Universal Adversarial Perturbations Through the Lens of Deep Steganography: Towards a Fourier Perspective

Chaoning Zhang*, Philipp Benz*, Adil Karjauv, In So Kweon
Korea Advanced Institute of Science and Technology (KAIST)
chaoningzhang1990@gmail.com, pbenz@kaist.ac.kr, mikolez@gmail.com, iskweon77@kaist.ac.kr

Abstract

The booming interest in adversarial attacks stems from a misalignment between human vision and a deep neural network (DNN), *i.e.* a human imperceptible perturbation fools the DNN. Moreover, a single perturbation, often called universal adversarial perturbation (UAP), can be generated to fool the DNN for most images. A similar misalignment phenomenon has also been observed in the deep steganography task, where a decoder network can retrieve a secret image back from a slightly perturbed cover image. We attempt explaining the success of both in a unified manner from the Fourier perspective. We perform task-specific and joint analysis and reveal that (a) frequency is a key factor that influences their performance based on the proposed entropy metric for quantifying the frequency distribution; (b) their success can be attributed to a DNN being highly sensitive to high-frequency content. We also perform feature layer analysis for providing deep insight on model generalization and robustness. Additionally, we propose two new variants of universal perturbations: (1) high-pass UAP (HP-UAP) being less visible to the human eye; (2) Universal Secret Adversarial Perturbation (USAP) that simultaneously achieves attack and hiding.

Introduction

Deep learning has achieved large success in a wide range of vision applications, such as recognition (Zhang et al. 2019, 2021a), segmentation (Vania, Mureja, and Lee 2019; Kim et al. 2020; Pan et al. 2020) as well as scene understanding (Lee et al. 2019b,a; Zhang et al. 2020d; Argaw et al. 2021b,a). Nonetheless, the vulnerability of deep neural networks (DNNs) to adversarial examples (Szegedy et al. 2013) has attracted significant attention in recent years. In machine learning, there is a surging interest in understanding the reason for the success of the adversarial attack (AA) (Szegedy et al. 2013; Zhang et al. 2020b). The root reason for this booming interest lies in the misalignment between human vision and DNN perception (see Figure 1). A similar misalignment phenomenon has also been observed in deep steganography (DS) (Baluja 2017; Zhang et al. 2020c), where a decoder network retrieves a secret image from a slightly perturbed cover image, often referred to as container image. In this work, for consistency, a small change to an image is termed perturbation (P) for both DS and AA. In both tasks, the original image I and perturbed image $I + P$ are nearly indistinguishable for the human vision system, given $||P|| \ll ||I||$ (see Figure 1). However, for a DNN, $M(I+P)$ is more similar to $M(P)$ than $M(I)$ where M indicates the model of interest as a function. For AA and DS, the DNN of interest is the target DNN and decoder network, respectively. For an instance-dependent perturbation (IDP) case, taking AA for example, this misalignment is relatively less surprising. We focus on the misalignment in "universal" scenario, with conflicting features in I and P, while I is dominated by P when they are summed, *i.e.* $I + P$, as the M input.

For both AA and DS, the misalignment constitutes the most fundamental concern, thus we deem it insightful to explore them together. We first attempt explaining its misalignment based on our adopted universal secret perturbation (USP) generation framework introduced in (Zhang et al. 2020c), where a secret image is hidden in a cover-agnostic manner. The success of DS has been attributed to the discrepancy between C and the encoded secret image (Zhang et al. 2020c). Inspired by the success of explaining the USP induced misalignment from the Fourier perspective, we explore the UAP induced misalignment in a similar manner.

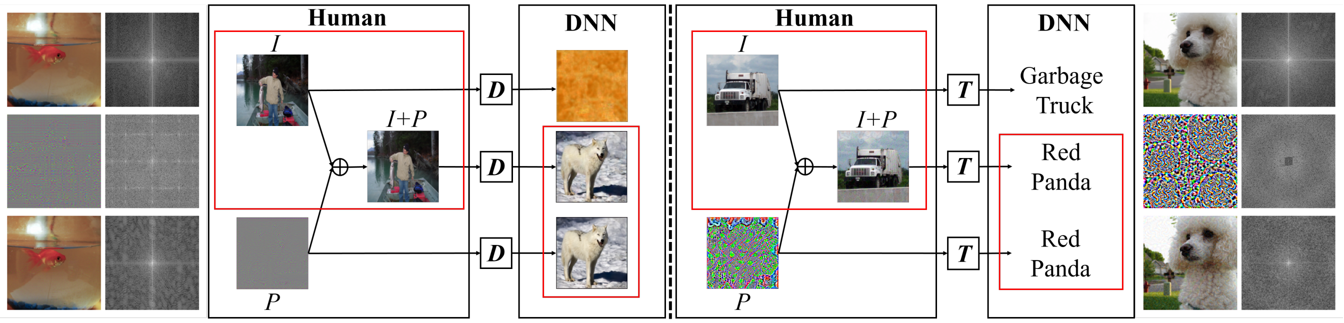
Our analysis shows that the influence of each input on the combined DNN output is determined by both frequency and magnitude, but mainly by the frequency. To quantitatively analyze the influence of image frequency on the performance of the two tasks, we propose a new metric for quantifying the frequency that involves no hyperparameter choices. Overall, our task-specific and cross-task analysis suggest that image frequency is a key factor for both tasks.

Contrary to prior findings regarding IDP in (Yin et al. 2019), we find that UAPs, which attack most images are a strictly high-frequency (HF) phenomenon. Moreover, we perform a feature layer analysis to provide insight on model generalization and robustness. With the frequency understanding, we propose two novel universal attack methods.

Related Work

Fourier Perspective on DNN. The behavior of DNNs has been explored from the Fourier perspective in multiple prior arts. Some works (Jo and Bengio 2017; Wang et al. 2020) analyze why the DNN has good generalization while be-

[*] Equal contribution



(a) USP induced misalignment

(b) UAP induced misalignment

Figure 1: Misalignment between human perception and DNN perception under the universal framework. D stands for Decoder DNN, while T is the Target DNN. (a) USP induced misalignment; (b) UAP induced misalignment. In both (a) and (b): given $||P|| \ll ||I||$, $H(I+P) \approx H(I)$ while $M(I+P) \approx M(P)$, where I indicates a clean image, P is an amplified perturbation, and I+P is a perturbed image. To both sides example images and their Fourier images for the respective task are shown. From top to bottom the images represent: clean image (I), amplified perturbation (P), and perturbed image ($I + P$). The corresponding Fourier images show that P has HF property contrary to that of I.

ing vulnerable to adversarial examples. Their results suggest that surface-statistical regularities, exhibiting HF property, are useful for classification. Similar findings have also been shown in (Ilyas et al. 2019) that human unrecognizable non-robust-features with HF property are sufficient for the model to exhibit high generalization capability. On the other hand, DNNs trained only on low-pass filtered images appearing to be simple globs of color are also found to be sufficient for generalizing with high accuracy (Yin et al. 2019). Overall, there is solid evidence that both HF features and LF features can be useful for classification. It is interesting to explore whether a DNN is more biased towards HF or LF features. One work (Geirhos et al. 2019) shows that DNNs are more biased towards texture than shape through a texture-shape cue conflict. Given that texture mainly has HF content and the shape can be seen to have LF content (most flat regions except the object boundary), it can be naturally conjectured that DNNs are more biased towards HF content. We verify this by presenting extensive analysis. We acknowledge that this does not constitute a major discovery, instead, we highlight that *we apply it to explain the model robustness to UAPs in the context of independent yet conflicting features in the* $I + P$. Regarding the Fourier perspective to model robustness, adversarial perturbations are widely known to have the HF property, motivated by which several defense methods (Aydemir, Temizel, and Temizel 2018; Das et al. 2018; Liu and JaJa 2019) have been explored. However, Yin et al. concluded that "*Adversarial examples are not strictly a high frequency phenomenon*", which echoed with explorations of LF perturbations (Guo, Frank, and Weinberger 2020; Sharma, Ding, and Brubaker 2019) as well as the finding in (Carlini and Wagner 2017) regarding false claims of detection methods that use PCA (Gong, Wang, and Ku 2017; Grosse et al. 2017; Metzen et al. 2017). Our claim that *UAPs attacking most images is a strictly HF phenomenon* does not conflict with the claim in (Yin et al. 2019) because they implicitly mainly discuss IDPs, not UAPs.

On Universal Adversarial Attack. The reason for the existence of IDP has been analyzed from various perspectives (Qiu et al. 2019), such as local linearity (Goodfellow, Shlens, and Szegedy 2015; Tabacof and Valle 2016), input high-dimension (Shafahi et al. 2019; Fawzi, Fawzi, and Fawzi 2018; Mahloujifar, Diochnos, and Mahmoody 2019; Gilmer et al. 2018), limited sample (Schmidt et al. 2018; Tanay and Griffin 2016), boundary tilting (Tanay and Griffin 2016), test error in noise (Fawzi, Moosavi-Dezfooli, and Frossard 2016; Gilmer et al. 2019; Cohen, Rosenfeld, and Kolter 2019), non-robust features (Bubeck et al. 2019; Nakkiran 2019; Ilyas et al. 2019), batch normalization (Benz et al. 2021; Benz, Zhang, and Kweon 2020) etc. These explanations for IDPs do not come to a consensus that can be directly used to explain the existence of UAPs. The image-agnostic nature of UAPs requires a specific explanation. Early works (Moosavi-Dezfooli et al. 2017a,b; Jetley, Lord, and Torr 2018; Moosavi-Dezfooli et al. 2019) focused on why a single UAP can fool most samples across the decision boundary and they attributed the existence of UAPs to the large curvature of the decision boundary. (Zhang et al. 2020b) shows that UAPs have independent semantic features that dominate the image features. Recently, class-wise UAPs (Zhang et al. 2020a) and double targeted UAPs (Benz et al. 2020) have also been investigated for making the universal attack more stealthy. Refer to (Zhang et al. 2021b) for a survey on universal attack.

When Adversarial Examples Meet Deep Steganography. Applying deep learning to steganography (Sharda and Budhiraja 2013; Shivaram et al. 2013) has recently become an active research field. Hiding binary messages has been explored in (Hayes and Danezis 2017; Zhu et al. 2018; Wengrowski and Dana 2019) and hiding image (or videos) has been explored in (Baluja 2017; Weng et al. 2018; Mishra et al. 2019). It is crucial to understand how the DNN works in DS. (Baluja 2017, 2019) disproves the possibility of the secret image being hidden in the least significant bit (LSB).

Recent work (Zhang et al. 2020c) shows that the success of DS can be attributed to the frequency discrepancy between cover image and encoded secret image. Refer to (Zhang et al. 2021c) for a survey on deep hiding. Joint investigation of AA and DS has also been investigated by proposing a unified notion of black-box attacks against both tasks (Quiring, Arp, and Rieck 2018), applying the lesson in multimedia forensics to detect adversarial examples (Schöttle et al. 2018). Our work differentiates by focusing on the "universal" property with a Fourier perspective.

Fourier Transform and Frequency

Since a large portion of the analysis in this work is dependent on the understanding of image frequency, here we summarize the main points regarding the Fourier transform which is one basic tool to perform image frequency analysis. Sharp contrast edges in the spatial image are considered as HF content, while smooth or constant patches are LF (Lim 1990). Natural images have the Fourier spectrum concentrated in low-medium frequency range that are in the center of the Fourier image. For performing frequency filtering, we define $X_f = \mathcal{F}^{-1}(f(\mathcal{F}(X), bw))$, where f indicates frequency filtering with the bandwidth bw. For high-pass (HP) filtering, $f(z(i,j), bw)$=$z(i,j)$ if $|i - W/2| >= bw/2$ or $|j - H/2| >= bw/2$, otherwise zero; for low-pass (LP) filtering, $f(z(i,j), bw)$=$z(i,j)$ if $|i - W/2| <= bw/2$ and $|j - H/2| <= bw/2$, otherwise zero. W and H are image width and height. Fourier images provide a qualitative presentation for the frequency analysis. No metric has been found to quantify the frequency distribution; to facilitate quantitative cosine similarity analysis in this work, we introduce one simple metric: entropy of the Fourier image z, *i.e.* $E(z) = -\sum_i \sum_j P(z(i,j))\log(P(z(i,j)))$ with $P(z(i,j))$ referring to element probability. Higher entropy indicates more energy being spread to HF regions of z, thus indicating the image has more HF content. Note that the entropy is calculated on the transform image $z(i,j)$ instead of the original image.

Methods for USP and UAP

Technically, UAPs are crafted to attack a target DNN while DS learns a pair of DNNs for encoding/decoding. Both tasks share a misalignment phenomenon between the human observer and the involved DNN. Specifically, in both cases, a human observer finds that the perturbed image looks natural, but the DNN gets fooled (for AA) or reveals a hidden image (for DS). Motivated by the observation of shared misalignment phenomenon, we deem it meaningful to study the two tasks in parallel to provide a unified perspective on this phenomenon.

Adopted USP Generation Method

Our adopted universal secret perturbation (USP) framework (Zhang et al. 2020c) is shown in Figure 2. Through a decoder DNN, a secret image S is transformed into a secret perturbation S_p, *i.e.* USP. This S_p can be randomly added to any cover C, resulting in container C'. From C', the decoder retrieves the hidden secret image S'. Following (Zhang et al. 2020c) we use the average pixel discrepancy (APD), defined as the L_1-norm of the gap between two images, to measure the hiding and revealing performance.

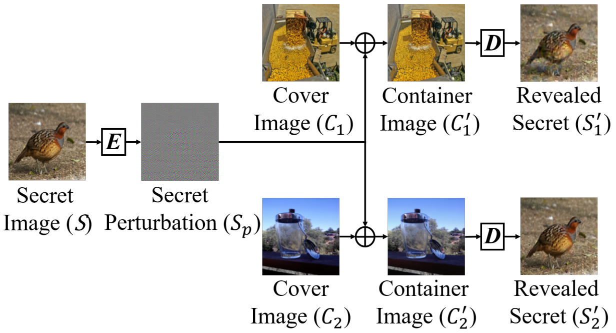


Figure 2: USP generation method. E indicates the encoder network, while D is the decoder network. A secret image is encoded to the secret perturbation S_p, which can be added to random cover images for hiding. We show two different cover images to indicate their random choice.

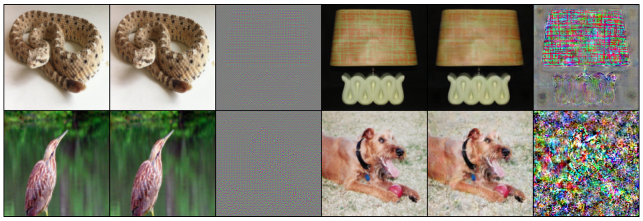


Figure 3: The first three columns indicate cover image C, container image C' and $C' - C$, *i.e.* S_p; the next three columns indicate secret image S, revealed secret imageS' and $S' - S$ respectively. Both $C' - C$ and $S' - S$ are amplified for visualization.

Quantitative results evaluated on the ImageNet validation dataset are shown in Table 1. The two scenarios of IDP and USP are performed with the same procedure as in (Zhang et al. 2020c). The qualitative results are shown in Figure 3, where the difference between C and C' as well as that between S and S' are challenging to identify.

Adopted UAP Generation Method

The adopted procedure for generating universal perturbation is illustrated in Algorithm 1, where a differentiable frequency filter $\mathcal{F}$ is adopted to control the frequency of the UAP. We treat the $\mathcal{F}$ as all-frequency pass at this stage, which makes it similar to the UAP algorithm introduced in (Zhang et al. 2020b,a). For $\mathcal{L}$, we adopt the widely used negative cross-entropy loss. Except for the image-agnostic nature, this algorithm can be seen adapted from the widely used PGD attack (Madry et al. 2018; Athalye, Carlini, and Wagner 2018). The vanilla UAP (Moosavi-Dezfooli et al. 2017a) generation process uses DeepFool (Moosavi-Dezfooli, Fawzi, and Frossard 2016) to generate a perturbation to push a single sample over the decision boundary and accumulates those perturbations to the final UAP. The

meta-archs	cAPD	sAPD (C')	sAPD (S_p)
IDP	2.44	3.42	N/A
USP	2.37	3.52	1.98

Table 1: Performance comparison for the IDP and USP generation frameworks. We report APD for both cover image (cAPD) and secret image (sAPD). For the secret image, we report the results with the container image (sAPD(C')) or only perturbation (sAPD(S_p)) as the input to the decoder network. N/A indicates revealing fails thus not available.

Method	AlexNet	GoogleNet	VGG16	VGG19	ResNet152
Our UAP	94.36	86.03	92.58	94.4	86.67
Our HP-UAP	91.1	84.4	92.3	90.1	78.4
Our targ. UAP	73.77	68.87	81.59	78.67	74.0

Table 2: Performance for untargeted (top) and targeted (bottom) attack with target class "red panda". The reported values are untargted and targeted fooling ration (%).

Algorithm 1: Universal attack algorithm

```
Input: Dataset $\mathcal{X}$, Loss $\mathcal{L}$, Target Model $M$,
       frequency Filter $\mathcal{F}$, batch size $b$
$v \leftarrow 0$                                  ▷ Initialization
for iteration $= 1, \ldots, N$ do
    $B \sim \mathcal{X}$: $|B| = b$               ▷ Randomly sample
    $g_v \leftarrow \mathop{\mathbb{E}}_{x,y \sim B}[\nabla_v \mathcal{L}(M(x + \mathcal{F}(v)), y)]$
    $v \leftarrow$ Adam($g_v$)                    ▷ Update perturbation
    $v \leftarrow$ Clamp($v, -\epsilon, \epsilon$)        ▷ Clamping
end
```

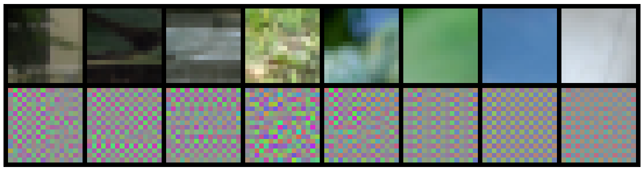


Figure 4: Local patch mapping from corresponding secret image S to secret perturbation S_p.

adopted algorithm is different from the vanilla UAP algorithm (Moosavi-Dezfooli et al. 2017a) by replacing the relatively cumbersome DeepFool (Moosavi-Dezfooli, Fawzi, and Frossard 2016) perturbation optimization with simple batch gradients. ADAM optimizer (Kingma and Ba 2015) is adopted for updating the perturbation values. A similar ADAM based approach has also been adopted for universal adversarial training (Shafahi et al. 2020).

Following (Moosavi-Dezfooli et al. 2017a; Poursaeed et al. 2018; Zhang et al. 2020b), we generate the perturbation with $\epsilon = 10/255$ on the ImageNet training dataset and evaluate it on the ImageNet validation dataset. The results for untargeted and targeted UAPs are shown in Table 2. Our simple algorithm achieves high (targeted) fooling ratio.

Explaining the USP Induced Misalignment

In the whole pipeline from S through S_p to S', in essence, the role of the C is just like noise. It is counter-intuitive that the pipeline still works well under such large disturbance($||I|| \gg ||P||$). Due to the independent property of S_p, we can visualize S_p directly, which is very crucial for qualitatively understanding how the secret image S is encoded in S_e (Zhang et al. 2020c). The visualization in Figure 4 clearly shows that S_p has very HF content.

Why does USP Have High Frequency? The decoder network recovers S' from S_p but with the existence of C as a disturbance. Intuitively its role can be decomposed into two parts: distinguishing S_p from C in C' and transforming S_p to S'. We conjecture that secret perturbation having high frequency mainly facilitates the role of distinguishing. To verify this, we design a toy task of scale hiding, where we assume/force the encoder to perform a trivial transformation as $S_p = Encoder(S) = S/10$. We then only train the decoder network to perform the inverse up-scaling transformation with the natural C as the disturbance. After the model is trained, we evaluate it in two scenarios: with and without the C. The revealing results are present in the supplementary [1]. We observe that the secret image can be recovered reasonably well without the C but fails to work with the C. This suggests the transformation S_p to S' has been trained well but still is not robust to the disturbance of C, which indicates trivial encoding just performing the magnitude change fails. Since natural images C mainly have LF content, it is not surprising that S_p is trained to have HF content, which significantly facilitates the decoder to distinguish S_p from C. The decoder network is implicitly trained to ignore LF content in C, while transforming the HF S_p back to S'. Thus, the revealing performance can be significantly influenced by the image frequency property.

Frequency: A Key Factor for Performance. We perform analysis with three types of images: artificial flat images with constant values in each RGB channel, natural images, and noise sampled from a uniform distribution of 0 to 1. The results are shown in Table 3. Note that flat images are LF while noise images have HF property. The secret APD performance decreases with the increase of frequency for both secret images and cover images. Since the secret perturbation S_p mainly has high frequency, the increase of frequency in the cover images will disrupt more on the S_p, resulting in the performance to decrease. The task complexity also increases with the increase in the frequency of secret images. Revealing fails when either S or C is random noise.

Explaining the UAP Induced Misalignment

Inspired by the above explanation of USP induced misalignment from the Fourier perspective, we extend it to understanding the success of UAP by showing that the target DNN is highly sensitive to HF content.

Disentangling Frequency and Magnitude. We explore

[1] Supplementary: https://arxiv.org/abs/2102.06479

	S_{Flat}	$S_{Natural}$	S_{Noise}
C_{Flat}	0.34	1.85	48.06
$C_{Natural}$	1.77	3.52	49.48
C_{Noise}	87.45	98.33	100.47

Table 3: Secret APD performance with three types of images. The rows and columns indicate cover images and secret images, respectively.

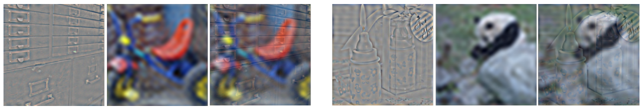


Figure 5: The columns for each image triplet indicate HF image, LF image and hybrid image, respectively.

the target DNN's sensitivity to features of different frequencies. Specifically, we analyze the dominance of two independent inputs on the combined output with the cosine similarity cos metric (Zhang et al. 2020b). I represents a natural image, while P is an image that extracts the content of a certain frequency range ω which is one control variable. We normalize P to have uniform magnitude and then multiply it by a new magnitude m which is another control variable. We then calculate $cos(M(I), M(I+P))$ and $cos(M(P), M(I+P))$. For a detailed result, refer to the supplementary, here we summarize the main findings: As expected, a higher magnitude m leads to higher dominance. On the other hand, we find that ω has an (even more) significant influence on the model prediction. Specifically, higher frequency leads to higher dominance.

Hybrid Images: HF vs. LF. The target DNN achieves high accuracy and we are interested in finding out whether HF content or LF content dominantly contributes to the success. Note that the targeted DNN has been trained on natural images containing both HF content and LF content and the learning algorithm does not involve any manual intervention to force the model to utilize high or low frequency. Manually forcing the model to specifically learn either LF or HF is possible as performed in (Yin et al. 2019). In contrast to their setup, we evaluate the performance of a normally trained model to filtered images. For a normally trained DNN, we show the usefulness of features with LF or HF content in the natural images as well as explore which side dominates in a hybrid image (Oliva, Torralba, and Schyns 2006), which combines the low frequencies of one image with the high frequencies of another. The qualitative results with bw of 20 are available in Figure 5. We observe that a hybrid image visually looks more similar to the LF image. The quantitative results of hybrid images are shown in Table 4. In a hybrid setup, the LF image feature is dominated by the HF one.

The hybrid setup is similar to the universal attack setup because the LF content image is not targeted for any specific HF content image and they are randomly combined. Overall, we observe that the LF image content dominates the human vision, while the HF image content dominates the DNN perception, *i.e.* prediction. Given the dominance of HF content on the DNN, it is not surprising that the optimization-based UAP with HF property can dominate most natural images.

	24	20	16	12
HF	23.13	31.07	41.79	53.31
LF	16.07	10.62	6.14	3.04
Hybrid HF	15.95	20.39	26.54	34.31
Hybrid LF	0.87	0.52	0.32	0.21

Table 4: Top1 accuracy (%) for LF, HF, and hybrid images on the ImageNet val dataset evaluated on the VGG19 network. Hybrid HF indicates the accuracy when the HF images labels are chosen as the ground-truth for the Hybrid images. Parallel reasoning applies to Hybrid LF. The columns indicate the bandwidth.

Class-Wise Robustness Imbalance. We randomly choose a targeted class *"red panda"* for performing a universal attack on VGG19. We find that robust classes have a targeted attack success rate of around 40%, while that for non-robust classes is 100%. One interesting observation from the qualitative results with Fourier analysis in Figure 7 is that all the classes with high robustness have repetitive semantic feature patterns, *i.e.* , HF features, such as the patterns on the feathers of a peacock. The classes with low robustness have LF feature patterns, such as the monotone color of a white washbasin. A Fourier analysis of samples from these classes confirms that robust classes have more HF features, making them more robust against UAP. This analysis shows that there are significant class-wise robustness disparity and the key factor that influences its robustness is their frequency. This also provides extra evidence that the DNN is biased towards HF features. Our work is the first to report and analyze this class-wise imbalanced robustness against UAP.

Feature Layer Analysis. In contrast to prior works performing dominance analysis on the DNN output (Zhang et al. 2020b), we provide fine-grained feature layer analysis with cos to provide deep insight on generalization and robustness of a target DNN, *e.g.* VGG19. In Figure 8, a high cos indicates a dominant contribution to the DNN response. First, with the introduced entropy metric, we explore the influence of the image frequency property on its robustness against UAP. We reveal that images of high entropy (HE) (indicating more HF content) are much more robust to UAP on all feature layers, especially on latter layers (see Figure 8 left). For example, at layer of $conv6$, $cos(M_i(I), M_i(I+P))$ is around 0.9 and 0 for images of HE and LE, respectively. The results clearly show that images with more HF content are more robust, which aligns well with our finding on class-wise robustness imbalance. Second, comparing $cos(M_i(I), M_i(HP(I)))$ and $cos(M_i(I), M_i(LP(I)))$ shows $cos(M_i(I), M_i(LP(I)))$ is higher only in the first two layers and then significantly lower in latter layers (see Figure 8 right). In other words, except for the very first few layers, all layers of DNN are much more sensitive to HF instead of LF content. When P is noise, $cos(M_i(I), M_i(I+P))$ first decreases and then increases again, with the $conv3$ being the most vulnerable to noise. The influence of random

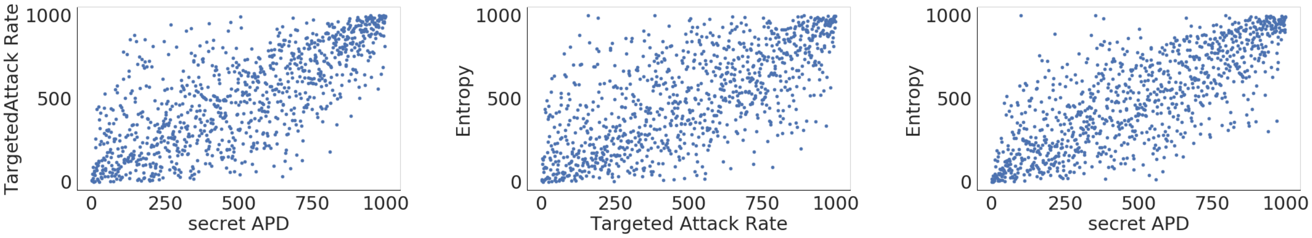


Figure 6: Ranking correlation with three ranking metrics.

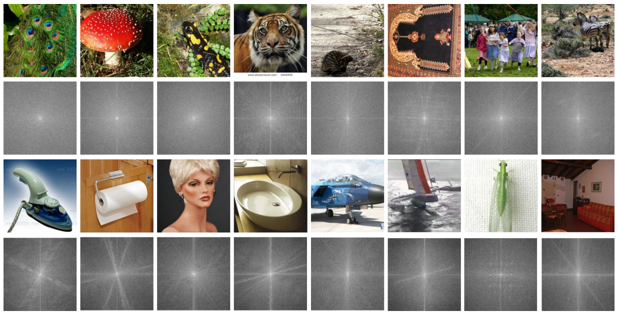


Figure 7: Fourier analysis of representative samples. We randomly choose one sample from 8 top robust classes and non-robust classes to perform Fourier analysis.

noise on latter layers is very limited, which provides insight on why DNN is robust against noise but not UAP.

Joint Analysis for Two Tasks

Cross-Task Cosine Similarity Analysis for Class Ranking We perform a *cos* analysis between two seemingly unrelated tasks, DS and AA. Specifically, the 1000 ImageNet classes were ranked along the attack success rate metric (R_1), secret APD metric (R_2) and the Fourier image entropy metric (R_3). The ranking plots of R_1 over R_2, R_3 over R_1, and R_3 over R_2 are shown in Figure 6. We find that $cos(R_1, R_2)$ is 0.74, indicating high linear correlation for two seemingly unrelated tasks. The fact that class robustness is an indicator of the revealing performance in DS task clearly shows that a certain factor exists to link them and we identify this factor to be *frequency*. Note that R_3 is our proposed metric for quantifying the energy distribution (corresponding to each frequency) of Fourier images. $cos(R_1, R_3)$ and $cos(R_2, R_3)$ are 0.68 and 0.77, respectively, attributing the high correlation between R_1 ranking and R_2 ranking to the *frequency*.

Investigation of LF Perturbation. To investigate the behavior of perturbations containing LF features we explore two methods: loss regularization and low-pass filtering. Similar to (Mahendran and Vedaldi 2015) we add a regularization term to the loss function during universal perturbation generation to force the perturbation to be smooth for both tasks. The results are shown in Figure 9. The results show that regularizing the perturbation to enforce smoothness results in a significant performance drop. An LF perturbation can also be enforced by performing an LP filtering to the perturbation before adding the perturbation to the image, for which $\mathcal{F}$ is a differentiable LPF (LP filter) in Algorithm 1. Smoothing the perturbations with an LPF and lead to lower attack success rates, see Figure 10 (top). Overall, regarding model robustness, we find that *UAP that attacks most images is a strictly high-frequency (HF) phenomenon*. Inspired by the above results, we propose a novel high-pass (HP) universal attack, *i.e.* **HP-UAP**, by setting $\mathcal{F}$ to be a differentiable

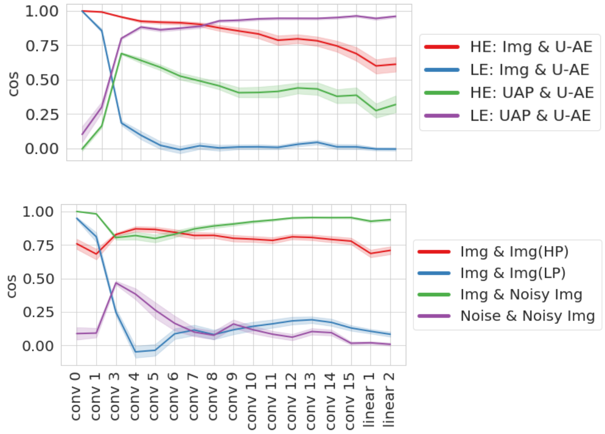


Figure 8: *cos* analysis on feature layers evaluated on 100 images. The abbreviations in the legends refer to: clean image (img), universal adversarial example (U-AE), universal adversarial perturbation (UAP), high/low entropy (HE/LE), high/low pass (HP/LP) filtered with the *bw* of 16.

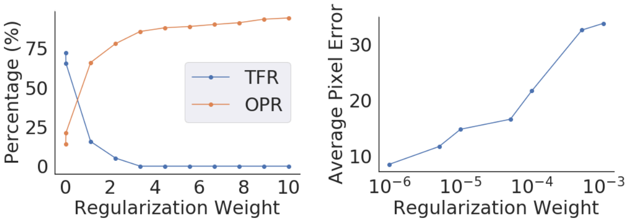


Figure 9: Regularization effect on UAP (left). TFR: Targeted Fooling Ratio; OPR: Original Prediction Ratio which indicates a ratio of samples keeping the same prediction. Regularization effect on USP (right). Secret APD is reported.

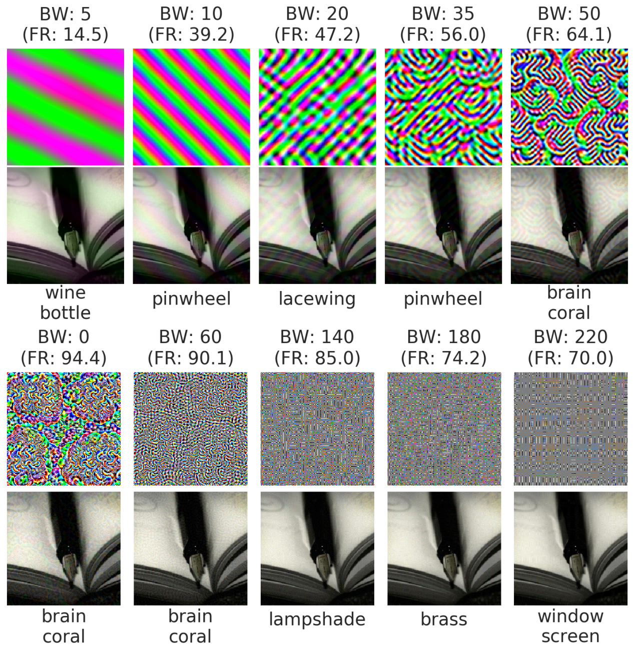


Figure 10: Examples for LP UAPs (top) and HP UAPs (bottom). The first row shows the perturbations for different bandwidths. The bandwidth (BW) and the achieved fooling ratio (FR) are written above the corresponding perturbation. The second row shows the adversarial example with the predicted class of VGG19 written below. The originally predicted and ground truth class is "fountain pen".

Metric	AlexNet	GoogleNet	VGG16	VGG19	ResNet152
Fooling Ratio	93.8	85.0	92.7	95.8	90.3
sAPD	13.6	8.9	14.2	11.1	11.9

Table 5: Performance evaluation of the proposed USAP.

HPF (HP filter) in Algorithm 1. Overall we observe a performance drop in fooling ratio with increasing bw. Results for the HP UAP generated for VGG19 are shown in Figure 10 (bottom). With bw 60, the perturbation is less visible to the human vision and still achieves a fooling ratio of 90.1%, with only a moderate performance drop compared with the 94.4% for bw 0 without filtering.

Universal Secret Adversarial Perturbation

We explore whether a single perturbation can fool the DNN for most images while simultaneously containing the secret information, termed universal secret adversarial perturbation (USAP). As shown in Figure 11, the secret image S is fed to the encoder DNN to produce the USAP S_p. It is then added to a random cover image C to produce the container image C', which fools the target DNN and reveals the secret S' if it is fed to the decoder DNN. Note, that during training, only encoder and decoder DNNs are trained, while the target DNN is a fixed pre-trained classifier. Therefore, after training, the encoder DNN can turn any image into USAP, so the approach is different from the original UAP generation where perturbation's pixels are updated. We adopt the same USP generation network, while adding an additional loss term $NCE(M(C'), y)$ resulting in the loss function: $\mathcal{L}(S_p, S, S', C') = ||S_p||+\beta||S-S'||+\gamma NCE(M(C'), y)$, where NCE indicates the negative cross-entropy loss and y indicates the ground-truth label. We set β and γ to 0.75 and 0.001, respectively. The USAP is constrained to be in $L_\infty = 10/255$. The results in Table 5 and Figure 12 demonstrate a high fooling ratio while containing secret information that can successfully be revealed by the decoder. We are the first to show the existence of such perturbation.

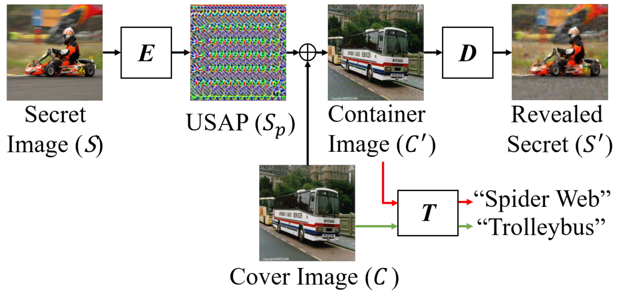


Figure 11: Universal secret adversarial perturbation (USAP) setup. E stands for the Encoder DNN, D stands for the Decoder DNN, and T is the Target DNN. The USAP S_p is shown in its amplified form for better visibility.

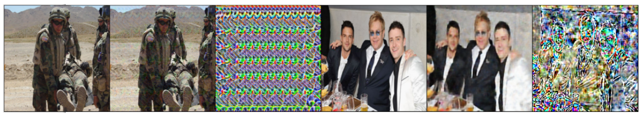


Figure 12: Qualitative result of USAP. The column order is the same as that in Fig 3. The container is misclassified from "military uniform" to "spider web".

Conclusion

This work has jointly analyzed AA and DS for the observed misalignment phenomenon and explained their success from the Fourier perspective. With the proposed metric for quantifying frequency distribution, extensive task-specific and cross-task analysis suggest that frequency is a key factor that influences their performance and their success can be attributed to the DNN being highly sensitive to HF content. Our feature layer analysis sheds new light on model generalization and robustness: (a) Images with more high-frequency content are more robust against UAP; (b) the influence of LF features on the DNN diminishes in the later layers. We also proposed two new variants of universal attacks: HP-UAP that is less visible to the human and USAP that simultaneously achieves attack and hiding.

Ethics Statement

Due to security concerns, adversarial attack and deep steganography have become popular topics in recent years.

We hope that our work will raise awareness of this security concern to the public.

References

Argaw, D. M.; Kim, J.; Rameau, F.; Cho, J. W.; and Kweon, I. S. 2021a. Optical Flow Estimation from a Single Motion-blurred Image. In *AAAI*.

Argaw, D. M.; Kim, J.; Rameau, F.; and Kweon, I. S. 2021b. Motion-blurred Video Interpolation and Extrapolation. In *AAAI*.

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*.

Aydemir, A. E.; Temizel, A.; and Temizel, T. T. 2018. The effects of JPEG and JPEG2000 compression on attacks using adversarial examples. *arXiv preprint arXiv:1803.10418* .

Baluja, S. 2017. Hiding images in plain sight: Deep steganography. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Baluja, S. 2019. Hiding images within images. *T-PAMI* .

Benz, P.; Zhang, C.; Imtiaz, T.; and Kweon, I. S. 2020. Double Targeted Universal Adversarial Perturbations. In *ACCV*.

Benz, P.; Zhang, C.; Karjauv, A.; and Kweon, I. S. 2021. Revisiting Batch Normalization for Improving Corruption Robustness. *WACV* .

Benz, P.; Zhang, C.; and Kweon, I. S. 2020. Batch Normalization Increases Adversarial Vulnerability: Disentangling Usefulness and Robustness of Model Features. *arXiv preprint arXiv:2010.03316* .

Bubeck, S.; Lee, Y. T.; Price, E.; and Razenshteyn, I. 2019. Adversarial examples from computational constraints. In *PMLR*.

Carlini, N.; and Wagner, D. 2017. Adversarial examples are not easily detected. In *ACM Workshop on Artificial Intelligence and Security-AISec'17*.

Cohen, J. M.; Rosenfeld, E.; and Kolter, J. Z. 2019. Certified adversarial robustness via randomized smoothing. In *Proceedings of Machine Learning Research*.

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 *KDD*.

Fawzi, A.; Fawzi, H.; and Fawzi, O. 2018. Adversarial vulnerability for any classifier. In *NeurIPS*.

Fawzi, A.; Moosavi-Dezfooli, S.-M.; and Frossard, P. 2016. Robustness of classifiers: from adversarial to random noise. In *NeurIPS*.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*.

Gilmer, J.; Ford, N.; Carlini, N.; and Cubuk, E. 2019. Adversarial examples are a natural consequence of test error in noise. In *ICML*.

Gilmer, J.; Metz, L.; Faghri, F.; Schoenholz, S. S.; Raghu, M.; Wattenberg, M.; and Goodfellow, I. 2018. Adversarial spheres. *arXiv preprint arXiv:1801.02774* .

Gong, Z.; Wang, W.; and Ku, W.-S. 2017. Adversarial and clean data are not twins. *arXiv preprint arXiv:1704.04960* .

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *ICLR*.

Grosse, K.; Manoharan, P.; Papernot, N.; Backes, M.; and McDaniel, P. 2017. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280* .

Guo, C.; Frank, J. S.; and Weinberger, K. Q. 2020. Low Frequency Adversarial Perturbation. In *PMLR*.

Hayes, J.; and Danezis, G. 2017. Generating steganographic images via adversarial training. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; and Madry, A. 2019. Adversarial examples are not bugs, they are features. In *NeurIPS*.

Jetley, S.; Lord, N.; and Torr, P. 2018. With friends like these, who needs adversaries? In *NeurIPS*.

Jo, J.; and Bengio, Y. 2017. Measuring the tendency of CNNs to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561* .

Kim, D.; Woo, S.; Lee, J.-Y.; and Kweon, I. S. 2020. Video Panoptic Segmentation. In *CVPR*.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Lee, S.; Kim, J.; Oh, T.-H.; Jeong, Y.; Yoo, D.; Lin, S.; and Kweon, I. S. 2019a. Visuomotor Understanding for Representation Learning of Driving Scenes. *BMVC* .

Lee, S. L.; Im, S.; Lin, S.; and Kweon, I. S. 2019b. Learning Residual Flow as Dynamic Motion from Stereo Video. *IROS* .

Lim, J. S. 1990. *Two-Dimensional Signal and Image Processing*. Prentice-Hall, Inc.

Liu, C.; and JaJa, J. 2019. Feature Prioritization and Regularization Improve Standard Accuracy and Adversarial Robustness. In *IJCAI*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. In *ICLR*.

Mahendran, A.; and Vedaldi, A. 2015. Understanding deep image representations by inverting them. In *CVPR*.

Mahloujifar, S.; Diochnos, D. I.; and Mahmoody, M. 2019. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. In *AAAI*.

Metzen, J. H.; Genewein, T.; Fischer, V.; and Bischoff, B. 2017. On detecting adversarial perturbations. In *ICLR*.

Mishra, A.; Kumar, S.; Nigam, A.; and Islam, S. 2019. VStegNET: Video Steganography Network using Spatio-Temporal features and Micro-Bottleneck. *BMVC* .

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017a. Universal adversarial perturbations. In *CVPR*.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; Frossard, P.; and Soatto, S. 2017b. Analysis of universal adversarial perturbations. *arXiv preprint arXiv:1705.09554* .

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Uesato, J.; and Frossard, P. 2019. Robustness via curvature regularization, and vice versa. In *CVPR*.

Nakkiran, P. 2019. A Discussion of 'Adversarial Examples Are Not Bugs, They Are Features': Adversarial Examples are Just Bugs, Too. *Distill* Https://distill.pub/2019/advex-bugs-discussion/response-5.

Oliva, A.; Torralba, A.; and Schyns, P. G. 2006. Hybrid images. *TOG* .

Pan, F.; Shin, I.; Rameau, F.; Lee, S.; and Kweon, I. S. 2020. Unsupervised Intra-domain Adaptation for Semantic Segmentation through Self-Supervision. In *CVPR*.

Poursaeed, O.; Katsman, I.; Gao, B.; and Belongie, S. 2018. Generative adversarial perturbations. In *CVPR*.

Qiu, S.; Liu, Q.; Zhou, S.; and Wu, C. 2019. Review of artificial intelligence adversarial attack and defense technologies. *Applied Sciences* .

Quiring, E.; Arp, D.; and Rieck, K. 2018. Forgotten siblings: Unifying attacks on machine learning and digital watermarking. In *EuroS&P*.

Schmidt, L.; Santurkar, S.; Tsipras, D.; Talwar, K.; and Madry, A. 2018. Adversarially robust generalization requires more data. In *NeurIPS*.

Schöttle, P.; Schlögl, A.; Pasquini, C.; and Böhme, R. 2018. Detecting adversarial examples-A lesson from multimedia forensics. *arXiv preprint arXiv:1803.03613* .

Shafahi, A.; Huang, W. R.; Studer, C.; Feizi, S.; and Goldstein, T. 2019. Are adversarial examples inevitable? In *ICLR*.

Shafahi, A.; Najibi, M.; Xu, Z.; Dickerson, J. P.; Davis, L. S.; and Goldstein, T. 2020. Universal Adversarial Training. In *AAAI*.

Sharda, S.; and Budhiraja, S. 2013. Image steganography: A review. *IJETAE* .

Sharma, Y.; Ding, G. W.; and Brubaker, M. A. 2019. On the Effectiveness of Low Frequency Perturbations. In *IJCAI*.

Shivaram, H.; Acharya, D.; Adige, R.; and Kamath, P. 2013. A Secure And High Capacity Image Steganography Technique. *Signal & Image Processing : An International Journal* .

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* .

Tabacof, P.; and Valle, E. 2016. Exploring the space of adversarial images. In *IJCNN*.

Tanay, T.; and Griffin, L. 2016. A boundary tilting persepective on the phenomenon of adversarial examples. *arXiv preprint arXiv:1608.07690* .

Vania, M.; Mureja, D.; and Lee, D. 2019. Automatic spine segmentation from CT images using convolutional neural network via redundant generation of class labels. *Journal of Computational Design and Engineering* .

Wang, H.; Wu, X.; Yin, P.; and Xing, E. P. 2020. High frequency component helps explain the generalization of convolutional neural networks. In *CVPR*.

Weng, X.; Li, Y.; Chi, L.; and Mu, Y. 2018. Convolutional video steganography with temporal residual modeling. *arXiv preprint arXiv:1806.02941* .

Wengrowski, E.; and Dana, K. 2019. Light Field Messaging With Deep Photographic Steganography. In *CVPR*.

Yin, D.; Lopes, R. G.; Shlens, J.; Cubuk, E. D.; and Gilmer, J. 2019. A fourier perspective on model robustness in computer vision. In *NeurIPS*.

Zhang, C.; Benz, P.; Argaw, D. M.; Lee, S.; Kim, J.; Rameau, F.; Bazin, J.-C.; and Kweon, I. S. 2021a. ResNet or DenseNet? Introducing Dense Shortcuts to ResNet. In *WACV*.

Zhang, C.; Benz, P.; Imtiaz, T.; and Kweon, I.-S. 2020a. CD-UAP: Class Discriminative Universal Adversarial Perturbation. In *AAAI*.

Zhang, C.; Benz, P.; Imtiaz, T.; and Kweon, I.-S. 2020b. Understanding Adversarial Examples from the Mutual Influence of Images and Perturbations. In *CVPR*.

Zhang, C.; Benz, P.; Karjauv, A.; Sun, G.; and Kweon, I. 2020c. UDH: Universal Deep Hiding for Steganography, Watermarking, and Light Field Messaging. *NeurIPS* .

Zhang, C.; Benz, P.; Lin, C.; Karjauv, A.; Wu, J.; and Kweon, I. S. 2021b. A Survey On Universal Adversarial Attack. *arXiv preprint arXiv:2103.01498* .

Zhang, C.; Lin, C.; Benz, P.; Chen, K.; Zhang, W.; and Kweon, I. S. 2021c. A Brief Survey on Deep Learning Based Data Hiding, Steganography and Watermarking. *arXiv preprint arXiv:2103.01607* .

Zhang, C.; Rameau, F.; Kim, J.; Argaw, D. M.; Bazin, J.-C.; and Kweon, I. S. 2020d. DeepPTZ: Deep Self-Calibration for PTZ Cameras. In *WACV*.

Zhang, C.; Rameau, F.; Lee, S.; Kim, J.; Benz, P.; Argaw, D. M.; Bazin, J.-C.; and Kweon, I. S. 2019. Revisiting Residual Networks with Nonlinear Shortcuts. In *BMVC*.

Zhu, J.; Kaplan, R.; Johnson, J.; and Fei-Fei, L. 2018. Hidden: Hiding data with deep networks. In *ECCV*.