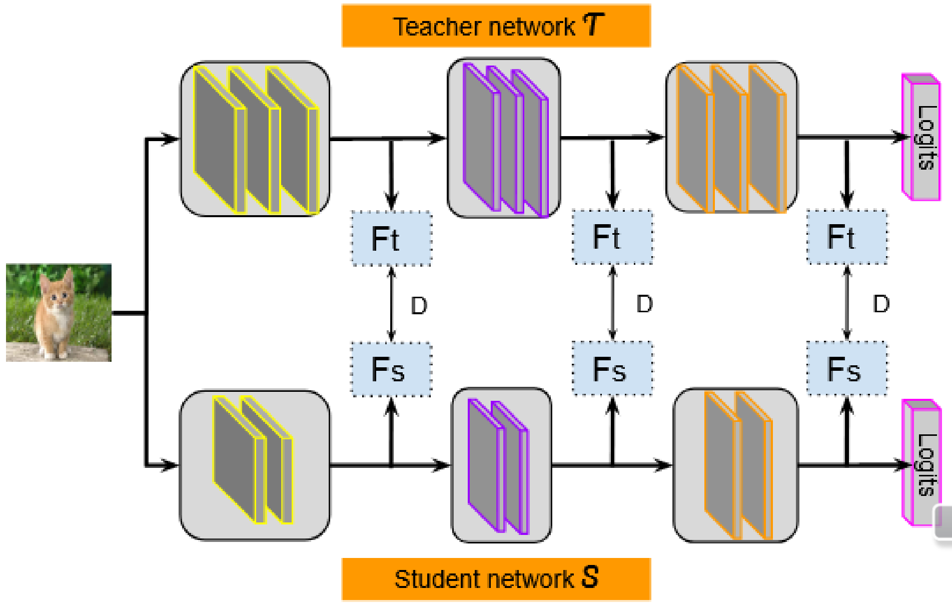
Peijie Dong1† Lujun Li2†* Zimian Wei1†

1 National University of Defense Technology, 2 Chinese Academy of Sciences

DisWOT: Background

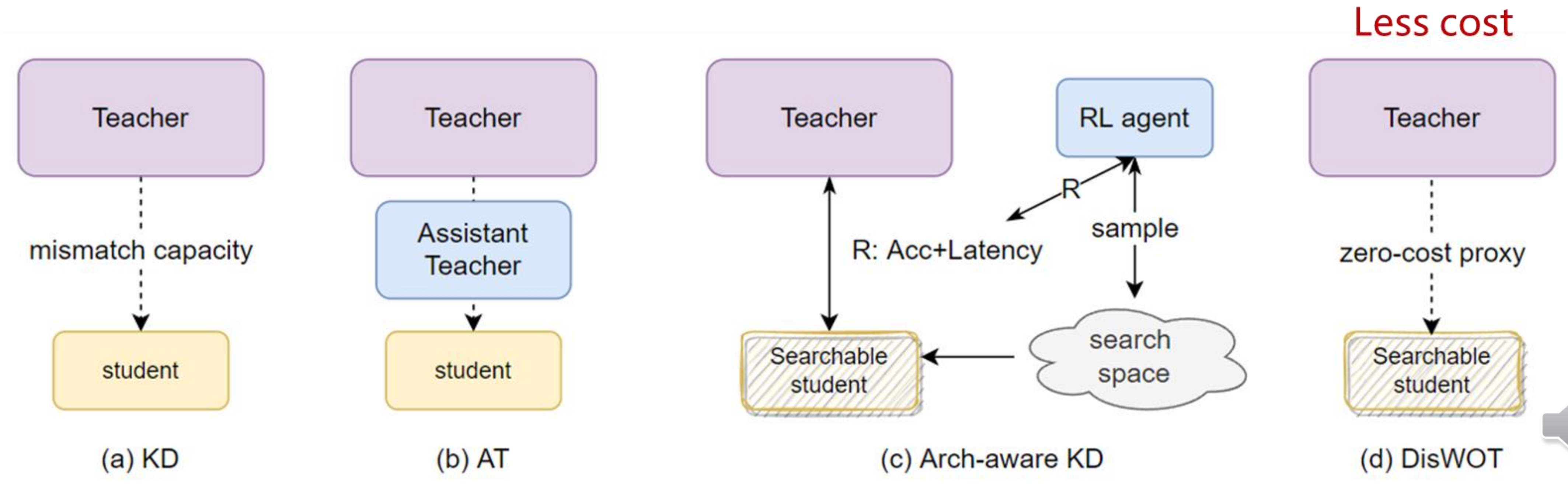
- Knowledge distillation is effective training strategy using the logits, feature of the teacher model.





DisWOT: Background

- Distillation gap: bigger model is NOT the better teacher model
- Some studies tackling this issue bring lots of extra training-costs



DisWOT: Methodology

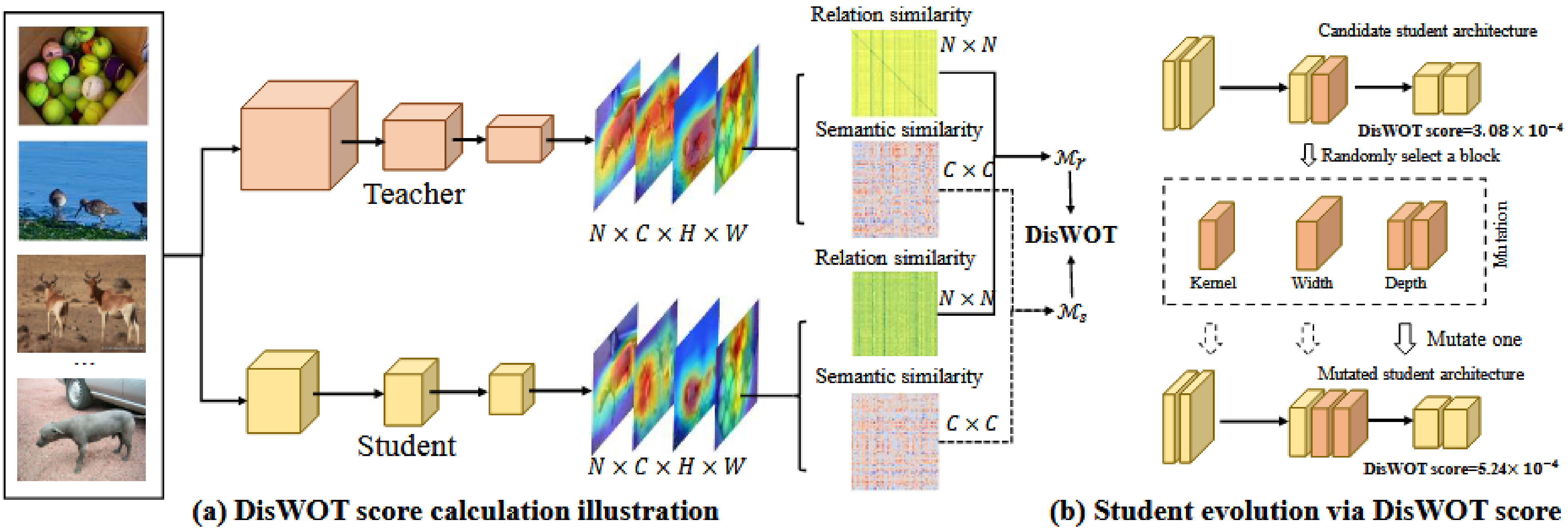


Figure 3. A schematic overview of our DisWOT, including (a) detailed calculation of the DisWOT scores and (b) evolution of the student architecture via the DisWOT scores. In search phase, DisWOT use semantic similarity metrics and relations similarity metrics to select good student for a given teacher. The semantic similarity metric is measured by l_2 distance of the channel-wise correlation matrix for Grad-cam activation maps. Similarly, the relation similarity matrix statistics the sample-wise correlation matrix distance of the randomly initialized teacher-student pairs. With the feedback from these metrics, the evolutionary search in DisWOT automatically imitates good student from weak ones. In distillation phase, this searched student is distilled via teacher model and achieves superior gains.

Dong, Peijie and Li, Lujun and Wei, Zimian, DisWOT: Student Architecture Search for Distillation WithOut Training, CVPR2023

DisWOT: Methodology

- Semantic Similarity Metric： Correlation matrix on Grad-CAM Maps

$$\mathcal{G}^T = \frac{(G_T)\cdot(G_T)^\top}{\|(G_T)\cdot(G_T)^\top\|_2}, \mathcal{G}^S = \frac{(G_S)\cdot(G_S)^\top}{\|(G_S)\cdot(G_S)^\top\|_2}$$

$$\mathcal{M}_s = \left\|\mathcal{G}^T - \mathcal{G}^{S_i}\right\|_2$$

- Relation Similarity Metric: Correlation matrix on Simple Relation

$$\mathcal{A}^T = \frac{(\tilde{A}_T)\cdot(\tilde{A}_T)^\top}{\left\|(\tilde{A}_T)\cdot(\tilde{A}_T^\top)\right\|_2}, \mathcal{A}^{S_i} = \frac{(\tilde{A}_S)\cdot(\tilde{A}_S)^\top}{\left\|(\tilde{A}_S)\cdot(\tilde{A}_S^\top)\right\|_2};$$

$$\mathcal{M}_r = \left\|\mathcal{A}^T - \mathcal{A}^{S_i}\right\|_2$$

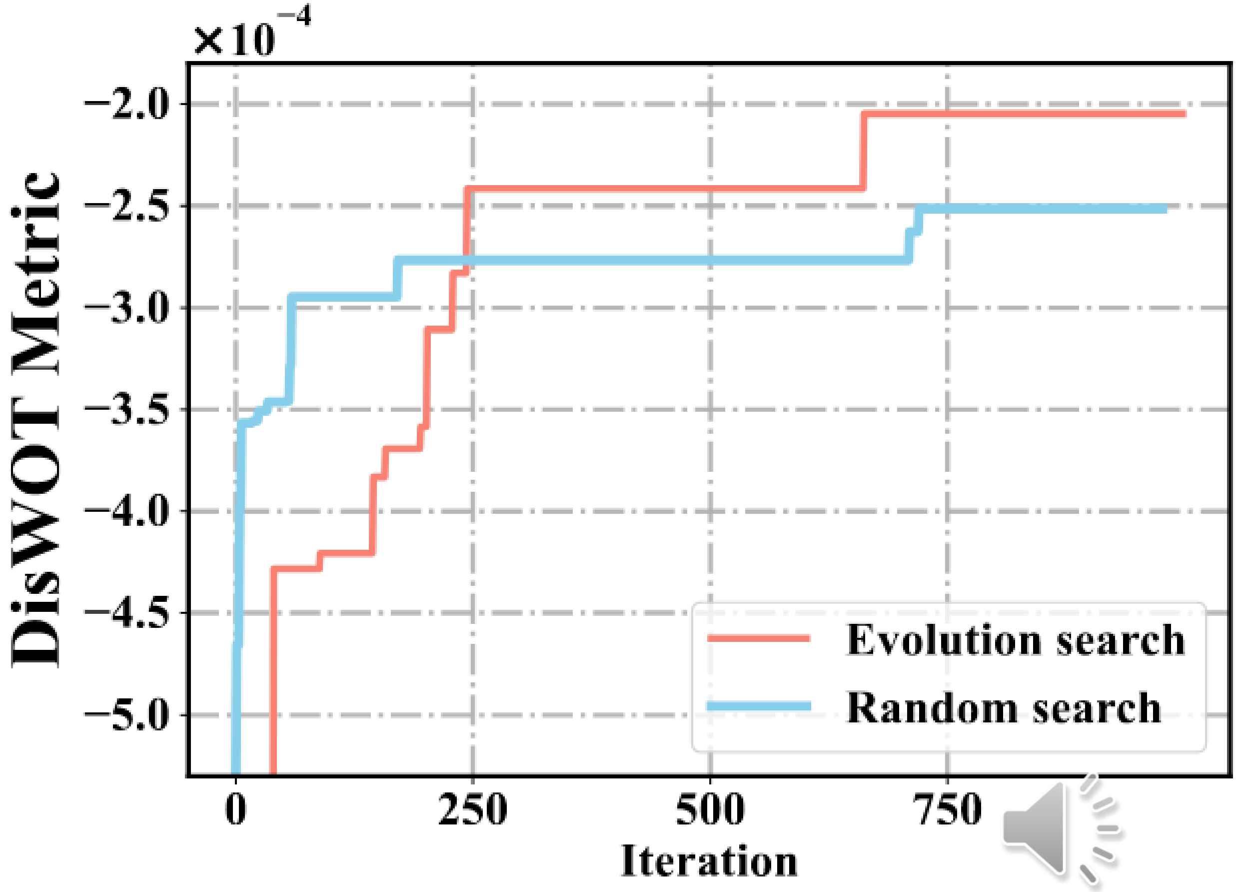
Dong, Peijie and Li, Lujun and Wei, Zimian, DisWOT: Student Architecture Search for Distillation WithOut Training, CVPR2023

DisWOT: Methodology

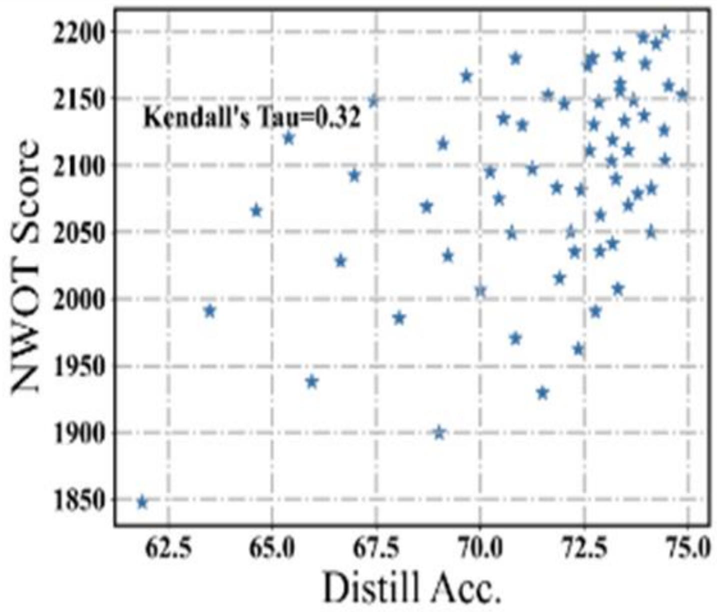
Algorithm 1 Evolution Search for DisWOT

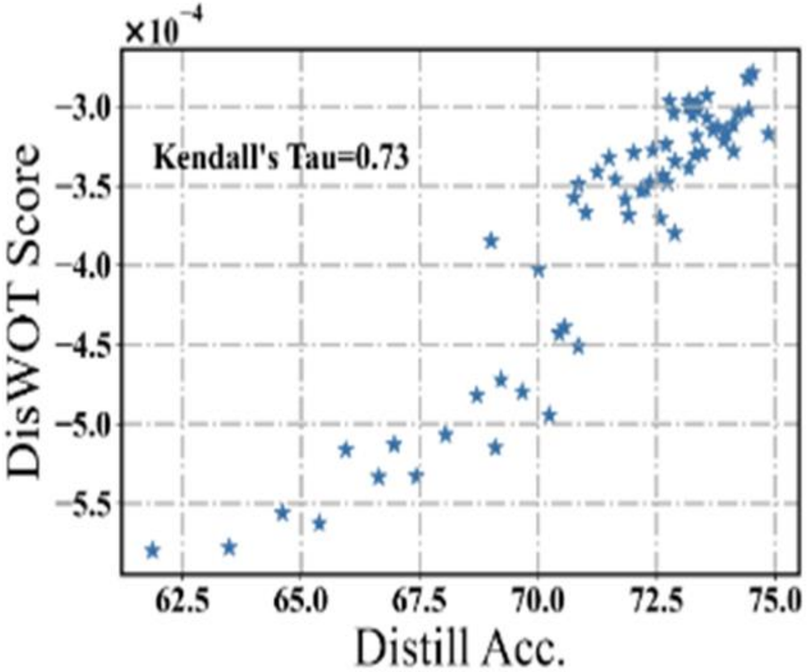
Input: Search space $\mathcal{S}$, population $\mathcal{P}$, architecture constraints $\mathcal{C}$, max iteration $\mathcal{N}$, sample ratio r, sampled pool $\mathcal{Q}$, topk k, teacher network $\mathcal{T}$.
Output: Highest DisWOT score architecture.

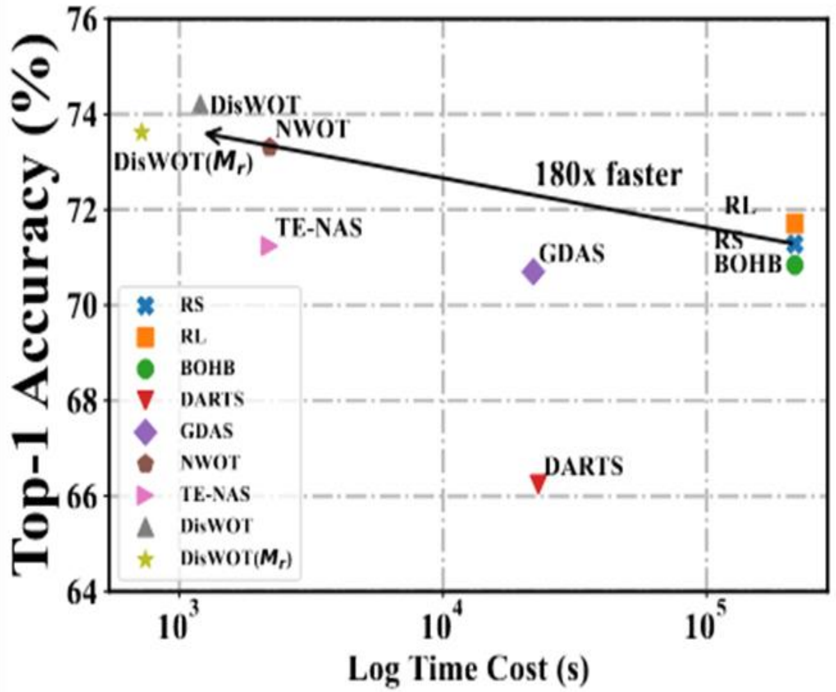
$\mathcal{P}_0$:= Initialize population$(P_i, \mathcal{C})$;
sample pool $\mathcal{Q} := \emptyset$;
for $i = 1 : \mathcal{N}$ **do**
    Clear sample pool $\mathcal{Q} := \emptyset$;
    Randomly select $r \times \mathcal{P}$ subnets $\hat{P}_i \in \mathcal{P}$ to get $\mathcal{Q}$;
    Candidates $\{A_i\}_k$:= GetTopk($\mathcal{Q}$, k);
    Parent A_i := RandomSelect($\{A_i\}_k$);
    Mutate $\hat{P}_i$:= MUTATE(A_i);
    if $\hat{P}_i$ do not meet the constraints $\mathcal{C}$ **then**
        Do nothing;
    else
        Get DisWOT-Score z := DisWOT($\hat{P}_i$, $\mathcal{T}$);
        Append $\hat{P}_i$ to $\mathcal{P}$;
    end if
    Remove network of smallest DisWOT-score;
end for



DisWOT: Methodology







DisWOT: Methodology

- DisWOT+: Distillation with Semantic Similarity & Relation Similarity knowledge
- Bridging KD losses and Zero-cost proxies

$$\mathcal{L}_{\mathcal{M}_s} = \frac{1}{c^2}\left\|\mathcal{G}^T - \mathcal{G}^S\right\|_2, \mathcal{L}_{\mathcal{M}_r} = \frac{1}{b^2}\left\|\mathcal{A}^T - \mathcal{A}^S\right\|_2$$

$$\mathcal{L}_{\text{DisWOT}} = \mathcal{L}_{CE}(f_S, Y) + \mathcal{L}_{KL}(f_S, f_T)$$

$$\mathcal{L}_{\text{DisWOT}\dagger} = \mathcal{L}_{\text{DisWOT}} + \mathcal{L}_{\mathcal{M}_s} + \mathcal{L}_{\mathcal{M}_r}.$$

Type	Method	ρ	Method	ρ
Zero-cost Proxies	Grad_Norm [1]	58.70%±0.11	Synflow [66]	**74.61%±0.08**
	SNIP [35]	58.17%±0.15	Jacob [64]	73.42%±0.03
	Fisher [1]	35.91%±0.09	Zen-NAS [38]	41.36%±0.06
	NWOT [47]	64.41%±0.08	FLOPs [1]	63.38%±0.06
KD-based Proxies	KD [27]	54.43%±0.09	PKT [53]	52.65%±0.09
	FitNet [61]	56.18%±0.09	CC [54]	65.90%±0.08
	SP [70]	51.24%±0.08	NST [31]	72.35%±0.09
	RKD [50]	25.71%±0.17	DisWOT(ours)	**72.36%±0.02**

DisWOT: Experiments

Table 8: Distillation results on CIFAR-10, CIFAR-100, and ImageNet-16. The proposed approach, DisWOT, achieves competitive results with the lowest costs. The results of NWOT and TE-NAS come from their original papers.

Type	Model	CIFAR-10			CIFAR-100			ImageNet-16-120		
		Dis. Acc(%)	Time (s)	Speed-up	Dis.Acc(%)	Time (s)	Speed-up	Dis. Acc(%)	Time (s)	Speed-up
Multi-trial	RS	93.63	216K	1.0×	71.28	460K	1.0×	44.88	1M	1.0×
	RL [4]	92.83	216K	1.0×	71.71	460K	1.0×	44.35	1M	1.0×
	BOHB [18]	93.49	216K	1.0×	70.84	460K	1.0×	44.33	1M	1.0×
	RSPS [37]	91.67	10K	21.6×	57.99	46K	21.6×	36.87	104K	9.6×
Weight-sharing	GDAS [16]	93.39	22K	12.0×	70.70	39K	11.7×	42.35	130K	7.7×
	DARTS [40]	89.22	23K	9.4×	66.24	80K	5.8×	43.18	110K	9.1×
Training-free	NWOT [47]	93.73	2.2K	100×	73.31	4.6K	100×	45.43	10K	100×
	TE-NAS [11]	93.92	2.2K	100×	71.24	4.6K	100×	44.38	10K	100×
DisWOT	$\mathcal{M}_s$ & $\mathcal{M}_r$	93.55	1.2K	180×	74.21	9.2K	180×	47.30	20K	180×
	$\mathcal{M}_r$	93.49	0.72K	**300×**	73.62	18.4K	**300×**	45.63	40K	**300×**

Table 9: Top-1 accuracy of ResNet18 w.r.t. various teachers on ImageNet-1k. Different from the baseline model, our the method shows better performance and improves students' performance positively correlated with that of the teacher.

Teacher	Student	Acc	Teacher	Student	KD [27]	ESKD [12]	ATKD [49]	ONE [34]	DML [80]	CRD [68]	DisWOT
ResNet34	ResNet18	Top-1	73.40	69.75	70.66	70.89	70.78	70.55	71.03	71.17	**72.08**
		Top-5	91.42	89.07	89.88	90.06	89.99	89.59	90.28	90.32	**90.38**
Teacher	Student	Acc	Teacher	Student	KD [27]	ATKD [49]	Review [10]	OFD [25]	DML [80]	CRD [67]	DisWOT
ResNet50	ResNet18	Top-1	76.16	69.75	70.68	70.72	71.32	71.25	71.13	71.20	**72.30**
		Top-5	92.86	89.07	90.30	90.03	90.62	90.34	90.22	90.22	**90.51**

DisWOT: Experiments

- Correlation comparison

Method	Kendall's Tau	Spearman	Pearson
FLOPs [1]	51.61	72.92	76.40
Fisher [1]	62.86	81.37	20.90
Grad_Norm [1]	63.75	82.35	39.35
SNIP [35]	67.22	85.07	51.09
NWOT [47]	31.87	45.66	48.99
DisWOT (ours)	**73.98**	**91.38**	**84.83**

- Different metrics

Knowledge	Metric	Spearman (%)
$\mathcal{M}_s$	FitNet [60]	64.06±6.11
$\mathcal{M}_s$	Similarity matrix	73.68±5.45
$\mathcal{M}_r$	RKD [67]	13.52±11.51
$\mathcal{M}_r$	Similarity matrix	72.36±3.42
$\mathcal{M}_s$ & $\mathcal{M}_r$	Similarity matrix	**77.51±2.76**

- Different Initialization

Method	Kaiming	Gaussian
DisWOT ($\mathcal{M}_s$)	82.96	77.24
DisWOT ($\mathcal{M}_r$)	29.92	49.38
DisWOT	36.55	77.51

Thanks