Appendix

A. Super-class constructions

Table 5 shows the construction of the 30 super-classes used in our Disjoint ImageNet Subsets (DINS) Test 1 and Test 2 experiments. Each super-class is the composition of 5 individual ImageNet classes, based on the WordNet [26] hierarchy. We take both the training and validation data from each of the ImageNet classes, so in total this dataset has about $1300 \times 5 \times 30 = 195000$ training images and $50 \times 5 \times 30 = 7500$ validation images (we say about because not all ImageNet classes have exactly 1300 training images).

Importantly, we do not propose that this is the only way to subset ImageNet for the purposes of constructing interesting and challenging transfer scenarios. We encourage future research to use these classes as a starting point, and to continue to build additional test environments using a similar methodological process.

Table 5: Construction of super-classes used in Disjoint ImageNet Subsets Tests 1 and 2

Super Class	ImageNet Components	ImageNet Class Names				
fish	[391, 392, 393, 394, 395]	coho salmon	rock beauty	anemone fish	sturgeon	garfish
bird	[10, 11, 12, 13, 14]	brambling	goldfinch	house finch	junco	indigo bunting
lizard	[38, 39, 40, 41, 42]	banded gecko	iguana	American chameleon	whiptail	agama
snake	[60, 61, 62, 63, 64]	night snake	boa constrictor	rock python	Indian cobra	green mamba
spider	[72, 73, 74, 75, 76]	black/gold garden spider	barn spider	garden spider	black widow	tarantula
dog-hound	[160, 161, 162, 163, 164]	Afghan hound	basset hound	beagle	bloodhound	bluetick
dog-terrier	[179, 180, 181, 182, 183]	Staffordshire bullterrier	American Staffordshire terrier	Bedlington terrier	Border terrier	Kerry blue terrier
dog-spaniel	[215, 216, 217, 218, 219]	Brittany spaniel	clumber spaniel	English springer spaniel	Welsh springer spaniel	English cocker spaniel
dog-retriever	[205, 206, 207, 208, 209]	flat-coated retriever	curly-coated retriever	golden retriever	Labrador retriever	Chesapeake Bay retriever
house-cat	[281, 282, 283, 284, 285]	tabby cat	tiger cat	Persian cat	Siamese cat	Egyptian cat
big-cat	[286, 287, 289, 290, 292]	cougar	lynx	snow leopard	jaguar	tiger
insect	[308, 309, 310, 311, 312]	fly	bee	ant	grasshopper	cricket
boat	[510, 554, 724, 814, 871]	container ship	fireboat	pirate ship	speedboat	trimaran
small-vehicle	[436, 511, 609, 717, 817]	station wagon	convertible	jeep	pickup	sports car
large-vehicle	[407, 779, 803, 864, 867]	ambulance	school bus	snowplow	tow truck	trailer truck
turtle	[33, 34, 35, 36, 37]	loggerhead turtle	leatherback turtle	mud turtle	terrapin	box turtle
big-game	[347, 348, 349, 350, 351]	bison	ram	bighorn sheep	ibex	hartebeest
drinkware	[572, 737, 898, 901, 907]	goblet	pop bottle	water bottle	whiskey jug	wine bottle
train	[466, 547, 565, 820, 829]	bullet train	electric locomotive	freight car	steam locomotive	trolley car
fungus	[992, 993, 994, 995, 996]	agaric	gyromitra	stinkhorn	earthstar	hen-of-the-woods
crab	[118, 119, 120, 121, 125]	Dungeness crab	rock crab	fiddler crab	king crab	hermit crab
mustelids	[356, 357, 359, 360, 361]	weasel	mink	black-footed ferret	otter	skunk
instrument	[402, 420, 486, 546, 594]	acoustic guitar	banjo	cello	electric guitar	harp
computer	[508, 527, 590, 620, 664]	computer keyboard	desktop computer	hand-held computer	laptop	monitor
fruit	[948, 949, 950, 951, 954]	Granny Smith	strawberry	orange	lemon	banana
monkey	[371, 372, 373, 374, 375]	hussar monkey	baboon	macaque	langur	colobus monkey
sports-ball	[429, 430, 768, 805, 890]	baseball	basketball	rugby ball	soccer ball	volleyball
clothing	[474, 617, 834, 841, 869]	cardigan	lab coat	suit	sweatshirt	trench coat
beetle	[302, 303, 304, 305, 306]	ground beetle	long-horned beetle	leaf beetle	dung beetle	rhinoceros beetle
butterfly	[322, 323, 324, 325, 326]	ringlet butterfly	monarch butterfly	cabbage butterfly	sulphur butterfly	lycaenid butterfly

** The horizontal dashed lines are for visual clarity purposes only.

B. Experimental setup details

For reproducibility, we include some additional details of our experimental setup here. In this section we primarily discuss the setup as it pertains to the Disjoint ImageNet Subsets Test 1 and Test 2 environments. See Appendix E for details regarding the ImageNet to Places365 tests.

Model Training. The first critical step in the setup, after constructing the Test 1.A/B and Test 2.A/B dataset splits, is to train the DNN models that we will transfer attacks between. On Test 1.A and 2.A data we train RN50 and DN121 models to be used as whiteboxes. On Test 1.B and 2.B data we train RN34, RN152, DN169, VGG19bn, MNv2 and RXT50 models to be used as blackboxes. All models are trained using the official PyTorch ImageNet example code from `https://github.com/pytorch/examples/blob/master/imagenet/main.py`. Table 6 shows the accuracy of each model after training, as measured on the appropriate in-distribution validation split.

Table 6: Test accuracy of models used in the DINS Test 1 and 2 experiments.

Train Data	Model	Accuracy
Test-1.A	RN50	94.2
Test-1.A	DN121	96.1
Test-1.B	RN34	95.7
Test-1.B	RN152	96.2
Test-1.B	DN169	96.8
Test-1.B	VGG19bn	96.7
Test-1.B	MNv2	95.5
Test-1.B	RXT50	96.0
Test-2.A	RN50	92.0
Test-2.A	DN121	95.0
Test-2.B	RN34	93.6
Test-2.B	RN152	95.1
Test-2.B	DN169	95.2
Test-2.B	VGG19bn	95.3
Test-2.B	MNv2	94.0
Test-2.B	RXT50	93.7

Attack Configurations. For all attacks, we use a standard configuration of L_∞ $\epsilon = 16/255$, $\alpha = 2/255$, $\text{perturb\_iters} = 10$, $\text{momentum} = 1$ when optimizing the adversarial noise [6, 35, 14]. As described in [13] and [14] the Feature Distribution Attacks (FDA) require a tuning step to find a good set of attacking layers (shown as $\mathcal{L}$ in eqn. (3)). We follow the greedy layer optimization procedure from [14] and find the following set of 4 attacking layers per whitebox model to work well:

- RN50-Test1.A = [(3,4,5), (3,4,6), (3,4,6,1), (3,4,6,2)]
- DN121-Test1.A = [(6,6), (6,12,10), (6,12,14), (6,12,24,12)]
- RN50-Test2.A = [(3,4,5), (3,4,6), (3,4,6,1), (3,4,6,2)]
- DN121-Test2.A = [(6,6), (6,12,2), (6,12,22), (6,12,24,12)]

This notation comes from the way the RN50 and DN121 models are implemented in code (see `http://pytorch.org/vision/stable/models.html`). The full RN50 model has 4 layer groups with (3,4,6,3) blocks in each, and DN121 has 4 layer groups with (6,12,24,16) blocks in each. So, for example, attacking at RN50 layer (3,4,5) means we are using the feature map output from the 5^{th} block in the 3^{rd} layer group. Further, we use $\eta_{RN50} = 1\text{e}{-6}$, $\eta_{DN121} = 1\text{e}{-5}$ to weight the contribution of the feature distance term in eqn. (3). As emphasized in the text, all of these hyperparameter settings are tuned by attacking between RN50-Test1.A↔DN121-Test1.A and RN50-Test2.A↔DN121-Test2.A. So, there is no dependence on querying a Test 1.B or 2.B model for hyperparameter tuning.

Attack Procedure. Because the images considered for attack are originally from the ImageNet validation set, each super-class has $5 \times 50 = 250$ test images. Since there are 15 super-classes in each split, and we do not consider a clean image for attack that is from the target class, we have a set of $14 \times 250 = 3,500$ images that are eligible for attack in any given (target, proxy) pair. As noted, we enforce that all clean starting images prior to attack are correctly classified by the blackbox models. From Table 6, these models operate at about $\sim 95\%$ accuracy, meaning each error/tSuc number reported in Tables 2, 3 and 7 is averaged over $\sim 3,300$ adversarially attacked images.

C. Full DINS Test 2 transfer results

Table 7 shows the full transfer results in the Disjoint ImageNet Subsets (DINS) Test 2 environment, and is supplemental to Table 3 in the manuscript. Note, only the numbers in the "avg." column are shown in the main document (Table 3), so the purpose of this table is to display the individual transfer rates to each blackbox model architecture. Please refer to Section 5.3 for further discussion and analysis of these results.

Table 7: Transfers in the DINS Test 2 environment (notation = error / tSuc)

			Blackbox Models (Split B)						
Target (Split B)	Proxy (Split A)	Attack	RN34	RN152	DN169	VGG19bn	MNv2	RXT50	avg.
large-vehicle	**train**	TMIM	29.7 / 7.0	23.4 / 5.3	27.1 / 8.3	27.1 / 6.4	33.3 / 11.7	36.5 / 10.0	29.5 / 8.1
		FDA	55.4 / 35.2	50.7 / 39.6	63.9 / 52.9	56.5 / 41.0	69.9 / 57.7	68.1 / 42.4	**60.8 / 44.8**
spider	**beetle**	TMIM	26.8 / 5.3	20.5 / 3.0	23.5 / 4.2	25.4 / 5.2	30.0 / 6.9	32.3 / 5.3	26.4 / 5.0
		FDA	44.9 / 23.2	43.6 / 27.1	55.6 / 21.2	53.6 / 36.0	53.5 / 31.9	52.9 / 27.7	**50.7 / 27.8**
large-vehicle	boat	TMIM	27.6 / 3.3	22.2 / 2.2	24.7 / 1.9	25.7 / 2.0	31.1 / 5.2	33.6 / 3.6	27.5 / 3.0
		FDA	48.4 / 6.9	40.4 / 4.1	49.8 / 3.0	41.4 / 4.2	51.9 / 13.8	55.9 / 5.9	**48.0 / 6.3**
spider	**insect**	TMIM	29.6 / 7.5	22.6 / 4.4	25.3 / 5.8	27.0 / 7.0	31.5 / 9.3	33.3 / 8.1	28.2 / 7.0
		FDA	49.8 / 38.4	46.8 / 37.1	54.2 / 38.1	59.4 / 50.6	63.5 / 52.8	60.7 / 49.4	**55.7 / 44.4**
fungus	fruit	TMIM	24.8 / 1.5	21.6 / 2.1	22.7 / 2.5	23.7 / 1.9	28.1 / 2.3	32.1 / 2.8	25.5 / 2.2
		FDA	49.8 / 4.2	55.7 / 3.4	59.1 / 3.1	43.3 / 8.1	49.0 / 4.7	53.1 / 10.2	**51.7 / 5.6**
mustelids	**monkey**	TMIM	27.6 / 8.9	22.5 / 5.0	27.2 / 11.8	25.7 / 7.0	30.7 / 10.7	30.0 / 9.7	27.3 / 8.9
		FDA	58.3 / 28.4	52.5 / 21.6	63.3 / 29.5	60.5 / 17.0	59.6 / 22.3	59.4 / 27.0	**58.9 / 24.3**
house-cat	**dog-spaniel**	TMIM	26.2 / 4.8	20.1 / 3.5	26.1 / 5.3	26.3 / 8.4	29.6 / 5.8	27.9 / 2.9	26.0 / 5.1
		FDA	58.1 / 15.0	52.2 / 26.4	63.1 / 21.8	60.2 / 25.8	65.2 / 32.1	59.1 / 13.8	**59.7 / 22.5**
clothing	instrument	TMIM	25.0 / 8.7	21.1 / 7.9	25.6 / 9.3	26.3 / 8.5	29.9 / 8.7	35.0 / 7.8	27.2 / **8.5**
		FDA	41.4 / 9.9	24.9 / 4.7	39.4 / 7.5	43.1 / 12.9	55.8 / 9.3	58.0 / 4.9	**43.8** / 8.2
fungus	crab	TMIM	28.8 / 1.9	25.8 / 1.9	26.6 / 3.1	28.9 / 3.3	32.7 / 2.3	33.6 / 2.7	29.4 / 2.5
		FDA	52.1 / 6.6	53.5 / 4.2	58.8 / 13.6	54.7 / 16.4	60.1 / 4.2	55.2 / 7.9	**55.7 / 8.8**
big-game	**dog-retriever**	TMIM	25.0 / 1.5	19.3 / 0.8	24.5 / 2.7	23.7 / 0.8	27.9 / 2.5	25.7 / 0.7	24.3 / 1.5
		FDA	56.7 / 18.7	52.8 / 13.0	61.7 / 24.0	56.7 / 16.4	64.6 / 14.4	66.3 / 23.9	**59.8 / 18.4**

D. Additional query attack results

Figure 5 is an extension of Figure 4 in the main document, and is produced under the same experimental conditions described in Section 5.5. The top row of Figure 5 subplots shows the targeted success rates (tSuc) versus query counts (q) for the (`snake`, `lizard`), (`big-cat`, `house-cat`), (`large-vehicle`, `small-vehicle`) and (`beetle`, `insect`) scenarios from DINS Test 1. The bottom row of subplots shows tSuc vs q for the (`spider`, `insect`), (`mustelids`, `monkey`), (`house-cat`, `dog-spaniel`) and (`large-vehicle`, `train`) scenarios from DINS Test 2. *RGF* represents the query-only Random Gradient Free [3] baseline attack; *TMIM+RGF* represents the *RGF* attack warm-started with the *TMIM* transfer attack direction; and *FDA+RGF* represents the *RGF* attack warm-started with the *FDA* transfer attack direction. Finally, all results are averaged over the six individual blackbox models in the corresponding test environment.

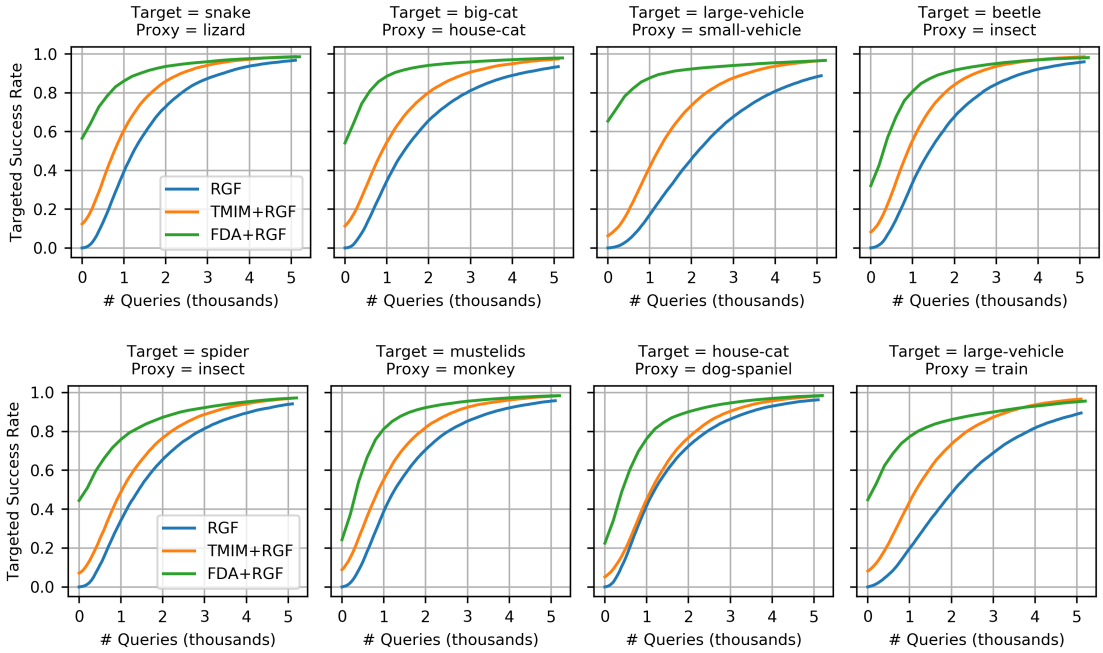

Figure 5: Targeted attack success when integrated with the RGF query attack method.

Importantly, the results in Figure 5 are consistent with those discussed in Section 5.5. Both transfer attacks provide useful prior adversarial directions when integrated with the *RGF* query-based attack. In practical terms, this means that using a transfer direction as a prior on the blackbox model's gradient signal, even if the source model on which the transfer direction is computed does not have the true target class in its label space, can significantly boost the query-efficiency of the RGF attack. To supplement Figure 5, Table 8 shows the tSuc results of the three attacks at a query budget of $q = 0/500/1000$.

Table 8: Summary of query results at 0 / 500 / 1000 queries. This table directly supplements Figure 5.

Test	Target	Proxy	RGF	TMIM+RGF	FDA+RGF
1	`snake`	`lizard`	0 / 14 / 41	12 / 36 / 61	**56 / 76 / 86**
1	`big-cat`	`house-cat`	0 / 13 / 35	11 / 32 / 55	**54 / 78 / 89**
1	`large-vehicle`	`small-vehicle`	0 / 4 / 17	6 / 20 / 42	**65 / 80 / 88**
1	`beetle`	`insect`	0 / 10 / 33	8 / 28 / 55	**32 / 62 / 81**
2	`spider`	`insect`	0 / 11 / 34	7 / 24 / 49	**44 / 63 / 76**
2	`large-vehicle`	`train`	0 / 7 / 20	8 / 23 / 44	**45 / 66 / 77**
2	`mustelids`	`monkey`	0 / 12 / 30	9 / 31 / 55	**24 / 61 / 81**
2	`house-cat`	`dog-spaniel`	0 / 15 / 43	5 / 21 / 45	**22 / 55 / 76**

Notice the $q = 0$ results for the *TMIM+RGF* and *FDA+RGF* attacks match the numbers in Tables 2 and 3, as this is equivalent to the transfer-only (i.e., no-query) setting. Also, notice that all $q = 0$ results for the RGF attack are 0% tSuc. Finally, we remark that the *FDA+RGF* attack is the top performer across all scenarios and can reach up to 80% tSuc at $q = 500$ and nearly 90% tSuc at $q = 1000$, depending on the particular (target, proxy) pair. Compare this to the *RGF* attack alone, which can only reach about 15% tSuc at $q = 500$ and 43% tSuc at $q = 1000$ in the best case that we examine.

E. Extra materials for ImageNet to Places365 transfers

To supplement the ImageNet to Places365 transfer results in Section 5.6 of the main paper, here we describe some additional experimental setup details and show the expanded transfer results.

Setup. For all attacks, we use a standard configuration of L_∞ $\epsilon = 16/255$, $\alpha = 2/255$, $\mathrm{perturb\_iters} = 10$, $\mathrm{momentum} = 1$ when optimizing the adversarial noise [6, 35, 14]. Similar to the FDA layer tuning process described in Appendix B and in [14], we tune the FDA layers for the RN50, DN121 and VGG16bn ImageNet whitebox models by attacking only amongst themselves (i.e., no queries to Places365 models required). We find the following FDA layer sets to be powerful:

- RN50-ImageNet = [(3,3), (3,4), (3,4,2), (3,4,5), (3,4,6)]
- DN121-ImageNet = [(6,10), (6,12), (6,12,10), (6,12,18), (6,12,24,8)]
- VGG16bn-ImageNet = [6, 9, 11]

See Appendix B for the RN50 and DN121 notation. For VGG16bn layers, the numbers indicate which convolutional layers we take the output feature maps from. We use $\eta_{RN50} = 1\mathrm{e}{-6}$, $\eta_{DN121} = 1\mathrm{e}{-5}$, $\eta_{VGG16bn} = 1\mathrm{e}{-6}$ for the feature distance weights (see eqn. (3)) of the RN50, DN121 and VGG16bn models, respectively. Finally, all transfer statistics in our tests are averaged over 5000 adversarial examples, where the clean images are randomly sampled from the Places365 validation set and all are correctly classified by the Places365 blackbox models.

Expanded Results. Table 9 shows the full transfer results for the ImageNet to Places365 transfers, and is meant to supplement Table 4 in the main paper. We include this table to show the transferability results to each individual Places365 blackbox model, separately. See Section 5.6 for a discussion on these results.

Table 9: Attacking Places365 models via ImageNet models (notation = error / tSuc)

			Blackbox Model (Places365)		
Target (Places365)	Proxy (ImageNet)	Attack	WRN18	RN50	DN161
83:Carousel	476:Carousel	TMIM	65.3 / 5.4	64.7 / 5.4	60.6 / 4.0
		FDA	95.1 / 74.7	92.4 / 48.8	93.5 / 65.5
154:Fountain	562:Fountain	TMIM	63.8 / 11.6	59.6 / 7.6	57.2 / 7.6
		FDA	92.6 / 75.9	89.4 / 65.0	91.1 / 73.2
40:Barn	425:Barn	TMIM	61.4 / 2.9	57.6 / 1.3	55.3 / 1.7
		FDA	87.6 / 27.3	83.6 / 14.6	84.1 / 20.9
300:ShoeShop	788:ShoeShop	TMIM	59.5 / 2.0	59.4 / 6.0	55.0 / 2.6
		FDA	94.9 / 76.9	95.4 / 82.3	94.2 / 83.3
59:Boathouse	449:Boathouse	TMIM	58.9 / 1.8	55.4 / 0.4	51.7 / 0.8
		FDA	88.9 / 39.5	84.8 / 20.1	84.2 / 32.2
350:Volcano	980:Volcano	TMIM	56.7 / 0.8	54.0 / 0.5	51.5 / 0.8
		FDA	80.2 / 37.0	78.6 / 19.3	79.6 / 40.3
72:ButcherShop	467:ButcherShop	TMIM	62.9 / 2.1	58.5 / 1.2	56.6 / 1.3
		FDA	92.1 / 22.1	88.6 / 13.8	89.6 / 27.9
60:Bookstore	454:Bookshop	TMIM	59.4 / 2.3	57.5 / 2.9	53.9 / 1.9
		FDA	92.5 / 31.7	90.3 / 28.9	88.5 / 28.3
342:OceanDeep	973:CoralReef	TMIM	61.6 / 4.7	57.5 / 3.0	55.4 / 1.4
		FDA	87.1 / 32.3	84.8 / 34.9	86.6 / 44.8
76:Campsite	672:MountainTent	TMIM	58.7 / 1.9	56.0 / 1.7	53.4 / 2.3
		FDA	90.4 / 61.4	86.2 / 43.6	90.2 / 73.5
6:AmusementArcade	800:Slot	TMIM	64.0 / 2.9	63.7 / 7.6	60.3 / 4.9
		FDA	93.1 / 30.7	93.4 / 43.3	91.6 / 30.6
214:Lighthouse	437:Beacon	TMIM	56.5 / 0.5	53.9 / 0.2	51.8 / 0.9
		FDA	79.4 / 17.0	74.9 / 5.6	74.5 / 15.4
147:FloristShop	985:Daisy	TMIM	59.0 / 0.2	55.5 / 0.0	52.8 / 0.0
		FDA	84.5 / 18.5	80.6 / 10.3	79.1 / 10.6
278:RailroadTrack	565:FreightCar	TMIM	60.6 / 2.1	58.4 / 1.6	55.3 / 2.0
		FDA	82.3 / 30.3	78.2 / 20.7	80.6 / 35.0
90:Church	406:Altar	TMIM	64.9 / 0.9	62.0 / 0.5	60.1 / 0.4
		FDA	93.9 / 20.0	92.7 / 7.4	93.6 / 14.8
196:jail_cell	743:prison	TMIM	56.2 / 3.0	53.6 / 0.6	50.3 / 2.1
		FDA	78.4 / 25.0	76.3 / 7.4	77.0 / 26.2
180:hot_spring	974:geyser	TMIM	56.3 / 0.1	53.6 / 0.1	49.6 / 0.2
		FDA	84.5 / 4.4	83.2 / 5.6	83.5 / 9.9
51:bedchamber	564:four-poster	TMIM	64.7 / 9.4	63.5 / 10.3	63.0 / 10.9
		FDA	92.7 / 63.3	91.2 / 46.1	92.1 / 56.6
268:playground	843:swing	TMIM	62.6 / 0.3	58.8 / 2.8	55.4 / 2.2
		FDA	77.0 / 12.1	75.1 / 10.3	74.7 / 16.3
42:baseball_field	981:ballplayer	TMIM	56.8 / 0.0	54.2 / 0.2	51.7 / 0.2
		FDA	77.7 / 15.3	77.0 / 23.6	75.1 / 27.0