In this supplementary material, we report the comparison results of DHF and other baselines and the results under different perturbation budgets. Besides, we present some visualization results of benign images and adversarial examples generated by DHF.

A Additional Experiment Results

A.1 Comparison with More Baselines

DHF modifies the feature calculation in forward propagation. We group it into surrogate refinement attacks. Therefore, we compare it with other surrogate refinement attacks, *i.e.*, TAP [74], ILA [23] SGM [63], ghost network [31] in our paper.

	Method	Res-18	Res-50	Res-101	Res-152	IncRes-v2	DenseNet-121	MobileNet	ViT	Swin
MI-FGSM	AA	58.7	52.3	45.9	57.2	47.8	58.6	56.7	28.7	41.9
	FIA	67.3	60.0	47.3	58.0	49.4	66.1	57.6	33.9	45.0
	NAA	68.9	60.3	47.0	55.4	48.7	68.3	56.9	33.2	40.6
	LTAP	68.5	64.3	46.6	59.7	50.1	68.1	58.1	33.3	47.0
	BIA	61.3	58.9	44.6	59.2	48.8	62.3	55.1	30.6	44.2
	DHF	**71.9**	**76.7**	**47.9**	**70.2**	**57.5**	**74.7**	**62.9**	**35.2**	**53.2**

Table A: Average black-box attack success rates (%) on nine models. The adversarial examples are generated on Res-101, Res-152 and IncRes-v2, respectively.

Meanwhile, there are some similar but distinct methods: 1) AA [25], FIA [60] and NAA [72]. They are feature disruption attacks, which also adjust the features of adversarial images but focus on the feature distance when optimizing the perturbation; 2) LTAP [39] and BIA [73]. They have similar mechanism with DHF but they focus on cross-domain transferability (*e.g.*, Cartoon $\rightarrow$ ImageNet) using pretrained generators, while DHF focuses on cross-model transferability (*e.g.*, Inc-v3 $\rightarrow$ ResNet-18).

To help us better understand the mechanism of DHF and illustrate the effectiveness of DHF, we extend to compare DHF with these similar but different methods in Tab. A. We observe that DHF surpasses AA, FIA, NAA, LTAP and BIA by 11.3%, 7.3%, 7.8%, 6.0% and 9.5% on average, respectively. The results further validate the superiority of DHF.

A.2 Results when Perturbation Budget $\varepsilon = 8$

The setting of perturbation budget $\varepsilon = 16$ is general for transfer-based attacks. Some works also take the perturbation budget $\varepsilon = 8$ as an optional setting [31, 68]. To fully validate the effectiveness of DHF, we compare DHF with the 2 SOTA baselines, *i.e.*, SGM, and ghost network when $\varepsilon = 8$. The results are summarized in Tab. B. Despite of the reduced perturbation budget, DHF still outperforms the baselines by a significant margin, showing its high effectiveness.

B Visualization Reuslts

In Fig. A, we present some visualization results of the adversarial examples generated by DHF when $\varepsilon = 8$ and $\varepsilon = 16$, respectively. The adversarial examples exhibit a remarkable visual similarity to the benign images with high adversarial transferability.

	Method	Res-18	Res-50	Res-101	Res-152	IncRes-v2	DenseNet-121	MobileNet	ViT	Swin
MI-FGSM ($\varepsilon = 8$)	SGM	39.3	34.1	26.1	28.3	26.2	35.4	41.7	14.0	23.6
	Ghost	34.7	35.3	34.5	30.0	25.5	37.4	34.6	12.5	23.0
	DHF	**41.4**	**44.0**	**43.3**	**38.7**	**30.1**	**44.7**	**42.1**	**19.0**	**29.7**

Table B: Average black-box attack success rates (%) on nine models. The adversarial examples are generated on Res-101, Res-152 and IncRes-v2, respectively, when $\varepsilon = 8$.

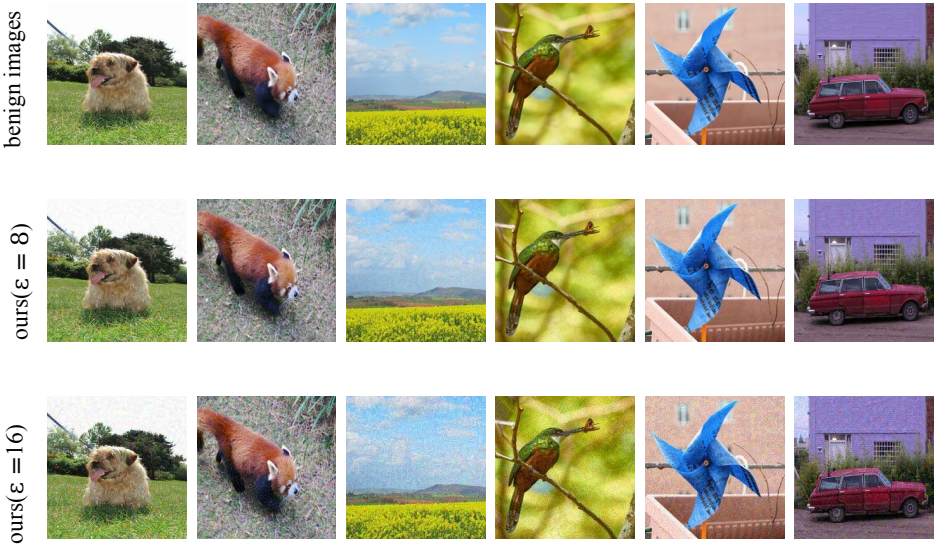

Figure A: Visualization of benign images and the adversarial examples generated by DHF when the perturbation budget $\varepsilon = 8$ and $\varepsilon = 16$, respectively.