Once and For All: Universal Transferable Adversarial Perturbation against Deep Hashing-based Facial Image Retrieval
Appendix

Anonymous submission

A Illustration of the UTAP Pipeline

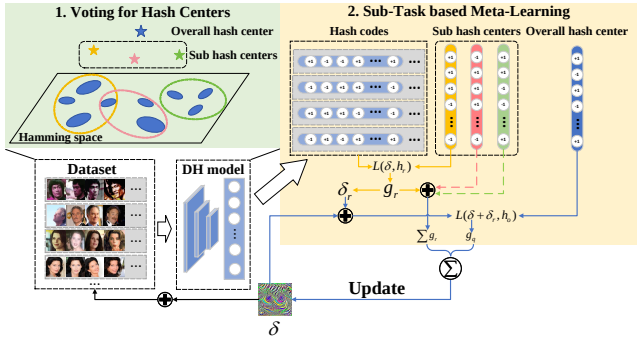

Figure 1: Illustration of the UTAP pipeline. Blue arrows represent the query process updating the main gradient with the overall hash center. Yellow arrows indicate the meta-learning support process applied to a sub hash center. The red and green dashed arrows depict the same process as the yellow arrows, but are applied to different sub hash centers.

We illustrate the general pipeline of Algorithm 1 in the main text as shown in Fig. 1. UTAP consists of two main components. It first utilizes the original dataset for voting to select the overall hash center and sub hash centers. Then it leverages these hash centers for sub-task based meta-learning, and eventually obtains the universal transferable adversarial perturbation.

B Additional Ablation Studies

Selecting Different Attack Methods

Since UTAP generates universal adversarial perturbation using gradient-based iterative adversarial attack, many methods used to enhance transferability in attacking classification models are also applicable to UTAP. We evaluate the effectiveness of four typical adversarial attack methods (I-FGSM (Kurakin, Goodfellow, and Bengio 2018), MI (Dong et al. 2018), DI (Xie et al. 2019), MIDI), and the results are presented in Table 1.

We conduct attacks on the ResNet34 and VGG19 models using the CASIA-WebFace dataset. Results show that UTAP still achieves better cross-model transferability compared to existing methods, regardless of the employed method. DI

Surrogate Model	Method	ResNet34		ResNet50		VGG16		VGG19		Avg↓
		O	Adv↓	O	Adv↓	O	Adv↓	O	Adv↓	
ResNet34	I-FGSM	91.13	***4.35****	91.06	36.20	89.50	42.39	86.33	30.26	28.30
	MI		*4.35**		34.58		42.43		32.19	28.39
	DI		*4.88**		**27.48**		**37.66**		**26.01**	**24.01**
	MIDI		*5.00**		29.60		38.60		26.33	24.88
VGG19	I-FGSM		42.35		41.39		29.79		*5.27**	29.70
	MI		45.25		40.03		29.08		*5.18**	29.88
	DI		33.97		31.01		25.04		*5.20**	23.81
	MIDI		**31.56**		**27.95**		**22.43**		***5.13****	**21.77**

Table 1: Ablation study for different adversarial attacks (mAP%). The default setting is (CSQ, 64-bit, CASIA). Avg is the average results of Adv. The white-box attacks are represented with italics and *, and the best results are emphasized in bold.

demonstrates better transferability on the ResNet34 model, while MIDI exhibits better transferability on the VGG19 model. Considering MIDI delivers stronger overall attack effectiveness according to **Avg**, we adopt MIDI as the baseline adversarial attack in the main text.

C More Visualizations

We present the additional visualization results, where we randomly select 5 images from the CASIA-WebFace and VGGFace2 datasets. We overlay the universal AdvHash and UTAP generated on ResNet34 onto these images. Subsequently, we employ ResNet34 (white-box) and VGG19 (black-box) model to perform image retrieval on these perturbed images (Query). UTAP's perturbation is constrained within the imperceptible range of $16/255$ in pixel value. The retrieval results are showcased in Fig. 2 and Fig. 3.

Notably, without any perturbations, both the ResNet34 and VGG19 models accurately return the retrieval results. In comparison to Fig. 5 in the main text, the outcomes depicted in Fig. 2 and Fig. 3 effectively demonstrate the universality of UTAP and AdvHash in white-box settings that a single adversarial patch/perturbation realizes effectiveness across all images and identities. In contrast, AdvHash exhibits near-complete failure in the black-box setting, while UTAP consistently achieves successful adversarial attacks, highlighting its superior transferability.

References

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In

Figure 2: Illustration of more universal and transferable retrieval effects of AdvHash and UTAP on CASIA-WebFace. The green box indicates the original identity and the red box indicates the wrong identity. The white-box model is emphasized with *.

Figure 3: Illustration of more universal and transferable retrieval effects of AdvHash and UTAP on VGGFace2. The green box indicates the original identity and the red box indicates the wrong identity. The white-box model is emphasized with *.

Proceedings of the IEEE conference on computer vision and pattern recognition, 9185–9193.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, 99–112. Chapman and Hall/CRC.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2730–2739.