A APPENDIX

A.1 ROBUST TEACHER LAYERS

In this section, we discuss robustness inducing capacity of teacher layers. We hypothesize that few teacher layers are more robust than others and thus should induce more robustness to the student models. In RNAS-CL, each student layer is associated with a teacher layer. Figures 6 and 8 plot the number of student layers connected to each robust teacher layer on the CIFAR-10 and ImageNet-100 datasets. For all student models on CIFAR-10, we observe that layers 15 and 21 of the robust teacher model have significantly more intermediate connections with the student models. Similarly, for ImageNet-100, layers 18, 32, and 40 are a few of the dominant robust layers. In Figures 7 and 9, we visualize the most robust teacher layers on CIFAR-10 and ImageNet, respectively.

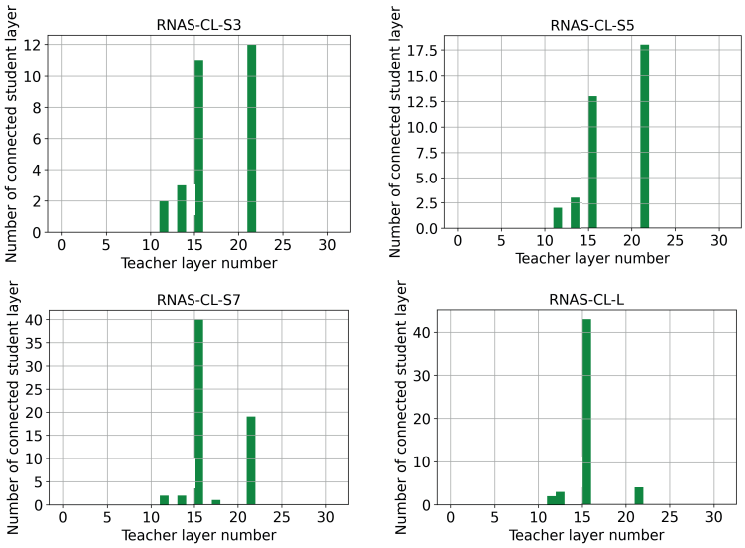

Figure 6: Illustrations of the number of student layers connected to each teacher layer in RNAS-CL for various student models on the CIFAR-10 dataset. We choose adversarially trained Wide-ResNet-34 as the robust teacher model for all the four student models, with one plot for each student model. All student architectures are described in Table 7.

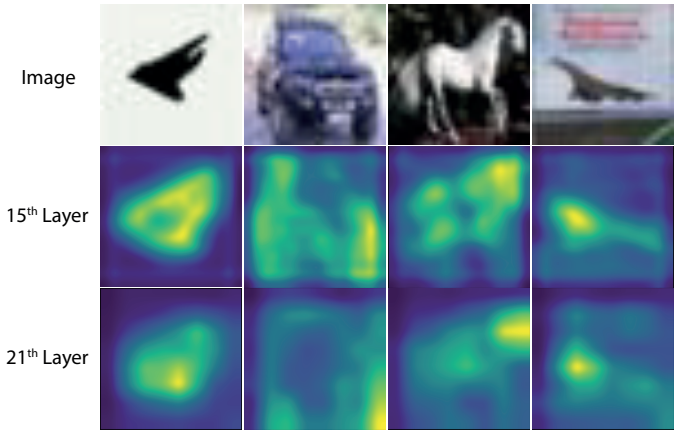

Figure 7: Attention map for most robust teacher layers on CIFAR-10 dataset. We chose the same robust teacher model as in Figure 6. The illustrated layers represent teacher layers with maximum number of intermediate connection for various RNAS-CL models (as described in Figure 6).

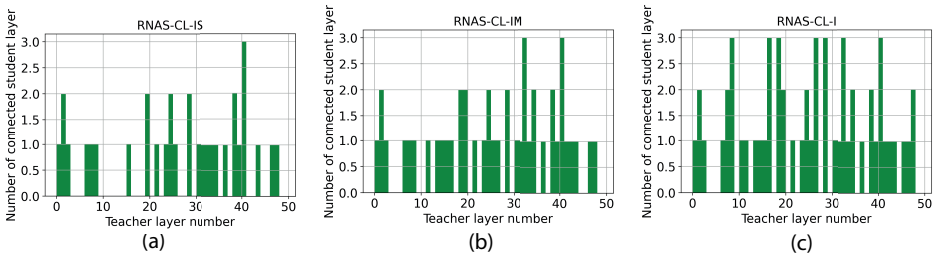

Figure 8: Illustrations of the number of student layers connected to each teacher layer in RNAS-CL for various student models on the ImageNet-100 dataset. We choose adversarially trained Wide-ResNet-50 as the robust teacher for all and three students models, with one plot for each student model. All RNAS-CL architectures are described in Table 8.

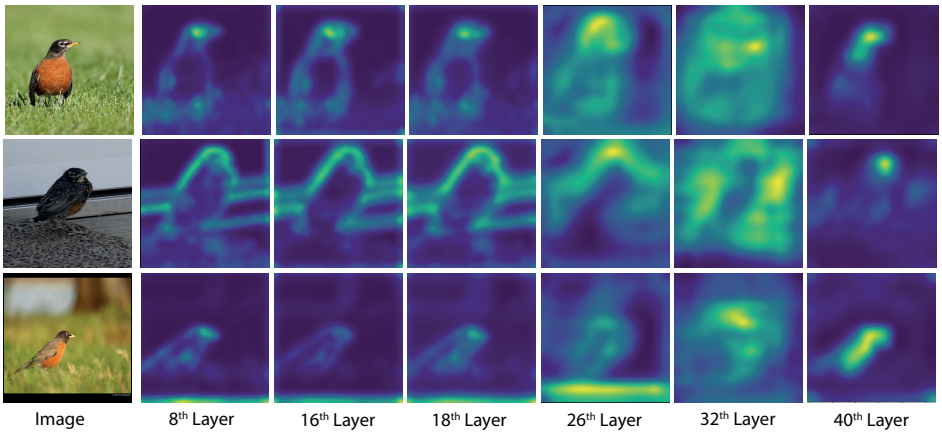

Figure 9: Attention maps for most robust teacher layers on ImageNet-100 dataset. We chose the same robust teacher model as in Figure 8. The illustrated layers represent teacher layers with maximum number of intermediate connection for various RNAS-CL models (as described in Figure 8).

A.2 More results on CIFAR-100

In this section, we conduct experiments using adversarially trained WRT-34 (Rice et al., 2020), ResNet-50 (Engstrom et al., 2019), and ResNet-18 (Sehwag et al., 2021) as the robust teacher models on the CIFAR-10 dataset. All RNAS-Cl models, while achieving similar clean accuracy, exceed its counterpart by more than 10% in PGD accuracy. RNAS-CL-R50 achieves higher robust accuracy than RNAS-CL-R18 and RNAS-CL-WRT-34. However, ResNet-50 has the lowest PGD accuracy among the teacher models, suggesting that the teacher's architecture has more influence on the student's performance than the teacher's performance. The higher number of teacher layers allows more options for the student layer to learn from, leading to better robustness. The teacher models' performance is reported in Table 6.

A.3 Compare Efficient and Robust ImageNet-100 models

We compare RNAS-CL to adversarially robust pruning methods on ImageNet-100 dataset, with results shown in Table 3. RNAS-CL models are trained with three different robust teachers, ResNet-18, ResNet-50, and WideResNet-50, with the ImageNet pre-trained (Engstrom et al., 2019) being the robust teacher. It is observed that RNAS-CL models consistently exceed other models by $\sim 25\%$ in terms of clean accuracy while exibiting adversarial robustness. In Table 3, both Hydra and LWM were adversarially trained using TRADES (Zhang et al., 2019a). For a fair comparison, after the regular training stage without TRADES, we retrain our RNAS-CL models with the TRADES optimization objective. We replace the cross-entropy term in (3) by the TRADES optimization

Method	Clean	PGD^{20}
Standard-S3	89.92	17.69
Standard-S5	90.76	18.44
Standard-S7	90.98	19.3
RNAS-CL-S3-WRT-34	89.4	34.3
RNAS-CL-S5-WRT-34	90.4	35.59
RNAS-CL-S7-WRT-34	90.62	37.24
RNAS-CL-S3-R50	89.39	35.76
RNAS-CL-S5-R50	90.53	37.32
RNAS-CL-S7-R50	90.41	37.98
RNAS-CL-S3-R18	88.47	26.35
RNAS-CL-S5-R18	88.77	25.49
RNAS-CL-S7-R18	89.47	27.96

Table 2: Performance of RNAS-CL method trained with various robust teacher models on the CIFAR-10 dataset. Standard represents models searched and trained by cross-entropy loss without any teacher model.

Method	Clean	PGD^{20}	# Params (M)	MACs (M)
Hydra (ResNet-18) - 90% (Sehwag et al., 2020)	59.96	29.79	**1.1**	1200
LWM (ResNet-18) - 90% (Han et al., 2015)	59.02	27.67	1.1	1200
RNAS-CL-I-R-18	85.22	8.3	3.94	**241.98**
RNAS-CL-I-R-50	**85.98**	5.08	3.96	244.76
RNAS-CL-I-WRT-50	85.46	3.36	4.01	255.37
RNAS-CL-I-R-18 + TRADES	78.94	29.02	3.94	**241.98**
RNAS-CL-I-R-50 + TRADES	79.95	**32.44**	3.96	244.76
RNAS-CL-I-WRT-50 + TRADES	79.42	28.06	4.01	255.37

Table 3: Performance of various efficient and robust methods on ImageNet-100 dataset. Clean Acc and Adv Acc are the same as that in Table 1. All MACs were calculated without special hardware (Han et al., 2016) or special software (Park et al., 2017)

objective. With such training, RNAS-CL achieves similar or higher adversarial accuracy while significantly outperforming Hydra and LWM in clean accuracy with only a fraction of MACs.

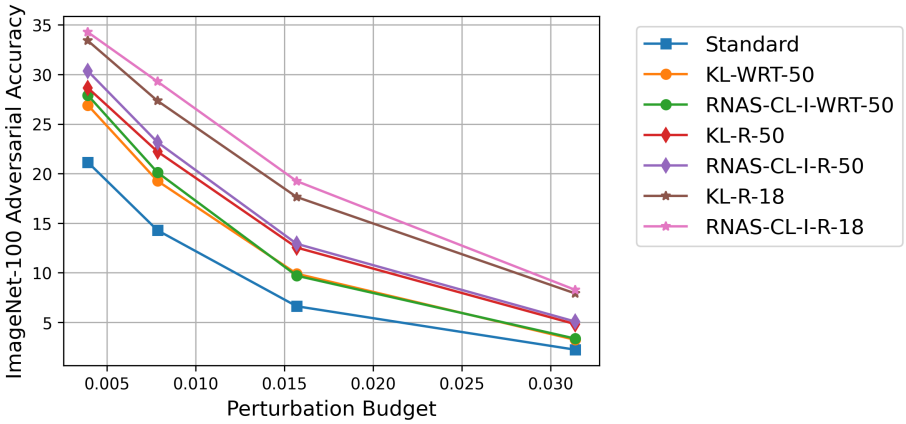

Figure 10: Adversarial accuracy of various models at various perturbation budgets on the ImageNet-100 dataset.

We further study adversarial accuracy at various perturbation budgets for three different teacher models. As illustrated in Figure 10, RNAS-CL exceeds its counterpart in adversarial accuracy at various perturbation budgets for all teacher models on the ImageNet-100 dataset. This demonstrates the significance of cross-layer connections in RNAS-CL.

A.4 COMPARE CIFAR-10 MODEL AGAINST CW AND AUTOATTACK

In this section, we compare RNAS-CL and (Huang et al., 2021) against recent attacks such as CW_∞ (Carlini & Wagner, 2017) and AutoAttack (Croce & Hein, 2020) on CIFAR-10 dataset. CW attacks were proposed to defeat defensive distillation. In Table 4, we use L_∞ version of CW attack optimized by PGD, with maximum perturbation budget set to $\epsilon = 8/255$. AutoAttack is a parameter-free ensemble attack currently considered one of the most reliable and widely acknowledged evaluation benchmark in Adversarial Defences.

Method	CW_∞	AA
VGG-R (Huang et al., 2021)	46.49	38.44
DN-121-R (Huang et al., 2021)	53.07	47.75
RNAS-CL-S3-WRT-34(Our)	47.07	37.17
RNAS-CL-S5-WRT-34(Our)	48.33	39.28
RNAS-CL-S7-WRT-34(Our)	47.91	38.36
RNAS-CL-M-WRT-34(Our)	**53.52**	46.89
RNAS-CL-L-WRT-34(Our)	52.63	**48.49**

Table 4: The table compared performance of (Huang et al., 2021) and RNAS-CL against CW_∞ (Carlini & Wagner, 2017) and AutoAttack (Croce & Hein, 2020) on CIFAR-10 dataset.

A.5 COMPARISON AGAINST KD VARIENTS

In this section, we compare our methods against various knowledge distillation methods Park et al. (2019); Ahn et al. (2019); Tung & Mori (2019); Tian et al. (2020b); Passalis & Tefas (2018). We use Robust WRT-34 as the teacher model for all KD methods and train three different student architectures: RNAS-CL-S3, RNAS-CL-S5, and RNAS-CL-S7. In Figure 11, models trained using our paradigm are explicitly on the upper right-most part of the graph. RNAS-CL-S3 architecture trained using RKD performs similarly to the model trained using our methods. Apart from this, all models trained using RNAS-Cl significantly outperform all other methods in terms of clean and adversarial accuracy.

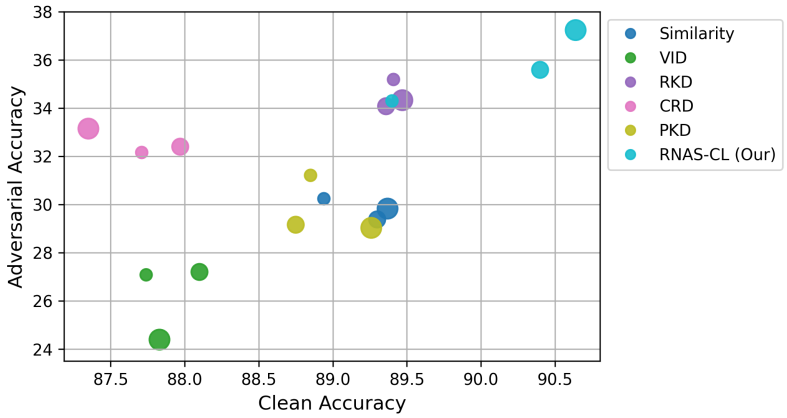

Figure 11: The figure compares various knowledge distillation variants (Similarity (Tung & Mori, 2019), VID (Ahn et al., 2019), RKD (Park et al., 2019), CRD (Tian et al., 2020b), PKD (Passalis & Tefas, 2018)) against RNAS-CL on the CIFAR-10 dataset. Adversarial Accuracy represents top-1 Accuracy on images perturbed by 20 step PGD attack. Clean Accuracy represents top-1 Accuracy on clean images. Larger marker size indicates larger architecture. For each method, RNAS-CL-S3, RNAS-CL-S5, and RNAS-CL-S7 are represented by increasing marker size.

A.6 RESULTS FOR IMAGENET

In this section, we compare our method on the ImageNet dataset. Standard represents the model searched and trained using cross-entropy loss without any teacher model. RNAS-Cl represents the

model trained using our training paradigm. Both models are further adversarially trained using FastAT (Wong et al., 2020). In Table 5, we evaluate the robustness against 10 step PGD attack with $\epsilon = 4/255$. Models trained with RNAS-CL exceed both in terms of clean and robust accuracy.

Method	Clean	PGD^{10}
Standard	53.92	25.45
RNAS-CL-WRT-50	56.1	29.78

Table 5: Robustness results on ImageNet dataset.

A.7 Robust teacher models

In this section, we report the robustness of adversarially trained teacher model used throughout the paper on CIFAR-10 dataset.

Model	Clean	PGD^{20}
WRT-34	86.07	58.33
ResNet 18	84.59	55.54
ResNet 50	87.03	49.25

Table 6: Robustness results for various teacher model on CIFAR-10 dataset.

A.8 Architecture

In this section, we discuss architectures for various proposed super-nets used in RNAS-CL for CIFAR-10 and ImageNet-100 datasets. Table 7 describes the super-nets used for CIFAR-10. We use super-nets with three blocks. Super-nets used for ImageNet-100 are described in Table 8. For ImageNet-100, the number of blocks varies from 3 to 5.

Search Space for CIFAR-10				
Search Space	Depth	Stage 1	Stage 2	Stage 3
RNAS-CL-S3	3-3-3	16, 12	32, 28, 24, 20	64, 60, 56, 52
RNAS-CL-S5	5-5-5	16, 12	32, 28, 24, 20	64, 60, 56, 52
RNAS-CL-S7	7-7-7	16, 12	32, 28, 24, 20	64, 60, 56, 52
RNAS-CL-M	9-7-1	80, 76	160, 156, 152, 148	128, 124, 120, 116
RNAS-CL-L	9-7-1	160, 156	320, 316, 312, 308	256, 252, 248, 244

Table 7: The table describes the search space for CIFAR-10. Depth represents the depth of each stage. For example, 3-3-3 represents three convolution blocks in each stage. All search spaces have three stages. Stage 1, Stage 2, and Stage 3 represent the filter choices for their respective stages. For example, at stage 3 of RNAS-CL-S3, for each convolution block, we search between 4 output channels (64, 60, 56, 52).

A.9 Architecture Search by FBNetV2

RNAS-CL builds both an efficient and adversarially robust deep learning model. In this work, we use the training paradigm of FBNetV2 to search for efficient models. In Figure 12, we illustrate the searching process for neural architecture at a single convolution layer. Each filter choice is attached with a Gumbel weight. These Gumbel weights are optimized to select an efficient model.

Search Space for ImageNet-100						
Search Space	Depth	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
RNAS-CL-IS	3-3-3	28, 24, 20, 16	40, 36, 32, 28	96, 88, 80, 72, 64, 56, 48		
RNAS-CL-IM	3-3-3-4	28, 24, 20, 16	40, 36, 32, 28	96, 88, 80, 72, 64, 56, 48	128 120, 108, 100, 92, 84, 76, 68	
RNAS-CL-I	3-3-3-4-4	28, 24, 20, 16	40, 36, 32, 28	96, 88, 80, 72, 64, 56, 48	128 120, 108, 100, 92, 84, 76, 68	216, 208, 200, 192, 184,176, 168, 160, 152, 144,136, 128, 120, 108

Table 8: The table describes the search space for ImageNet-100. Similar to Table 7, depth represents the depth of each stage. For ImageNet-100, we have up to 5 stages. Stage 1, Stage 2, Stage 3, Stage 4, and Stage 5 represent the filter choices for their respective stages. For example, in stage 1, for each convolution block, we search for its channel within 4 output channel options $(28, 24, 20, 16)$.

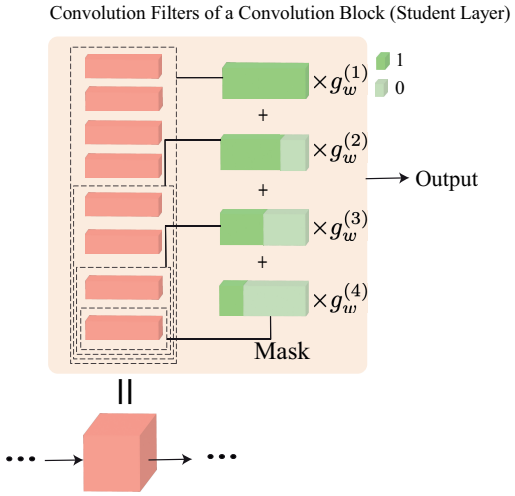

Figure 12: Illustration of searching for the neural architecture of each layer of student model using the searching mechanism in FBNetV2. g_w^i represents gumbel weights associated with each mask.