ADVERSARIAL IMAGE GENERATION BY SPATIAL TRANSFORMATIONS IN COLOR SPACE

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

AYBERK AYDIN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
MODELLING AND SIMULATION

AUGUST 2022

Approval of the thesis:

ADVERSARIAL IMAGE GENERATION BY SPATIAL TRANSFORMATIONS IN COLOR SPACE

submitted by **AYBERK AYDIN** in partial fulfillment of the requirements for the degree of **Master of Science in Modelling and Simulation Department, Middle East Technical University** by,

Prof. Dr. Sevgi Özkan Yıldırım
Dean, Graduate School of **Informatics** _____

Assoc. Prof. Dr. Elif Sürer
Head of Department, **Modelling and Simulation** _____

Prof. Dr. Alptekin Temizel
Supervisor, **Graduate School of Informatics, Middle East Technical University**
_____

Examining Committee Members:

Prof. Dr. Alptekin Temizel
Graduate School of Informatics, Middle East Technical University _____

Assoc. Prof. Dr. Elif Sürer
Graduate School of Informatics, Middle East Technical University _____

Assoc. Prof. Dr. Gökhan Koray Gültekin
Electrical&Electronics Engineering, Ankara Yildirim Beyazit University
_____

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Ayberk Aydin

Signature :

ABSTRACT

ADVERSARIAL IMAGE GENERATION BY SPATIAL TRANSFORMATIONS IN COLOR SPACE

Aydin, Ayberk
M.S., Department of Modelling and Simulation
Supervisor: Prof. Dr. Alptekin Temizel

August 2022, 34 pages

Deep neural networks are known to be vulnerable to adversarial perturbations. The amount of these perturbations are generally quantified using $\mathcal{L}_p$ metrics. However, even when the measured perturbations are small, they tend to be noticeable by human observers since $\mathcal{L}_p$ distance metrics are not representative of human perception. On the other hand, humans are less sensitive to changes in colorspace. In addition, pixel shifts in a constrained neighborhood are hard to notice. Motivated by these observations, we propose a method that creates adversarial examples by applying spatial transformations, which creates adversarial examples by changing the pixel locations independently to chrominance channels of perceptual colorspaces such as YC_bC_r and $CIELAB$, instead of making an additive perturbation or manipulating pixel values directly. In a targeted white-box attack setting, the proposed method is able to obtain competitive fooling rates with very high confidence. The experimental evaluations show that the proposed method has favorable results in terms of approximate perceptual distance between benign and adversarially generated images.

Keywords: A keyword, another keyword, some other keywords

ÖZ

UZAYSAL DÖNÜŞÜMLER İLE ÇEKİŞMELİ ÖRNEK ÜRETİLMESİ

Aydin, Ayberk
Yüksek Lisans, Bölümü
Tez Yöneticisi: Prof. Dr. Alptekin Temizel

Ağustos 2022 , 34 sayfa

TURKCE Deep neural networks are known to be vulnerable to adversarial perturbations. The amount of these perturbations are generally quantified using $\mathcal{L}_p$ metrics. However, even when the measured perturbations are small, they tend to be noticeable by human observers since $\mathcal{L}_p$ distance metrics are not representative of human perception. On the other hand, humans are less sensitive to changes in colorspace. In addition, pixel shifts in a constrained neighborhood are hard to notice. Motivated by these observations, we propose a method that creates adversarial examples by applying spatial transformations, which creates adversarial examples by changing the pixel locations independently to chrominance channels of perceptual colorspaces such as YC_bC_r and $CIELAB$, instead of making an additive perturbation or manipulating pixel values directly. In a targeted white-box attack setting, the proposed method is able to obtain competitive fooling rates with very high confidence. The experimental evaluations show that the proposed method has favorable results in terms of approximate perceptual distance between benign and adversarially generated images.

Anahtar Kelimeler: Bir anahtar kelime, başka bir anahtar kelime, başka anahtar kelimeler

To everyone

ACKNOWLEDGMENTS

This work has been funded by The Scientific and Technological Research Council of Turkey, ARDEB 1001 Research Projects Programme project no: 120E093

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ALGORITHMS

ALGORITHMS

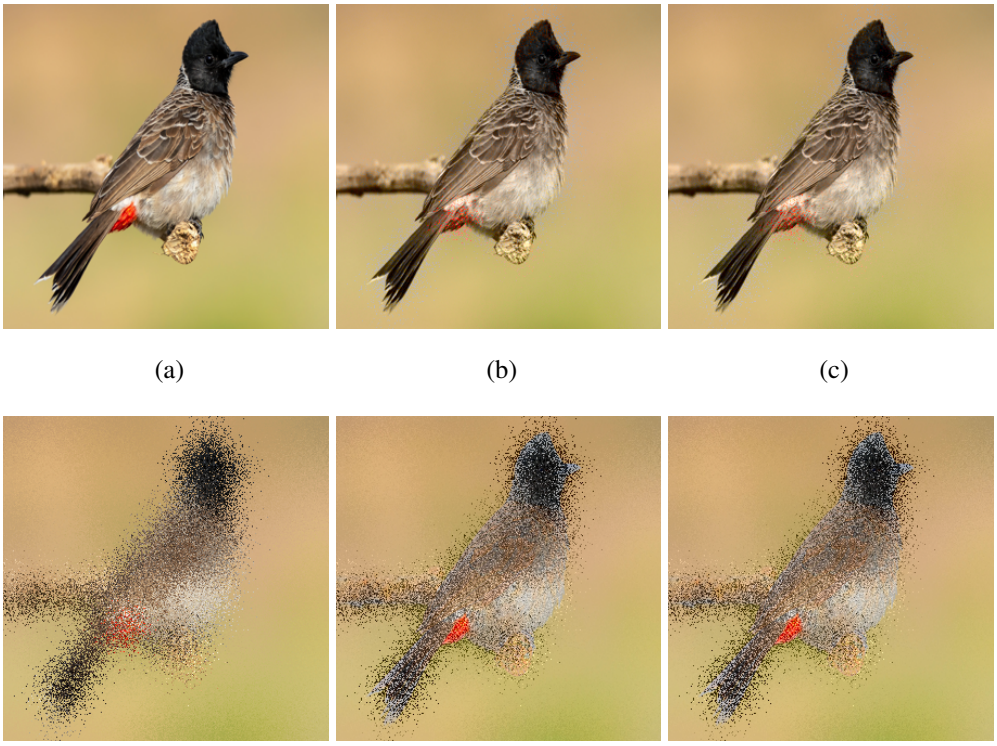
LIST OF ABBREVIATIONS

ESM	Experience Sampling Method
GLMM	Generalized Linear Mixed Model

CHAPTER 1

INTRODUCTION

In recent years, deep neural networks have shown impressive performance in many vision related tasks such as image classification [1], object detection [2] and image segmentation [3]. However, they are found to be vulnerable to intentionally crafted small perturbations called adversarial perturbations [4]. These small perturbations added to the input image successfully change the output of a trained classifier by altering the logits large enough to change its decision to a preferred class [5]. While these perturbations are optimized in $\mathcal{L}_p$ spaces [6], they are visible to human observers, since small $\mathcal{L}_p$ does not always correspond to small visible perturbations [7, 8]. There is an ongoing research on finding difference metrics over 2D images that aligns with human visual perception, which is challenging due to the nature and lack of knowledge about the human vision. Multimedia compression standards have been developed to compress visual multimedia such as images and videos to reduce the amount of data with minimum amount of distortion to the perceived output. One of the most fundamental ideas of visual multimedia compression is that human vision is much less sensitive to the information loss in color than the luminance. This observation is utilized in image compression as a technique known as "chroma subsampling". There are variants of chroma subsampling that only subsamples chrominance along horizontal axis (4:2:2) or both horizontal and vertical axes (4:2:0). Without further compression, (4:2:0) chroma subsampling reduces the size of an image effectively to half of its original size. Replacing the chroma components of the pixels in by neighboring chroma components does not yield visible artifacts. We employ this observation to derive a new type of adversarial attack based on spatial transformations in chroma channels of perceptual colorspaces. We apply spatial transformation only to the chroma components of input image while keeping the luminance component intact. Figure 1.1 shows the

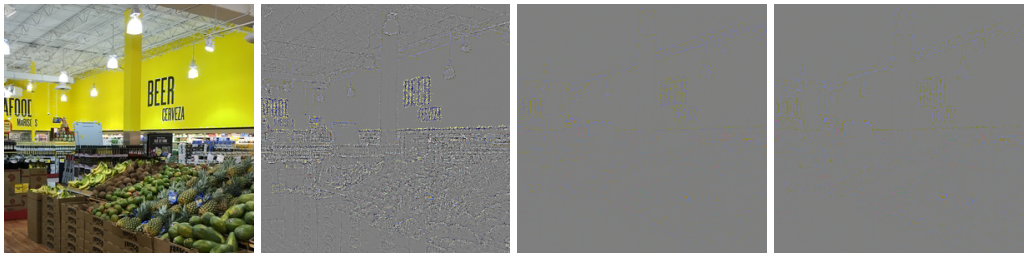


(a) (b) (c)

(d) (e) (f)

Figure 1.1: Effect of flow field applied to different channels, (a) original image, Images where flow field is applied to (b) C_bC_r, (c) a^*b^*, (d) RGB, (e) Y and (f) L channel. The magnitude of the flow is scaled up to emphasize the effect for illustration.

effect of a randomly initialized flow field applied to the luminance, chrominance and both set of channels. It is clear that spatial transformation in luminance channels causes visible distortions while chrominance only spatial transformations cause very subtle changes for human vision. This effect is much more highlighted when only the differences are observed after applying a flow field. Figure 5.4 shows the absolute pixel difference from the initial image when the same flow field is applied to RGB, C_bC_r and a*b* channels, respectively.



(a) (b) (c) (d)

Figure 1.2: Visual difference from flow field applied to different channels, (a) original image, Visualization of pixel differences where flow field is applied to (b) RGB, (c) C_bC_r, (d) a^*b^* channels. The magnitude of the flow is scaled up and contrast of the pixel differences is increased to increase the visibility for illustration.

1.1 Research Questions

1.2 Contributions of the Study

1.3 Organization of the Thesis

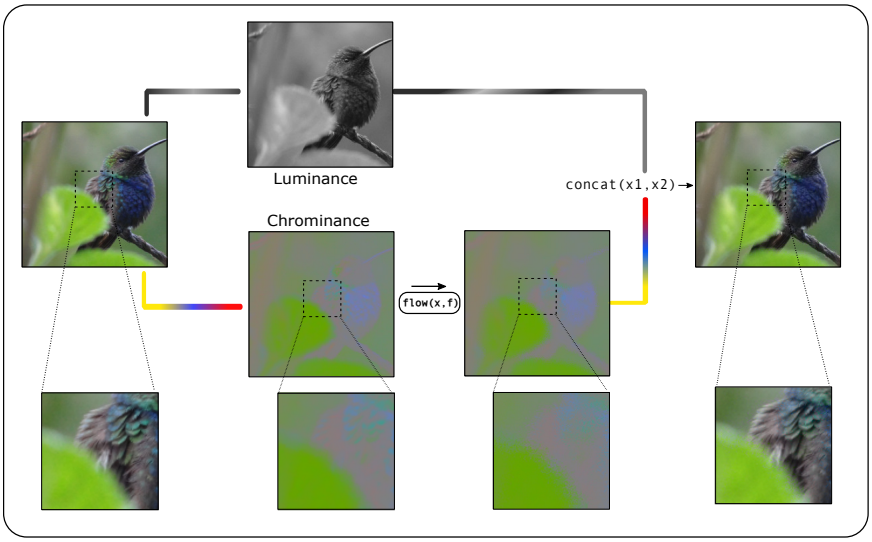


Figure 1.3: Visual illustration of the proposed adversarial example generation method. Luminance and chrominance channels are Y and C_bC_r when YC_bC_r colorspace and L and a^*b^* when CIELAB colorspace is used. Visual representation of flow field, subpixel restriction by $\tanh$ and conversion of concatenated image back to RGB colorspace is omitted for brevity.

CHAPTER 2

RELATED WORK

In this chapter, related studies are given in detail. Spatial transformations as a method for generating adversarial examples was first proposed in [9], where it is shown that small displacements applied to input pixels can successfully fool a target network. However, using this method, even small displacements could cause visible distortions when the adjacent pixels are drifted towards different directions. As a remedy to this problem, use of Total Variation (TV) regularization [10] was proposed. Application of TV regularization to the flow field pushes the neighboring displacement vectors to the same direction and, hence, produces smoother output. Similarly, Jordan et al. combined spatial transformations with l_∞ bounded attacks to forge stronger attacks with better perceptual quality. Croce et al. argued adding noise to smooth areas of an image causes visible artifacts and proposed "hiding" the perturbations at the locations with high spatial variations such as edges and corners [11]. As seen from Figure 1.2, perturbations made with our method naturally occurs in the places with high variations since it is based on local spatial transforms.

Utilizing perceptual colorspaces and metrics for imperceptible adversarial example generation is investigated in several studies. Aksoy et al. investigated additive noise based attacks on chrominance channels in YUV colorspace [12], which is the analog counterpart of YC_bC_r space. Despite Pestana et al. found that adversarial perturbations are more highlighted in luminance channels in terms of the magnitude [13], Aksoy et al. found that even suppressing the luminance perturbation, additive noise based attack on chrominance channels still successfully fool target networks, yet causes visible distortion. In our earlier work, we also explored spatial transformations to UV channels of YUV to generate imperceptible adversarial examples [14]

and we extend this work by exploring YC_bC_r space as well as perceptually uniform CIELAB space and measuring structured similarity metrics such as SSIM [15] and MS-SSIM [16] between benign images and adversarially generated images. Karli et al. leveraged perceptual metric LPIPS [17] to improve the quality of adversarial examples. Since LPIPS is a differentiable metric, they used gradient based optimization to minimize LPIPS alongside the adversarial loss. Similarly, Zhao et al. replaced CIEDE2000 perceptual distance metric [18] with $\mathcal{L}_p$ norm constraint in Carlini & Wagner attack to produce perceptually close adversarial examples.

Unlike these methods, the attack proposed in this paper does not rely on auxiliary losses or explicit perceptual distance terms in optimization process to produce examples with high perceptual quality. In addition, it does not require regularization, unlike spatial transformation based methods such as [9], due to its intrinsic imperceptibility. It should be noted that the existing spatial transformation based methods, as well as our work, does not utilize limited degree of freedom transformations such as rotation, translation or scaling that can be formulated as a 4×4 transformation matrix [19]. In that formulation, the flow field $f \in \mathbb{R}^{2\times H\times W}$ is calculated using the transformation matrix. Instead, we directly define and optimize flow field, where the number of parameters is equal to twice number of pixels in the input image since there is an x and y component for each pixel. Application and optimization of flow field is explained in the Section **??**.

2.1 Related Work Section I

CHAPTER 3

USER EXPERIMENT

In this chapter, the details of the user experiment are presented. In this work, we address the problem of creating targeted adversarial examples without adversarial perturbation being perceptible by human vision. To obtain this, we use a modified version of Spatially Transformed Adversarial Examples [9] that perturbs the input image only in the channels that human vision is not sensitive to the spatial information loss. For this purpose, we use YC_bC_r and CIELAB colorspace representations of the input image. The proposed adversarial example generation method is as follows. Let $x \in \mathbb{R}^{3\times H\times W}$ be the 3-channel input image, where H, W are the height and the width of the image, respectively. First, we randomly initialize a flow field $f \in \mathbb{R}^{2\times H\times W}$ where a two-dimensional vector exists for each pixel location of the adversarial image x_{adv}. Then, we apply the flow field to the benign image as explained below to obtain the adversarial image. Then, we feed the adversarial image to the target network and backpropagate the loss gradient to the flow field. Since the flow field application is a differentiable process, it can be optimized by stochastic gradient descent and variants such as Adam [20] or L-BFGS[21]. The optimization process is repeated until the attack is successful or the maximum iteration count is reached.

3.0.1 Application of flow field

Flow field is applied to the benign image following the methodology in [9]. For each pixel in adversarial image i_{adv}, corresponding flow field vector value $p_{i,j}$ is added to the pixel location. Then, the corresponding pixel at the added location is sampled. Since the added location is not an integer, bilinear interpolation is used to sample from the fractional pixel locations. Bilinear interpolation also makes the method end-to-

end differentiable, thus optimizable by gradient based optimizers. Chroma subsampling effectively causes the same chroma values to be used by the neighboring pixels, and it is widely accepted to cause negligible changes to the images. Accordingly, to exploit this fact, we can impose a restriction to the flow field to keep its values in the range $(-1, 1)$. We initialize a pre-flow field f_{pre} and calculate the applied flow field as $f = \tanh(f_{pre})$. This differentiable reparameterization [22] of flow field constraints the flow field magnitude to be smaller than 1 without inhibiting end-to-end differentiability so that chrominance value of each pixel of the adversarial image x_{adv} is only affected by the value of the pixel of the same location in x and its neighboring pixels.

3.0.2 Colorspace conversion

To make the adversarially perturbed images indistinguishable from their benign counterparts, the flow field is applied only to the channels that human vision is not very sensitive to [23]. Since widely used RGB colorspace is not designed to be a perceptual colorspace, even small spatial perturbations to any RGB channel creates visually distinguishable changes. Hence, we first convert the benign image to a perceptual colorspace such as YC_bC_r where human vision is not sensitive to the spatial perturbations in, which is C_b and C_r in YC_bC_r, and a^* and b^* in CIELAB colorspace. Then, we apply the flow field only to the channels Cb and Cr in YC_bC_r, and A and B in CIELAB colorspace.

YC_bC_r is a colorspace that is used in digital photography and visual media compression. In this space, luminance (brightness) and chrominance (color) is separated according to human visual perception. Y dimension of the space is the luminance information, or simply a grayscale representation of the image. C_b and C_r dimensions are the blue-difference and red-difference chroma components, respectively. The relation between RGB space and YC_bC_r space is modeled as Equation 3.1, which is a set of linear equations defined in ITU-T H.273 [24];

$$\begin{aligned} Y &= 0.299R + 0.587G + 0.114B \\ C_b &= 128 - (0.168736R) - (0.331264G) + (0.5B) \\ C_r &= 128 + (0.5R) - (0.418688G) - (0.081312B) \end{aligned} \tag{3.1}$$

CIELAB colorspace [25] defined by the International Commission on Illumination (CIE) has the following three components: L, a^* and b^*. L is perceptual lightness where $L = 0$ and $L* = 100$ define a black and a white pixel, respectively, regardless of the a^* and b^* values. a^* and b^* dimensions are the chroma components. They are designed to be perceptually uniform where a numerical change in pixel value corresponds to a similar change in human perception [26]. Both chroma components are in the range $[-127, 127]$. Unlike YC_bC_r, CIELAB space does not have a linear relationship with RGB space. In fact, conversion to an intermediary space CIEXYZ is needed to transform from RGB to CIELAB and there are different implementations of CIELAB conversion. We used the RGB to LAB implementation from Kornia library [27], which assumes D65 illuminant and Observer 2.

3.1 Research Method and Experiment Design

Algorithm 1: Adversarial example generation by spatial transformation in chrominance channels in a perceptual colorspace.

Input: x

Output: x_{adv}

Data: target_class, model, κ, colorspace, max_iters, is_restricted,

$f \leftarrow initializeflowfieldf$;

$i \leftarrow 0$;

while $i < max\_iters$ **do**

  if $colorspace == YC_bC_r$ **then**

    $x_{color} \leftarrow to\_ycbcr(x)$;

  end

  if $colorspace == CIELAB$ **then**

    $x_{color} \leftarrow to\_lab(x)$;

  end

  $x_{luma}, x_{chroma} \leftarrow splitchannels(x_{color})$;

  if *is_restricted* **then**

    $f \leftarrow \tanh(f)$

  end

  $x_{chroma} \leftarrow apply\_flow(x_{chroma}, f)$;

  $x_{adv} \leftarrow concat(x_{luma}, x_{chroma})$;

  $x_{adv} \leftarrow to\_rgb(x_{adv})$;

  $adv\_scores = model(x_{adv})$;

  $loss \leftarrow loss\_fn(adv\_scores, target\_class, \kappa)$;

  if $loss \leq \kappa$ **then**

    return x_{adv};

  else

    $backprop(loss)$;

    $update(f)$;

    $i \leftarrow i + 1$;

  end

end

Figure 3.1: Examples from the dataset and adversarial examples generated with their target class probabilities. From left to right; Benign image, adversarial image generated by attacking in and RGB.

CHAPTER 4

EXPERIMENTS

4.1 Experimental Evaluation

We conducted our experiments in a white-box setup where the gradients are fully available. Experiments have been done in a targeted attack setting with the dataset provided targets. We optimized using Adam [20] with the default settings and used Carlini & Wagner loss [6] with confidence margin $\kappa \in \{0, 10\}$.

We used the dataset and the provided model from NIPS 2017 Competition on Adversarial Attacks and Defenses [28] to evaluate our method. NIPS 2017 dataset is a collection of 1000 images curated by Google Brain with the resolution of 299×299 with their corresponding true and target classes from Imagenet [29] dataset. Alongside the dataset, an Imagenet trained Inception-v3 [30] model is provided.

We compared the success rate of our attack in CIELAB and YC_bC_r against stAdv in both restricted and unrestricted settings. An attack is considered successful if the Carlini & Wagner loss is less than $-\kappa$. We did not use the smoothness regularization term in stAdv for a fair comparison.

4.1.1 Analysis of the Results

Figure 3.1 shows the original images alongside with the adversarial images generated (with $\kappa = 10$) by attacking in a^*b^*, C_bC_r and RGB spaces. As can be observed from these images, perceptual distortions are much less pronounced for chrominance-only attacks. Attacking in RGB domain, which is the default approach in the literature, results in modification of the luminance channels, leading to much more visible arti-

Table 4.1: Attack success rates with $\kappa = 0$ and $\kappa = 10$ in not restricted and subpixel restricted settings for RGB, a^*b^* and C_bC_r attacks.

	RGB	C_bC_r	a^*b^*
Not Restricted			
κ = 0	100%	95.0%	95.7%
κ = 10	100%	83.8%	87.3%
Restricted to Subpixel			
κ = 0	99.8%	86.1%	89.2%
κ = 10	99.7%	47.0%	53.2%

facts.

Table 5.1 shows the attack success rates for attacks on different colorspaces. The results show that, adversarial images generated by attacks exclusively targeting the chrominance channels can fool the network with a high probability as well. On the other hand, they are less effective when restricted to operate in a subpixel-only setting. The fooling rate of a*b* attacks are slightly higher than C_bC_r attacks. We argue that this is due to many examples in the dataset being chroma subsampled in YC_bC_r space, as an indirect effect of image compression, restricting the search space for C_bC_r attacks.

We measured the amount of distortion required to generate confident ($\kappa = 10$) adversarial examples with the following perceptual metrics: Learned Perceptual Image Patch Similarity (LPIPS) [17], Structured Similarity Index (SSIM) [15] and Multi-Scale SSIM (MS-SSIM) [16]. Table 5.2 shows the average results over the successful attacks for each perturbation mode in terms of these metrics. Since SSIM and MS-SSIM are similarity metrics, values of $1-$SSIM and $1-$MS-SSIM are provided. Hence, for all metrics, lower values are better. According to these results, colorspace restricted attacks have much better scores in terms of perceptual metrics compared to RGB attacks, implying that there is significantly less perceptual difference between benign and adversarial examples. While C_bC_r attacks generally produce better images in terms of perceptual quality metrics than a*b* attacks, the difference is rel-

Table 4.2: Average amount of distortion required to fool the target network with very high confidence ($\kappa = 10$) in not restricted and subpixel restricted settings.

	RGB	C_bC_r	a^*b^*
Not Restricted			
LPIPS	0.327	**0.019**	0.022
SSIM	0.321	**0.067**	0.070
MS-SSIM	0.164	0.017	**0.016**
Restricted to Subpixel			
LPIPS	0.222	**0.012**	0.014
SSIM	0.220	**0.050**	0.056
MS-SSIM	0.037	**0.011**	0.013

atively low.

4.1.2 Analysis of Failure Cases

Experimental results show that there are two main restrictions of the proposed method: out of gamut values in the chrominance channels emerging during optimization leading to visible artifacts and failing to generate adversarial images when the original image has limited colorfulness.

Out of Gamut Values: Modifying the chrominance channels in YC_bC_r and CIELAB spaces may lead to improper values on individual RGB channels. This is also common in widely used chroma subsampling and mitigating this issue is an open research topic [31]. In our work, we clip the reconstructed RGB to the valid range and feed the target network with the clipped image at each iteration to prevent further change in the pixel values out of the gamut. Clipping also zeroes out the gradient and prevents further updates in gradient based optimization. However, we found that it still causes visible artifacts in the adversarial image, especially around the borders between red and gray tones. Figure 5.4 shows two examples where spatial transformation in red-gray borders yield out of gamut pixels and clipping the values still causes visible

artifacts since clipping in RGB space effectively changes the values of luminance channels.

Failed Attacks on Less Colorful Images: Results in Table 5.1, show that the attack success rate does not reach 100% when spatial transform attack is restricted to chrominance channels. This implies that the chrominance based attacks fail for a number of images in the dataset. Examples of such images are provided in Figure 5.3. We observed that these particular images are either monochromatic examples or have a uniform color pattern, for which spatial transformation in a neighborhood lead to little change.

To analyze the effect of colorfulness on the attack performance, we calculated the colorfulness index histogram of the images in the dataset (Figure 5.1) . We found that 3.2% of the dataset consists of grayscale images, for which our method would not be able to make any changes to the input image, inevitably resulting in a failed attack. Figure 5.2 shows the attack success rate using the subsets where colorfulness is lower-limited by filtering out examples having colorfulness index less than the x axis value. Although a*b* attacks are slightly more successful than C_bC_r in the low colorfulness regime $(<= 0.2)$, they have the same success rate of the attacks over higher colorfulness.

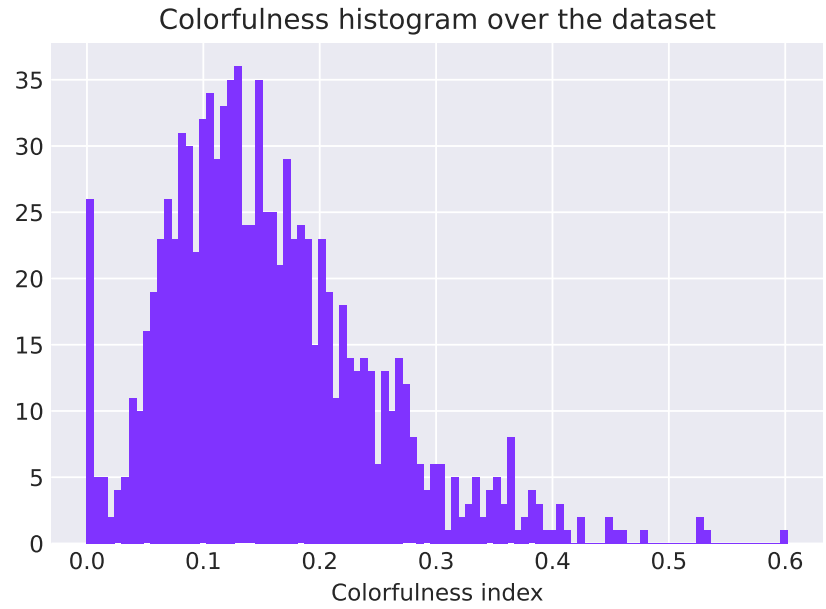


Figure 4.1: Colorfulness index histogram over NIPS2017 dataset.

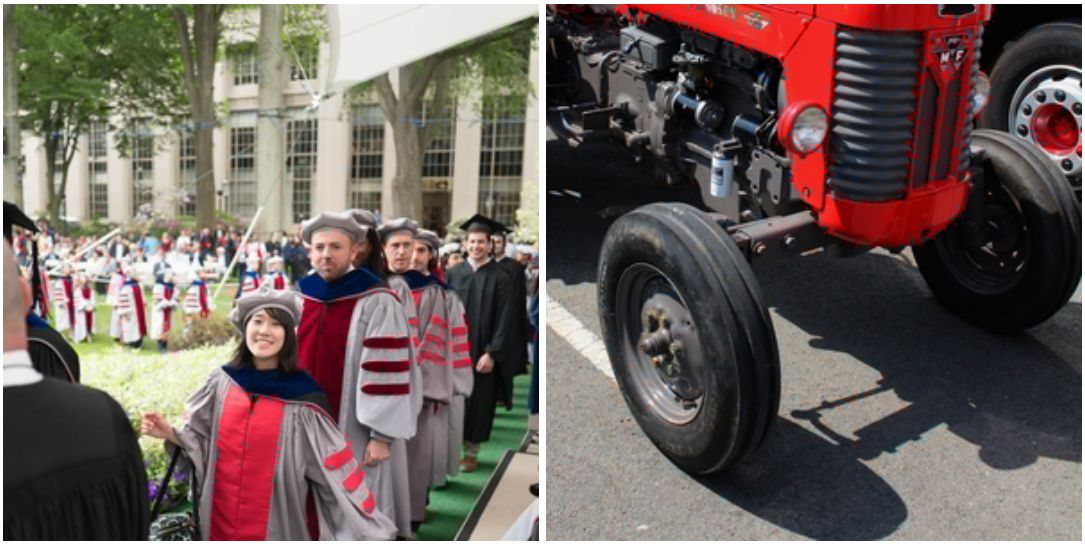


Figure 4.4: Examples of visible clipping artifacts of out-of-gamut pixels caused by spatial transform around red-gray borders. Flow magnitude has been scaled up to highlight the visible effects for illustration.

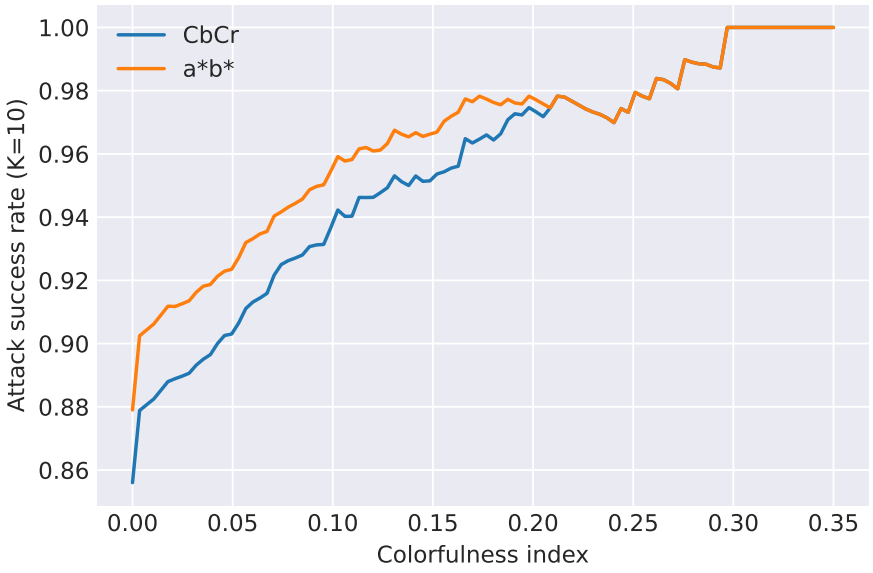


Figure 4.2: Attack success rate analysis with regards to colorfulness index with $\kappa = 10$ on $CbCr$ and a*b* channels. Images having colorfulness index less than the x axis value are excluded in calculation of the success rate. Note that both colorspaces attain very close success rates after around colorfulness index 0.2.

4.2 Discussion

As it can be seen by Figure 5.3, the input images that our method fails are generally grayscale or monochromatic images, which prevents chrominance spatial transforms from changing the pixel values due to the low magnitude of chrominance channel values. In addition, input images having a very limited local color variation negatively affect the performance by limiting the potential search space. We observed that there is a significant drop in the success rate with the setup confidence margin $\kappa = 10$ if the attack is restricted to subpixel changes in comparison to the unrestricted attacks. We argue that this performance drop is arising from the fact that the most examples are already JPEG compressed, which means chroma subsampling is applied to the benign examples, which restricts the subpixel restricted search space by dramatically reducing the local chrominance variation. This leads to the observation that chroma subsampling could be an effective defense method against our attack.

Moreover, the search space is further restricted in JPEG compressed images as the quantization step of JPEG compression attenuates high frequency information, especially in the chrominance channels. Nonetheless, we observed adversarial examples generated by spatial transforms in chrominance channels of perceptual colorspaces obtain competitive fooling rates without making perceptible changes to the image. This observation provides further evidence for the hypothesis that representation of deep neural networks does not necessarily align with human vision [32].

4.3 Method

Figure 4.3: Examples from the dataset that our method fails to generate successful adversarial examples from in both YC_bC_r and CIELAB spaces, sorted from top bottom by colorfulness amount.

CHAPTER 5

EXPERIMENTS

5.1 Experimental Evaluation

We conducted our experiments in a white-box setup where the gradients are fully available. Experiments have been done in a targeted attack setting with the dataset provided targets. We optimized using Adam [20] with the default settings and used Carlini & Wagner loss [6] with confidence margin $\kappa \in \{0, 10\}$.

We used the dataset and the provided model from NIPS 2017 Competition on Adversarial Attacks and Defenses [28] to evaluate our method. NIPS 2017 dataset is a collection of 1000 images curated by Google Brain with the resolution of 299×299 with their corresponding true and target classes from Imagenet [29] dataset. Alongside the dataset, an Imagenet trained Inception-v3 [30] model is provided.

We compared the success rate of our attack in CIELAB and YC_bC_r against stAdv in both restricted and unrestricted settings. An attack is considered successful if the Carlini & Wagner loss is less than $-\kappa$. We did not use the smoothness regularization term in stAdv for a fair comparison.

5.1.1 Analysis of the Results

Figure 3.1 shows the original images alongside with the adversarial images generated (with $\kappa = 10$) by attacking in a^*b^*, C_bC_r and RGB spaces. As can be observed from these images, perceptual distortions are much less pronounced for chrominance-only attacks. Attacking in RGB domain, which is the default approach in the literature, results in modification of the luminance channels, leading to much more visible arti-

Table 5.1: Attack success rates with $\kappa = 0$ and $\kappa = 10$ in not restricted and subpixel restricted settings for RGB, a^*b^* and C_bC_r attacks.

	RGB	C_bC_r	a^*b^*
Not Restricted			
κ = 0	100%	95.0%	95.7%
κ = 10	100%	83.8%	87.3%
Restricted to Subpixel			
κ = 0	99.8%	86.1%	89.2%
κ = 10	99.7%	47.0%	53.2%

facts.

Table 5.1 shows the attack success rates for attacks on different colorspaces. The results show that, adversarial images generated by attacks exclusively targeting the chrominance channels can fool the network with a high probability as well. On the other hand, they are less effective when restricted to operate in a subpixel-only setting. The fooling rate of a*b* attacks are slightly higher than C_bC_r attacks. We argue that this is due to many examples in the dataset being chroma subsampled in YC_bC_r space, as an indirect effect of image compression, restricting the search space for C_bC_r attacks.

We measured the amount of distortion required to generate confident ($\kappa = 10$) adversarial examples with the following perceptual metrics: Learned Perceptual Image Patch Similarity (LPIPS) [17], Structured Similarity Index (SSIM) [15] and Multi-Scale SSIM (MS-SSIM) [16]. Table 5.2 shows the average results over the successful attacks for each perturbation mode in terms of these metrics. Since SSIM and MS-SSIM are similarity metrics, values of $1-$SSIM and $1-$MS-SSIM are provided. Hence, for all metrics, lower values are better. According to these results, colorspace restricted attacks have much better scores in terms of perceptual metrics compared to RGB attacks, implying that there is significantly less perceptual difference between benign and adversarial examples. While C_bC_r attacks generally produce better images in terms of perceptual quality metrics than a*b* attacks, the difference is rel-

Table 5.2: Average amount of distortion required to fool the target network with very high confidence ($\kappa = 10$) in not restricted and subpixel restricted settings.

	RGB	C_bC_r	a^*b^*
Not Restricted			
LPIPS	0.327	**0.019**	0.022
SSIM	0.321	**0.067**	0.070
MS-SSIM	0.164	0.017	**0.016**
Restricted to Subpixel			
LPIPS	0.222	**0.012**	0.014
SSIM	0.220	**0.050**	0.056
MS-SSIM	0.037	**0.011**	0.013

atively low.

5.1.2 Analysis of Failure Cases

Experimental results show that there are two main restrictions of the proposed method: out of gamut values in the chrominance channels emerging during optimization leading to visible artifacts and failing to generate adversarial images when the original image has limited colorfulness.

Out of Gamut Values: Modifying the chrominance channels in YC_bC_r and CIELAB spaces may lead to improper values on individual RGB channels. This is also common in widely used chroma subsampling and mitigating this issue is an open research topic [31]. In our work, we clip the reconstructed RGB to the valid range and feed the target network with the clipped image at each iteration to prevent further change in the pixel values out of the gamut. Clipping also zeroes out the gradient and prevents further updates in gradient based optimization. However, we found that it still causes visible artifacts in the adversarial image, especially around the borders between red and gray tones. Figure 5.4 shows two examples where spatial transformation in red-gray borders yield out of gamut pixels and clipping the values still causes visible

artifacts since clipping in RGB space effectively changes the values of luminance channels.

Failed Attacks on Less Colorful Images: Results in Table 5.1, show that the attack success rate does not reach 100% when spatial transform attack is restricted to chrominance channels. This implies that the chrominance based attacks fail for a number of images in the dataset. Examples of such images are provided in Figure 5.3. We observed that these particular images are either monochromatic examples or have a uniform color pattern, for which spatial transformation in a neighborhood lead to little change.

To analyze the effect of colorfulness on the attack performance, we calculated the colorfulness index histogram of the images in the dataset (Figure 5.1) . We found that 3.2% of the dataset consists of grayscale images, for which our method would not be able to make any changes to the input image, inevitably resulting in a failed attack. Figure 5.2 shows the attack success rate using the subsets where colorfulness is lower-limited by filtering out examples having colorfulness index less than the x axis value. Although a*b* attacks are slightly more successful than C_bC_r in the low colorfulness regime $(<= 0.2)$, they have the same success rate of the attacks over higher colorfulness.

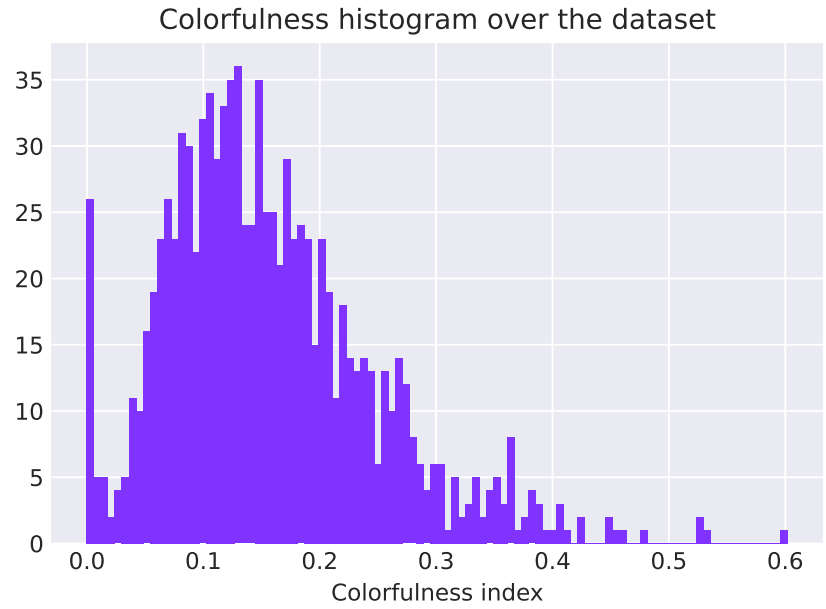


Figure 5.1: Colorfulness index histogram over NIPS2017 dataset.

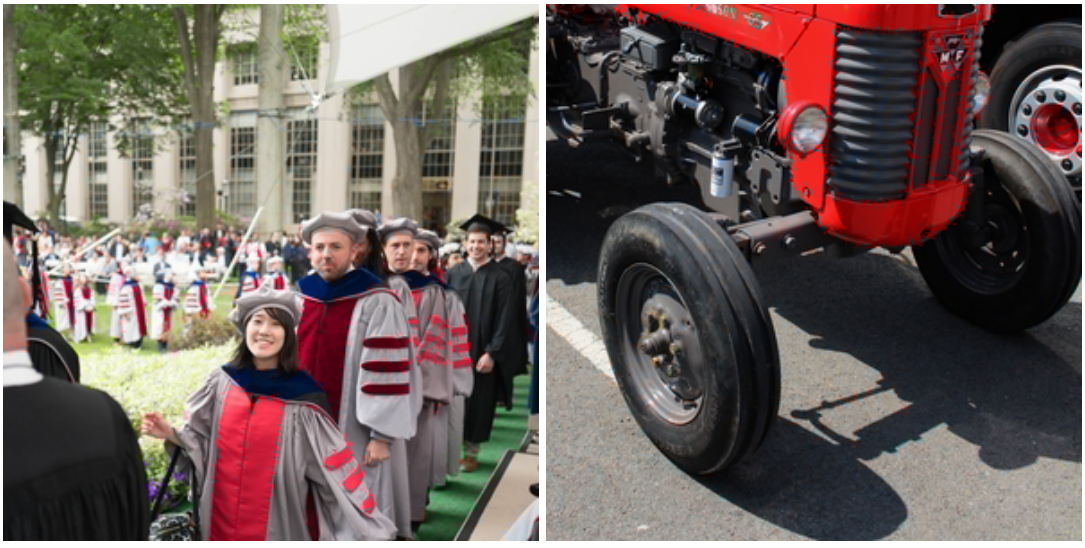


Figure 5.4: Examples of visible clipping artifacts of out-of-gamut pixels caused by spatial transform around red-gray borders. Flow magnitude has been scaled up to highlight the visible effects for illustration.

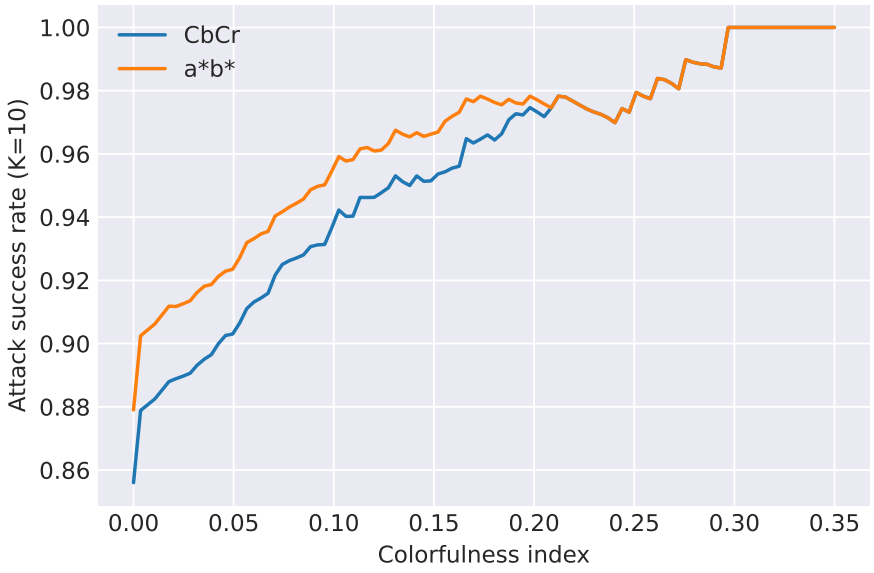


Figure 5.2: Attack success rate analysis with regards to colorfulness index with $\kappa = 10$ on $CbCr$ and a*b* channels. Images having colorfulness index less than the x axis value are excluded in calculation of the success rate. Note that both colorspaces attain very close success rates after around colorfulness index 0.2.

5.2 Discussion

As it can be seen by Figure 5.3, the input images that our method fails are generally grayscale or monochromatic images, which prevents chrominance spatial transforms from changing the pixel values due to the low magnitude of chrominance channel values. In addition, input images having a very limited local color variation negatively affect the performance by limiting the potential search space. We observed that there is a significant drop in the success rate with the setup confidence margin $\kappa = 10$ if the attack is restricted to subpixel changes in comparison to the unrestricted attacks. We argue that this performance drop is arising from the fact that the most examples are already JPEG compressed, which means chroma subsampling is applied to the benign examples, which restricts the subpixel restricted search space by dramatically reducing the local chrominance variation. This leads to the observation that chroma subsampling could be an effective defense method against our attack.

Moreover, the search space is further restricted in JPEG compressed images as the quantization step of JPEG compression attenuates high frequency information, especially in the chrominance channels. Nonetheless, we observed adversarial examples generated by spatial transforms in chrominance channels of perceptual colorspaces obtain competitive fooling rates without making perceptible changes to the image. This observation provides further evidence for the hypothesis that representation of deep neural networks does not necessarily align with human vision [32].

5.3 Method

Figure 5.3: Examples from the dataset that our method fails to generate successful adversarial examples from in both YC_bC_r and CIELAB spaces, sorted from top bottom by colorfulness amount.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

Adopting the techniques used in multimedia compression and using the idea that pixel shifts in a constrained neighborhood are hard to notice, we designed a method that applies local spatial transformations to chrominance channels of perceptual colorspaces. The proposed method results in adversarial images having imperceptible distortions without requiring any regularization. In addition to obtaining competitive fooling rates, restricting magnitude of the spatial transformations still yields successful attacks, when there is sufficient amounts of local chrominance variation in the input image.

In addition to the perceptual colorspaces investigated in this work, other perceptual colorspaces such as CIELUV, HSLuv and CIEXYZ [25] can also be utilized to create imperceptible adversarial examples. Out of gamut values at borders with red pixels may result in visible artifacts during the adversarial image generation and preventing such out-of-gamut values would result in better quality adversarial images. While our method does not require optimizing using a visual quality metric, it can be utilized along with our method to obtain a better visual quality.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *ComputerScience*, 2015.

[2] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

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

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[6] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*, pp. 39–57, Ieee, 2017.

[7] M. Jordan, N. Manoj, S. Goel, and A. G. Dimakis, "Quantifying perceptual distortion of adversarial examples," *arXiv preprint arXiv:1902.08265*, 2019.

[8] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "A rotation and a translation suffice: Fooling cnns with simple transformations," 2018.

[9] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song, "Spatially transformed adversarial examples," *arXiv preprint arXiv:1801.02612*, 2018.

[10] V. V. Estrela, H. A. Magalhães, and O. Saotome, "Total variation applications in computer vision," in *Handbook of Research on Emerging Perspectives in Intelligent Pattern Recognition, Analysis, and Image Processing*, pp. 41–64, IGI Global, 2016.

[11] F. Croce and M. Hein, "Sparse and imperceivable adversarial attacks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4724–4732, 2019.

[12] B. Aksoy and A. Temizel, "Attack type agnostic perceptual enhancement of adversarial images," in *International Workshop on Adversarial Machine Learning And Security (AMLAS), IEEE World Congress on Computational Intelligence (IEEE WCCI)*, 2019.

[13] C. Pestana, N. Akhtar, W. Liu, D. Glance, and A. Mian, "Adversarial perturbations prevail in the Y-Channel of the YCbCr color space," Feb. 2020.

[14] A. Aydin, D. Sen, B. T. Karli, O. Hanoglu, and A. Temizel, *Imperceptible Adversarial Examples by Spatial Chroma-Shift*, p. 8–14. New York, NY, USA: Association for Computing Machinery, 2021.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[16] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, pp. 1398–1402, Ieee, 2003.

[17] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

[18] M. R. Luo, G. Cui, and B. Rigg, "The development of the cie 2000 colour-difference formula: Ciede2000," *Color Research & Application*, vol. 26, no. 5, pp. 340–350, 2001.

[19] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[21] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical programming*, vol. 45, no. 1, pp. 503–528, 1989.

[22] A. Mordvintsev, N. Pezzotti, L. Schubert, and C. Olah, "Differentiable image parameterizations," *Distill*, vol. 3, no. 7, p. e12, 2018.

[23] M. Vorobyev, "Ecology and evolution of primate colour vision," *Clinical and Experimental Optometry*, vol. 87, no. 4-5, pp. 230–238, 2004.

[24] E. Hamilton, "Jpeg file interchange format," 2004.

[25] J. Schanda, *Colorimetry: understanding the CIE system*. John Wiley & Sons, 2007.

[26] M. Mahy, B. Van Mellaert, L. Van Eycken, and A. Oosterlinck, "The influence of uniform color spaces on color image processing: A comparative study of cielab, cieluv, and atd," *Journal of Imaging Technology*, vol. 17, pp. 232—-243, 1991.

[27] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, and G. Bradski, "Kornia: an open source differentiable computer vision library for pytorch," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3674–3683, 2020.

[28] A. Kurakin, I. Goodfellow, S. Bengio, Y. Dong, F. Liao, M. Liang, T. Pang, J. Zhu, X. Hu, C. Xie, *et al.*, "Adversarial attacks and defences competition," in *The NIPS'17 Competition: Building Intelligent Systems*, pp. 195–231, Springer, 2018.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[31] G. Chan, "Toward better chroma subsampling," *SMPTE motion imaging journal*, vol. 117, no. 4, pp. 39–45, 2008.

[32] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, “Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness,” *arXiv preprint arXiv:1811.12231*, 2018.