Table 1: We conduct a quantitative study on the SYNTHIA [2] to Cityscapes [1] domain adaptation task for semantic segmentation, using ResNet-101 as the backbone. The evaluation is performed under the Fast Gradient Sign Method (FGSM) attack across 13 common semantic classes shared between the two datasets.

ϵ	road	sidewalk	building	light	sign	vegetation	sky	person	rider	car	bus	motor cycle	bike	**mIoU**	**mIoU drop**	**mIoU***
0.00 (clean)	82.4	39.4	75.6	21.6	16.5	76.6	77.8	54.3	20.0	80.2	46.8	23.5	43.6	**50.64**	0.00	50.64
0.03	80.0	37.5	73.1	20.0	16.2	74.0	76.4	53.5	18.0	78.0	45.3	22.1	42.0	**48.93**	**1.71**	50.64
0.10	60.0	29.4	55.6	11.6	6.5	56.6	57.8	39.3	9.0	60.2	26.8	11.7	28.6	**34.85**	**15.78**	50.64
0.25	36.4	28.4	35.6	13.6	4.6	26.5	43.8	24.3	8.7	40.2	9.1	13.5	23.6	**23.71**	**26.93**	50.64

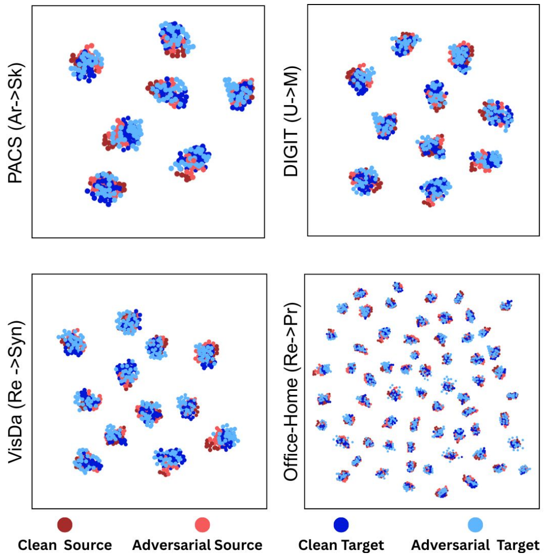

Figure 1: **t-SNE** We visualize the alignment of domains across both clean and adversarial inputs using features extracted from the shared feature extractor (i.e., the output of $\mathcal{F}$).

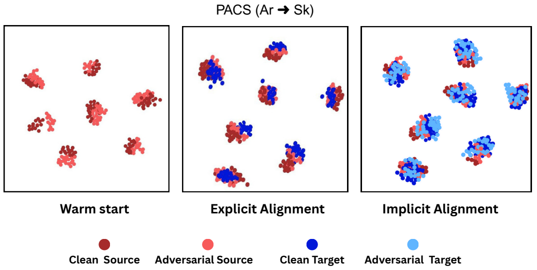

Figure 2: t-SNE visualization of feature across different domains at various training phases, illustrating the nature of progressive alignment strategy.

Table 2: Natural accuracy (Nat.) and robust accuracy under the PGD20 attack with $\epsilon = 2/255$ (PGD) are reported on the target test data across the DomainNet dataset, covering four subdomains(Clipart,Art,Real,Painting). † Results are not reported in the mean ± standard deviation format.

Method		**C→P**	**C→R**	**C→S**	**P→C**	**P→R**	**P→S**	**R→C**	**R→P**	**R→S**	**S→C**	**S→P**	**S→R**	**Avg.**
Ours	Nat.	35.7 ± 0.3	51.2 ± 0.2	42.4 ± 0.4	40.7 ± 0.3	52.1 ± 0.1	35.4 ± 0.4	49.1 ± 0.3	49.4 ± 0.2	37.5 ± 0.1	51.6 ± 0.3	42.9 ± 0.2	51.1 ± 0.4	44.9 ± 0.1
	PGD	31.9 ± 0.1	46.1 ± 0.3	36.5 ± 0.2	37.9 ± 0.1	48.2 ± 0.3	30.2 ± 0.2	44.9 ± 0.1	45.8 ± 0.4	31.3 ± 0.2	45.0 ± 0.2	35.8 ± 0.1	43.9 ± 0.3	39.7 ± 0.2

Table 3: Comparison of convergence speed over iterations for two datasets: Office-Home (Art → Clipart) and PACS (Photo → Sketch)

Dataset	**Method**	**Adversarial Sample**	**Iterations**	**Nat.(%)**	**PGD(%)**
Office-Home (Ar → Cl)	DANN	✗	20K	49.1±0.3	2.6±0.4
	SRoUDA	✓	25K	48.2±0.5	38.9±0.5
	DART	✓	25K	50.4±0.9	42.2±0.6
	OURS	✓	35K	**55.6±0.3**	**47.8±0.2**
PACS (Ph → Sk)	DANN	✗	20K	74.0±1.1	0.0±0.0
	SRoUDA	✓	25K	50.2±3.8	41.9±2.7
	DART	✓	25K	79.9±0.9	74.0±0.9
	OURS	✓	30K	**81.7±0.1**	**78.7±0.2**

Table 4: Natural accuracy (Nat.) and robust accuracy (PGD) on the VisDA dataset, evaluated under a PGD-20 attack with $\epsilon = \frac{2}{255}$, using ResNet-50 and ResNet-101 as backbone networks.

S → T	**Syn → Re**		**Re → Syn**	
Method	Nat.	PGD	Nat.	PGD
DANN	67.4 ± 0.2	0.5 ± 0.2	78.6 ± 0.9	0.8 ± 0.1
ARTUDA	45.2 ± 4.8	32.5 ± 2.7	72.5 ± 2.5	62.6 ± 0.3
SRoUDA	48.2 ± 2.7	33.4 ± 0.7	81.2 ± 1.4	72.9 ± 1.3
DART	69.5 ± 0.2	58.0 ± 0.5	87.3 ± 0.3	85.3 ± 0.2
CAM+SPLR†	72.8	65.9	89.5	87.1
Ours(Resnet50)	**75.3 ± 0.1**	**66.6 ± 0.2**	**89.7 ± 0.1**	**88.0 ± 0.2**
Ours(Resnet101)	**78.9 ± 0.1**	**69.5 ± 0.3**	**92.4 ± 0.1**	**89.3 ± 0.1**

Table 5: Natural and robust accuracy (Fast Gradient Sign Method, Projected Gradient Descent, and AutoAttack) evaluated using different methods and backbone networks (Resnet50, Resnet101. ViT-B/16).

Backbone	**Method**	**Pr→Re**				**Pr→Ar**				**Cl→Pr**				**Sk→Re**			
		Nat.	FGSM	PGD	AA	Nat.	FGSM	PGD	AA.	Nat.	FGSM	PGD	AA.	Nat.	FGSM	PGD	AA.
ResNet-50	DANN	60.0±0.6	12.2±0.4	0.3±0.1	0.0±0.0	49.1±0.3	11.7±0.2	0.2±0.1	0.0±0.0	47.9±0.8	9.4±0.3	3.6±1.0	1.1±0.3	67.4±0.2	13.5±0.1	0.5±0.2	0.0±0.0
	DART	63.5±0.8	54.7±0.2	43.6±0.5	42.6±0.5	43.7±2.5	34.5±0.3	21.5±0.8	20.0±1.0	57.0±0.3	51.7±0.1	45.5±0.6	44.8±0.5	69.5±0.2	62.4±0.3	58.0±0.5	55.7±0.1
	Ours	70.1±0.3	62.9±0.1	54.7±0.2	53.4±0.1	50.4±0.2	42.7±0.2	32.3±0.1	30.9±0.3	62.7±0.2	57.4±0.1	52.5±0.3	50.9±0.2	75.3±0.1	70.9±0.5	66.6±0.2	65.3±0.1
ResNet-101	Ours	75.9±0.3	69.2±0.2	59.6±0.1	56.8±0.1	55.7±0.4	51.2±0.1	37.4±0.1	35.2±0.3	67.1±0.2	63.3±0.1	58.1±0.3	56.8±0.1	78.9±0.1	73.4±0.1	69.5±0.2	67.5±0.3
ViT	Ours	82.3±0.2	79.4±0.1	68.6±0.2	65.2±0.2	64.1±0.1	59.3±0.1	46.9±0.2	43.8±0.3	75.3±0.1	72.4±0.3	66.1±0.2	64.9±0.2	82.1±0.3	78.5±0.1	73.4±0.1	71.5±0.2

Table 6: Comparison of natural and robust accuracy (PGD) between individual classifier heads and their average at inference on three dataset .

Dataset	**Head H1**		**Head H2**		**Avg(H1+H2)**	
	Nat. (%)	PGD (%)	Nat. (%)	PGD (%)	Nat. (%)	PGD (%)
Office-Home (Re→Cl)	59.5	54.3	59.2	54.2	59.6	54.5
PACS (Ca→Sk)	86.2	83.4	86.0	83.1	86.4	83.5
VisDA (Syn→Real)	75.3	66.6	75.4	66.5	75.6	66.8

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[2] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.