Towards Feature Space Adversarial Attack by Style Perturbation

{ Qiuling Xu, Guanhong Tao, Siyuan Cheng, Xiangyu Zhang }

Introduction

Most existing adversarial attacks focus on perturbing pixels, i.e., pixel space attack. Instead, we propose adversarial attacks by perturbing individual features, i.e., feature space attack. We demonstrate the followings in this work:

1. Feature space attack has large and semantically meaningful perturbation.
2. An effective algorithm with key insights is proposed for the feature space attack.
3. Subtle features play an improperly important role in model prediction.
4. Defense on the pixel space attack is not sufficient for the feature space attack, and vice versa.

Motivation : Large and Semantic Change

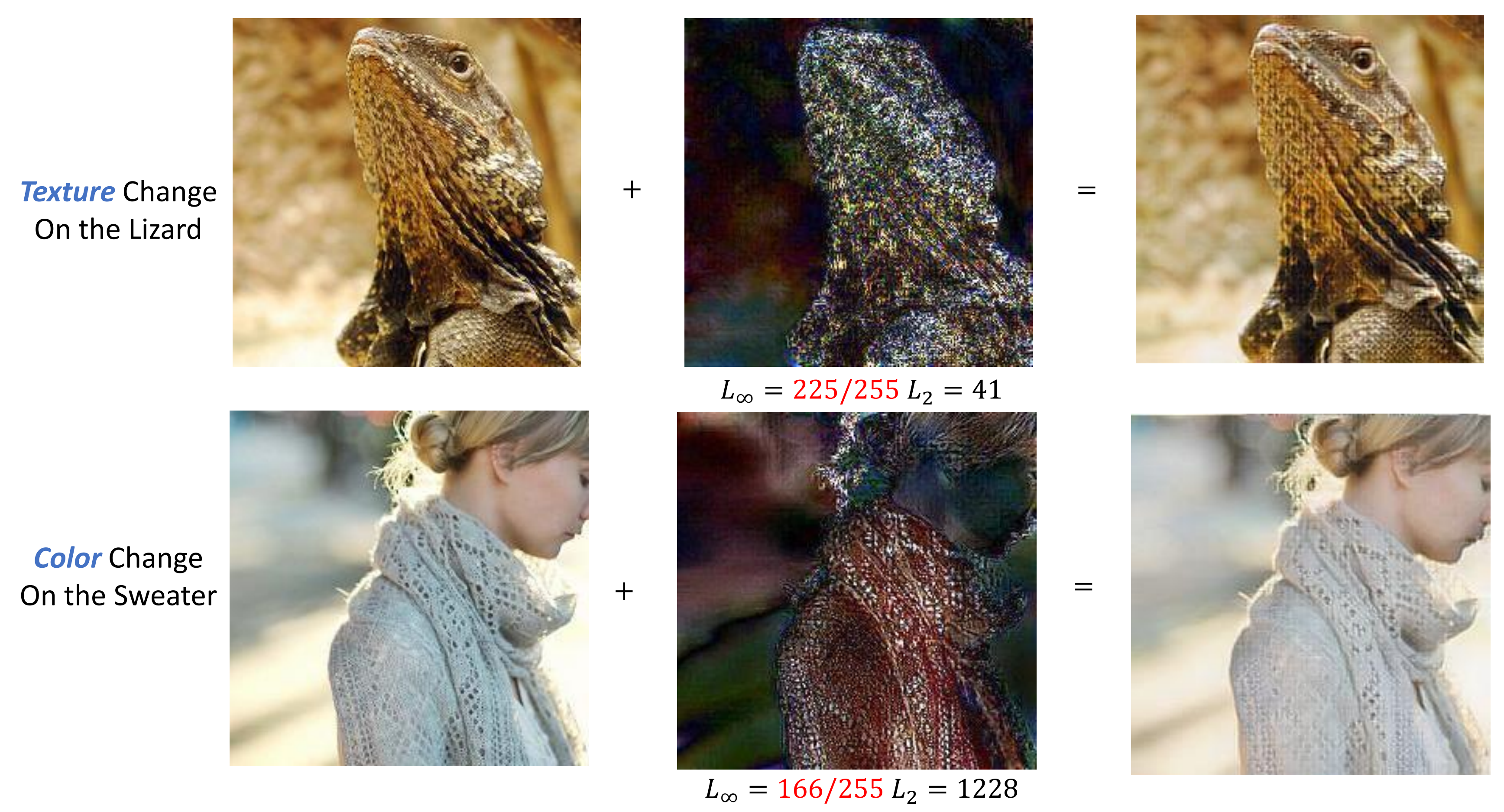

Texture Change
On the Lizard
+
=
$L_\infty = 225/255$ $L_2 = 41$
Color Change
On the Sweater
+
=
$L_\infty = 166/255$ $L_2 = 1228$

Figure 1: Feature Space Attack

Feature space attack has **large and semantically focused perturbation**. In contrast, the pixel space attack has small perturbation and looks more random.

Motivation : Subtle Features are Incorrectly Learned

During training, models utilize subtle features instead of apprehending the whole picture for classification. Attackers can thus hijack these inappropriately learned subtle features. This phenomenon makes feature space attack a better option in the black-box setting.

Table 1: Transferability

	ResNet′	VGG	MobileNet	DenseNet
Feature Space (Ours)	**68**%	**60**%	**48**%	**52**%
Pixel Space (PGD)	62%	40%	42%	30%

The table shows success rates under transfer attack. A higher value indicates the attack generalizes better.

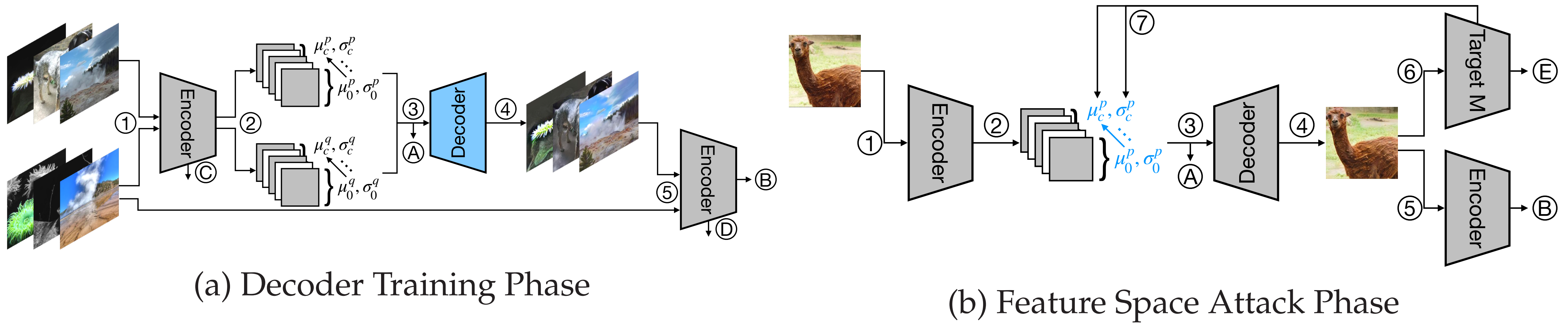
Future Work

In this work, we refrain ourselves from manipulating primary content features. Our follow-up work **[Q. Xu, G. Tao, and X. Zhang. D-square-b: Deep distribution bound for natural-looking adversarial attack, 2021]** further manipulates content feature by bounding deep features. And it achieves the state-of-the-art trade-off between attack success rate and naturalness.

Methods

We train an decoder to translate feature space perturbation to pixel level changes. Then we optimizes the feature space perturbations that causes misclassification to generate adversarial samples. Two key elements of the auto encoder are identified for the high-quality feature space attack:

1. Separation of the primary content feature and the secondary style feature.
2. Robust decoding in the presence of feature space perturbation. We add feature space perturbation in the training phase and penalize inaccurate translation in the attack phase.



(a) Decoder Training Phase

(b) Feature Space Attack Phase

Figure 2: Procedure of feature space adversarial attack. Two phases are involved during the attack generation process: (a) decoder training phase and (b) feature space attack phase.

Results

Attack	ImageNet		
	Denoise (t,1)	Denoise (u,1)	Denoise (u,5)
None	61.25	61.25	78.12
PGD	42.60	12.50	27.15
Decoder	64.68	64.00	82.37
Feature Space	**11.41**	**1.25**	**1.25**

Table 2: Evaluation of adversarial attacks against various defense approaches with a fair perturbation scale. State-of-the-art pixel space defenses are not sufficient for feature space attack.

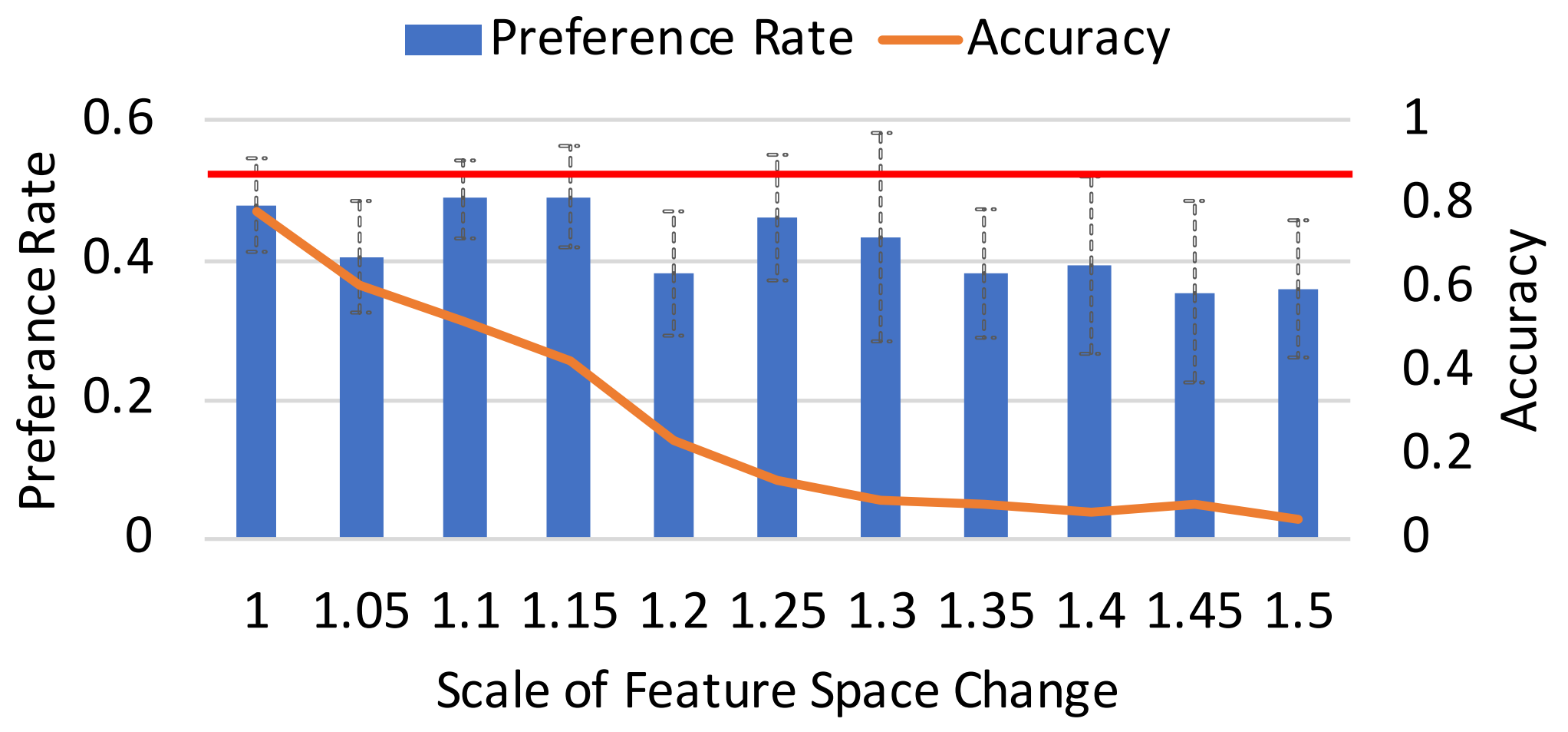


Figure 3: Human preference evaluation. The feature space adversarial samples are comparably favored by human.

Model	Accuracy	Success Rates			
		FS	HSV	SM	PGD
Normal	**92.1**%	100%	62.2%	100%	99.6%
PGD-Adv	78.1%	94.0 %	65.1 %	78.8 %	44.2 %
FS-Adv	82.4%	73.3 %	48.3 %	92.7 %	92.7 %

Figure 4: Adversarial training against the pixel space attack and the feature space attack. The results are colored blue when the defense and the attack are from the same family and are colored red when different. Feature space and pixel space are two different aspects of robustness.

Contact Information

Web qiulingxu.github.io **Email** xu1230@purdue.edu **Phone** +1 (765)-701-5968