# Adaptive Clustering of Robust Semantic Representations for Adversarial Image Purification on Social Networks

**Samuel Henrique Silva, Arun Das, Adel Alaeddini, Peyman Najafirad** [*]

Secure AI and Autonomy Laboratory
The University of Texas at San Antonio
San Antonio, Texas 78249
(samuelhenrique.silva, arun.das, adel.alaeddini, peyman.najafirad)@utsa.edu

## Abstract

Advances in Artificial Intelligence (AI) have made it possible to automate human-level visual search and perception tasks on the massive sets of image data shared on social media on a daily basis. However, AI-based automated filters are highly susceptible to deliberate image attacks that can lead to content misclassification of cyberbulling, child sexual abuse material (CSAM), adult content, and deepfakes. One of the most effective methods to defend against such disturbances is adversarial training, but this comes at the cost of generalization for unseen attacks and transferability across models. In this article, we propose a robust defense against adversarial image attacks, which is model agnostic and generalizable to unseen adversaries. We begin with a baseline model, extracting the latent representations for each class and adaptively clustering the latent representations that share a semantic similarity. Next, we obtain the distributions for these clustered latent representations along with their originating images. We then learn semantic reconstruction dictionaries (SRD). We adversarially train a new model constraining the latent space representation to minimize the distance between the adversarial latent representation and the true cluster distribution. To purify the image, we decompose the input into low and high-frequency components. The high-frequency component is reconstructed based on the best SRD from the clean dataset. In order to evaluate the best SRD, we rely on the distance between the robust latent representations and semantic cluster distributions. The output is a purified image with no perturbations. Evaluations using comprehensive datasets including image benchmarks and social media images demonstrate that our proposed purification approach guards and enhances the accuracy of AI-based image filters for unlawful and harmful perturbed images considerably.

## Introduction

The age of the internet brought rapid advances in the way we share and consume information. Today's connected world enables real-time high-definition communication with people across the globe, and social media is a huge catalyst of this digital revolution. However, the rise of fake news, and the wide adoption of easy services such as Fake TV News Maker, Fake Newspaper Maker, Journalist CreativeBot, etc. has reduced trust in social media. Further, with the introduction of deepfakes, the synthesis of convincing, highly detailed, and novel human faces is easier to access, provoking psychological dilemmas in discriminating the truth (Pantserev 2020). Numerous research has shown the importance of the impact of content and layout of social media posts for user engagement (Shahbaznezhad, Dolan, and Rashidirad 2021). The actual content in the form of audio, video, or text, is moderated by automated content filtering algorithms, mostly driven by either rule-based or deep learning (DL) algorithms. As a rapidly developing area of research, deep learning (DL) represents a change in the way we interpret and make decisions from data (LeCun, Bengio, and Hinton 2015).

Despite the use of content filtering algorithms, social media is plagued with fake audio-visual or text content, cyberbullying (Vishwamitra et al. 2021), as well as pornographic adult content due to the inability to screen content across social media platforms. For example, the spread of fake images generated for misinformation campaigns that were recently found in Brazil and India used messaging platforms such as Whatsapp (Reis et al. 2020) which are difficult to regulate. Visual analytic systems for investigating misinformation are driven by multimodal decision-making AI algorithms which scrub the internet for textual and image-based data (Karduni et al. 2018), which work well for platforms in which they can be deployed. Similarly, child sexual abuse material (CSAM) and online circulation of pornographic content is accelerated by the borderless nature of internet and the wide adoption of social media and messaging services (Lee et al. 2020). However, child sexual abuse (CSA) detection algorithms using deep learning methods utilize multimodal image or video descriptors which can tag and classify pornographic content with ease. Current deep learning algorithms are shown to outperform other machine learning based classifiers, image hash databases, filename or metadata related methods, image descriptors, or skin detectors for unknown CSAM detection.

These classifiers are known to perform well in an independent and identically distributed (*i.i.d.*) setting, when testing and training data are sampled from the same distribution (Schneider et al. 2020). However, in many such applications involving large-scale web datasets, the *i.i.d.* assumption does not hold (Chacon, Silva, and Rad 2019). Further-

[*] Corresponding Author

more, the shift in distribution can be artificially introduced by adversarial attacks. Initially discovered by (Szegedy et al. 2014), adversarial attacks are small additive perturbations that, when combined with the input data, cause models to generate wrong predictions with high confidence. The effects of these perturbations in DL models can generate catastrophic consequences in safety critical applications. Additionally, adversarial attacks are mostly imperceptible to the human eye, can be easily transferable across models, and depend on relatively little information on the target model, as seen in (Chen, Jordan, and Wainwright 2020; Ru et al. 2020). Hardly distinguishable from clean images, adversarial examples largely lie in the tail of the dataset distribution used in training (Song et al. 2018). As a consequence, there is a possibility that adversaries will attack CSAM or cyberbully images such that the classifiers fail to filter CSAM or cyberbully content (Vitorino et al. 2018; Castrillón-Santana et al. 2018). In this paper, we study the impact of adversarial attacks on social media content filtering algorithms and propose a countermeasure purification methodology for robust semantic representations of adversarially attacked images on social networks.

The methods and effects of adversarial attacks are thoroughly detailed in literature (Silva and Najafirad 2020). Similarly, techniques to enhance robustness, either through adversarial training (AT) or input transformation (IT), have recently received considerable attention (Madry et al. 2018; Wong, Rice, and Kolter 2019; Guo et al. 2018). The emergence of almost imperceptible perturbations capable of changing the model's output, as well as the constant emergence of new attack methods indicates that current supervised methods lack model features that determine a causal relation between input and labels without failure. One of these critical features is generalizability to unseen attacks, that is, the model's ability to memorize the attack pattern and choose representations that are robust against small perturbations. To develop this feature, we go beyond standard AT and combine it with IT to learn semantic representations that demonstrate robustness against small input variations, capture the semantic correlations of the clean dataset, and purify adversarial images by removing perturbations. We address these tasks in a three-stage algorithm that achieves performance on par with state-of-the-art (SOTA) AT methods, generalizes for unseen attacks, and is task and model agnostic.

In our algorithm, we initially train a baseline model with clean images. The baseline model is used to extract the latent representations for all images in the training dataset. Even within the corresponding classes, the latent representations present high variability that influence the model to learn complex and susceptible to manipulation representations. We adaptively cluster the latent representations to create clusters of semantically similar representations. Moreover, we extract the distribution of the latent representations, and from the originating images, we learn semantic reconstruction dictionaries. On the second stage, we train a semantic robust model by modifying the standard adversarial training. We constrain the latent representations to minimize the distance between adversarial representations and the distribution of the clean cluster to which that sample should belong. By constraining the latent representations, we enable the model to extract similar features for clean and adversarial samples. Finally, in the third stage, we purify the adversarial images by reconstructing the high-frequency components of this image using the reconstruction dictionaries obtained from the clean samples in stage one. Once the robust semantic feature purification method is trained, we adversarially attack the Not-Safe-for-Work (NSFW) (Alagiri 2021) image classification dataset of adult content and study the performance loss due to adversarial attacks and the performance improvements due to image purification using our proposed algorithm which is trained on ImageNet. More specifically, this paper contributes the following:

- We propose a new adversarial training schema, minimizing the distance between the distribution of features extracted from the clean dataset and features from adversarial inputs, thereby improving generalization for multiple unseen attacks.
- We propose an Adaptive Clustering algorithm for robust Semantic Representation (ACSR) of image latent space in order to generate multiple reconstruction dictionaries within a class of images based on their similar semantic features to achieve adversarial image purification.
- We evaluate the efficacy of ACSR using quantitative and qualitative methods on comprehensive datasets such as CIFAR-10, ImageNet-10, NSFW dataset, and social network images. We also show that ACSR significantly outperforms other input transformation defenses and can return the classification accuracy of AI-based content filters against adversarially attacked images to roughly the same level of accuracy as a clean image.

## Related Work

**Adversarial Data Attacks:** Since the introduction of adversarial attacks in (Szegedy et al. 2014; Biggio et al. 2013), the field has been improved with a vast series of important contributions (Goswami et al. 2018; Jia et al. 2019; Cao et al. 2019; Xiao et al. 2018). Adversarial attacks are small, norm-constrained perturbations injected in the test data at inference time, capable of fooling a target model. These perturbations can be generated in a white box setting (Papernot et al. 2016; Morgulis et al. 2019), a black box setting (Chen, Jordan, and Wainwright 2020) or a gray-box setting(Silva and Najafirad 2020). We classify our image purification experimental setting as gray box. We assume the attacker has full knowledge of the target model, but cannot extract gradient information from our transformation due to the nature of the Convolutional Basis Pursuit Denoising (CBPDN) algorithm. Our Robust Semantic Training is trained and tested in a white box setting. More specifically, we evaluate our defense capabilities against FGSM (Goodfellow, Shlens, and Szegedy 2015), BIM, DeepFool (Moosavi-Dezfooli, Fawzi, and Frossard 2016), and CW (Carlini and Wagner 2017) attacks. Our robust model is trained with a projected gradient descent (PGD) attack (Madry et al. 2018).

**Social Media Classifiers:** The openness, ease of access, and ease of sharing content in social media platforms have

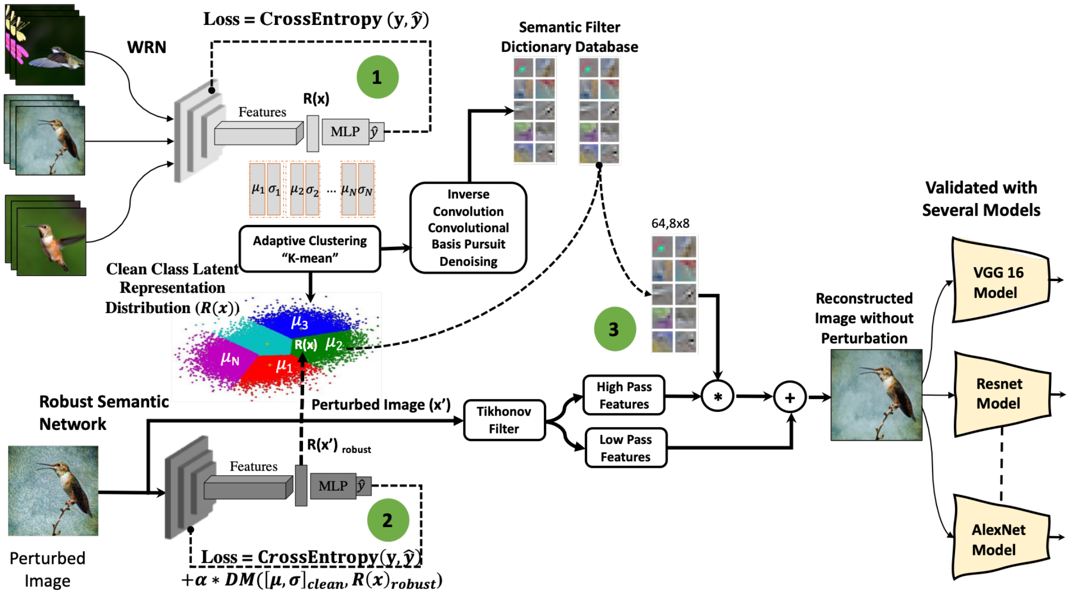

Figure 1: Our purification method combines adversarial training with input transformation. We initially train $f_{bsl}$ in the clean dataset. Based on $f_{bsl}$, we extract clean reference distributions to train our robust model $f_{rob}$. We cluster the clean input images based on their latent representations, and train semantic reconstruction dictionaries based on these clusters. We purify the adversarial inputs with the dictionary from a cluster that minimizes the distance from extracted latent representation and cluster distributions.

enabled adversaries to share content which are not suitable for a general audience. This problem has disrupted the cyber security and emotional integrity of individuals, leading to the rise in the focus of sensitive media analysis. Several deep learning and non-deep learning based content classifiers were proposed to screen content flowing to social media. Some of the most prominent and highly researched areas are in CSAM, NSFW, FakeNews, and DeepFake content. In (Pantserev 2020), authors laid out the importance of preventing content such as DeepFakes which includes face swapping. This task often interests the entertainment and pornography industries, but it also has business use cases. In (Vitorino et al. 2018), authors introduced an adult content detector which can not only screen sexual content, but also screens suggestive materials and images of child pornography. In (Moreira et al. 2016), authors introduced a non-deep learning based, temporally robust feature extractor and a bag of visual words method to classify pornographic videos with considerable accuracy. Content modified images, which skew perceptions of viewers (Reis et al. 2020), and unfair classifiers, which are biased to different target populations (Kyriakou et al. 2019) are also topics concerning social media image analytics.

**Adversarially Robust Classifier:** Adversarial training techniques modify the optimization objective of the objective model by incorporating an empirical maximum loss term in the objective, turning the training process into a min-max optimization. Several researchers have contributed empirical methods to calculate this loss (Goodfellow, Shlens, and Szegedy 2015; Madry et al. 2018; Kannan, Kurakin, and Goodfellow 2018). Moreover, Song proposed variations in the feature mapping to improve the accuracy of (re)trained models (Song et al. 2019). A tighter upper bound for the max in the loss was proposed by (Zhang et al. 2019). Although shown to improve robustness even for large models (ImageNet (Xie et al. 2019)), these models come with a drawback: when instantiated in practice with the approximation heuristics, these models are unable to provide robustness guarantees, certifications, or even generalize to unseen attacks. This class of defenses, even though very practical to implement, are model and attack dependent.

**Input Transformation Defenses:** Input transformation methods propose composing a set of transformations to adversarial images, such that the output has minimum influence of the adversarial perturbation. (Das et al. 2018) used JPEG compression as a countermeasure to the pixel displacement generated by the adversarial attacks. (Guo et al. 2018) proposed a combination of total variation minimization and image quilting to defend against adversarial attacks. Even though these transformations are nonlinear, a neural network was used to approximate the transformations, making them differentiable and consequently easier to obtain the gradient. (Raff et al. 2019) proposed an ensemble of weak transformation defenses to improve the robustness of the models. Among the transformations in the defense are color precision reduction, JPEG noise, Swirl, Noise Injection, FFT perturbation, Zoom Group, Color Space Group, Contrast Group, Grey Scale Group, and Denoising Group. Input transformation methods are, by nature, model independent, which improves transferability but limits efficiency;

the general model is unaware of the important input features that need restoration to improve the target model's accuracy.

# Methodology

## Method Overview

Our method uses a model class $\mathcal{F}$ and a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^{n}$ of feature vectors $x_i \in \mathcal{X} \subseteq \mathbb{R}^m$, where $\mathcal{X}$ is the feature space, and labels $y_i$ from some label set $\mathcal{Y}$, assuming up to $c$ possible values to $y_i$, corresponding to the class set $\mathcal{C}$. This dataset is typically assumed to be sampled *i.i.d* from the distribution $\mathcal{P}$, which is unknown. A classification model is a probability estimator, which maximizes the true class probability $p(y_i|x_i)$. On the contrary, an adversarial perturbation is an additive noise $\delta$, which maximizes another class probability such that the model's decision changes:

$$d(x_i, x_i + \delta) \leq \epsilon, \ and \ f(x) \neq f(x')$$

in which $\epsilon$ is the maximum allowed perturbation, and $d(.,.)$ is some specified distance. Our proposed method purifies adversarial inputs $x'_i = x_i + \delta$, such that $f(T(x + \delta)) = f(T(x)) = f(x)$. Our method is a composition of adversarial training and input transformation, and is divided in three stages: *i)* Baseline Training; *ii)* Robust Semantic Training; and *iii)* Robust Semantic Feature Purification. Algorithm 1 describes our approach. All details and notations are presented in the following sections of this section. Figure 1 represents all stages of our algorithm in detail. In a social media context, we then evaluate the performance of our generalized purification DL model on a not-safe-for-work (NSFW) classifier for adversarially attacked inputs. The manuscript does not include any illustrations from the NSFW dataset or any unlawful or harmful social media images due to ethical concerns.

In *i)* Baseline Training, we train a Wide-Resnet model (Zagoruyko and Komodakis 2016) that achieves high classification accuracy on the ImageNet dataset. We use this model to extract all the latent representations $R(x) \in \mathbb{R}^k$. For each class, we adaptively cluster all $R(x)$ that share semantic similarities, generating a cluster set $\Psi^c = \{\Psi_1^c, \Psi_2^c, ...\}$, with an optimal number of clusters for class $c$. Based on the cluster of $R(x)$ we cluster the originating inputs $x$ keeping track of images and clusters associations. For each $\Psi_i^c$, we calculate the parameters defining their distribution $(\mu_i^c, \sigma_i^c)$. In addition, for each cluster, based on the originating images $x$, we generate a sparse code reconstruction dictionary, using Convolutional Dictionary Learning (CDL). The details on the adaptive clustering, distribution calculation, and dictionary learning are discussed in the following sub-sections.

In *ii)* Robust Semantic Training, we train a robust model $f_{rob}$, which extracts robust latent representations that lie on the same distribution, independent of whether $x_i$ is an adversarial input or a clean input. We group samples which share semantic similarity. To achieve this similarity, we constrain the latent representation to minimize the distance between adversarial latent representation and clean sample cluster distribution. The robust model can classify the input, and more importantly, generate a latent representation close to that which would be generated by a clean sample.

Algorithm 1: Adaptive Clustering of Robust Semantic Representations (ACSR) Image Purification Algorithm

```
**Input:** $\mathcal{D}_{tr}$
*Stage i* - Baseline Training
$f_{bsl} \leftarrow$ standard-training($\mathcal{D}_{tr}$)
$\mathcal{R} \leftarrow$ extract-latent-representation($f_{bsl}$, $\mathcal{D}_{tr}$)
for $c \in \mathcal{C} = \{C_1, C_2, ...\}$ do
    $\psi_c$ = elbow($R(x_i)$), $\forall x_i \in c$
    $\Psi_c$ = k-means($R(x_i)$, $\psi_c$), $\forall x_i \in c$
    for $j \leq \psi_c$ do
        $\mu_c^j = \frac{1}{size(\Psi_c^j)} \sum x_i, \ \forall x_i \in \Psi_c^j$
        $\sigma_c^j = \mathbb{E}[(\Psi_c^j - \mathbb{E}[\Psi_c^j])(\Psi_c^j - \mathbb{E}[\Psi_c^j])^T], \ \forall x_i \in \Psi_c^j$
        $\Phi_c^j$ =concat(CDL($x_i$), ($x_i$), ($\mu_c^j, \sigma_c^j$)), $\forall x_i \in \Psi_c^j$
    end for
end for
*Stage ii* - Robust Semantic Training
for $(x, y) \in \Phi$ do
    $x' =$ adversarial($x$)
    $\hat{y}, R(x') \leftarrow f_{rob}(x')$
    $l = loss(\hat{y}, y) + \lambda * dist(R(x'), (\mu, \sigma))$
    $f_{rob} \leftarrow update(f_{rob}, l)$
end for
*Stage iii* - Robust Semantic Feature Purification
$x_{low}, x_{high}$ = tikhonov($x_i$)
$R(x_i)$ = $f_{rob}(x_i)$
$\Phi_{rec}$ = $argmin_{\Phi} \ dist(\Phi, R(x_i))$
$x_{high}^{rec}$ = CBPDN($x_{high}$, $\Phi_{rec}$)
$x_{pur}$ = $x_{high}^{rec} + x_{low}$
```

In *iii)* Robust Semantic Feature Purification, we purify $x_i$ by reconstructing the high frequency components of $x_i$ with its semantic reconstruction dictionary learned in *(i)*. We use $f_{rob}$ to extract robust latent representations from $x_i$, and match them with the semantic reconstruction dictionary that minimizes the distance between $R(x_i)$ and $C_i$ obtained in *(i)*. Following the semantic dictionary matching, we decompose $x_i$ into its low and high-frequency components. We use convolutional sparse coding to reconstruct the high-frequency components of $x_i$ and combine with the low-frequency to generate the purified and transformed version of $x_i$, $T(x_i)$.

## Problem Definition

The objective of supervised methods is to find a model $f \in \mathcal{F}$, such that:

$$\mathbb{E}_{(x,y)\sim\mathcal{P}}[l(f(x), y)] \leq \mathbb{E}_{(x,y)\sim\mathcal{P}}[l(f'(x), y)] \ \forall \ f' \in \mathcal{F},$$

where the loss function, $l(f(x), y)$, measures the error that $f(x)$ makes in predicting the true label $y$. In practice, $\mathcal{P}$ is unknown, and in replacement we use a training data $\mathcal{D}_{tr}$ in order to find a candidate $f$ that is a good approximation of the labels actually observed in this data. This raises the problem known as *empirical risk minimization* (ERM):

$$|s|\theta \sum_{i \in \mathcal{D}_{tr}} l(f(x_i; \theta), y_i) + \lambda\rho(\theta), \tag{1}$$

in which $\theta$ are the model parameters and $\rho(\theta)$ is a regularization function to constrain the changes of the model parameters at each learning step. We refer to Equation 1 as the baseline model training. The standard premise in DL is to find the optimum parameters in Equation 1 in order to deliver high performance on unseen data draw from the same distribution. On the contrary, the baseline objective from (Equation 1) is highly vulnerable to small perturbations, crafted by adversarial algorithms. In a general formulation, these perturbations are generated such that:

$$\max_{\|\delta\|_2 \leq \epsilon} \quad l(f(x_i + \delta; \theta), y_i). \tag{2}$$

By introducing the perturbation $\delta$ in the test data $\mathcal{D}_{te}$, the actual test distribution is shifted to the tail of the training distribution, effecting the performance of $f(x)$ when evaluated in $\mathcal{D}_{te}$. A natural approach to mitigate the effects of these manipulations is to introduce adversarial examples in Equation 1, known as the *robust optimization framework*:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \max_{\delta \in \Delta} \ l(f(x + \delta), y) + \lambda\rho(\theta), \tag{3}$$

Equation 3 is the *standard adversarial training*. The standard adversarial training addresses the immediate issue of samples generated by the technique used to empirically approximate Equation 2 for model $f(.)$. It has been shown by many publications that this formulation does not generalize to unseen attacks (Silva and Najafirad 2020). Moreover, the need to retrain every model can make this formulation very expensive.

In the baseline model (Equation 1), and adversarial model (Equation 3), the convolutional layers learn to extract latent representations $R(x) \in \mathbb{R}^k$, which are meaningful to the Fully Connected layers of the model. $R(x)$ is the output of the last layer before the fully connected layers of the model. As shown in (Engstrom et al. 2019), considerably different inputs can generate fairly similar latent representations. This evidence shows that even though latent representations are relevant condensed features for the model's Fully Connected layers, the similarity between latent representations of different classes is the source of the model's susceptibility to adversarial attacks.

## Baseline Training

In the baseline training, we construct all the references for robust model training and image purification. As mentioned in the method overview, we train a Wide Residual Network (WRN) for a classification task. The WRN model has been used in several publications as a benchmark for adversarial training in classification models and for its feature extraction capability. We refer to the baseline model $f_{bsl}$, a model trained without any adversarial training or robustness technique (except standard techniques), such as: Batch Normalization, Dropout, and Parameter Regularization.

The baseline model $f_{bsl}$ is trained on $\mathcal{D}_{tr}$. It is required that $f_{bsl}$ accurately models the distribution of $\mathcal{D}_{tr}$, and consequently achieves evaluation accuracy on $\mathcal{D}_{te}$. This premise allows us to assume a good class separation in the feature space, and as a consequence, a well-defined set of class distributions, which will be used for constructing the references for stages *(ii)* and *(iii)*. We initially construct a set $\mathcal{R} = \{R(x_i), x_i, y_i\}$ of the latent representations extracted from dataset $\mathcal{D}_{tr}$ by model $f_{bsl}$, and its originating images-labels pair. The latent representations $R(x_i)$ correspond to the set of features the model originally distilled from the input, allowing the FC layers to generate the class probabilities. We use such information as reference.

The set $\mathcal{R}$ contains the latent representations for all samples of all classes in $\mathcal{D}_{tr}$. We divide $\mathcal{R}$ in $C$ sub-sets, one for each class, and we refer to each sub-set as $\mathcal{R}_c$. The high variability within each class image generates high variability within $\mathcal{R}_c$. Fitting such high-dimensional data, with high-variability to a distribution, would generate meaningless parameters. We address the variability by adaptively clustering semantically similar latent representations within each $\mathcal{R}_c$. We are searching for features that gravitate around a mean value and tolerate certain dispersion around this center. Given the nature of the data, we are not fixing the number of centers in the data, nor the radius of dispersion. Instead, we semantically cluster our data based on the representations.

Using K-means clustering (Arthur and Vassilvitskii 2007), we cluster the data from each $\mathcal{R}_c$. We chose to cluster within each class to guarantee separation between classes. Given the set $\mathcal{R}_c$, composed of all $x_i \in \mathcal{R}_c$, we want to minimize the within cluster variance:

$$\arg\min_{\Psi} \quad \sum_{i=1}^{\psi} \sum_{R(x)\in\Psi_i} \|R(x) - \mu_i\|_2 \tag{4}$$

where $\Psi \subset \mathcal{R}_c$, and $\psi$ is the number of cluster centers. We iteratively search for the best value for $\psi$, taking into consideration that a higher value would reduce variability within clusters, but would also reduce the distance between cluster centers. The objective is to find a number of clusters that balances these two factors, such that the distance between samples within the same cluster is minimized and the distance between clusters is maximized.

In Equation 4, the objective is to reduce the Within Cluster Sum of Squares (WCSS). We employ the elbow method to balance the cost of increasing $\psi$ with respect to the variance reduction. The WCSS is used as a performance indicator. We iterate over the value of $\psi$, smaller values on the WCSS indicate greater homogeneity within clusters, yet indefinitely increasing $\psi$ will eventually reduce the separation of the clusters. When the the value of $\psi$ is closer to the optimal number of clusters, the WCSS curve shows a rapid decline, which reduces significantly as $\psi$ increases. It is important to highlight that this process is calculated independently for each $\mathcal{R}_c$. Datasets composed of multiple classes, as seen in real applications, would not affect the performance of this algorithm.

For each cluster $\Psi_j$, we obtain the mean $\mu_{\Psi_j} \in \mathbb{R}^k$ as the average of each individual component of each $R(x) \in \Psi_j$, and the covariance:

$$\sigma_{\Psi_j} = \mathbb{E}[(\Psi_j - \mathbb{E}[\Psi_j])(\Psi_j - \mathbb{E}[\Psi_j])^T]$$

where T is the transpose operator.

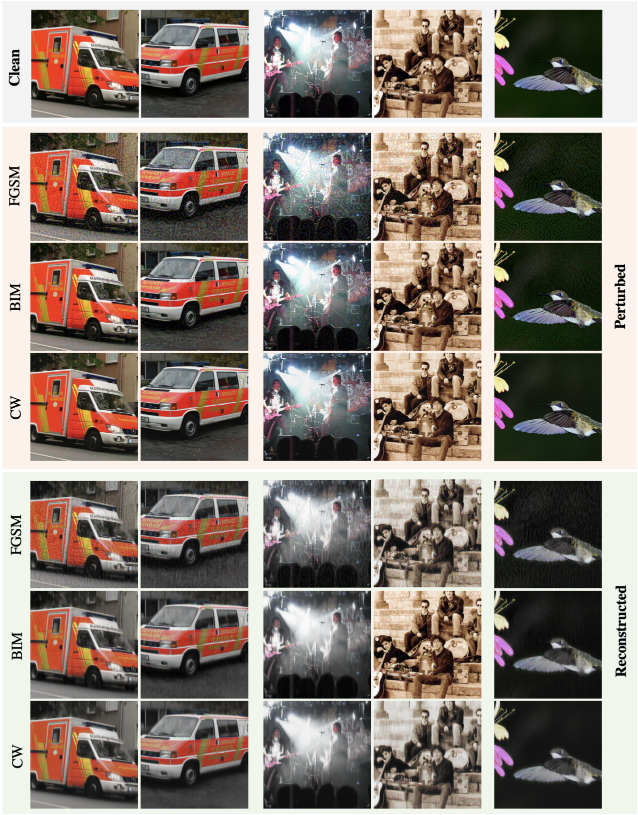

Figure 2: The reconstruction output of ACSR on ImageNet-10 images. From left to right: the first column shows the original clean image; the third, fourth and fifth columns show the output of FGSM ($l_2 \leq 0.08$), BIM and CW ($l_2 \leq 0.04$). The last three columns show the reconstruction output of the respective attacks. The images are better viewed in color and zoomed in to visualize better, with FGSM attacks the most noticeable for the human eyes.

Each cluster represents a set of semantic features within each of the classes. These semantic features are translated from the originating images $x_i$. In stage *iii*, we propose reconstructing the images with dictionaries based on these clean images. We generate reconstruction dictionaries based on the known problem *Convolutional Dictionary Learning* (CDL). Specifically, given a set of images $x_i \in \Psi_j$ composed of $S$ training images $\{x_t\}_{s=1}^{S}$, CDL is implemented through minimizing:

$$\begin{aligned} \min_{\{d_m\},\{r_{s,m}\}} \quad & \frac{1}{2}\sum_{1}^{S}\left\|\sum_{1}^{M} d_m * r_{s,m} - x_s\right\|_2^2 \\ & + \lambda \sum_{1}^{S}\sum_{1}^{M}\|r_{s,m}\|_1 \\ \text{s.t.} \quad & \|d_m\|_2 \leq 1, \forall m \in 1, ..., M \end{aligned} \tag{5}$$

where $d_m$ are the $M$ atoms that comprise the dictionary $\Omega$, and $r_{s,m}$ are a set of coefficient maps, defined as:

$$\min_{\{r_m\}} \quad \frac{1}{2}\left\|\sum_{1}^{M} d_m * r_m - x\right\|_2^2 + \lambda \sum_{1}^{M}\|r_m\|_1 \tag{6}$$

We observe that CDL is a computationally expensive algorithm that does not scale well to larger images and datasets. To overcome scalability issues in our method, we implement the optimizations proposed by (Liu et al. 2018) in an algorithmic level. Currently, we use ADMM (Boyd, Parikh, and Chu 2011) to solve the minimization problem. The clusters, cluster distributions, and cluster reconstruction dictionaries generated for all classes are utilized for the semantic reconstruction dictionary, $\Phi = \{D, \Psi, (\mu_\Psi, \sigma_\Psi)\}$.

## Robust Semantic Training

While the adversarial attack strategy of a min-max optimization shown in Equation 3 has shown very successful results, it fails in generalizing the method to unseen attacks. This issue results from the empirical solution provided to the maximization term. Since no closed form solution can be derived for such complex functions, it is often approximated by the chosen adversarial attack algorithm. While it is very effective in adding resistance to this specific algorithm, it often fails to generalize to other attacks. We argue that the network does not learn to extract robust latent representations, but rather learns to change the FC layers to classify latent representations extracted from adversarial and clean samples in the same class.

In our proposed solution, we address this issue by introducing a constraint in the representation space. Our objective is not to change the boundaries of the decision on the FC layer, but rather extract robust semantic representations from the adversarial and clean samples that lie on the same distribution. We modify the standard adversarial training equation, adding an extra constraint in the objective function:

$$\begin{aligned} \min_{\theta} \; \mathbb{E}_{(x,y)\sim\mathcal{D}} \max_{\delta\leq\epsilon} l(f(x'), y) + \lambda \left\|\theta)\right\|_2^2 + \\ \alpha(R(x') - \mu)\sigma^{-1}(R(x') - \mu))^{\frac{1}{2}} \end{aligned} \tag{7}$$

where the last term, the *Mahalanobis Distance* (MD), minimizes the distance between the extracted adversarial latent representations $R(x')$ and the cluster distribution, following the association in $\Phi$. By minimizing the distance between the clean distribution clusters and the adversarial latent representations, the model, instead of learning the adversarial attack pattern, learns to extract meaningful representations, ignoring the added noise in the input. We refer to the robust semantic model as $f_{rob}$, and latent representations extracted from input, $x_i$, with $f_{rob}$ as $R_{rob}(x)$.

Along with our empirical evaluation, we noticed a data-dependent constraint to our MD formulation: some clusters present covariance matrices that are not full rank. This constraint indicates near perfect correlation among some features in the latent representation. In this case, finding the exact inverse of the covariance matrix is not possible. For those specific cases, we applied the Moore-Penrose pseudo-inverse (Barata and Hussein 2012).

| Defense | No-attack | PGD | CW | BIM | TPGD |
|---|---|---|---|---|---|
| No Defense | 86.36 | 27.68 | 9.91 | 42.50 | 30.37 |
| PGDAT+RST (PGDAT) | 79.79 | 67.43 | 50.74 | 62.59 | 75.25 |
| PGDAT (PGDAT) | 79.06 | 66.74 | 50.46 | 61.25 | 74.69 |
| PGDAT+RST (Random) | 77.36 | 66.99 | 39.66 | 60.52 | 73.92 |
| PGDAT (Random) | 75.63 | 65.55 | 52.26 | 57.86 | 73.57 |
| PGDAT+RST (Trades) | 87.56 | 73.60 | 56.30 | 75.27 | 78.18 |

Table 1: CIFAR-10 Classification accuracy (in %) using WRN32-10 trained with PGDAT + RST.

### Robust Semantic Feature Purification

The input purification occurs only at inference time. For each input $x'_i$, we extract $R_{rob}(x'_i)$. At inference time, no label with respect to class or cluster is available. Moreover, since the input can be adversarially manipulated, it is highly important that $f_{rob}$ extracts latent representations which lie on the same distribution for both clean and adversarial images. Recognizing clean and adversarial inputs is beyond the scope of this work. For the purposes of this study, we purify all inputs.

Based on $R_{rob}(x'_i)$, we select the semantic reconstruction dictionary which best reconstructs the high-frequency components of $x'_i$. We compute the MD between $R_{rob}(x'_i)$ and all clusters in $\Phi$. The cluster with minimum distance is selected as the reconstruction dictionary. In parallel, we decompose $x'_i$ into a high-frequency component, $x'_{high}$, and a low frequency component, $x'_{low}$, using the Tikhonov filter (Garcia-Cardona and Wohlberg 2018):

$$\underset{x_{low}}{\arg\min} \quad \frac{1}{2}\|x_{low} - x\|_2^2 + \frac{\lambda}{2}\sum_j \|G_j x_{low}\|_2^2$$

where $G_j$ is an operator that computes the discrete gradient along image axis $j$. Therefore, $x'_{high} = x' - x'_{low}$.

The reconstruction of $x'_{high}$ follows the standard sparse coding representation:

$$x^{rec}_{high} \approx Dr = d_1 r_1 + \cdots + d_M r_M,$$

in which $D$ is the dictionary learned only from patches of clean images. Under such circumstances, we have a high frequency component formed by patches of clean images, and are therefore free from adversarial manipulation. The full image purification follows from adding the low and high-frequency components:

$$x_{pur} = x_{low} + x^{rec}_{high} \tag{8}$$

## Experiments

### Experimental Settings

| Model | No attack | FGSM 0.08 | FGSM 0.04 | BIM | DF | CW |
|---|---|---|---|---|---|---|
| AlexNet | 91.07 | 73.07 | 76.78 | 83.92 | 82.69 | 86.53 |
| VGG-16 | 94.23 | 78.57 | 75.00 | 76.78 | 83.92 | 87.05 |
| ResNet50 | 95.19 | 76.78 | 86.53 | 84.61 | 90.38 | 90.78 |
| GoogleNet | 90.38 | 79.80 | 83.65 | 87.5 | 88.5 | 85.57 |

Table 2: CIFAR-10 classification accuracy (in %) against adversarial attacks across different models when the input is purified with ACSR. DF denotes DeepFool algorithm.

| Defense | Clean | FGSM 0.08 | FGSM 0.04 | BIM | DF | CW |
|---|---|---|---|---|---|---|
| No Defense | 94.23 | 58.16 | 65.23 | 18.03 | 17.60 | 9.36 |
| MagNet | 90.35 | 61.45 | 65.21 | 43.12 | 65.35 | 48.45 |
| PixelDefend | 85.26 | 68.10 | 73.29 | 77.29 | 74.14 | 75.79 |
| STL | 83.60 | 71.03 | 75.47 | 75.31 | 79.59 | 79.06 |
| ACSR | 94.23 | 78.57 | 75.00 | 76.78 | 83.92 | 87.50 |

Table 3: CIFAR-10 classification accuracy (in %) using VGG-16 on images reconstructed with ACSR. 'No Defense' indicates no image reconstruction was applied. DF denotes DeepFool algorithm.

| **Resolution 64x64** | | | | | | |
|---|---|---|---|---|---|---|
| Defense | Clean | FGSM 0.08 | FGSM 0.04 | BIM | DF | CW |
| No Defense | 86.65 | 28.16 | 30.8 | 18.83 | 8.11 | 7.51 |
| TVM | 75.55 | 59.97 | 69.3 | 71.56 | 72.1 | 71.87 |
| Quilting | 77.41 | 73.04 | 74.18 | 76.42 | 76.46 | 76.62 |
| Crop-Ens | 75.08 | 69.68 | 72.21 | 73.69 | 74.01 | 73.04 |
| PD-Ens | 82.5 | 66.34 | 76.07 | 79.03 | 79.55 | 78.13 |
| STL | 84.21 | 75.14 | 80.38 | 81.03 | 82.21 | 81.22 |
| ACSR | 87.50 | 84.37 | 78.12 | 87.50 | 84.37 | 81.25 |

| **Resolution 128x128** | | | | | | |
|---|---|---|---|---|---|---|
| Defense | Clean | FGSM 0.08 | FGSM 0.04 | BIM | DF | CW |
| No Defense | 89.91 | 21.23 | 24.09 | 17.90 | 5.84 | 5.04 |
| TVM | 85.91 | 25.68 | 43.86 | 65.86 | 63.60 | 61.29 |
| Quilting | 81.49 | 39.03 | 58.89 | 64.34 | 62.42 | 59.22 |
| Crop-Ens | 77.30 | 46.22 | 64.47 | 68.76 | 70.60 | 68.88 |
| PD-Ens | 87.89 | 23.33 | 42.86 | 72.21 | 73.59 | 72.72 |
| STL | 86.54 | 47.33 | 66.06 | 73.23 | 73.01 | 74.32 |
| ACSR | 87.50 | 87.25 | 89.28 | 87.5 | 90.62 | 88.25 |

Table 4: ImageNet-10 classification accuracy (in %) using VGG-16. In images with resolution 64x64 and resolution 128x128. DF denotes DeepFool algorithm.

We evaluate our proposed model on two main datasets, CIFAR-10 (Krizhevsky, Hinton et al. 2009), and ImageNet (Russakovsky et al. 2015), from which we extract 10 classes to compose ImageNet-10. CIFAR-10 is composed of 60000 images (50000 for training and 10000 for testing), uniformly distributed among 10 classes. ImageNet is composed of 1000 classes with roughly 1300 images per class for training and 50 samples for testing per class.

To evaluate the efficiency of our model, we use VGG16 (Simonyan and Zisserman 2014) and ResNet-50 (He et al. 2016) as classification models for comparison with other defense methods. We attack our model with FGSM(Goodfellow, Shlens, and Szegedy 2015), BIM (Kurakin, Goodfellow, and Bengio 2017), DeepFool (Moosavi-Dezfooli, Fawzi, and Frossard 2016), and CW (Carlini and Wagner 2017). For FGSM, we evaluate under the $l_2 norm < 0.04$ and $l_2 norm < 0.08$, and for BIM, DeepFool, and CW, we restrict to $l_2 norm < 0.04$ and 100 iteration steps. The images in the dataset are normalized in the range between 0 and 1. For all experiments, a single model is robustly trained to extract robust features for each dataset evaluated.

We provide a qualitative evaluation of the effects of dictionary selection based on the semantic cluster. Even though PSNR can measure the signal to noise ratio, we have observed that this is not a good metric for adversarial comparison, since attacks like CW and DeepFool are effective in adding perturbations without changing PSNR significantly. In all tables, we refer to our image purification algorithm as (ACSR), and the Robust Semantic Training as RST.

To study the impact of our proposed ACSR purification algorithm on social media content moderation neural networks, we adversarially perturb the NSFW dataset (Alagiri 2021) to fool classifiers. Adversarial attacks on the social media data could spoof and bypass the content moderation classifiers. The NSFW dataset consists of 334327 images in total from five classes including neutral, drawing, hentai, porn, and sexy. A classification model based on ResNet-50 model was trained on a 14:3:3 split for training, validation, and testing data. All further evaluations were carried out on the test set.

## Robust Semantic Training

In evaluating the effectiveness of our adversarial training, we first use the adaptive semantic clusters as a reference for the cluster approximation and the generalization capabilities of our method. To achieve this, we use CIFAR-10 dataset, and a WideResNet32-10 (WRN32-10), as proposed in (Song et al. 2019). Moreover, we adversarially train WRN32-10 using the standard adversarial training method approximating the maximization with Projected Gradient Descent (PGD) attack (Madry et al. 2018), in which 10 steps are used, with $l_2$ norm perturbation limit of $\delta = 0.3$.

Table 1 shows the accuracy of our model minimizing the distance between the distribution of reference clusters and the model's extracted latent representation. We used PGD Adversarial Training (PGDAT) in composition with Robust Semantic Training (RST). In parenthesis, we indicated if we used any pre-trained weights set as initialization for our model parameters, *(Random)* indicates no pre-training.

As seen in Table 1, we have evaluated our model under several conditions, including transfer learning from other techniques (we restricted our study to only using other techniques as weight initialization, no modifications to Equation 7 was used). We observed that when combined with transfer learning from other techniques such as TRADES (Zhang et al. 2018), we achieve high performance; however, we achieve relatively low performance when training from random initial weights. It is important to highlight that this mechanism is meant to extract robust latent representations, it is not meant to be a standalone defense.

## Robust Semantic Feature Purification Evaluation

Our defense method is advantageous in its transferability among multiple attacks and across multiple models. Figure 2 shows the reconstruction capabilities of our model under different attacks for samples in ImageNet-10. Table 2 shows a quantitative evaluation of our model to defend AlexNet, VGG16, GoogleNet, and ResNet-50 attacked by FGSM, BIM, DeepFool, and CW. We train a single model $f_{rob}$ to extract the robust latent representations and defend any of the models in Table 2.

We compare our model to other robust defenses that involve input transformation. In Table 3, we use a pre-trained VGG-16, adjusting the parameters and output layer for CIFAR-10, and obtain the accuracy against the mentioned attacks and compare against defenses such as MagNet(Meng and Chen 2017), PixelDefend (Song et al. 2018), and STL (Sun et al. 2019), following the experimental setting of (Sun et al. 2019). Table 4 summarizes our results against the defenses proposed in (Guo et al. 2018) (TVM, Quilting, Crop-Ens), (Prakash et al. 2018) (PD-Ens) and (Sun et al. 2019) (STL), in ImageNet-10 for image resolutions of 64x64 and 128x128.

Table 5 summarizes the improvement in accuracy of the NSFW classifiers on the attacks explored in Table 4. Here, we reuse our pretrained robust cluster embeddings to purify novel images from social media. Hence, the NSFW classifier does not need any retraining to purify novel images from the web. This significantly improves the wider use of our proposed algorithm for social media images in the wild.

For CW and DF attacks which considerably reduced the accuracy of the adult content classifier, our ACSR purification algorithm provided a 14.75% and 28.75% increase in accuracy respectively. Additionally, we see a marginal improvement of performance for FGSM (with epsilon of 0.08) and BIM methods with 1.94% and 5.84% accuracy.

**Ablation Studies:** We've evaluated the efficacy of our model with and without the influence of transfer learning. As seen on Table 1, our adversarial model achieves better results when trained with already saturated models, providing an improvement in accuracy. Consequently, we improve the clean data accuracy which is lost on standard adversarial

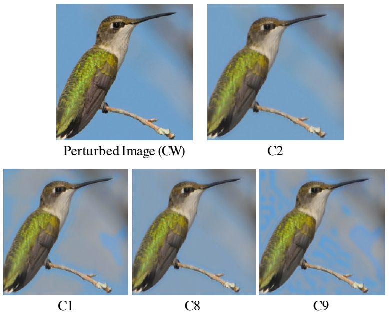

Figure 3: Qualitative analysis of the reconstruction based on the cluster selection. Top images are attacked and best reconstruction. C1, C8, and C9 are reconstructions from manually selected clusters.

| AI-based Content Filters | Image Purification Method | Image w/o Perturbations | CW | BIM | DF |
|---|---|---|---|---|---|
| Robust Classifier (NSFW-ResNet-50) | w/o ACSR | **86.87** | 61.82 | 69.23 | **53.84** |
| | w/ ACSR | **83.89** | 76.34 | 75.07 | **82.69** |
| Classifer (NSFW-AlexNet) | w/o ACSR | **94.23** | 69.23 | 71.15 | **5.76** |
| | w/ ACSR | **94.21** | 82.75 | 82.69 | **80.76** |

Table 5: Performance statistics (classification accuracy in %) of our proposed algorithm in purifying social media content classifiers. Our purification algorithm significantly increases the classification accuracy of different NSFW classifiers for strong attacks like DeepFool. The NSFW-AlexNet classifier that is not robustly trained sees a performance drop of nearly 88% when evaluated against DeepFool perturbations. With the use of ACSR image purification we are able to recover 75% of the accuracy that was lost due to perturbations.

training.

Moreover, we evaluated the effects of using our current setting, pre-trained (ImageNet) classification models, against the models trained from scratch. For CIFAR-10, clean images attacked with CW, classified with vanilla VGG-16, and trained from scratch, these models achieve an accuracy of 61.53%, in contrast to our reported 87.5%. This shows that transfer learning also contributes to our defense. Without our defense, in both cases, the models performed poorly, achieving only 9.36% accuracy for the transfer learning model. Qualitative evaluation reveals that the reconstruction quality is dependent on the cluster selection, as illustrated in Figure 3.

## Discussion

### ACSR Purification Algorithm

Comparison results show that our method outperforms current state-of-the-art input transformation methods based on image transformation and on sparse code image reconstruction at defending models against gray box attacks. We assume the attacker has full knowledge of the model, but no awareness of the transformation itself.

We have shown that our defense is model agnostic and is able to maintain accuracy across different models, unseen attacks, and on different applications. We credit this to two features of our design: our model's ability to approximate the robust latent representations to the clean distribution; and our sparse dictionaries made from clean images and used for reconstructions, allowing our model to be minimally dependent on the attack's empirical approximation of equation 3.

The assumption of the class separation based on the accuracy has proven reasonable within the set of experiments proposed in this manuscript. With the increase in the number of classes, it would be expected that a reduction in the distance between the distributions would occur, leading to a reduction of the accuracy of the algorithm. But in a high-dimensional setting, common in the SOTA models used for this vision task, $R(x) \in \mathbb{R}^d$ generally has a high-dimensional feature space ($d \geq 2048$). With this high-dimensional feature space, we can enforce the class separation through the training loss, hence the effect of increasing the number of classes is almost indifferent for the overall accuracy. Moreover, the increase in resolution of the images, as discussed in Table 4, could lead Equation 6 to be unsolvable due to the exponential computational cost. But our implementation of the CBPDN problem, based on the optimizations proposed in (Liu et al. 2018), allowed a scalable solution to this problem.

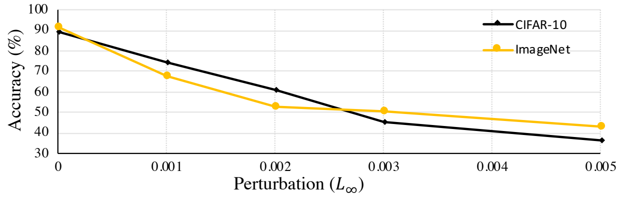

Figure 4: Accuracy evaluation on CIFAR-10 (yellow miters) of ResNet-50 and VGG-16 on ImageNet-10 FGSM under $l_\infty$ norm.

All the experiments we have presented demonstrate the effectiveness of our method to attacks in which the space norm is defined by $l_2$ norm. We have observed that $l_2$ bounded attacks generate less visible corruption on the input. We have also evaluated our model under $l_\infty$ norm bounded FGSM. In Figure 4, we show the efficacy of our defense method under different levels of $l_\infty$ attack. We observed that as the perturbation level increases, it affects the low-frequency components of the image, and our method cannot purify these images. Therefore, it would be necessary to supplement our method with those involving GANs and Autoencoders in order to complete the full reconstruction.

### Social Media Image Purification for Perturbed Images

Social media cyber-security is prone to adversarial attacks and deepfakes (Lago et al. 2021; Korshunov and Marcel 2018; Neekhara et al. 2021) while being an avid source of pornographic content (Lee et al. 2020) and cyber-bullying (Vishwamitra et al. 2021). We saw from Table 5 that adversarially attacking images uploaded to social media, for example, can fool and bypass social media content regulation deep learning algorithms. As a response, we used our proposed ACSR purification algorithm as a counter measure to clean adversarially attacked pornographic images such that the classifier's accuracy improves considerably.

We tested the accuracy of this task on two models. The first model was an adversarially robust (Madry et al. 2017) ResNet-50 classifier that was trained from ImageNet-50 weights and fine-tuned to NSFW classification, which we name NSFW-ResNet-50. The second model we tested was an AlexNet classifier that was trained from scratch on the NSFW dataset, which we name NSFW-AlexNet. Table 5 shows that the NSFW-ResNet-50 classifier's accuracy suffers when adversarially attacked. When this model takes in particularly effective image attacks, such as the DeepFool attack, we find that our image purification method improved classification accuracy by more than 28%, compared to the adversarially attacked baseline. The NSFW-AlexNet classifier benefited even more from the use of ACSR image purification. The NSFW-AlexNet classifier showed an accuracy greater than 94% with an AUC of 0.89 and F1-score of 0.88 when tested against images without pertubations. However, when the NSFW-AlexNet model was tested against images perturbed by the DeepFool algorithm, the classification accuracy drops to less than 6% with an AUC of 0.1. A user of an AI image classifier for social media content could be misled to believe that they have a highly accurate classifier due to the 94% accuracy they found in testing, however, the danger of image perturbation attacks shows that these classifiers will not be as effective in practice as they may seem. On the other hand, when these perturbed images are processed by our ACSR image purifier, and then passed to NSFW-AlexNet model, we are able to recover 75% of the classification accuracy, resulting in an accuracy of 80.76% with an AUC of 0.85 and F1-score of 0.80.

The difference between the gains in accuracy between the robust NSFW-ResNet-50 model and the NSFW-AlexNet model shows us in what circumstances our ACSR image purifier provides the most value. Large organizations with the resources to train a robust classifier may see less benefit to using image purification methods. On the other hand, as the NSFW-AlexNet model demonstrates, non-robust models can retain most of their accuracy even when facing perturbed images after these images are processed by our ACSR image purifier.

We reuse the robust semantic reconstruction dictionary initially trained for CIFAR-10 dataset for NSFW purification. We find that the semantic dictionaries are generalized and hence, we do not have to retrain any part of the algorithm or neural network parameters for social media image purification. Thus, our proposed ACSR works as a generalized method to purify social media content, even though the input images were out of distribution.

## Conclusion

In this paper, we have presented Adaptive Clustering of Robust Semantic Representations Algorithm, our novel method of image purification that shows state-of-the-art results against $l_2$ bounded adversarial attacks, unseen at training time. We designed a new methodology for input transformation which creates semantic reconstruction dictionaries for high-frequency components of each cluster of latent representations of the images in our dataset. We evaluate our proposed methodology on CIFAR-10, ImageNet, NSFW dataset, and social media images. We have also proven that our defense data purification method achieves robustness against several unseen attacks and different target models such as AlexNet, VGG-16, GoogleNet, and ResNet-50 trained on disturbing social media images. We have also evaluated our model against $l_\infty$ bounded perturbations and have observed a less effective transformation. Our qualitative and quantitative data on $l_\infty$ perturbations indicate that when the corruption achieves lower frequency portions of the image, the image needs to be regenerated rather than purified, suggesting a need for generative approaches.

## Acknowledgements

This material is based on research sponsored by the Department of Homeland Security (DHS), United States Secret Service, National Computer Forensics Institute (NCFI) via contract number 70US0920D70090004.

## References

Alagiri, K. 2021. Not-Safe-For-Work (NSFW) Pornographic Image Classification Dataset. *Kaggle.com/krishnaalagiri/nsfw-image-classification-resnet50*.

Arthur, D.; and Vassilvitskii, S. 2007. k-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 1027–1035.

Barata, J. C. A.; and Hussein, M. S. 2012. The Moore–Penrose pseudoinverse: A tutorial review of the theory. *Brazilian Journal of Physics*, 42(1-2): 146–165.

Biggio, B.; Corona, I.; Maiorca, D.; Nelson, B.; Šrndić, N.; Laskov, P.; Giacinto, G.; and Roli, F. 2013. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, 387–402. Springer.

Boyd, S.; Parikh, N.; and Chu, E. 2011. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc.

Cao, Y.; Xiao, C.; Cyr, B.; Zhou, Y.; Park, W.; Rampazzi, S.; Chen, Q. A.; Fu, K.; and Mao, Z. M. 2019. Adversarial sensor attack on lidar-based perception in autonomous driving. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2267–2281.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 Ieee symposium on security and privacy (sp)*, 39–57. IEEE.

Castrillón-Santana, M.; Lorenzo-Navarro, J.; Travieso-González, C. M.; Freire-Obregón, D.; and Alonso-Hernandez, J. B. 2018. Evaluation of local descriptors and CNNs for non-adult detection in visual content. *Pattern Recognition Letters*, 113: 10–18.

Chacon, H.; Silva, S.; and Rad, P. 2019. Deep Learning Poison Data Attack Detection. In *2019 IEEE 31st Interna-*

*tional Conference on Tools with Artificial Intelligence (ICTAI)*, 971–978. IEEE.

Chen, J.; Jordan, M. I.; and Wainwright, M. J. 2020. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 IEEE Symposium on Security and Privacy (SP)*, 1277–1294. IEEE.

Das, N.; Shanbhogue, M.; Chen, S.-T.; Hohman, F.; Li, S.; Chen, L.; Kounavis, M. E.; and Chau, D. H. 2018. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 196–204.

Engstrom, L.; Ilyas, A.; Santurkar, S.; Tsipras, D.; Tran, B.; and Madry, A. 2019. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*.

Garcia-Cardona, C.; and Wohlberg, B. 2018. Convolutional dictionary learning: A comparative review and new algorithms. *IEEE Transactions on Computational Imaging*, 4(3): 366–381.

Goodfellow, I.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*.

Goswami, G.; Ratha, N.; Agarwal, A.; Singh, R.; and Vatsa, M. 2018. Unravelling robustness of deep learning based face recognition against adversarial attacks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Guo, C.; Rana, M.; Cisse, M.; and van der Maaten, L. 2018. Countering Adversarial Images using Input Transformations. In *International Conference on Learning Representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jia, Y.; Lu, Y.; Shen, J.; Chen, Q. A.; Chen, H.; Zhong, Z.; and Wei, T. 2019. Fooling detection alone is not enough: Adversarial attack against multiple object tracking. In *International Conference on Learning Representations*.

Kannan, H.; Kurakin, A.; and Goodfellow, I. 2018. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*.

Karduni, A.; Wesslen, R.; Santhanam, S.; Cho, I.; Volkova, S.; Arendt, D.; Shaikh, S.; and Dou, W. 2018. Can you verifi this? studying uncertainty and decision-making about misinformation using visual analytics. In *Twelfth international AAAI conference on web and social media*.

Korshunov, P.; and Marcel, S. 2018. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning Multiple Layers of Features from Tiny Images. *University of Toronto*.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2017. Adversarial examples in the physical world. *ICLR Workshop*.

Kyriakou, K.; Barlas, P.; Kleanthous, S.; and Otterbacher, J. 2019. Fairness in proprietary image tagging algorithms: A cross-platform audit on people images. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 313–322.

Lago, C.; Romón, R.; López, I. P.; Urquijo, B. S.; Tellaeche, A.; and Bringas, P. G. 2021. Deep Learning Applications on Cybersecurity. In *International Conference on Hybrid Artificial Intelligence Systems*, 611–621. Springer.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature*, 521(7553): 436–444.

Lee, H.-E.; Ermakova, T.; Ververis, V.; and Fabian, B. 2020. Detecting child sexual abuse material: A comprehensive survey. *Forensic Science International: Digital Investigation*, 34: 301022.

Liu, J.; Garcia-Cardona, C.; Wohlberg, B.; and Yin, W. 2018. First-and second-order methods for online convolutional dictionary learning. *SIAM Journal on Imaging Sciences*, 11(2): 1589–1628.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Meng, D.; and Chen, H. 2017. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 135–147.

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

Moreira, D.; Avila, S.; Perez, M.; Moraes, D.; Testoni, V.; Valle, E.; Goldenstein, S.; and Rocha, A. 2016. Pornography classification: The hidden clues in video space–time. *Forensic science international*, 268: 46–61.

Morgulis, N.; Kreines, A.; Mendelowitz, S.; and Weisglass, Y. 2019. Fooling a Real Car with Adversarial Traffic Signs. *arXiv preprint arXiv:1907.00374*.

Neekhara, P.; Dolhansky, B.; Bitton, J.; and Ferrer, C. C. 2021. Adversarial threats to deepfake detection: A practical perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 923–932.

Pantserev, K. A. 2020. The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability. In *Cyber defence in the age of AI, smart societies and augmented humanity*, 37–55. Springer, Cham.

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

Prakash, A.; Moran, N.; Garber, S.; DiLillo, A.; and Storer, J. 2018. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8571–8580.

Raff, E.; Sylvester, J.; Forsyth, S.; and McLean, M. 2019. Barrage of random transforms for adversarially robust defense. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6528–6537.

Reis, J. C.; Melo, P.; Garimella, K.; Almeida, J. M.; Eckles, D.; and Benevenuto, F. 2020. A dataset of fact-checked images shared on whatsapp during the brazilian and indian elections. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 903–908.

Ru, B.; Cobb, A.; Blaas, A.; and Gal, Y. 2020. Bayesopt adversarial attack. In *International Conference on Learning Representations*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252.

Schneider, S.; Rusak, E.; Eck, L.; Bringmann, O.; Brendel, W.; and Bethge, M. 2020. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33.

Shahbaznezhad, H.; Dolan, R.; and Rashidirad, M. 2021. The role of social media content format and platform in Users' engagement behavior. *Journal of Interactive Marketing*, 53: 47–65.

Silva, S. H.; and Najafirad, P. 2020. Opportunities and Challenges in Deep Learning Adversarial Robustness: A Survey. *arXiv preprint arXiv:2007.00753*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, C.; He, K.; Lin, J.; Wang, L.; and Hopcroft, J. E. 2019. Robust Local Features for Improving the Generalization of Adversarial Training. In *International Conference on Learning Representations*.

Song, Y.; Kim, T.; Nowozin, S.; Ermon, S.; and Kushman, N. 2018. PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples. In *International Conference on Learning Representations*.

Sun, B.; Tsai, N.-h.; Liu, F.; Yu, R.; and Su, H. 2019. Adversarial defense by stratified convolutional sparse coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11447–11456.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Estrach, J. B.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*.

Vishwamitra, N.; Hu, H.; Luo, F.; and Cheng, L. 2021. Towards Understanding and Detecting Cyberbullying in Realworld Images. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*.

Vitorino, P.; Avila, S.; Perez, M.; and Rocha, A. 2018. Leveraging deep neural networks to fight child pornography in the age of social media. *Journal of Visual Communication and Image Representation*, 50: 303–313.

Wong, E.; Rice, L.; and Kolter, J. Z. 2019. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*.

Xiao, C.; Zhu, J. Y.; Li, B.; He, W.; Liu, M.; and Song, D. 2018. Spatially transformed adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018*.

Xie, C.; Wu, Y.; Maaten, L. v. d.; Yuille, A. L.; and He, K. 2019. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 501–509.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zhang, H.; Weng, T.-W.; Chen, P.-Y.; Hsieh, C.-J.; and Daniel, L. 2018. Efficient neural network robustness certification with general activation functions. In *Advances in neural information processing systems*, 4939–4948.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. I. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *ICML*.