Supplementary Material:
A Self-supervised Approach for Adversarial Robustness

Muzammal Naseer[*†‡], Salman Khan[†], Munawar Hayat[†], Fahad Shahbaz Khan[†§], Fatih Porikli[*]
[*]Australian National University, Australia, [‡]Data61-CSIRO, Australia
[†]Inception Institute of Artificial Intelligence, UAE, [§]CVL, Linköping University, Sweden
{muzammal.naseer,fatih.porikli}@anu.edu.au
{salman.khan,munawar.hayat,fahad.khan}@inceptioniai.org

We first explore why Self-supervised Perturbation (SSP) attack works in Appendix A. In Appendix B, we compare NRP with conventional adversarial training (AT) method known as feature denoising [17] in terms of adversarial robustness and defense training time. Differences of our proposed attack and defense from feature scattering [19] method are discussed in Appendix C. Ability of SSP to fool object detectors is compared against CDA [14] in Appendix D. We show that different transformation based defenses, JPEG, total variation minimization (TVM) and median filtering (MF) are not effective against SSP in Appendix E. Attack parameters against which our defense is evaluated are provided in Appendix F. Finally, we visually demonstrate NRP's ability to remove different kinds of adversarial perturbations in Appendix G.

Appendix A. Why Self-supervision Works?

Here, we highlight our intuition to create adversarial examples using feature space of VGG model [15].

- **Neural Style Transfer:** [5, 13] observed that the ability to transfer styles improves with AT, a phenomenon often related to VGG models [15] . On the other hand, VGG networks are more vulnerable to adversarial attacks [5]. A hypothesis was presented in [5] that perhaps VGG initial layers are as robust as adversarially trained models which allows better style transfer without AT.

- **Transferability of Natural vs. Robust Layers:** In addition to style transfer hypothesis [5], we explore the connection between layers of VGG and adversarially trained models in the context of adversarial attacks:

  - **Maximum Distortion of Non-Robust Features:** Datasets containing natural images contain both robust and non-robust features [9]. Robust features can be described by high level concepts like shape *e.g.* ear or noise etc., while non-robust features can arise from background or texture [7]. Ilyas *et al.* [9] argues that neural networks can pick-up on non-robust features to minimize the empirical risk over the given the data distribution and the transferability of adversarial examples can be explained by these non-robust features in different networks.

  - **Transferability:** VGG's ability to destroy non-robust features translates to better transferability even without any AT as compared to ResNet models (see Figures 1 and 2).

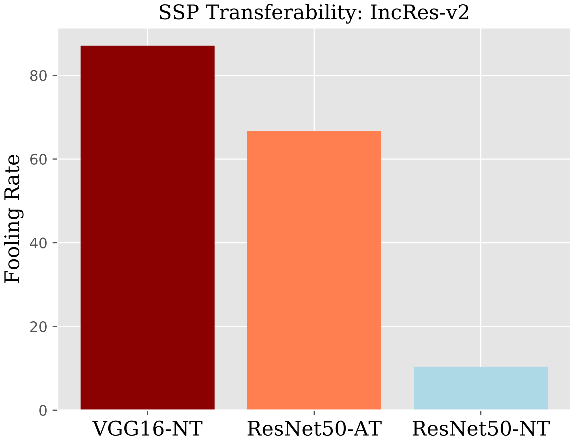


Figure 1: Fooling rate comparison is presented. NT and AT represent naturally and adversarially trained models, respectively. VGG16-NT and ResNet50-NT are trained on ImageNet while ResNet50-AT [6] is adversarially trained on a subset of ImageNet. Adversaries are created by applying distortion to the feature space of each model on NeurIPS dataset and then transferred to naturally trained IncRes-v2 [16]. Adversaries found in VGG space have higher transferability. In comparison, transferability of feature space of ResNet50 increases after adversarially training.

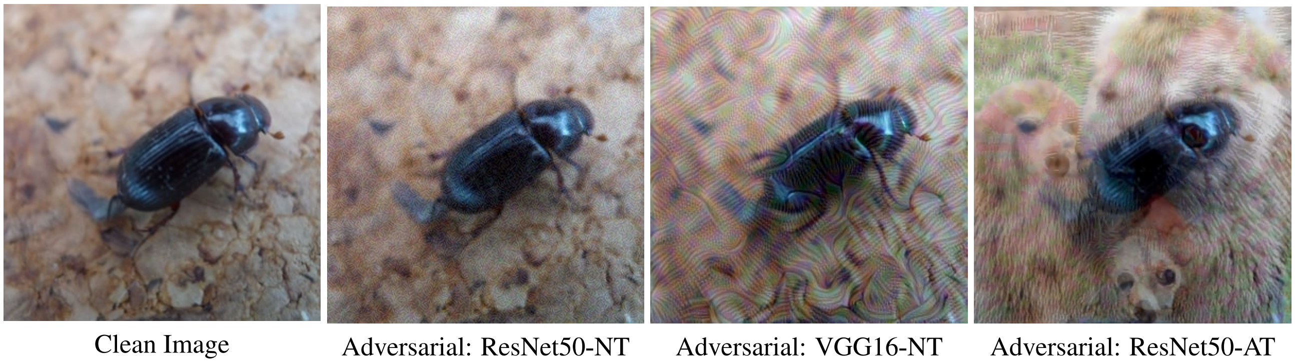
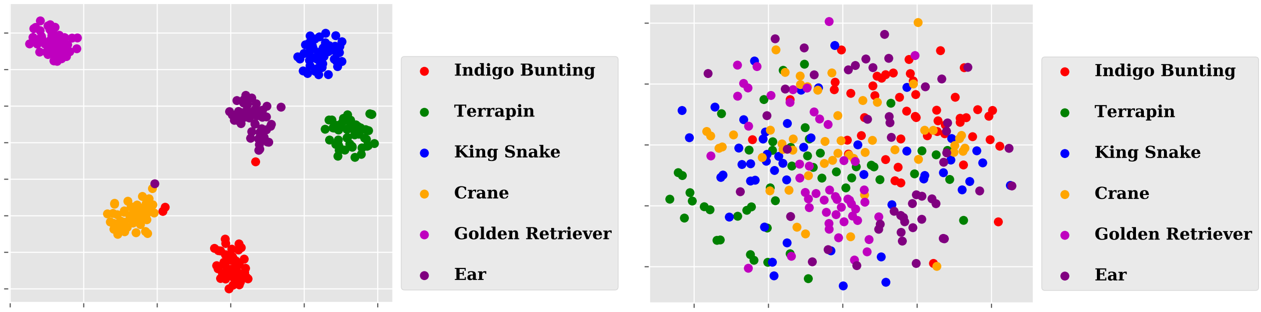


Figure 2: A visual demonstration of adversaries found by SSP in the feature space of diffrent networks. Perturbation buget is set to $l_\infty \leq 16$. NT and AT represent naturally and adversarially trained models, respectively.



(b) Logits

(b) Features

Figure 3: t-SNE [11] visualization of logits vs. feature representation of randomly selected classes from ImageNet validation set. Logits are computed from VGG16 [15] last layer while features are extracted from ”Block3-Conv3” of the same model. Our intuition is based on the observation that features space is shared among input samples rather than the logit space. Attacking such shared representation space removes task dependency constraint during adversary generation optimization and produces generalizable adversarial examples.

- **Shared Representation Space:** Our objective is to find adversarial patterns that can generalize across different network architectures trained for different tasks (*e.g.* classification, objection detection or segmentation). These are diverse tasks that do not share loss functions, dataset or training mechanism. Decision-boundary based attacks use model final response (*e.g.* logits in the case of classification) that is specific to input sample which leads to task-specific perturbations. A network's feature space, however, is shared regardless the input category. Therefore, perturbations found in such a space are highly generalizable (see Figure 3).

Appendix B. Comparison with AT

Conventional AT methods, such as [17], lose clean accuracy to gain adversarial robustness. Take an example of ResNet152 adversarially trained by [17]. In order to gain 55.7% robustness ($\epsilon \leq 16$) against targeted PGD attacks with ten number of iterations, the model clean accuracy drops from 78% to 65.3% which is even lower than VGG11. In contrast, our approach does not suffer from performance degradation on clearn samples.

Appendix B.1. Defense Results

To compare against [17], we ran ten number of PGD attack iterations. Labels for this targeted attack were chosen randomly as suggested by [17]. It is important to note that NRP can be turned into a dynamic defense, for example by first taking a random step in the input space and then projecting the modified input sample onto the perceptual space using our NRP. This way, NRP can be used to defend against attacks that try to incorporate NRP during attack optimization (a white box setting). We demonstrate this behavior in Table 1 by incorporating NRP in PGD attack using backpass approach introduced in [1]. Even for this challenging scenario, NRP shows significantly higher robustness than [17] while maintaining a higher clean accuracy. This highlights the benefit of self-supervision in AT.

Method	Clean	Adversarial
		$\epsilon \le 16/255$
Original	**78.31**	0.66
Feature Denoising[17]	65.3	55.7
NRP	73.5 ±1.5	**63.0 ±2.0**

Table 1: Defense success in terms of accuracy on ImageNet validation set (50k images). Higher is better.

Appendix B.2. Training Cost

Conventional AT methods like [17] depend on number of classes, dataset and task. In contrast, our defense is independent of such constraints. We describe the computational benefits of our defense with feature denoising based AT [17] in Table 2. Training time of our defense remains the same regardless of the backbone model while training time for [17] increases with the model size. In conclusion, conventional AT requires large amount of labelled data (*e.g.*, [17] is trained on 1.3 million images of ImageNet), while our defense can be trained on small unlabelled data (*e.g.*, 25k unlabelled MS-COCO images).

Method	No. of GPUs	Training Time	Task/Label Dependency	Dataset Specific
[17]	128	52	Yes	Yes
NRP	4	28	No	No

Table 2: Comparison of training time (hours) between NRP and AT on ResNet152 model [17].

Appendix C. Comparison with [19]

Defense comparison: Feature Scattering (FS) [19] based AT remains model and task-specific. Instead, our defense is independent to the target model and task, thereby providing better generalizability.

Attack comparison: The proposed attack FSA [19] operates in logit space in an unsupervised way by maximizing Optimal Transport distance, as compared to our SSP which operates in perceptual feature space (*e.g.*, VGG features). we compare the transferability of their attack with our SSP. As demonstrated in Table 3, SSP performs favorably well against FSA.

Appendix D. SSP vs. CDA

We compare our SSP with a recent transferable attack [14] in Table 4 on MS-COCO validation set using Mask-RCNN. mAP is reported with IoU = 0.5.

Attack	Inc-v3	Inc-v4	Res-152	IncRes-v2	Adv-v3	IncRes-v2$_{\text{ens}}$
FSA [19]	60.4	64.2	68.8	71.0	72.2	88.6
SSP	**5.3**	**5.9**	**16.5**	**14.1**	**25.9**	**58.0**

Table 3: Transferability ($\epsilon \le 16$) comparison of FSA with our attack (SSP). Results are reported for ImageNet-NeurIPS dataset. Lower is better.

Attack	$\epsilon \le 8/255$	$\epsilon \le 16/255$
CDA-ImageNet	35.2	**8.1**
CDA-Comics	40.5	16.8
CDA-Paintings	41.7	14.8
SSP	**31.8**	9.7

Table 4: SSP is compared with CDA [14]. Lower is better.

Appendix E. Effect of Input Transformations on SSP Attack

Different input transformations have been proposed to mitigate the adversarial effect. We have tested strength of SSP attack against well studied transformations including:

- *JPEG*: This transformation reduces adversarial effect by removing high frequency components in the input image.
- *Total Variation Minimization (TVM)*: TVM measures small variations thus it can be effective against relatively smaller adversarial perturbations.
- *Median Filtering (MF)*: This transformation filters out the input image by replacing each pixel with the median of its neighboring pixels.

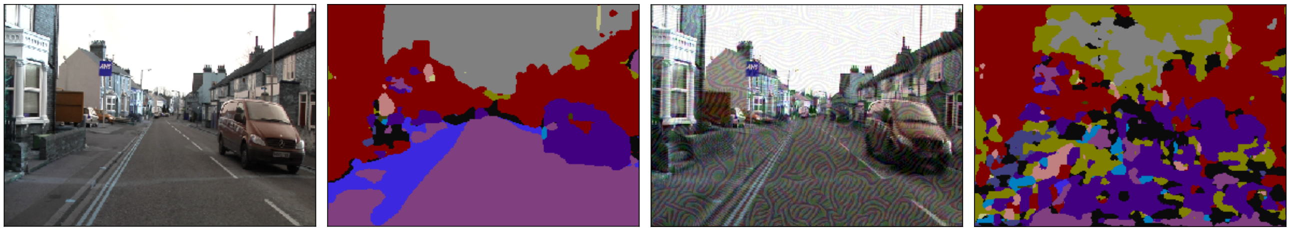
We report our experimental results on segmentation and object detection tasks.

Segmentation: SSP attack created on CAMVID [2] was able to bring down per pixel accuracy of Segnet-Basic by 47.11% within $l_\infty \le 16$ (see Table 6 and Figure 4). JPEG and TVM transformations are slightly effective but only at the cost of drop in accuracy on benign examples.

Object Detection: RetinaNet [10] collapses in the presence of adversaries found by SSP on MS-COCO validation set. Its mean average precision (mAP) with 0.5 intersection over union (IOU) drops from 53.78% to 5.16% under perturbation budget $l_\infty \le 16$ (see Table 7 and Figure 5). TVM is relatively more effective compared to other transforms against the SSP.

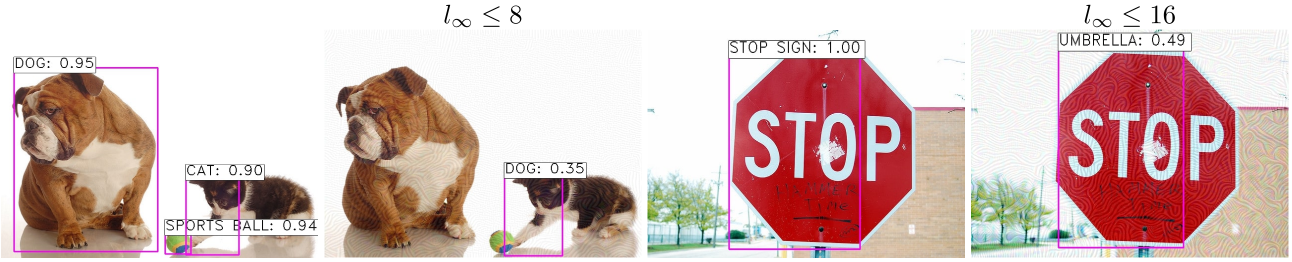
Appendix F. Attack Parameters

For FGSM, we use a step size of 16. For R-FGSM, we take a step of size α=16/3 in a random direction and then a gradient step of size 16$-\alpha$ to maximize model loss. The attack methods, I-FGSM, MI-FGSM and DIM, are run for 10 iterations. The step size for these attacks is set to 1.6, as per the standard practice. The momentum decay factor for MI-FGSM is set to 1. This means that attack accumulates



(a) Original (b) Prediction for Original (c) Adversarial (d) Prediction for Adversarial

Figure 4: Segnet-Basic output is shown for different images. (a) is the original image, while (b) shows predictions for the original image. (c) is the adversary found by SSP attack, while (d) shows predictions for the adversarial image. Perturbation budget is $l_\infty \leq 16$.



(a) Original (b) Adversarial (c) Original (d) Adversarial

Figure 5: RetinaNet detection results are shown for different images. (a) and (c) show detection for the original images, while (b) and (d) show detection for adversaries found using SSP attack.

all the previous gradient information to perform the current update and is shown to have the best success rate [3]. For DIM, the transformation probability is set to 0.7. In the case of FFF [12], we train the adversarial noise for 10K iterations to maximize the response at the activation layers of VGG-16 [15]. For the SSP, we used VGG-16 [15] conv3-3 feature map as the feature loss. Since SSP generation approach maximizes loss w.r.t a benign example, it does not suffer from the over-fitting problem. We run SSP approach for the maximum number of 100 iterations. The transferability of different attacks is compared against the number of iterations in Figure 6. MI-FGSM and DIM quickly reach to their full potential within ten iterations. The strength of I-FGSM strength decreases, while feature distortion strength (SSP) increases with the number of attack iterations. Top-1 (T-1) and Top-5 (T-5) accuracies of Imagenet trained models on NeurIPS dataset are reported in Table 5.

Appendix G. Generalization to Unseen Attacks

We show visual demonstration (see Figures 7, 8, 9 and 10) of how our defense, NRP, trained using SSP attack is able to generalize on the variety of unseen perturbations created by different attack algorithms. NRP successfully removes the perturbations that it never saw during training.

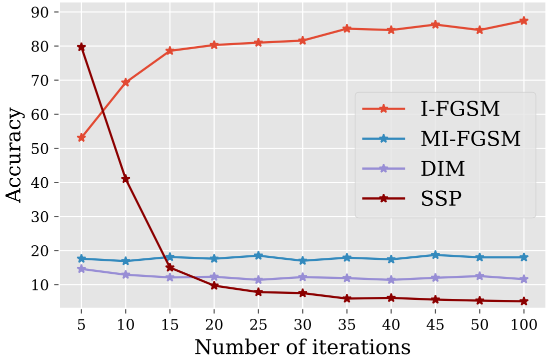


Figure 6: Accuracy of Inc-v3 for adversaries created on VGG-16 by different attacks. SSP's strength increases with number of iterations, in contrast to MI-FGSM and DIM.

- Figure 7 shows adversaries coming from adversarially robust model. It's the most difficult case as perturbations does not resemble to a noisy patter rather represent meaningful structured pattern that are in-painted into the clean image. NRP's ability to remove such difficult patterns shows that our defense can separate the original signal from the adversarial one.
- NRP has no difficulty in removing thick patterns introduced by DIM or smooth perturbations of DIM-TI attacks (Figure 8).

Table 5: Model accuracies are reported on original data set ImageNet-NIPS containing benign examples only. T-1: top-1 and T-5: top-5 accuracies. Best performances are shown in bold.

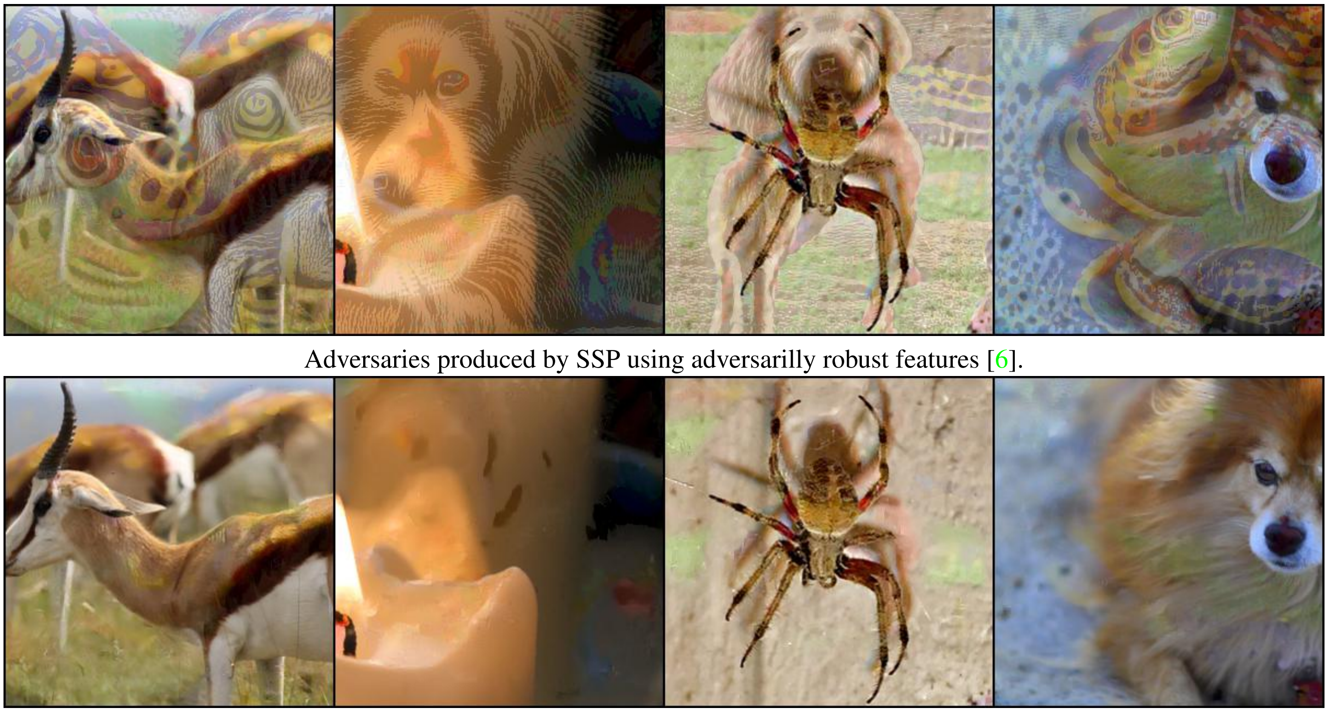
Accuracy	Naturally Trained					Adv. Trained		
	Inc-v3	Inc-v4	Res-152	IncRes-v2	VGG-19	Adv-v3	Inc-v3$_{\text{ens3}}$	IncRes-v2$_{\text{ens}}$
T-1	95.3	97.7	96.1	**100.0**	85.5	95.1	93.9	97.8
T-5	99.8	99.8	99.9	**100.0**	96.7	99.4	98.1	99.8

Table 6: Segnet-Basic accuracies on CAMVID test set with and without input transformations against SSP. Best performances are shown in bold.

Method	No Attack	SSP	
		$l_\infty \leq 8$	$l_\infty \leq 16$
No Defense	**79.70**	52.48	32.59
JPEG (quality=75)	77.25	51.76	32.44
JPEG (quality=50)	75.27	52.45	33.16
JPEG (quality=20)	68.82	53.08	35.54
TVM (weights=30)	73.70	55.54	34.21
TVM (weights=10)	70.38	**59.52**	**34.57**
MF (window=3)	75.65	49.18	30.52

Table 7: mAP (with IoU = 0.5) of RetinaNet is reported on MS-COCO validation set with and without input transformations against SSP. Best performances are shown in bold.

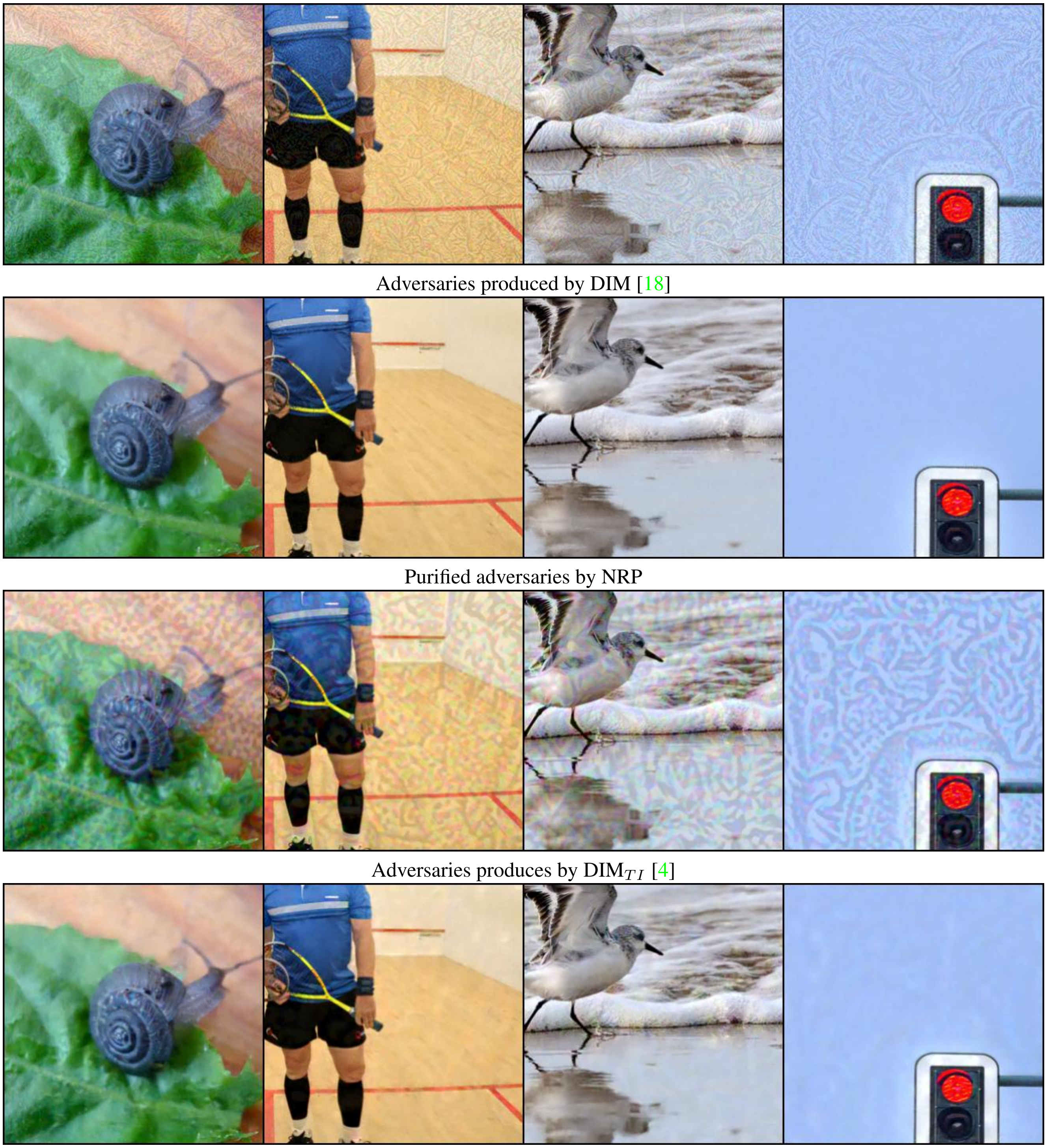
Method	No Attack	SSP	
		$l_\infty \leq 8$	$l_\infty \leq 16$
No Defense	**53.78**	22.75	5.16
JPEG (quality=75)	49.57	20.73	4.7
JPEG (quality=50)	46.36	19.89	4.33
JPEG (quality=20)	40.04	19.13	4.58
TVM (weights=30)	47.06	27.63	6.36
TVM (weights=10)	42.79	**32.2**1	**9.56**
MF (window=3)	43.48	19.59	5.05



Adversaries produced by SSP using adversarilly robust features [6].

Purified adversaries by NRP.

Figure 7: NRP is capable to remove these difficult adversaries where adversarial image is in-painited into the clean image. Untargetted adversaries are created by applying SSP to feature space of adversarially trained ResNet50 [6]. Perturbation budget is set to $l_\infty \leq 16$.



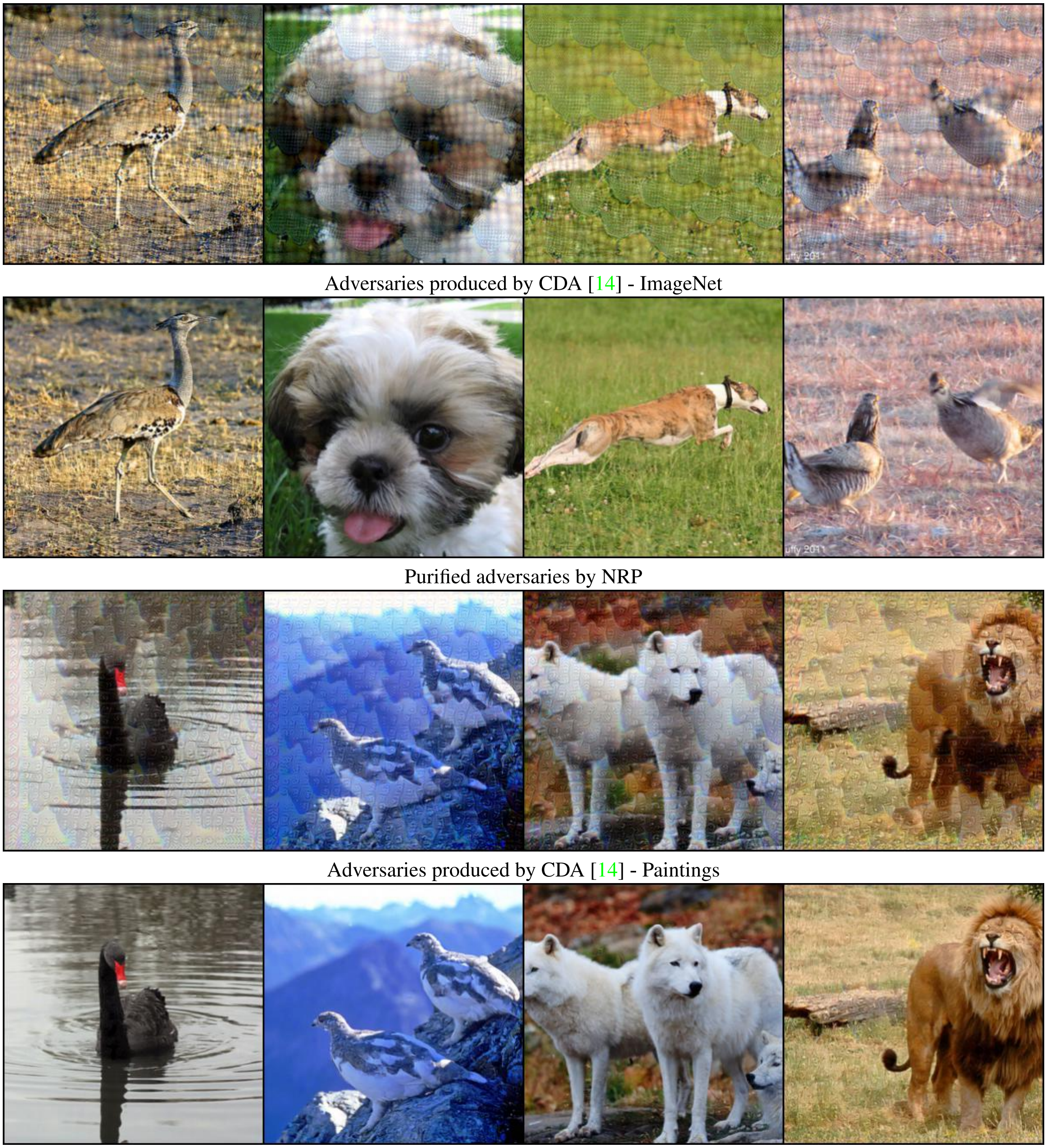
Adversaries produced by DIM [18]

Purified adversaries by NRP

Adversaries produces by DIM$_{TI}$ [4]

Purified adversaries by NRP

Figure 8: NRP removes diverse patterns produces by DIM [18] and translation-invariant attacks [4] to a great extent. Untargetted adversaries are created by ensemble of **ensemble** of Inc-v3, Inc-v4, IncRes-v2, and Res-152. Perturbation budget is set to $l_\infty \leq 16$.



Adversaries produced by CDA [14] - ImageNet

Purified adversaries by NRP

Adversaries produced by CDA [14] - Paintings

Purified adversaries by NRP

Figure 9: Our defense successfully able to recover original samples from unseen adversarial patterns. These are untargeted adversaries produced by CDA [14] trained against Inc-v3 on ImageNet and Paintings. Perturbation budget is set to $l_\infty \leq 16$.

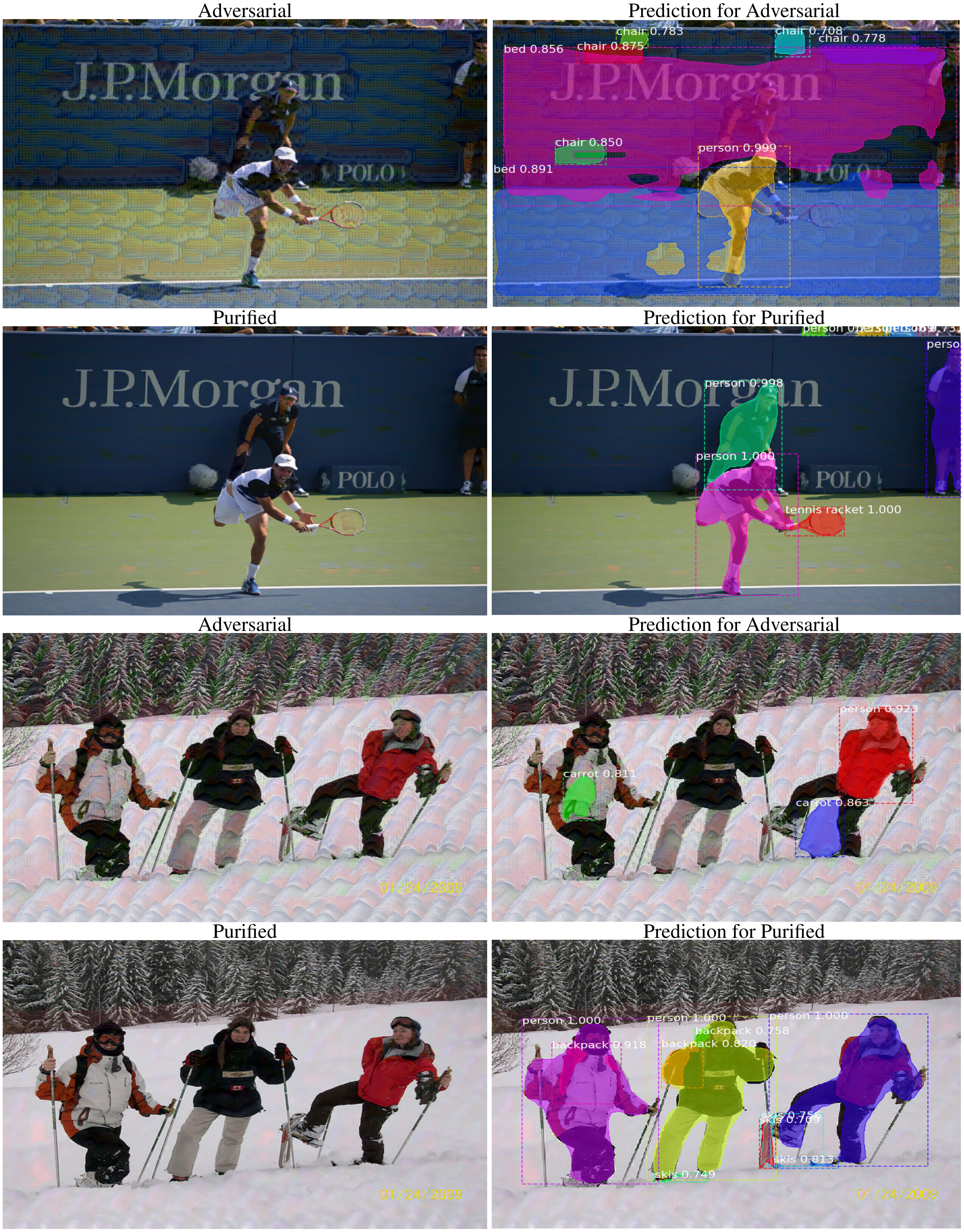


Figure 10: Adversaries generated by CDA [14] reduce Mask-RCNN [8] performance. NRP successfully removes adversarial perturbations and greatly stabilizes Mask-RCNN predictions.

References

[1] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018. 2

[2] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. 3

[3] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 4

[4] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019. 6

[5] Logan Engstrom, Justin Gilmer, Gabriel Goh, Dan Hendrycks, Andrew Ilyas, Aleksander Madry, Reiichiro Nakano, Preetum Nakkiran, Shibani Santurkar, Brandon Tran, Dimitris Tsipras, and Eric Wallace. A discussion of 'adversarial examples are not bugs, they are features'. *Distill*, 2019. 1

[6] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations, 2019. 1, 5

[7] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 1

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 8

[9] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019. 1

[10] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 3

[11] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 2

[12] Konda Reddy Mopuri, Utsav Garg, and R Venkatesh Babu. Fast feature fool: A data independent approach to universal adversarial perturbations. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017. 4

[13] Reiichiro Nakano. Neural style transfer with adversarially robust classifiers, Jun 2019. 1

[14] Muzammal Naseer, Salman H Khan, Harris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-domain transferability of adversarial perturbations. *Advances in Neural Information Processing Systems*, 2019. 1, 3, 7, 8

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. 1, 2, 4

[16] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017. 1

[17] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 501–509, 2019. 1, 2, 3

[18] Cihang Xie, Zhishuai Zhang, Jianyu Wang, Yuyin Zhou, Zhou Ren, and Alan Yuille. Improving transferability of adversarial examples with input diversity. *arXiv preprint arXiv:1803.06978*, 2018. 6

[19] Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. *arXiv preprint arXiv:1907.10764*, 2019. 1, 3