Increasing Adversarial Robustness Around Uncertain Boundary Regions with Amodal Segmentation

Sheila Alemany*, Niki Pissinou†, Brian Yang‡
*†School of Computing and Information Sciences, Florida International University
‡College of Engineering, University of Michigan
{*salem010, †pissinou}@fiu.edu, ‡byyang@umich.edu

2024 International Conference on Machine Learning and Applications (ICMLA) | 979-8-3503-7488-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICMLA61862.2024.00290

***Abstract*—Adversarial perturbations in object recognition and image classification tasks impact a learning model's ability to perform accurately and increase the safety risks of deployed machine learning models. During the adversarial example generation process, adversaries approach areas most prone to model uncertainty. Identifying partially occluded items, especially without understanding general object shapes, contributes to significant model uncertainty since object boundaries are not inherently at the forefront of the feature generalization process in deep learning models. Thus, this work aims to reduce model uncertainty surrounding partially occluded boundaries and increase adversarial robustness by augmenting the training dataset with amodal segmentation boundary masks. By observing performance degradation, robust sensitivity, and loss sensitivity, we show how including these masks during training impacts an adversary's ability to generate effective adversarial examples on the versatile MS COCO dataset. Lastly, we observe how including these masks during training influences the performance of adversarial training.**

I. Introduction

Adversaries have been found to exploit image boundaries during computer vision and image processing tasks since object boundaries are not at the forefront of feature generalization in learning models [1]. These boundaries refer to the edges or transitions between different objects, regions, or features in an image. Various adversarial attacks have been shown to modify image pixels, changing boundaries to represent other classes in the dataset [2].

Amodal segmentation is the result of the estimation of the boundary shape of an object beyond a visible region and the mask for the occluded region [3]. These boundaries are often challenging to decipher in images because strong contextual knowledge is needed to correctly interpret one or more occluded items in images. This can be challenging even in human interpretation contexts because our environment is dynamic, complex, and unorganized. The capability of humans to perceive incomplete objects is called amodal completion, allowing humans to have an easier time classifying partially occluded items [4]. Unfortunately, this task is not as straightforward for computers since occlusion can happen in various ratios, angles, viewpoints, and lighting environments [4]. For a simple example, suppose an apple partially covers the center of a banana. In that case, it is challenging to identify the separate disjointed pixels that correspond to the singular banana without understanding the general shape of the banana and, thus, the areas where the banana ends and the apple starts are the areas that are prone to the highest uncertainty.

During the adversarial generation process, adversaries tend to iteratively approach areas most prone to uncertainty. Visually, the areas of uncertainty in an image have been visualized through a saliency map, where the areas of highest uncertainty contain the most significant increase in slope surrounding those pixels. Gradient- and Jacobian-based attacks using this information to optimize their adversarial examples [5]. Identifying partially occluded items, especially without a proper understanding of the general object shapes, contributes to significant model uncertainty. As a result, this work reduces the impact of minor adversarial perturbations by augmenting an image dataset with boundary masks to strengthen the contextual information around the boundaries of images.

This research effort shows how including these masks during the image classification training phase impacts an adversary's ability to generate effective adversarial examples on the Microsoft Common Objects in COntext (MS COCO) [6] dataset. We observe the attack success (i.e., how much performance accuracy degrades), the robust sensitivity [7], and loss sensitivity [8] to analyze how the trained generalized information model changes when augmenting the training dataset with amodal boundary masks. Lastly, we also observe how including these masks influences the performance of adversarial training.

II. Related Work

The research field of adversarial machine learning is vast. As a result, we include the relevant literature regarding the current state of knowledge surrounding the vulnerable input spaces and the current state of amodal segmentation models. This work builds on these two research areas to implement and discuss the impact of including amodal segmentation contextual information on adversarial vulnerability.

A. Vulnerable Input Spaces: Occluded Items

The study of machine learning tasks with occluded items has lied mainly in the research field of generative learning, where learning models are tasked to generate images containing various partially occluded items [4]. However, these efforts provide insight into the vulnerable input spaces in images. For instance, the quality of images through resolution enhancement and occlusion handling has been improved by authors in [9] to

1946-0759/24/$31.00 ©2024 IEEE
DOI 10.1109/ICMLA61862.2024.00290

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on March 13,2025 at 19:15:47 UTC from IEEE Xplore. Restrictions apply.

overcome attribute classification challenges (e.g., telling two items of the same type but with different unique attributes). Multiple inpainting approaches also exist to increase the quality of images to be used in image classification tasks [10], [11], [12], [13]. These techniques aim to restore the corrupted regions of occluded objects due to low image resolution, extreme lighting variations, occlusion, or even disguise. They highlight how contextual information influences the generalization of information in image classification tasks and makes it more challenging for learning models to classify occluded items correctly.

Autonomous driving applications of machine learning also highlight how occluded items contribute to high model uncertainty since they have to make decisions with incomplete information about a particular object [14], [15]. Specifically regarding adversarial perturbations, authors in [2] deterministically evaluated how much a region of input space is vulnerable to adversarial perturbations relative to other areas of the complete input space. They observed that occluding input spaces where boundary masks are most similar to other classes are the input spaces most vulnerable to perturbations. Overall, occluded items in natural environments contain too much or too little information regarding the context surrounding an object in an image. Thus, we evaluate whether amodal segmentation can decrease uncertainty by highlighting the shape of the *classifying items* (e.g., including only masks of animals if the tasks is to classify animals). Through this approach, we augment the training dataset with boundaries between multiple occluded items, reducing unnecessary context that may be present in an image and, consequently, reducing a learning model's sensitivity to adversarial perturbations.

B. Amodal Segmentation Models

Three main methods have been used for image segmentation: traditional methods, CNN-based methods, and weakly supervised methods [3]. Traditional methods were the initial methods proposed for amodal shape completion [3]. These techniques assume the most likely shape or curve of an invisible region and have limited abilities to generate accurate masks. CNN-based methods use deep learning for their occlusion handling, but the ability to predict an amodal mask is inversely proportional to how many occluded objects are included. Thus, various techniques use image patching and heavily labeled datasets to output accurate amodal masks. For instance, authors in [16] use a Bayesian approach for their amodal segmentation model that shows to be more robust to occluders and could successfully generalize out-of-distribution items when trained with non-occluded objects. However, this approach relies on 2D shape priors for its high accuracy. This work highlights the benefits of including complete information regarding non-occluded items for lower uncertainty and high-quality generalization. Similarly, Mask R-CNN [17], [18] has also been a successful technique to generate segmentation masks building on a Region-based CNN (R-CNN) that provides a class label and a bounding-box offset to allow for precise extraction of the mask and spatial layout of an object.

Weakly supervised techniques have been proposed to explore ways to relieve supervision demand for amodal mask completion. For example, the authors in [19] proposed a weakly supervised method for estimated amodal segmentation where they specifically estimated the occlusion boundary in addition to the occlusion mask. The occlusion boundary highlights how the areas of uncertainty lie in the pixels where two or more objects overlap. However, this approach has issues separating occlusions from the foreground and background, sometimes merging two items. In general, amodal segmentation models are computationally expensive models that aid in object detection computer vision tasks with limitations regarding the accuracy of the occlusion handling and identifying where they are relatively located in the input space [3].

III. Evaluation Methodology

A. Contextual Dataset

Due to the challenging nature of amodal segmentation, various datasets for amodal segmentation and object recognition are limited in their dataset evaluations. We utilize the Microsoft Common Objects in COntext (MS COCO) [6] dataset, which contains various items and contexts. The MS COCO dataset benefits image object recognition tasks where there is superfluous contextual information surrounding the classifying objects [6]. This was achieved by including images of complex everyday scenes containing common items in their natural real-world context. The dataset contains photos of 90 objects with 2.5 million labeled instances in 328k images. The 2014 release originally included 82,783 training images and 40,775 testing images. There are approximately 270,000 segmented people and 886k segmented object instances in the 2014 train and validation sets, including 80 objects corresponding to 12 superclasses.

This work used a subset of 92,544 images containing 4 superclasses (e.g., 'animal', 'food', 'vehicle', and 'person') for the training and validation sets to simplify the original learning task without reducing contextual information. Table I includes the sub-class labels in the superclasses used for this evaluation to provide insight into the types of items included in this dataset. We specifically selected a subset of the dataset with superclasses whose amodal shapes are generally distinct between classes. Lastly, significant existing amodal segmentation of occluded items focuses on the occlusion handling of 2D images and excludes 3D and video data [1]. Thus, the dataset used in this work follows the same assumptions.

B. Amodal Segmentation Mask Generation

Considering that amodal segmentation mask generation models are generally computationally expensive, we selected the CNN-based Mask R-CNN model for instance segmentation. Mask R-CNN is an improved version of Fast R-CNN to take advantage of its quick training speeds. It classifies individual objects and localizes each using a bounding box and semantic segmentation, where this approach compartmentalizes each pixel into a fixed set of categories without differentiating object instances with relatively fewer computational resources

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on March 13,2025 at 19:15:47 UTC from IEEE Xplore. Restrictions apply.

TABLE I: Subset of the MS COCO superclass labels and their corresponding sub-class labels, enumerating the items included in images for our evaluation.

Superclass Label	**Class Labels**
'animal'	'bird', 'cat', 'dog', 'horse', 'sheep', 'cow', 'elephant', 'bear', 'zebra', 'giraffe'
'food'	'banana', 'apple', 'sandwich', 'orange', 'broccoli', 'carrot', 'hot dog', 'pizza', 'donut', 'cake'
'vehicle'	'bicycle', 'car', 'motorcycle', 'airplace', 'bus', 'train', 'truck', 'boat'
'person'	'person'

Fig. 1: Snapshot of training data samples for the augmented MS COCO dataset.

to run compared to other existing object handling models. This framework can run at 200ms per frame on a GPU, and training on MS COCO takes one to two days on a single 8-GPU machine. Overall, Mask R-CNN provides state-of-the-art amodal masks to include in this work.

We include generated amodal segmentation masks in the training dataset and then train a learning model to classify the 4 superclasses. The boundary masks focus on the boundaries of the class labels (e.g., the images in the 'animal' class include shapes corresponding to birds, cat, etc.). Figure 1 includes a snapshot of the training data samples for the augmented MS COCO dataset. The augmented MS COCO dataset contained 113,214 total samples spanning those 4 superclasses, meaning we generated a total of 22,642 masks using the Mask R-CNN approach across all categories. We did not use a 1-to-1 ratio of original images to masks to minimize the computational resources necessary and ensure that the classification accuracy did not degrade. Minimizing the required computational resources includes less mask generation and less training time. We observed approximately 15-20% of the training and validation sets with masks increase robustness against adversarial examples in this evaluation, even when using adversarial training, as shown in Section IV. The results in Section IV reflect the average across training sets including 15-20% of the training and validation sets with masks. The number of masks to include in a training set will vary per dataset and learning model.

C. Learning Model

The training architecture for this work was ResNet-50 [20], a convolutional neural network architecture that was developed for large image datasets with high performance accuracy. For the MS COCO dataset, we achieved over 95% accuracy with a learning rate of 0.001 in both the augmented training dataset and the original training dataset. To achieve this level of accuracy, both models were equally trained for 80 epochs with a batch size of 32. We selected this training environment to ensure the highest performance across all the models that were trained without modifying any training parameters outside of our control variable of the augmented training dataset with the amodal masks. Lastly, all images, including the amodal masks, were shaped into squares of size 180x180 for training to add homogeneity since the original MS COCO dataset and the generated masks were all varying shapes and sizes.

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on March 13,2025 at 19:15:47 UTC from IEEE Xplore. Restrictions apply.

D. Adversarial Example Generation

Assuming a threat model with full-white box access, the adversarial examples are created to cause the machine learning model to produce a specific incorrect output, known as a targeted attack. Specifically, for a correct label for $f(x) \in M$, we use attack strategies that generate a perturbed input $\hat{x} = x + \eta$ such that $f(x) \neq f(\hat{x})$ where η is the imperceptible adversarial noise. These malicious perturbations can be optimized using a set of different strategies.

To generate the adversarial examples, we used the Fast Gradient Sign Method (FGSM) [21], and Projected Gradient Descent (PGD) [22] from the Adversarial Robustness Toolbox (ART) by IBM Research [23] with no changed hyperparameters outside of the varying perturbation budgets. These attacks are both gradient-based attacks that will optimize adversarial examples around the pixels with the highest uncertainty using the gradient of the loss function.

E. Metrics Used

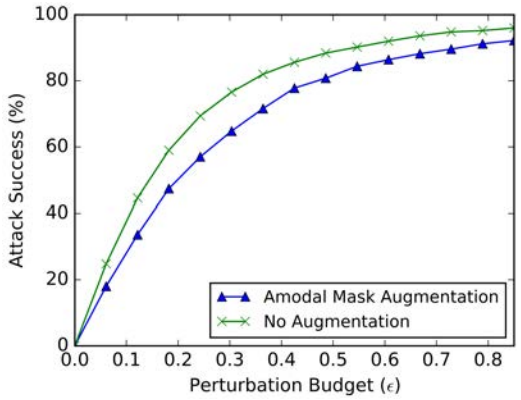
1) Attack Success: A standard metric for evaluating model robustness is the performance accuracy as a function of the perturbation budget ϵ [24]. Attack success is defined as an adversary's ability to reduce model accuracy: attack success $= (1-\text{model accuracy}) * 100$, meaning that 100% attack success signifies that the adversarial examples brought the performance accuracy down to 0%. This metric verifies whether a gradient-based adversarial attack is properly configured for a classification task by observing an attack success that approaches 100% as the perturbation budget increases [24]. To ensure that we evaluate an adversary's impact on the accuracy, we only perturb and evaluate 500 random benign input samples that were correctly classified before the model was attacked.

2) Empirical Robustness: Authors in [7] proposed the measure of empirical robustness by evaluating the minimum perturbation necessary to craft a successful attack (e.g., cause a misclassification). We estimated the minimal perturbation by calculating the sign of the gradient:

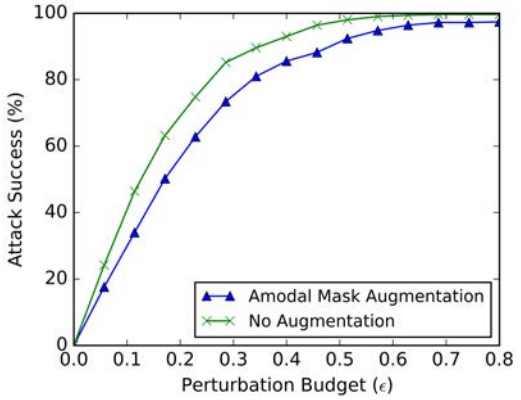
$$\hat{\rho}_{adv}(f) = \epsilon \text{sign}(\nabla_x J(\theta, x, y)) \tag{1}$$

where J is the cost used to train the learning model, θ are the model parameters, and y is the classification label of the input x. The method used to approximate this value is FGSM.

We employed the empirical robustness measure implementation by IBM's ART [25] with no changed hyperparameters. We calculated the empirical robustness values over 500 random benign input samples correctly classified before the model was attacked. We compared the distribution of the empirical robustness values of the learning models trained with the original, unmodified training data, the augmented dataset, and adversarial training. Increasing the empirical robustness is an indication of increased robustness. However, a value of 0 (e.g., no perturbation change) implies that in the no perturbation less than ϵ was reached that caused a misclassification.



(a) FGSM



(b) PGD

Fig. 2: Comparing performance of the FGSM attack with the MS COCO dataset on the ResNet-20 architecture

3) Loss Sensitivity: Authors in [8] proposed a loss sensitivity metric that measures the impact of each sample on the average loss. They accomplish this by measuring the norm of the loss gradient for a previous input x after t number of stochastic gradient descent (SGD) updates. The loss sensitivity function is defined as:

$$g_x^t = ||\partial L_t / \partial x||_1 \tag{2}$$

where L_t is the loss after t updates. The average g_x^t over a T set of SGD updates is the loss sensitivity value.

Similarly to the empirical robustness metric, we utilized the loss sensitivity implementation by IBM's ART [25] and calculated the loss sensitivity values over each of the 500 random benign input samples correctly classified before the model was attacked and compared against our various configurations. Analyzing the density function of this loss sensitivity allows us to observe how much adversarial perturbations impact performance accuracy across the different training scenarios.

IV. Results & Discussion

Figures 2(a) and 2(b) compare the two neural networks, one trained on an augmented training dataset with boundary masks

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on March 13,2025 at 19:15:47 UTC from IEEE Xplore. Restrictions apply.

and the other without the training dataset augmentations. Figure 2(a) includes the attack success against the FGSM attack for perturbation budgets $\epsilon \in (0, 0.8]$. Under the FGSM attack, the model trained with amodal mask augmentation consistently accomplished a lower attack success for all perturbation budgets, an average robustness improvement of approximately 7.28% performance accuracy.

Similarly, Figure 2(b) includes the attack success against the PGD attack for perturbation budgets $\epsilon \in (0, 0.8]$. Under this attack, the model trained with amodal mask augmentation also consistently accomplished less attack success for all perturbation budgets with an average improvement of approximately 6.65% performance accuracy. For both attacks, we achieved this reduction of attack success only augmenting the training data with 18% of added amodal boundary masks. Since both of these models were trained using 80 epochs, the only added computational costs were those of generating the masks on the previously trained Mask R-CNN model.

A. Stacking with Adversarial Training

Adversarial training generates and includes adversarial examples during training to encourage learning models to familiarize themselves with worst-case, malicious perturbation inputs [22]. This results in a shift in the distribution of the training dataset with heavy cross-over between benign and malicious examples [26], [27]. Since adversarial attacks tend to exploit the context and boundaries surrounding partially occluded objects, the adversarially trained learning model should gain awareness surrounding these areas prone to high uncertainty and increase overall robustness. Thus, we evaluate the ResNet-50 models similarly to Figure 2. However, we include adversarial training during the training phase and observe whether adversarial training would benefit from having additional object boundary context for a more considerable increase in robustness.

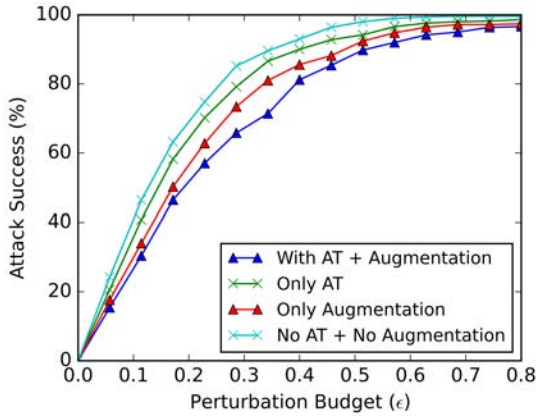


Fig. 3: Comparing performance of the PGD attack with the MS COCO dataset on the ResNet-20 against our augmented training dataset and adversarial training (AT).

Figure 3 shows that adversarial training benefits from the added information surrounding the image's occluded and non-occluded items for the PGD attack. Interestingly, we can see that the model trained with only the amodal augmentation performed better than the model with only adversarial training, with an average difference of 3.6%. Adversarial training has been attributed to not prioritizing high-quality generalization, which leads to areas of high uncertainty despite the significant increase in robustness and results in perhaps fewer adversarial examples that cause misclassification but ones with significant attack success [28]. On the other hand, amodal mask augmentation prioritizes minimizing the uncertainty in the areas prone to the highest uncertainty, resulting in adversarial examples of less magnitude. When adversarial training is the only defense deployed, there is a difference of 7% when compared to both defenses are stacked together. Overall, these complementing features of each defense seem to be highlighted when joined together for the highest level of robustness, with an average reduction of attack success of 10% compared to no implemented defenses.

B. Empirical Robustness

Table II summarized the empirical robustness metric values for the varying training configurations for the 500 input samples. When we calculated the empirical robustness metric, the stacking of the both defenses (i.e., adversarial training and boundary mask augmentations) resulted in the most samples that could not be perturbed to successfully cause a misclassification. Only including the augmentations performed comparably with about 9.6% more empirical robustness than adversarial training alone. The results are consistent for successful adversarial perturbations with a higher average perturbation signifying that larger perturbations were more often necessary to successfully attack.

Figure 4(a) shows the density distribution of the empirical robustness metric values to observe the distribution of perturbation magnitudes for a successful attack for all 500 samples. We can see that using no defense had the highest amount of samples successfully attacked with minimal perturbations and adversarial training, having only slightly improved performance from that baseline. These results are consistent with the understanding that adversarial training does not reduce the areas of high uncertainty that adversaries can exploit for stealthy perturbations [28]. Overall, stacking both defenses contributes to the highest levels of robustness, but when selecting between adversarial training or boundary mask augmentations, the masks reliably provide higher adversarial robustness.

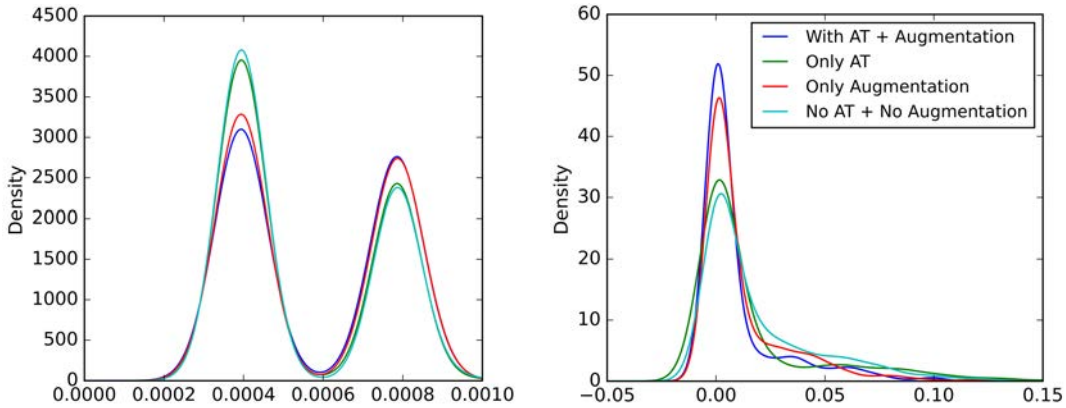
C. Loss Sensitivity

Figure 4(b) shows the density distribution of the loss sensitivity metric. We can see that adversarial training with boundary mask augmentations has the least number of samples with loss sensitivity, meaning that a random sample x is more likely to have a loss sensitivity value close to 0 than the other training configurations. Using both adversarial training and boundary mask augmentations decreases the overall loss sensitivity by 47.6%. Only using the amodal boundary masks decreases the overall sensitivity by 37.7%. In this figure, we also

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on March 13,2025 at 19:15:47 UTC from IEEE Xplore. Restrictions apply.

Training Configuration	$\hat{\rho}_{adv} = 0$ (%)	Average $\hat{\rho}_{adv}$ ($\hat{\rho}_{adv} \neq 0$)	Maximum $\hat{\rho}_{adv}$
With AT & Augmentation	52.8%	0.000581	0.00117
Only AT	40.8%	0.000545	0.00095
Only Augmentation	49.6%	0.000575	0.00089
No AT & No Augmentation	36.2%	0.000542	0.00096

TABLE II: Insight into empirical robustness metric values for the varying training configurations.



(a) Empirical Robustness (b) Loss Sensitivity

Fig. 4: Comparing the empirical robustness and loss sensitivity of the varying training configurations. The legend corresponds to both subfigures.

see that only using the amodal boundary masks contributes to a model with less overall loss sensitivity than adversarial training alone, with a 27.7% decrease in loss sensitivity. These results show the significant influence of model uncertainty around objects' boundary context in image classification. Without providing boundary masks for all the training inputs, we could reduce the overall loss sensitivity, contributing to a learning model with more predictable outputs and less susceptible to minor adversarial perturbations than the outstanding adversarial training approach.

V. Conclusions & Future Work

Our results conclude that increasing the data context around the boundaries of objects in images through amodal masks, partially occluded or not, forces the generalized information model to better understand the shapes of the items that correspond to the correct classification. Amodal segmentation was previously a task often used for object recognition tasks outside of the research area of adversarial ML to encourage learning models to focus on the item in question and not the distracting environmental information surrounding the object. The applications of the amodal masks in the training process highlight how improving the quality of the data context in learning tasks can significantly suppress the impact of minor adversarial perturbations on feature characteristics that adversaries often exploit. Our data augmentation was also evaluated with and against adversarial training. The results show that the model trained with only the amodal augmentation performed better than the model with only adversarial training, highlighting how prioritizing high-quality data context in the generalized information model can impact robustness more than the generalization of adversarial examples during training. Our approach, in conjunction with adversarial training, resulted in the highest robustness levels since the two priorities of each defense are complementary and result in the highest level of robustness. Overall, this shows how amodal segmentation in the training dataset reduces a learning model's sensitivity to adversarial perturbations.

The main limitation of this work is that amodal symbol systems can only be applied to computer vision tasks; thus, we cannot evaluate this concept with datasets that correspond to other domains. Additionally, we rely on the Mask R-CNN technique to generate the masks we used for training. However, as with any deep neural network, the generated output has a margin of error, resulting in potentially inaccurate masks for multiple or occluded items. To address this limitation, we manually reviewed a randomly selected sample set from the generated masks to verify the quality of the amodal mask. The masks were highly accurate since Mask R-CNN was optimized for MS COCO. However, for other image datasets, the generated masks may not be as precise, and thus, the impact on those masks is unclear concerning adversarial robustness. Overall, future work can extend this application to spatiotemporal datasets (e.g., video) with accurate amodal segmentation techniques designed specifically for that domain and increase the robustness of the

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on March 13,2025 at 19:15:47 UTC from IEEE Xplore. Restrictions apply.

highly complex and inherently more vulnerable spatiotemporal datasets.

REFERENCES

[1] F. Alrasheedi and X. Zhong, "Imperceptible adversarial attack on deep neural networks from image boundary," *arXiv preprint arXiv:2308.15344*, 2023.

[2] K. Sooksatra and P. Rivas, "Evaluation of adversarial attacks sensitivity of classifiers with occluded input data," *Neural Computing and Applications*, vol. 34, no. 20, pp. 17615–17632, 2022.

[3] J. Ao, Q. Ke, and K. A. Ehinger, "Image amodal completion: A survey," *Computer Vision and Image Understanding*, p. 103661, 2023.

[4] K. Saleh, S. Szénási, and Z. Vámossy, "Generative adversarial network for overcoming occlusion in images: A survey," *Algorithms*, vol. 16, no. 3, p. 175, 2023.

[5] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European symposium on security and privacy (EuroS&P)*, pp. 372–387, IEEE, 2016.

[6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755, Springer, 2014.

[7] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

[8] D. Arpit, S. Jastrzkebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, *et al.*, "A closer look at memorization in deep networks," in *International conference on machine learning*, pp. 233–242, PMLR, 2017.

[9] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[10] Y.-A. Chen, W.-C. Chen, C.-P. Wei, and Y.-C. F. Wang, "Occlusion-aware face inpainting via generative adversarial networks," in *2017 IEEE International conference on image processing (ICIP)*, pp. 1202–1206, IEEE, 2017.

[11] S. Ge, C. Li, S. Zhao, and D. Zeng, "Occluded face recognition in the wild by identity-diversity inpainting," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3387–3397, 2020.

[12] S. Liu, K. Luo, N. Ye, C. Wang, J. Wang, and B. Zeng, "Oiflow: Occlusion-inpainting optical flow estimation by unsupervised learning," *IEEE Transactions on Image Processing*, vol. 30, pp. 6420–6433, 2021.

[13] L. Huang and Y. Huang, "Drgan: A dual resolution guided low-resolution image inpainting," *Knowledge-Based Systems*, vol. 264, p. 110346, 2023.

[14] J. Higgins and N. Bezzo, "Negotiating visibility for safe autonomous navigation in occluding and uncertain environments," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4409–4416, 2021.

[15] H. Ryu, M. Yoon, D. Park, and S.-E. Yoon, "Confidence-based robot navigation under sensor occlusion with deep reinforcement learning," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 8231–8237, IEEE, 2022.

[16] Y. Sun, A. Kortylewski, and A. Yuille, "Amodal segmentation through out-of-task and out-of-distribution generalization with a bayesian model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1215–1224, 2022.

[17] K. Dollár and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[18] W. Abdulla, "Mask r-cnn for object detection and instance segmentation on keras and tensorflow." https://github.com/matterport/Mask_RCNN, 2017.

[19] K. Nguyen and S. Todorovic, "A weakly supervised amodal segmenter with boundary uncertainty estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7396–7405, 2021.

[20] B. Koonce and B. Koonce, "Resnet 50," *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*, pp. 63–72, 2021.

[21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *stat*, vol. 1050, p. 20, 2015.

[22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[23] M.-i. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. B. Angel, B. Chen, H. Ludwig, *et al.*, "Adversarial robustness toolbox v1. 0.0," *arXiv*, 2019.

[24] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," *arXiv preprint arXiv:1902.06705*, 2019.

[25] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. Molloy, and B. Edwards, "Adversarial robustness toolbox v1.2.0," *CoRR*, vol. 1807.01069, 2018.

[26] X. Liang, Y. Qian, J. Huang, X. Ling, B. Wang, C. Wu, and W. Swaileh, "Towards the desirable decision boundary by moderate-margin adversarial training," *arXiv preprint arXiv:2207.07793*, 2022.

[27] H. Salman, S. Jain, A. Ilyas, L. Engstrom, E. Wong, and A. Madry, "When does bias transfer in transfer learning?," *arXiv preprint arXiv:2207.02842*, 2022.

[28] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," *International Joint Conference on Artificial Intelligence (IJCAI-21)*, 2021.

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on March 13,2025 at 19:15:47 UTC from IEEE Xplore. Restrictions apply.