Appendices

A. RQ1 Density Plots

The density plots in Figure 8 shows the probability density plots of exit numbers in DyNN that are used to generate black-box adversarial inputs using PGD and FGSM attacks.

B. Early Attack Success Rate with Different α values

C. Transferability

Table 3 shows the $T1$ and $T2$ values of all three models on two datasets.

D. RQ1 results based on MI-FGSM attack

Here, through Figure 13, we show the adversarial transferability results between DyNNs and SDNNs using MI-FGSM attack. These results again confirm that adversarial examples from DyNN to SDNN are more transferable than adversarial examples from SDNN to DyNN.

E. Transferability experiments on MI-FGSM attack

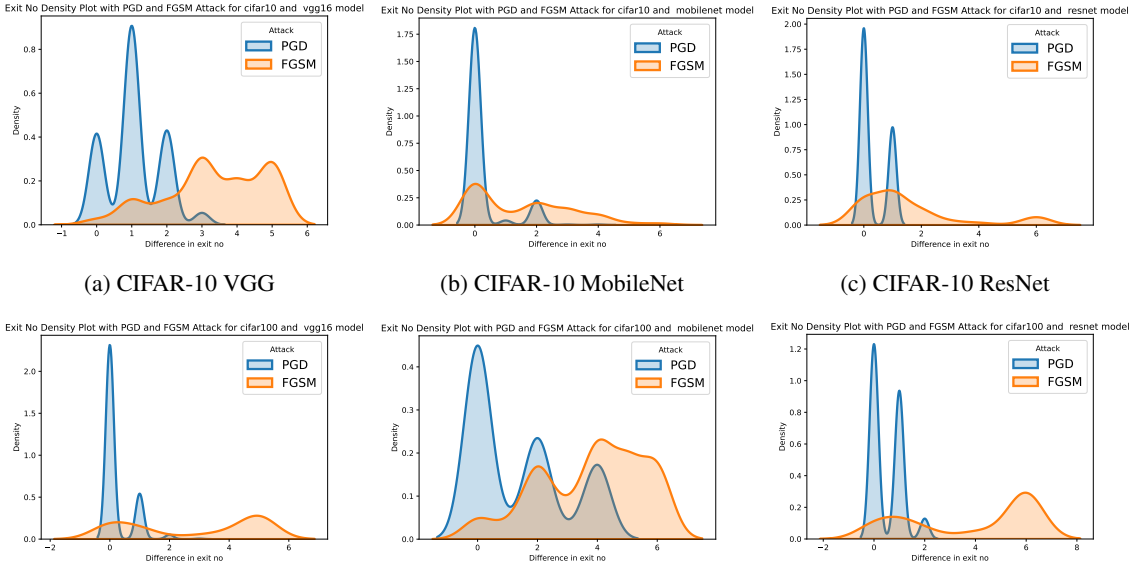
Through Figure 13, we show the S2D and D2S transferability with MI-FGSM attack [10]. The results confirm our claim that D2S transferability is higher than S2D transferability.

F. Comparing different adversarial images

In this section, we show original images, adversarial images generated through DyNNs and adversarial images generated through SDNNs through Figure 14,Figure 16 and Figure 15. We find that in terms of quality, images generated through SDNNs (Average PSNR [11] = 23.20) are slightly better than images generated through DyNNs (Average PSNR = 23.19).

G. RQ1 results based on Tiny Imagenet Images

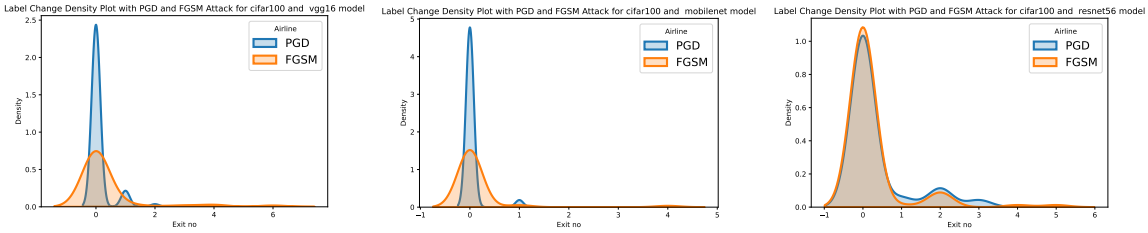
Here, through Figure 17, we show the adversarial transferability results between DyNNs and SDNNs for Tiny Imagenet [8] datasets. These images are larger in size (64×64) than CIFAR images (size 32×32). These results reconfirm that adversarial examples from DyNN to SDNN are more transferable than adversarial examples from SDNN to DyNN, even if the input feature space is larger. Although, we can notice a slight decrease in transferability for both D2S and S2D.



(a) CIFAR-10 VGG (b) CIFAR-10 MobileNet (c) CIFAR-10 ResNet

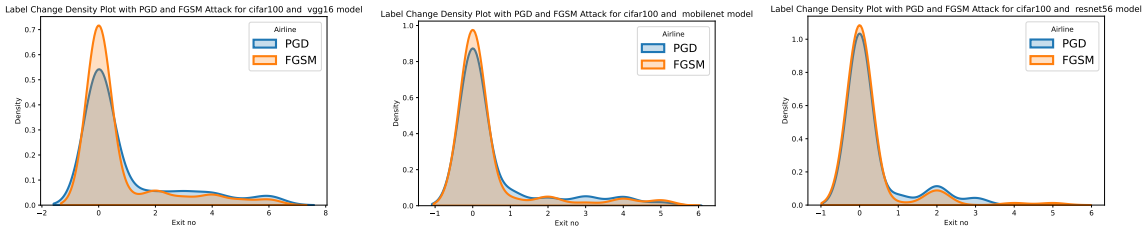
(d) CIFAR-100 VGG (e) CIFAR-100 MobileNet (f) CIFAR-100 ResNet

Figure 8: Density plots of exit numbers in DyNN that are used to generate black-box adversarial inputs using PGD and FGSM attacks. The x axis represents the exit number while y axis represents the density.



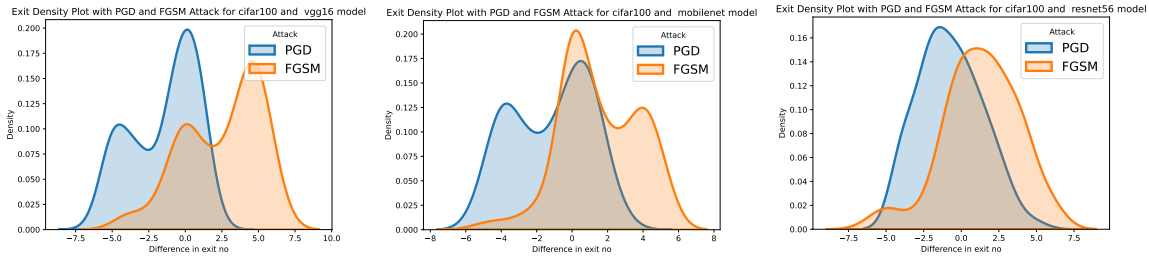
(a) CIFAR-100 VGG (b) CIFAR-100 MobileNet (c) CIFAR-100 ResNet

Figure 9: Density plots representing during which exit number output label is changed because of PGD and FGSM attack (For CIFAR-100 data). The x axis represents the exit number while y axis represents the density.



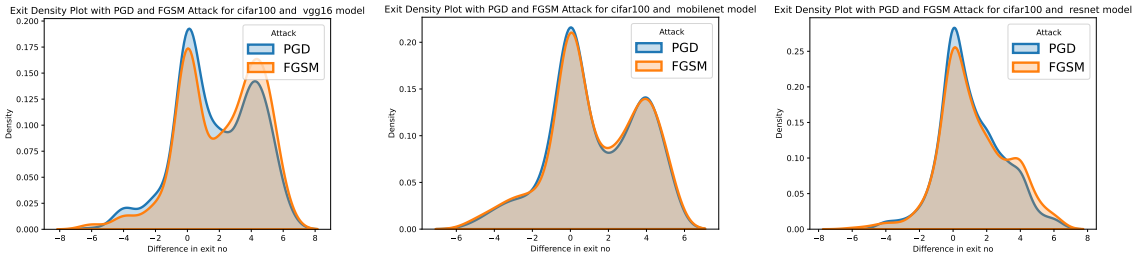
(a) CIFAR-100 VGG (b) CIFAR-100 MobileNet (c) CIFAR-100 ResNet

Figure 10: Density plots representing during which exit number output label is changed because of PGD and FGSM **black-box** attack (For CIFAR-100 data). The x axis represents the exit number while y axis represents the density.



(a) CIFAR-100 VGG (b) CIFAR-100 MobileNet (c) CIFAR-100 ResNet

Figure 11: Density plots of change in exit numbers because of PGD and FGSM attack (For CIFAR-100 data). The x axis represents the change in exit number while y axis represents the density.

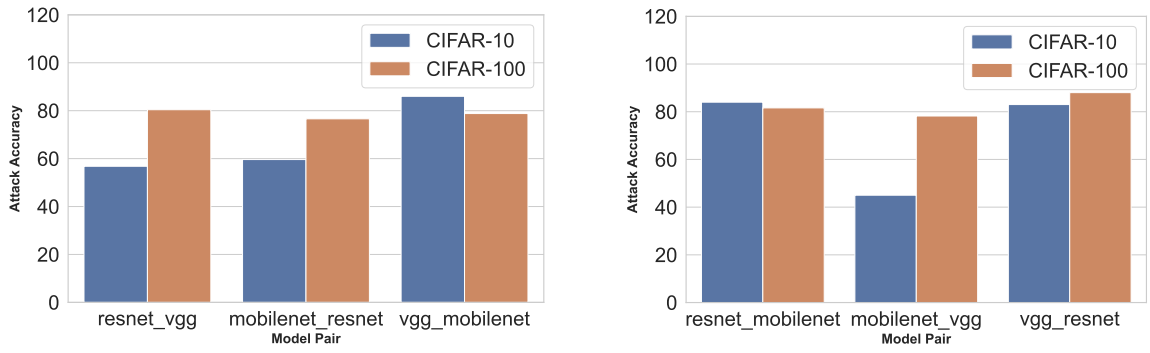


(a) CIFAR-100 VGG (b) CIFAR-100 MobileNet (c) CIFAR-100 ResNet

Figure 12: Density plots of change in exit numbers because of PGD and FGSM **black-box** attack (For CIFAR-100 data). The x axis represents the change in exit number while y axis represents the density.

Table 3: **Attack accuracy percentage of Early Attack with different α values**

Dataset	Model	EA(α=0.001)	EA(α=0.01)	EA(α=0.1)	EA(α=1)	EA(α=20)	EA(α=40)
CIFAR-10	VGG	0	0	0	35	10	4
	MobileNet	0	0	11	4	0	0
	ResNet	7	32	73	81	49	32
CIFAR-100	VGG	1	5	48	82	86	70
	MobileNet	0	16	74	97	92	77
	ResNet	46	74	92	96	96	93



(a) Target DyNN MI-FGSM (b) Target SDNN MI-FGSM

Figure 13: Transferable Attack Success Rate for MI-FGSM attack

Figure 14: Original Images

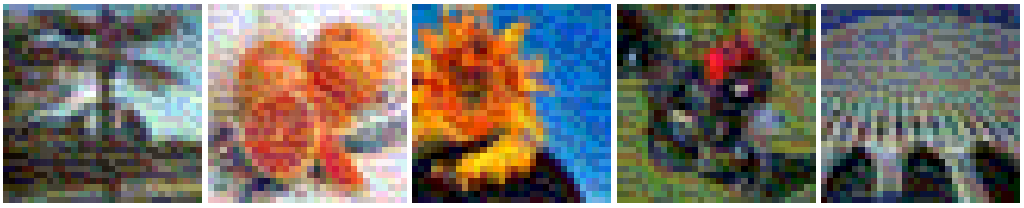


Figure 15: Adversarial Images generated on SDNNs

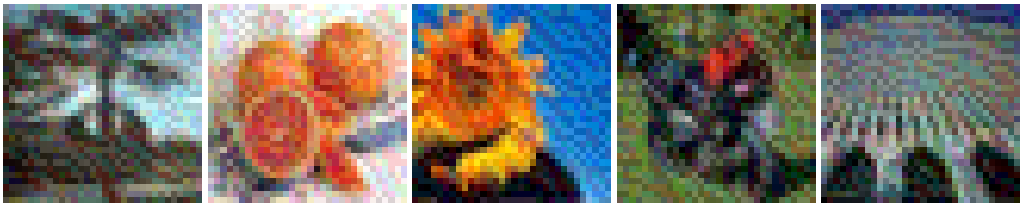
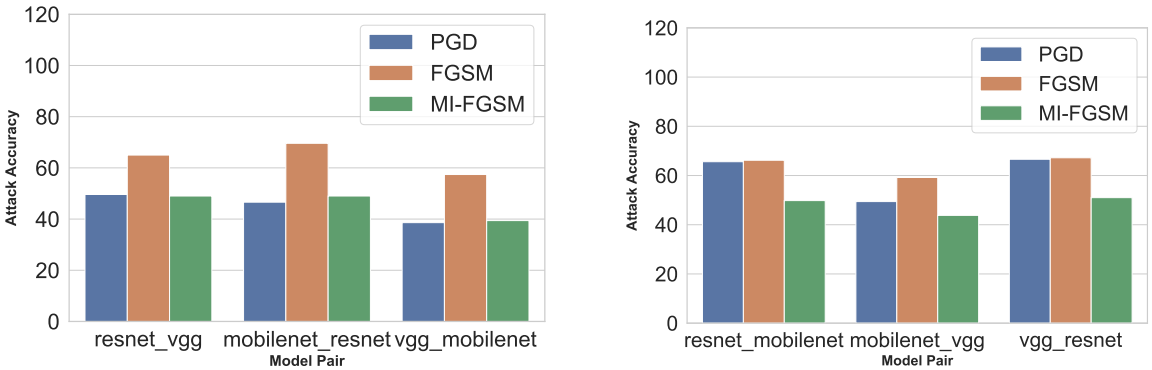


Figure 16: Adversarial Images generated on DyNNs



(a) Target DyNN (Tiny Imagenet)

(b) Target SDNN (Tiny Imagenet)

Figure 17: Transferable Attack Success Rate for Tiny Imagenet Data