A. Experimental Details

In this section, we disclose the details of our experimental evaluations regarding the specific computational resources utilized, including hardware, memory, and time consumption. All our experimental evaluations are all conducted on GPU compute units equipped with an 11th Gen Intel(R) Core(TM) i9-11900K CPU, a single NVIDIA GeForce RTX 4090 GPU, and 128 GB of onboard memory.

For dSVA with DINO, MAE, and the vanilla supervised ViT-B/16 at a stride of $s = 16$, as well as for all compared generative attacks (CDA, BIA), generator $\mathcal{G}_\theta$ is trained on the entirely of the ImageNet training set for one epoch with a batch size of 32. Under this setup, single model variants of dSVA require up to 4 hours of training, *a duration comparable to previous methods.* For the joint variant, i.e., dSVA (Joint), batch size is set to 22, where its training takes up to 7 hours to complete. Our proposed additional exploit of self-attention (which is optional) in dSVA does not increase the training time. The inference time for the adversarial generator is comparable to, if not faster than, that of gradient-based iterative adversarial attacks. For all settings, GPU memory utilization approximates to over 90%. We organize the rest of the experimental details in Tab. 3, which includes ViTs with stride of $s = 8$ that we use in sections that report results of cross-domain transferability.

Attack	Stride s	Batch Size	GPU Memory	Training Time
dSVA (DINO)	16	32	> 90%	~4 hours
dSVA (DINO)	8	12	> 90%	~13 hours
dSVA (MAE)	16	32	> 90%	~4 hours
dSVA (MAE)	8	12	> 90%	~13 hours
dSVA (Joint)	16	22	> 90%	~7 hours
dSVA (Joint)	8	6	> 90%	~25 hours

Table 3. **Computational resource details of our experiments.** We report the computational details of all variants of dSVA with different ViT configurations that we evaluate.

B. Results of Cross-domain Transferability

In this section, we provide supplemental experimental results on the cross-domain transferability of dSVA in both coarse and fine-grained classification tasks. The evaluations follow the baseline settings specified in previous work [71]. For coarse-grained classification, we evaluate both attacks on target black-box domains, namely, CIFAR-10, CIFAR-100 [32], SVHN [42], and STL-10 [12], with the same models. For fine-grained classification, we report black-box transferability across three fine-grained domains: CUB-200-2011 [59], Stanford Cars [31], and FGVC Aircraft [37]. For each domain, we evaluate against three black-box ConvNets with ResNet-50 (Res-50), SENet154, and SE-ResNet101 (SE-Res-101) backbones, trained using the DCL framework [10].

Attack	s	A	Domain			
			CIFAR-10	CIFAR-100	SVHN	STL-10
CDA (VGG-19)	/	/	12.65	30.79	3.36	7.56
CDA (Res-152)	/	/	10.34	28.23	5.49	6.15
CDA (Den-169)	/	/	27.42	53.22	6.84	10.31
BIA (VGG-19)	/	/	**39.04**	**68.25**	6.38	9.84
BIA (Res-152)	/	/	26.24	49.36	3.75	7.35
BIA (Den-169)	/	/	22.05	45.82	**12.79**	**10.75**
dSVA (DINO)	16	w/o	13.98	37.67	**12.88**	11.07
dSVA (DINO)	8	w/o	24.05	53.00	6.54	11.18
dSVA (DINO)	16	w/	13.34	37.42	9.30	**12.66**
dSVA (DINO)	8	w/	21.94	48.94	7.53	10.70
dSVA (MAE)	16	w/o	16.89	35.80	6.80	10.41
dSVA (MAE)	8	w/o	24.77	41.15	9.13	10.26
dSVA (MAE)	16	w/	17.47	34.32	4.91	9.31
dSVA (MAE)	8	w/	24.30	44.61	6.74	11.44
dSVA (Joint)	16	w/o	23.64	50.28	8.94	11.04
dSVA (Joint)	8	w/o	**26.87**	**55.53**	8.83	12.42
dSVA (Joint)	16	w/	21.56	43.25	8.82	11.89
dSVA (Joint)	8	w/	24.13	46.73	11.73	11.95

Table 4. **Transferability towards coarse-grained classification domains.** We report transferability (%) towards domains CIFAR-10, CIFAR-100, SVHN, and STL-10. s is the stride of ViT-B/16. A denotes whether attention regularization in dSVA is activated.

Table 4 showcases our findings on coarse-grained classification domain transferability. With the target models in CIFAR-10 and CIFAR-100 being VGG-like architectures, the BIA attack using a VGG-19 surrogate model unsurprisingly yields superior results. Among the dSVA variants, dSVA (Joint) with DINO and MAE at stride $s = 8$ excels, closely matching the baseline performance in these domains. In contrast, for the SVHN and STL-10 domains, dSVA variants outperform the baseline, with dSVA (DINO) surpassing dSVA (Joint) in SVHN due to DINO's sensitivity to global shape and structure, which aligns with the focus of the SVHN domain on *house numbers* (digit classification). Interestingly, self-attention exploitation in dSVA does not enhance performance in this coarse-grained context.

Turning to fine-grained classification transferability in Tab. 5, dSVA (Joint) with active self-attention exploitation leads in most scenarios, outperforming nearly all baselines except when the target model is *Res-50*. Notably, dSVA (DINO) outperforms the otherwise dominant dSVA (Joint) variant in a specific case: attacking the *Stanford Cars* domain's *SE-Res-101* model.

Aggregating the results, we conclude that dSVA (Joint) variant remains the most robust attack overall for even most challenging cross-domain transfer scenarios, with the self-attention exploitation proving beneficial in most cases.

Attack	s	A	CUB-200-2011			Stanford Cars			FGVC Aircraft		
			Res-50	SENet154	SE-Res-101	Res-50	SENet154	SE-Res-101	Res-50	SENet154	SE-Res-101
CDA (VGG-19)	/	/	29.49	29.94	20.79	21.84	20.95	10.42	24.81	40.91	23.02
CDA (Res-152)	/	/	49.85	48.77	34.77	48.08	37.91	21.60	33.80	48.01	36.19
CDA (Den-169)	/	/	39.55	29.52	36.40	42.16	25.26	19.22	30.61	32.92	33.77
BIA (VGG-19)	/	/	62.21	52.78	36.84	70.93	37.01	29.86	82.61	51.17	51.27
BIA (Res-152)	/	/	63.53	68.15	38.92	56.91	58.49	19.03	41.52	77.61	42.33
BIA (Den-169)	/	/	**83.36**	65.75	45.77	**91.67**	51.75	52.57	**96.16**	59.78	65.22
dSVA (DINO)	16	w/o	38.86	51.65	43.66	**53.57**	59.22	50.79	**72.52**	81.45	64.73
dSVA (DINO)	8	w/o	71.18	61.15	59.57	49.39	59.76	**56.23**	54.38	77.71	67.96
dSVA (DINO)	16	w/	41.55	49.48	47.75	47.01	51.25	47.23	53.57	61.83	66.10
dSVA (DINO)	8	w/	33.68	40.99	38.12	33.78	37.92	29.92	37.12	46.25	55.68
dSVA (MAE)	16	w/o	42.93	51.81	37.56	28.80	47.10	20.24	34.13	50.62	43.86
dSVA (MAE)	8	w/o	37.38	58.97	36.44	44.28	38.30	26.74	29.70	50.10	36.58
dSVA (MAE)	16	w/	60.08	63.80	42.42	41.22	62.48	26.79	38.81	72.95	57.45
dSVA (MAE)	8	w/	42.38	62.11	41.99	46.04	38.99	29.33	30.41	52.90	43.73
dSVA (Joint)	16	w/o	**78.77**	79.62	66.11	48.67	**68.47**	51.97	65.65	89.24	**83.15**
dSVA (Joint)	8	w/o	62.58	72.17	59.11	41.42	55.68	41.17	46.76	75.07	63.62
dSVA (Joint)	16	w/	76.44	**79.64**	**69.72**	47.29	67.91	50.99	68.94	**89.93**	77.37
dSVA (Joint)	8	w/	70.88	78.85	68.24	47.25	66.30	50.12	68.15	87.97	74.10

Table 5. **Transferability towards fine-grained classification domains.** We report transferability (%) towards domains CUB-200-2011, Stanford Cars, and FGVC Aircraft. s is the stride of ViT-B/16. A denotes whether attention regularization in dSVA is activated.

Attack	Res-18 [48]	Res-50 [63]	ViT-B [39]	Swin-B [39]	XCiT-S12 [13]	ViT-S +ConvStem [51]	ConvNeXt +ConvStem [51]	ConvNeXt-v2+Swin-L [3]
CDA (VGG-19)	7.13	8.25	6.09	10.15	7.91	6.69	4.96	5.68
CDA (Res-152)	12.56	11.39	12.31	13.20	10.74	7.39	7.04	7.07
CDA (Den-169)	11.21	12.54	9.96	16.38	13.93	10.33	8.19	8.89
BIA (VGG-19)	12.05	11.22	8.85	12.96	11.22	9.51	7.50	7.50
BIA (Res-152)	16.13	15.35	14.52	19.32	16.06	11.97	10.61	8.24
BIA (Den-169)	14.09	14.19	18.95	22.62	16.65	10.92	9.80	9.42
CDA (ViT-B/16)	12.39	13.04	8.85	18.70	14.52	11.39	9.00	8.67
BIA (ViT-B/16)	10.70	9.90	12.86	12.47	8.97	8.10	7.50	5.03
MI (ViT-B/16)	7.81	7.92	11.62	12.96	8.26	7.51	6.46	6.96
PNA (ViT-B/16)	7.13	8.58	10.79	14.06	8.03	7.98	6.11	7.71
TGR (ViT-B/16)	12.73	11.55	16.18	18.34	12.16	11.50	8.88	9.32
ATT (ViT-B/16)	12.22	12.05	17.70	19.19	12.04	11.27	8.65	10.49
dSVA (DINO)	**20.88**	19.47	**23.93**	**26.28**	21.49	**15.96**	**12.80**	11.67
dSVA (MAE)	15.11	14.69	14.52	18.46	15.94	11.50	10.04	10.39
dSVA (Joint)	19.19	**19.64**	21.44	24.45	**22.31**	14.79	12.11	**11.99**

Table 6. **Additional transferability comparisons against models with defenses.** We include additional comparisons in defense evasion against various robust ConvNets, ViTs, and hybrid models equipped with state-of-the-art adversarial defenses.

C. Additional Comparisons of Transferability to Defense Models

In this section, we present additional comparisons on the transferability of dSVA to robust ConvNets, ViTs, and hybrid models with state-of-the-art defenses, which are lacking in prior work. We report the results in Tab. 6, where the citations accompanying the model names refer to the respective state-of-the-art adversarial defenses employed on the model itself. Note that we here use the same experimental setups as in Sec. 4, except for employing a larger $\varepsilon = 16$ constraint, otherwise the transferability across all evaluated attacks would be too low to be comparable.

We observe that dSVA still consistently outperforms the baselines across all models, averaging 17.04% black-box transferability, even against the most resilient defenses. dSVA (DINO) outperforms the joint variant in some cases, indicating that the shape/structural features are more adversarially impactful for robust models with smooth decision boundaries. These remarkable results once again underscore the robustness and effectiveness of our dSVA.

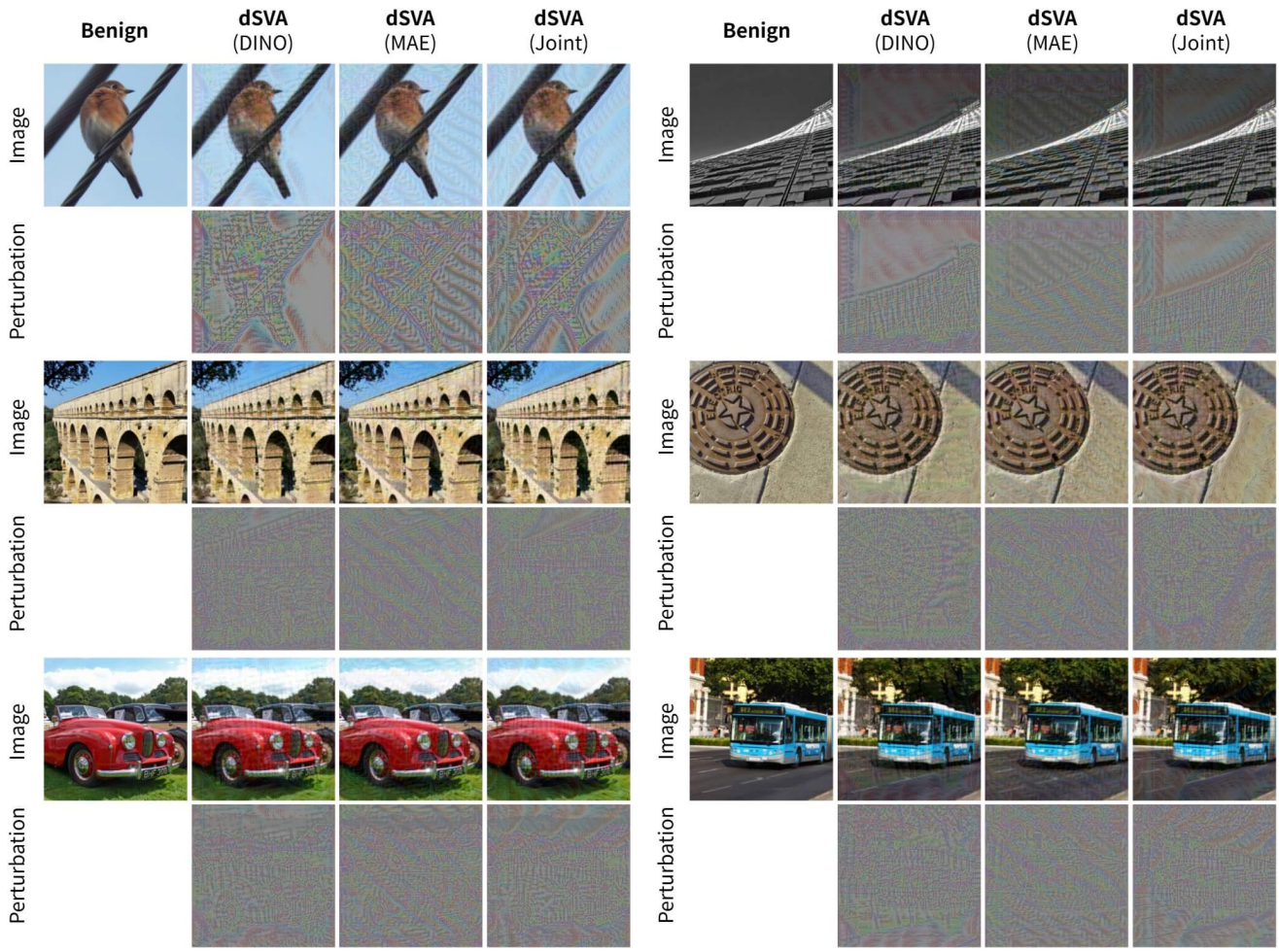

Figure 9. **Visualizations of adversarial examples.** We provide a few examples of side-by-side comparisons of the benign image, and adversarial examples generated by the 3 variants of dSVA (DINO, MAE, Joint). Perturbation is scaled and normalized for better visualization.

D. Visualization of Adversarial Examples

In this section, we provide a few visual examples of the adversarial examples and perturbations generated by dSVA. Figure 9 showcases several instances of successful attacks by the 3 variants of dSVA, namely, dSVA (DINO), which emphasizes structural features; dSVA (MAE), which emphasizes textural features; and dSVA (Joint), which successfully attends to both aspects, from left to right respectively. These visualizations highlight the rich, impactful perturbations crafted by our method, demonstrating its remarkable ability to exploit model vulnerabilities effectively.

E. Limitations and Future Work

While dSVA demonstrates impressive black-box transferability by exploiting self-supervised ViT features, we acknowledge certain limitations in our current work and outline potential avenues for future work.

Although dSVA shows strong transferability in a digital settings, our current work lacks full-scale physical world experiments. The potential of adopting generative adversarial attacks for physical real-world scenarios is a complex, challenging, yet valuable direction for future work.

Self-supervised methods with scaled training setups, such as DINOv2, may offer potentially improved transferability for dSVA. Additionally, investigating the use of ViTs with registers, and considering the use of multiple layers during adversarial optimization, could further enhance the effectiveness and robustness of dSVA. These approaches could lead to more effective adversarial attacks and are crucial directions for future work.

We acknowledge the importance of ethical implications of our work, as with all research in adversarial machine learning. Future research will continue to explore the broader societal impacts of adversarial attacks and contribute to the development of more robust and secure AI systems.