Journal of Imaging Science and Technology® 66(5): 050401-1–050401-10, 2022.
© Society for Imaging Science and Technology 2022

Exploring Effects of Colour and Image Quality in Semantic Segmentation by Deep Learning Methods

Kanjar De
Embedded Intelligent Systems Laboratory, Luleå University of Technology, 97187 Luleå, Sweden
E-mail: kanjar.de@associated.ltu.se

Abstract. *Recent advances in convolutional neural networks and vision transformers have brought about a revolution in the area of computer vision. Studies have shown that the performance of deep learning-based models is sensitive to image quality. The human visual system is trained to infer semantic information from poor quality images, but deep learning algorithms may find it challenging to perform this task. In this paper, we study the effect of image quality and color parameters on deep learning models trained for the task of semantic segmentation. One of the major challenges in benchmarking robust deep learning-based computer vision models is lack of challenging data covering different quality and colour parameters. In this paper, we have generated data using the subset of the standard benchmark semantic segmentation dataset (ADE20K) with the goal of studying the effect of different quality and colour parameters for the semantic segmentation task. To the best of our knowledge, this is one of the first attempts to benchmark semantic segmentation algorithms under different colour and quality parameters, and this study will motivate further research in this direction.* © *2022 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2022.66.5.050401]

1. INTRODUCTION

After the success of AlexNet [1] in ImageNet [2] Challenge 2012, deep convolutional neural networks have become an indispensable tool in computer vision. From the early successes in image classification, these networks have been used in different computer vision tasks such as object detection, object tracking [3, 4] and image segmentation [5–8]. Image classification is the task of assigning a class label to a particular image, but merely assigning a class label does not provide information on the understanding of the entire scene. Human visual systems are naturally trained to develop a deeper understanding simply by looking at the image. Semantic segmentation is the task of assigning a class label to a particular pixel and grouping together similar pixels. Semantic segmentation provides a deeper understanding of the context of the entire image. Several architectures based on convolutional neural networks have been developed for image classification over the last decade. An example of semantic segmentation is shown in Figure 1 (Ref. [9]). Some of the most popular ones are ResNet [10], DenseNet [11], VGGNets [12], GoogleNet [13], EfficientNets [14] among others. Recently, next-generation ConvNeXts [15] have been proposed to take on transformers, where ResNet architectures have made some design changes to mimic patchifying. In these architectures, depth-wise convolutions and inverted bottlenecks, larger kernel sizes permit global receptive power and micro-designs like fewer activations and normalization layers.

Transformers have been popular in the field of natural language processing. Recently, computer vision researchers have achieved state-of-the-art results on the task of image classification outperforming most CNN based architectures. Popular architectures include the vision transformer (VIT) [16] and the shifted window transformer (SWIN) [17]. The core idea of VIT is to split the input image into patches followed by vectorization. The vectors are then followed by dense layers with shared parameters. The next step is positional encoding, which is used to represent structural information. The vectors along with a token for classification are passed through a series of multihead self-attention layers followed by dense layers, which constitute the transformer encoder network. The SWIN transformer also transforms the image into patches that are passed through a linear transform. The SWIN transformer uses small patches in the first transformer layer and merges into larger patches in the deeper transformer layers. Then the concept of shifted window-based self-attention is applied followed by a series of transformers with limited attention and merging layers followed by a linear dimensionality reducers. One of the recent advances is the data-efficient image transformer (DEIT), which uses the concept of distillation and the attention mechanism [18].

Generally, deep learning-based models are data-dependent and need a large amount of data to develop robust models, and as a result, the quality of data is an important parameter when training these models. One of the most critical challenges in the deployment of machine learning-based systems in the real world is that during testing time, if there is a shift in distribution of the data, the system is vulnerable to failures. Recent work [19–21] has demonstrated that image quality is a very important attribute in developing a machine learning-based system involving images. Multiple studies by Hendrycks et al. [22–24] in the ImageNet database have demonstrated that perturbations and distribution shifts in images have a significant impact on the performance of deep learning-based computer vision models.

One of the lesser explored areas in deep learning, in computer vision systems is the impact of colour on the per-

Received May 15, 2022; accepted for publication July 16, 2022; published online Aug. 10, 2022. Associate Editor: Marius Pedersen.
1062-3701/2022/66(5)/050401/10/$25.00

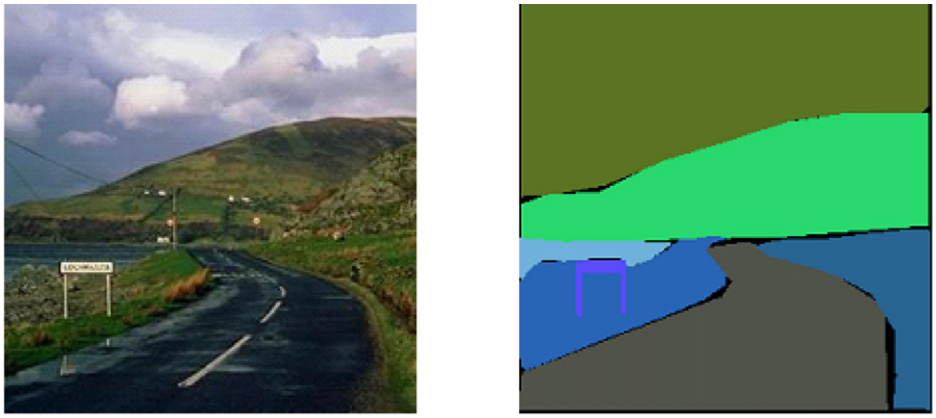

Figure 1. Example from ADE20K dataset [9] with segmentation map.

formance of these networks. Deep learning networks such as CNNs [25] and generative adversarial networks (GANs) [26, 27] have shown promising results in converting grayscale images into colour images, and also some approaches have been proposed in the area of demosaicing [28, 29]. Recent studies [30–32] have shown that colour information has a significant impact on image classification tasks. Colour parameters such as the hue angle shift [33] have shown a significant impact on the performance of state-of-the-art deep neural networks trained on pristine ImageNet data. Colour information has been exploited successfully in the past by image segmentation algorithms [34–36]. Kantipudi [37] et al. have shown that colour channels can be exploited to attack deep learning systems. Previous studies have shown that CNNs trained on ImageNet are biased towards texture [38]. Much current research is focused on the robustness [39–43] of deep learning research; therefore,it is important to investigate the effect of colour and image quality on the robustness of these deep learning methods. To the best of our knowledge, there has been very little work exploring the impact of colour information on modern deep learning-based semantic segmentation networks.

In this paper, we try to study the robustness of deep learning-based semantic segmentation models [44]. One of the first challenges is to identify and inspect quality and colour based parameters which are likely to have an impact on the performance of deep learning based semantic segmentation models and generate a dataset to bench-mark the performance. The parameters used for our study include color space information, ISONoise, gamut, hue angle shift, saturation, contrast, brightness, etc. to name a few. The proposed dataset is built using the standard ADE20K [9, 45]. We tested some of the CNN- and transformer-based methods for semantic segmentation to gain insight on how these models respond to inputs which have been perturbed from the distribution on which they are trained. With more and more real-life tasks being deployed based on deep learning trained computer vision models, understanding the robustness parameters of these models is of paramount importance. The rest of the paper is organized as follows: we briefly describe the methods and architectures used in this study, and then we describe in detail the dataset generation and investigations conducted.

2. ARCHITECTURES AND METHODS

For this work, we have studied a few state-of-the-art semantic segmentation networks and their backbones. We have included CNN- and transformer-based methods for this study. For our analysis, we have used models pre-trained in ADE20K from the MMSegementation [46] model zoo.

- Fully convolutional networks (FCN) [47] were one of the first techniques to explore the use of convolutional networks to perform semantic segmentation. They transformed the classification network by adding upsampling into a segmentation network. For this particular study, we have included an FCN with a ResNet-based backbone, namely ResNet-50 and ResNet-101 backbones.
- Pyramid Scene Parsing networks (PSPNets) [48] is the next image segmentation method included in our study. PSPNets were designed to extract context information and improve the quality of segmentation. Like FCN we tested our augmented data for PSPNets and also the backbones used are ResNet-50 and ResNet-101 backbones for a fair comparison.
- Unified perceptual parsing networks (UPerNet) [49] are networks designed based on unified perceptual parsing, which involves learning multiple possible visual concepts from a given image. We have conducted extensive experiments using UPerNets and also we have performed experiments on models which are combination of UPerNets with the latest backbones namely ConvNeXt, VIT, DEIT, and SWIN transformers in addition to the ResNet-50 and ResNet-101 backbones and found some interesting results, which give us an idea about robustness of these models.
- With current advances in the applications of transformers in computer vision tasks, Strude et al. [50] proposed a semantic segmentation method using transformers,

where the authors have extended the VIT architecture for the task of segmentation. The method uses the output embeddings corresponding to the patches of the images, and these embeddings yield the class labels using a mask transformer decoder or a pointwise linear decoder.

3. PIXEL ACCURACY

In this section, we describe the performance measure used for our analysis. Common problems in semantic segmentation include mismatched class labels, getting inconspicuous classes, mismatched relationships, among others to name a few. For semantic segmentation, each and every pixel in the image is assigned a class label. To measure performance, we calculate the ratio of the number of pixel labels identified by the semantic segmentation algorithm to the pixel annotations from the ground truth. Let *GT* and *Pred* be the ground truth and the predicted segmented maps, respectively. Let T_N be the total number of pixels in the ground truth and T_{matched} be the total number of pixels where the ground truth and the predicted segmentation map labels are in a pair. Pixel accuracy (PA) is the ratio between T_{matched} and T_N. For this analysis, we report the mean pixel accuracy for 2000 images considered the dataset and the values are reported as a percentage.

4. DATASET GENERATION AND ANALYSIS

One of the key challenges, to our knowledge, is that there are no data to support the study of how image quality and colour affect the process of semantic segmentation in images. ADE20K is one of the most challenging benchmarks in semantic segmentation, so we decided to build our image quality and colour dataset on the classes of the ADE20K dataset by using 2000 of their labeled validation datasets, and we only just modified those colour and quality parameters so that the boundaries of segmentation are not altered. We have used the pixel-level accuracy between the predicted and ground-truth maps to evaluate different semantic segmentation techniques. Table I shows the average pixel classification accuracy for all 2000 pristine images in the dataset. The main observation is that transformer-based models have better accuracy than CNN models, but the UPerNet-ConvNeXt combination has competitive performance. For all the methods, as expected, ResNet-101 performs better than the ResNet-50 backbones.

Table I. Average Pixel Classification Accuracy (PA) for the pristine images in the dataset.

Method	Backbone	Pristine PA
FCN	ResNet-50	78.2
	ResNet-101	80.1
PSP	ResNet-50	81.0
	ResNet-101	81.8
UPerNet	ResNet-50	80.8
	ResNet-101	81.5
	Next	83.4
	VIT-B	83.1
	DEIT	82.9
	SWIN-B	83.0
Segmenter	VIT-B	83.8

Table II. Average Pixel Classification Accuracy (PA) for the colour space-modified images in the dataset.

Method	Backbone	Hue PA	Sat PA	Cb PA	Cr PA	A PA	B PA
FCN	ResNet-50	73.5	70.1	65.9	62.1	77.0	75.1
	ResNet-101	77.5	73.6	72.6	65.2	79.1	78.1
PSP	ResNet-50	78.9	75.8	74.0	66.1	80.3	79.4
	ResNet-101	79.5	76.7	76.2	68.0	80.1	80.4
UPerNet	ResNet-50	78.3	75.4	73.0	50.8	79.9	78.7
	ResNet-101	79.3	76.9	75.7	65.5	80.9	80.1
	Next	81.5	80.7	79.4	77.3	82.3	81.8
	VIT-B	80.9	80.7	79.4	74.9	81.8	81.3
	DEIT	80.7	79.2	79.0	77.3	81.4	80.7
	Swin-B	80.5	79.1	78.1	76.1	81.5	81.2
Segmenter	VIT-B	82.1	82.0	82.0	79.8	82.7	82.7

4.1 *Colour Space based Distortions*

Colour information from an image can be modeled using different colour spaces. During the pre-deep learning era, colour space information was used for segmentation [51]. Some of the popular colour spaces are red-green-blue (RGB), hue saturation value (HSV), luminescence chrominance (YCbCr), and CIE-Lab colour space. Generally, deep neural networks are trained on the RGB colour space. To study the impact of colour information encoded in color spaces, we modified (setting to zero) the hue, saturation, Cb, Cr and A, B components of the HSV, YCbCr, and CIE-Lab spaces, respectively, and created six subsets of images as shown in Figure 2. The images were generated by converting the RGB image into HSV [52], YCbCr [53] and CIE-Lab spaces and then the hue (H), saturation (S), Cb, Cr, a and b channels were set to 0 to generate the images. All the colour space based distortions were implemented using Matlab 2021a.

One of the key observations from Table II is that in the YCbCr space, the Cb and Cr components have a significant effect on the performance of deep learning-based models. The methods with CNN backbones perform significantly worse than those with transformer backbones. The perturbation of the Cb and Cr components in the image has created a maximum reduction in performance in comparison to any other component in other colour spaces.

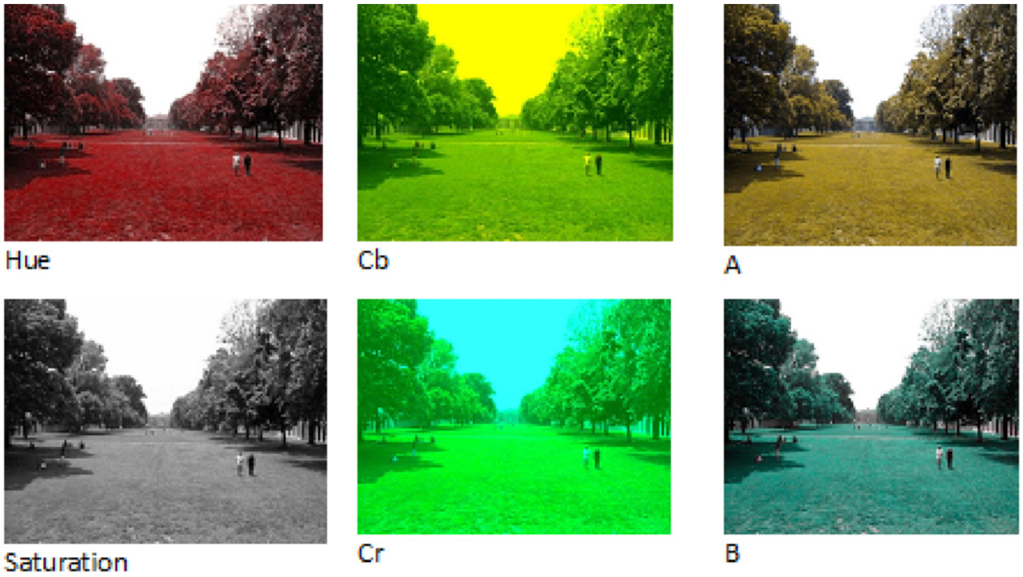

Figure 2. Example from the generated dataset the component not seen is mentioned in the labels.

4.2 *Hue Angle Shift*

For image classification, studies have shown [33] that changing the hue angle to red or blue has a significant impact on the performance of deep convolutional networks. Similarly to the previous study, here we shift the hue angle by 60 degrees to create a subset of five classes of images (60,120,180,240,300) as depicted in Figure 3. The aim is to study the effect of hue shift on semantic segmentation models trained on pristine images with a few augmentations. The main observation from Table III is that when we change the hue angle, there is a change in the distribution and there is a drop in performance. The networks with transformer backbones perform better than the earlier methods with ResNet-50 or ResNet-101 backbones. In addition, the performance of UPerNet with the next-generation ConvNeXt backbone is also very competitive. The networks perform the worst when the hue-angle shift is around 180 degrees. Fully convolutional networks (FCN) are seen to be more sensitive to hue-angle shifts compared to the competitors PSPNets and UPerNets. The UPerNets that are combined with transformers are found to be more robust compared to the ones with CNN backbones.

Table III. Average Pixel Classification Accuracy (PA) for different hue-angle shifts (in degrees) for images in the dataset.

		60	120	180	240	300
Method	Backbone	PA	PA	PA	PA	PA
FCN	ResNet-50	76.8	72.9	69.1	70.7	76.4
	ResNet-101	79.1	76.2	74.3	75.8	79.1
PSP	ResNet-50	80.0	76.1	76.1	77.4	80.2
	ResNet-101	80.9	78.5	76.8	78.1	81.1
UPerNet	ResNet-50	79.5	76.7	74.8	76.6	79.7
	ResNet-101	80.6	78.1	76.1	77.8	80.7
	Next	82.7	80.9	79.7	80.9	82.8
	VIT-B	82.6	80.3	79.3	80.1	82.5
	DEIT	82.4	80.0	79.6	79.9	82.5
	Swin-B	82.3	80.0	78.9	79.7	82.4
Segmenter	VIT-B	83.4	81.8	81.1	81.6	83.4

4.3 *Saturation*

We have varied the saturation levels in the images into four levels in order of distortion (refer to Figure 4 reduction in this case) to study the impact of saturation on the semantic segmentation task performed by deep learning models. The four levels were created by scaling the values of the S channel in the HSV colour space. For the experiments in this paper, the scaling factors used were 0.8, 0.6, 0.4, 0.2, respectively. which are termed Levels 1 to 4, respectively. The modification was performed using MATLAB 2021a where the saturation channel was scaled using these scaling factors. The observation from Table IV is that, as expected, the transformer models tend to perform better than the CNN based models. There is a drop in performance when the saturation is reduced in the images, but the transformer-based models show better resistance to these changes.

4.4 *Brightness and Contrast*

Brightness and contrast are two fundamental parameters of image quality. The goal is to study how the semantic segmentation of images works when the input image has poor contrast or it is too dark or too bright. In real-life applications, sometimes brightness and contrast may get affected due to uncontrollable situations, and thus a deep learning-based model must be robust to brightness and contrast variations. For this study, we used four levels of

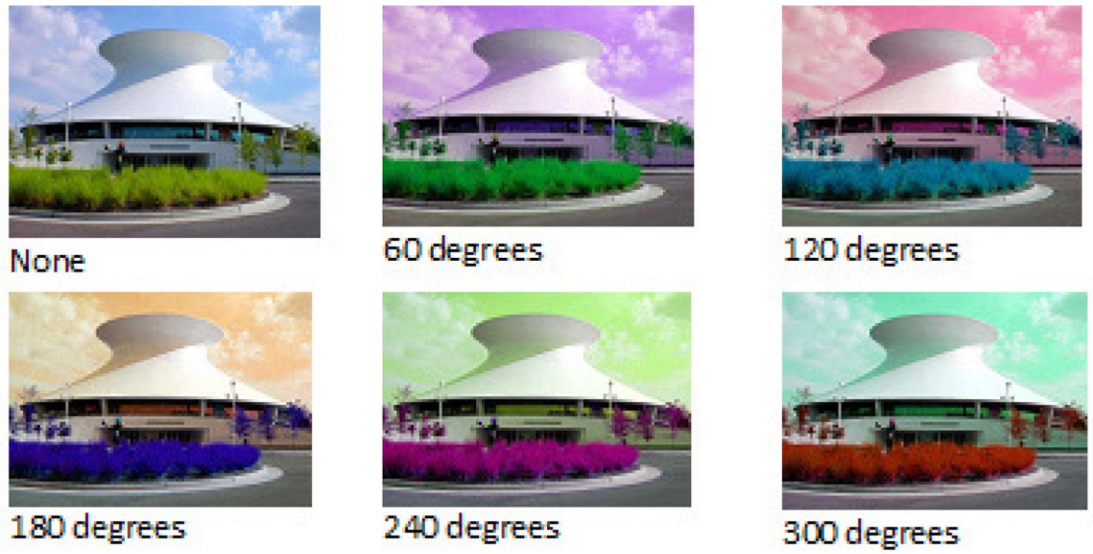

Figure 3. Example from the generated dataset with the hue shift mentioned in the labels.

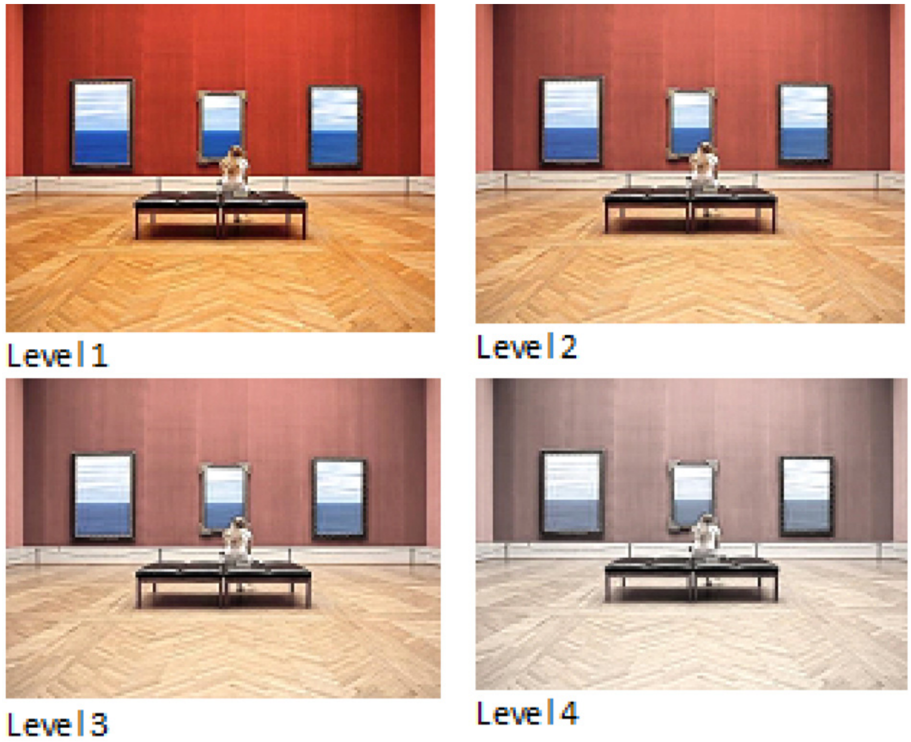

Figure 4. Example from the generated dataset with different saturation levels mentioned in the labels.

contrast and brightness and implemented them using the Augly library [54] which is used for adversarial robustness. Examples of brightness modification are shown in Figure 5. Four levels were chosen for the final analysis that have poor visual contrast for the final reporting of the results to demonstrate their effect on semantic segmentation. Images with poor visual quality showed poorer performance for semantic segmentation. Table V shows that darker images(B1) degrade the performance of methods with ResNet backbones compared to methods based on transformers. A similar behavior is observed for too bright images (B4).

Contrast is an important parameter of image quality, and it is very important to study its effects. We have generated four levels of contrast (examples shown in Figure 6) and their corresponding average pixel accuracy for each of the models is reported in Table VI. For poorer images (C1), CNN-based methods show a performance drop compared to pristine images in Table I and the transformer based models appear to be more robust.

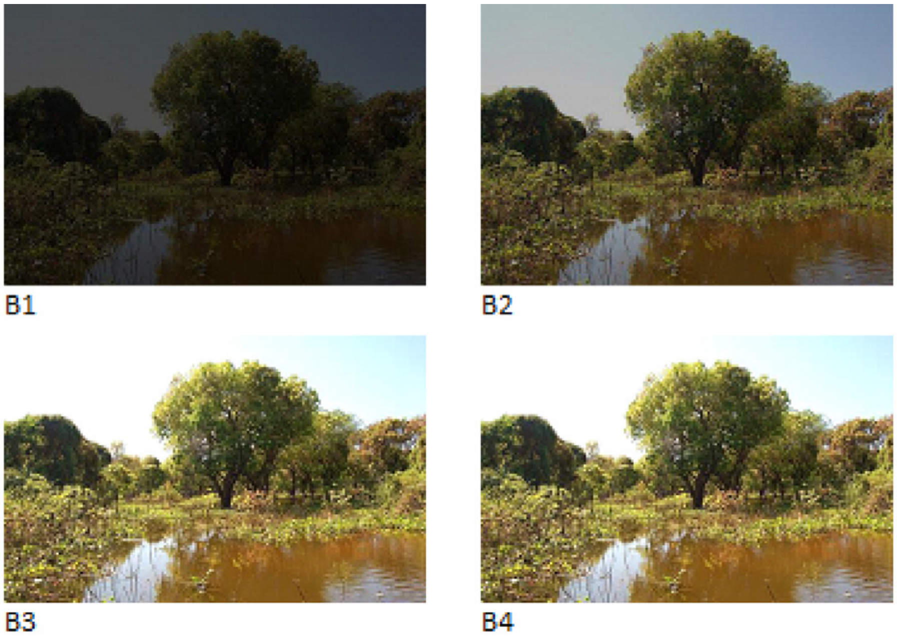

Figure 5. Example from the generated dataset with different brightness levels.

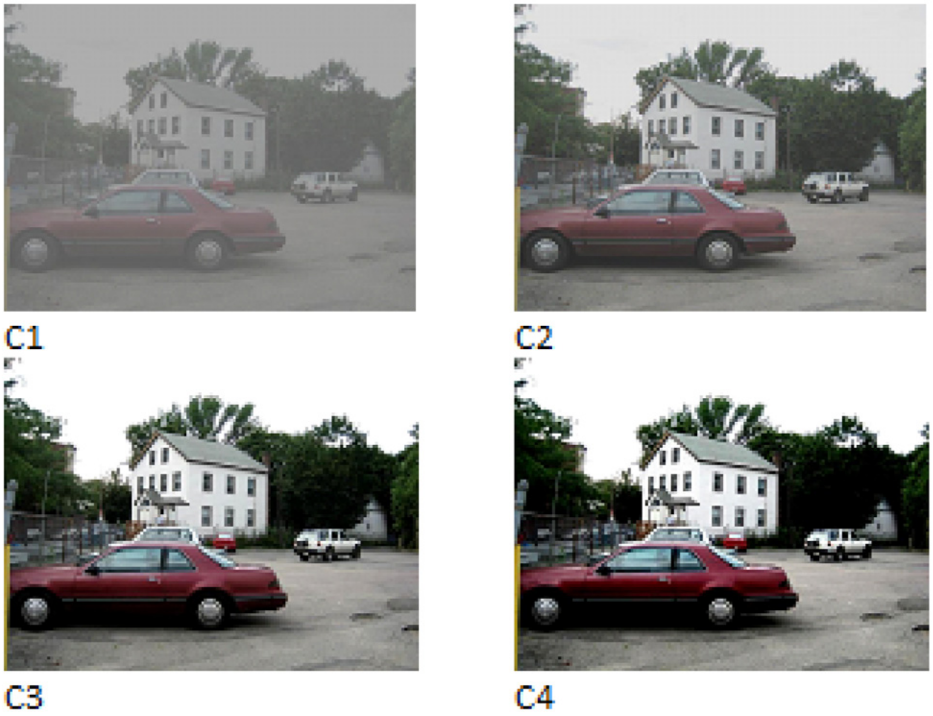

Figure 6. Example from the generated dataset with the different contrast mentioned in the labels.

4.5 *Colour Gamut*

Colour gamut [55] comprises the total subset of colors that the display device can represent and is one of the key considerations in imaging and display technologies. We conducted a set of experiments on different synthetically generated images using ICC profiles (www.color.org) which are generally used for printing, and ran semantic segmentation models on these images. An example of a newsprint gamut is shown in Figure 7. Table VII shows that the newspaper gamut has a reducing effect on the performance of semantic segmentation networks. This modification shows that there is a distribution shift between the training and testing data. The ResNet-50 FCN network has the greatest performance drop, and even transformer-based models have shown a performance drop. Similar trends were observed in experiments carried out on other print gamuts.

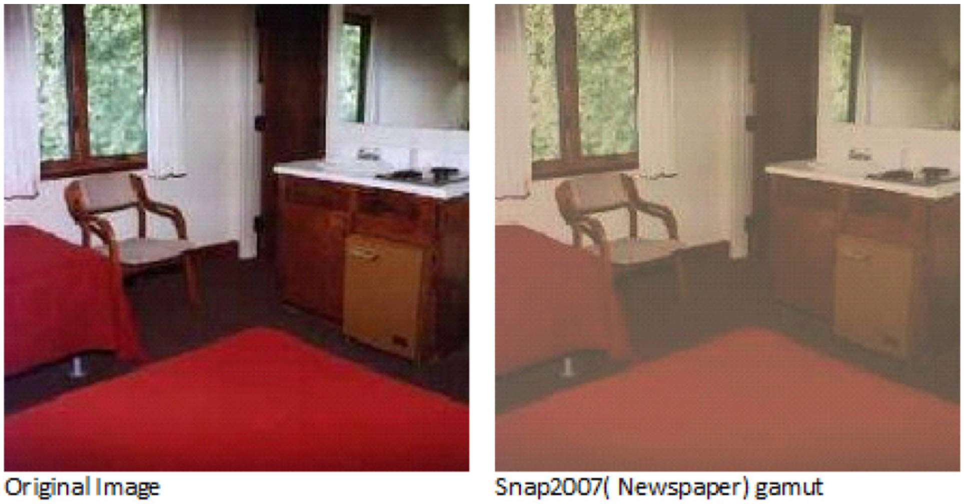

Figure 7. Example from the generated dataset with the different gamut mentioned in the labels.

Table IV. Average Pixel Classification Accuracy (PA) for different saturation levels for images in the dataset.

		Level 1	Level 2	Level 3	Level 4
Method	Backbone	PA	PA	PA	PA
FCN	ResNet-50	78.1	77.9	77.2	75.0
	ResNet-101	80.1	80.0	79.4	77.8
PSP	ResNet-50	81.0	80.8	80.4	79.1
	ResNet-101	81.8	81.7	81.2	79.9
UPerNet	ResNet-50	80.7	80.5	79.9	78.5
	ResNet-101	81.5	81.4	81.0	79.8
	Next	83.4	83.2	83.1	82.6
	VIT-B	83.1	83.0	82.8	82.2
	DEIT	82.9	82.8	82.5	82.0
	Swin-B	83.0	82.8	82.7	82.0
Segmenter	VIT-B	83.7	83.7	83.6	83.3

Table V. Average Pixel Classification Accuracy (PA) for different brightness levels for images in the dataset.

		B1	B2	B3	B4
Method	Backbone	PA	PA	PA	PA
FCN	ResNet-50	76.2	77.9	77.8	76.2
	ResNet-101	78.8	79.9	79.8	78.6
PSP	ResNet-50	79.7	80.9	80.7	79.5
	ResNet-101	80.6	81.6	81.5	80.5
UPerNet	ResNet-50	79.2	80.5	80.4	79.2
	ReNets-101	80.2	81.4	81.2	80.1
	Next	82.2	83.2	83.1	82.0
	VIT-B	82.1	82.9	82.8	81.8
	DEIT	81.8	82.7	82.7	81.8
	SWIN-B	82.0	82.7	82.7	81.7
Segmenter	VIT-B	83.2	83.7	83.5	82.9

4.6 *ISO Noise*

To study the sensitivity of semantic segmentation to image sensor noise, we have created two subsets with two levels of noise where level 2 indicates the presence of more noise than level 1 as shown in Figure 8. The ISO noise model is based on the Poisson distribution implemented in the Albumentations [56] library (https://albumentations.ai/docs/api_reference/augmentations/transforms/.) From Table VIII we can infer that, as expected, there is a difference in the performance of semantic segmentation when the level of noise increases. Transformer-based methods are more immune to noise compared to competitors with ResNet-50 and ResNet-101 backbones. The accuracy of the ConvNeXt architecture backbone-based UPerNet is on par with transformer-based architectures. The level 2 images are visibly more distorted than level 1 images and the experimental results also show that the models perform worse on noisier images.

4.7 *Colour Modification*

We included two colour modifications that converted the colour image to grayscale and sepia (refer Figure 9) using the Albumentations. The main aim was to check how these semantic segmentation methods behave in the absence of colour information.

Table IX shows that colour information plays a significant information in computer vision tasks. There is a significant decrease in accuracy for CNN-based networks, and transformer-based methods have also shown a performance dip. The patch-based nature of transformers most likely have enabled them to perform better in comparison with CNN based models.

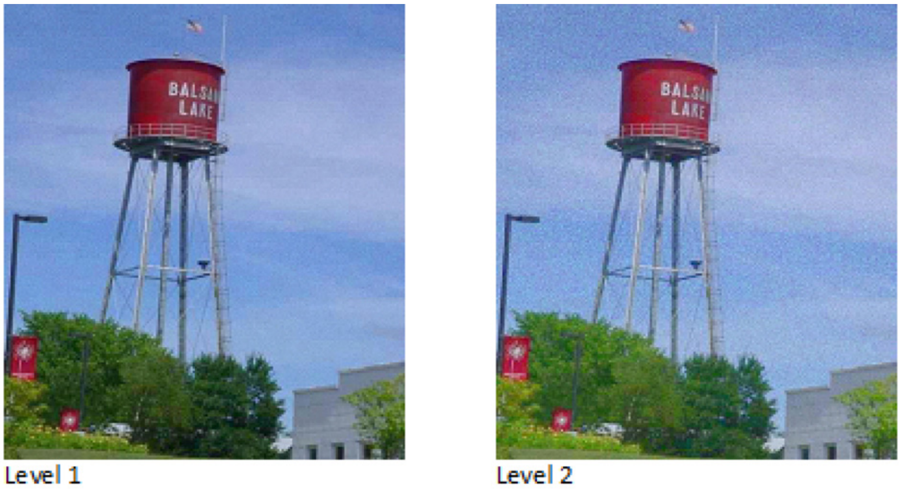

Figure 8. Example from the generated dataset with two levels of ISO Noise (Level 2 more noisier than Level 1).

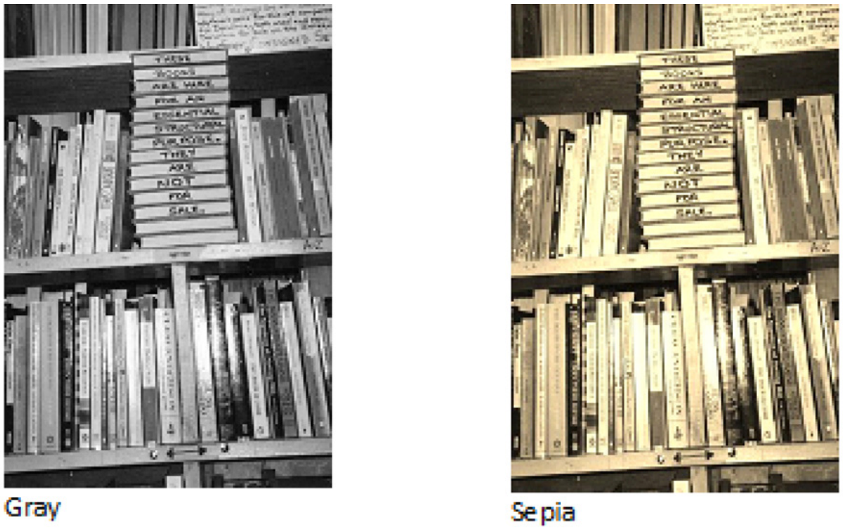

Figure 9. Example from the generated dataset with grayscale and sepia mentioned in the labels.

5. RESULTS AND DISCUSSION

In this paper, we have created synthetic data based on the popular ADE20K dataset to understand the impact of colour information and quality parameters on the performance of semantic segmentation networks. We have performed comparative studies ranging from basic FCNs to the most recent state-of-the-art transformer-based methods and attempted to get an estimate of robustness of these methods under different colour and image quality-based modifications. The general observation is that transformer-based methods show more robustness to poor-quality data compared to their CNN-based counterparts. However, transformer-based methods can be difficult to train and may require significant computing resources. There has to be a trade-off between robustness and model complexity. One consistent observation is that pyramid scene parsing networks (PSPNets) and unified perceptual parsing networks (UPerNets) with ResNet-50 and ResNet-101 backbone perform significantly better than FCN on these backbones, with ResNet-101 outperforming ResNet-50. Colour channel information is not widely studied in the context of deep neural networks and the current era is based on deep learning, and one of the main objectives of this study is to find ways in which these modern techniques match to quality and colour parameters and motivate researchers to incorporate colour channels into the architecture engineering process. Real-time computer vision systems have been deployed in critical areas such as surgery, autonomous driving, etc. to name a few, errors in scene understanding can lead to catastrophe. To the best of our knowledge, we have listed some of the key quality and colour parameters which need to be investigated in detail by both the colour imaging and deep learning community to build safer systems.

Table VI. Average Pixel Classification Accuracy (PA) for different contrast levels for images in the dataset.

		C1	C2	C3	C4
Method	Backbone	PA	PA	PA	PA
FCN	ResNet-50	73.6	77.8	77.8	76.9
	ResNet-101	76.8	79.8	79.8	79.0
PSP	ResNet-50	77.9	80.8	80.8	80.0
	ResNet-101	79.1	81.5	81.5	80.8
UPerNet	ResNet-50	77.1	80.4	80.5	79.7
	ResNet-101	78.5	81.3	81.3	80.5
	Next	81.8	83.1	83.2	82.6
	VIT-B	81.7	82.9	82.9	82.3
	DEIT	81.7	82.7	82.8	82.3
	SWIN-B	81.3	82.6	82.9	82.3
Segmenter	VIT-B	83.0	83.6	83.7	83.2

Table VII. Average Pixel Classification Accuracy (PA) for images with reduced (newspaper) gamut.

		Gamut
Method	Backbone	PA
FCN	ResNet-50	72.6
	ResNet-101	76.2
PSP	ResNet-50	77.6
	ResNet-101	78.4
UPerNet	ResNet-50	76.9
	ResNet-101	78.3
	Next	81.3
	VIT-B	81.6
	DEIT	81.3
	SWIN-B	80.2
Segmenter	VIT-B	82.8

Table VIII. Average Pixel Classification Accuracy (PA) with Two Levels of ISO Noise.

		Level 1	Level 2
Method	Backbone	PA	PA
FCN	ResNet-50	77.6	75.3
	ResNet-101	79.6	77.7
PSP	ResNet-50	80.6	78.8
	ResNet-101	81.3	79.4
UPerNet	ResNet-50	80.3	78.7
	ResNet-101	81.1	79.5
	Next	83.1	82.0
	VIT-B	82.8	81.9
	DEIT	82.6	81.5
	SWIN-B	82.6	81.0
Segmenter	VIT-B	83.6	83.1

Table IX. Average Pixel Classification Accuracy (PA) for a subset of images with sepia and grayscale filters.

		Sepia	Gray
Method	Backbone	PA	PA
FCN	ResNet-50	70.9	70.7
	ResNet-101	75.7	74.3
PSP	ResNet-50	77.4	76.6
	ResNet-101	78.1	77.4
UPerNet	ResNet-50	76.6	76.1
	ResNet-101	78.0	77.4
	Next	80.2	81.0
	VIT-B	79.8	80.9
	DEIT	78.9	79.5
	SWIN-B	78.8	79.4
Segmenter	VIT-B	81.6	82.3

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” Adv. Neural Inf. Process. Syst. **25**, 1097–1105 (2012).

[2] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” *2009 IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2009), pp. 248–255.

[3] R. Girshick, “Fast r-cnn,” *Proc. IEEE Int’l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2015), pp. 1440–1448.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2016), pp. 779–788.

[5] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *Int’l. Conf. on Medical Image Computing and Computer-assisted Intervention* (Springer, Cham, 2015), pp. 234–241.

[6] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, “The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation,” *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops* (IEEE, Piscataway, NJ, 2017), pp. 11–19.

[7] S. Rajpal, D. Sadhya, K. De, P. P. Roy, and B. Raman, “Eai-net: Effective and accurate iris segmentation network,” *Int’l. Conf. on Pattern Recognition and Machine Intelligence* (Springer, Cham, 2019), pp. 442–451.

[8] M. Osadebey, H. K. Andersen, D. Waaler, K. Fossaa, A. Martinsen, and M. Pedersen, “Three-stage segmentation of lung region from ct images using deep neural networks,” BMC Med. Imaging **21**, 1–19 (2021).

[9] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, “Semantic understanding of scenes through the ade20k dataset,” Int. J. Comput. Vis. **127**, 302–321 (2019).

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2016), pp. 770–778.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2017), pp. 4700–4708.

[12] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv Preprint arXiv:1409.1556, (2014).

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,”

Proc. IEEE Conf. on Computer Vision and Pattern Recognition (IEEE, Piscataway, NJ, 2015), pp. 1–9.

14 M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *Int'l. Conf. on Machine Learning* (PMLR, Long Beach, California, 2019), pp. 6105–6114.

15 Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, A convnet for the 2020s. arXiv Preprint arXiv:arXiv:2201.03545, (2022).

16 A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and SJ. Uszkoreit, An image is worth 16×16 words: Transformers for image recognition at scale. arXiv Preprint arXiv:arXiv:2010.11929, (2020).

17 Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *Proc. IEEE/CVF Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2021), pp. 10012–10022.

18 H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *Int'l. Conf. on Machine Learning* (PMLR, Cambridge, MA, 2021), pp. 10347–10357.

19 S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," *2016 Eighth Int'l. Conf. on Quality of Multimedia Experience (QoMEX)* (IEEE, Piscataway, NJ, 2016), pp. 1–6.

20 S. Dodge and L. Karam, "A study and comparison of human and deep learning recognition performance under visual distortions," *2017 26th Int'l. Conf. on Computer Communication and Networks (ICCCN)* (IEEE, Piscataway, NJ, 2017), pp. 1–7.

21 P. Roy, S. Ghosh, S. Bhattacharya, and U. Pal, Effects of degradations on deep neural network architectures. arXiv Preprint arXiv:1807.10108 (2018).

22 D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *Proc. of the Int'l. Conf. on Learning Representations* (2019), arXiv Preprint arXiv:1903.12261.

23 D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples.," *Proc. IEEE/CVF Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2021), pp. 15262–15271.

24 D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, and D. Song, "The many faces of robustness: A critical analysis of out-of-distribution generalization.," *Proc. IEEE/CVF Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2021), pp. 8340–8349.

25 R. Zhang, P. Isola, and A. Efros, "Colorful image colorization," *European Conf. on Computer Vision* (Springer, Cham, 2016), pp. 649–666.

26 K. Nazeri, E. Ng, and M. Ebrahimi, "Image colorization using generative adversarial networks," *Int'l. Conf. on Articulated Motion and Deformable Objects* (Springer, Cham, 2018), pp. 85–94.

27 S. Halder, K. De, and P. Roy, "Perceptual conditional generative adversarial networks for end-to-end image colourization," *Asian Conf. on Computer Vision* (Springer, Cham, 2018), pp. 269–283.

28 I. Shopovska, L. Jovanov, and W. Philips, "Rgb-nir demosaicing using deep residual u-net," *2018 26th Telecommunications Forum (TELFOR)* (IEEE, Piscataway, NJ, 2018), pp. 1–4.

29 J. Luo and J. Wang, "Image demosaicing based on generative adversarial network," Math. Proble. Eng. **2020** (2020).

30 V. Buhrmester, D. Münch, D. Bulatov, and M. Arens, "Evaluating the impact of color information in deep neural networks," *Iberian Conf. on Pattern Recognition and Image Analysis* (Springer, Cham, 2019), pp. 302–316.

31 K. De and M. Pedersen, "Impact of colour on robustness of deep neural networks," *Proc. IEEE/CVF Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2021), pp. 21–30.

32 A. Flachot and K. Gegenfurtner, "Color for object recognition: hue and chroma sensitivity in the deep features of convolutional neural networks," Vis. Res. **182**, 89–100 (2021).

33 K. De and M. Pedersen, "Effect of hue shift towards robustness of convolutional neural networks," *Proc. IS&T Electronic Imaging: Color Imaging XXVII: Displaying, Processing, Hardcopy, and Applications* (IS&T, Springfield, VA, 2022), pp. 156-1–156-6.

34 Y. Deng, B. Manjunath, and H. Shin, "Color image segmentation," *Proc. 1999 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 1999), 2, pp. 446–451.

35 T. Q. Chen and Y. Lu, "Color image segmentation–an innovative approach," Pattern Recognit. **35**, 395–405 (2002).

36 D. Khattab, H. M. Ebied, A. S. Hussein, and M. F. Tolba, "Color image segmentation based on different color space models using automatic grabcut," Sci. World J. **2014** (2014).

37 J. Kantipudi, S. R. Dubey, and S. Chakraborty, "Color channel perturbation attacks for fooling convolutional neural networks and a defense against such attacks," IEEE Trans. Artif. Intell. **1**, 181–191 (2020).

38 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," *Proc. Int'l. Conf. on Learning Representations* (ICLR, New Orleans, LA, 2019).

39 N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," *2016 IEEE European Symposium on Security and Privacy (EuroS&P)* (IEEE, Piscataway, NJ, 2016), pp. 372–387.

40 N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," *2017 IEEE Symposium on Security and Privacy (sp)* (IEEE, Piscataway, NJ, 2017), pp. 39–57.

41 R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, "Measuring robustness to natural distribution shifts in image classification.," Adv. Neural Inf. Process. Syst. **33**, 18583–18599 (2020).

42 J. Rauber, R. Zimmermann, M. Bethge, and W. Brendel, "Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and jax," J. Open Source Softw. **5**, 2607 (2020).

43 C. Kamann and C. Rother, "Benchmarking the robustness of semantic segmentation models," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2020), pp. 8828–8838.

44 S. Minaee, Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence* (IEEE, Piscataway, NJ, 2021).

45 B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2017), pp. 633–641.

46 MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

47 J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2015), pp. 3431–3440.

48 H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2017), pp. 2881–2890.

49 T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," *Proc. European Conf. on Computer Vision (ECCV)* (Springer, Cham, 2018), pp. 418–434.

50 R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," *Proc. IEEE/CVF Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2021), pp. 7262–7272.

51 S. Sural, G. Qian, and S. Pramanik, "Segmentation and histogram generation using the hsv color space for image retrieval," *Proc. Int'l. Conf. on Image Processing* (IEEE, Piscataway, NJ, 2002), Vol. 2, pp. II–II.

52 A. R. Smith, "Color gamut transform pairs," ACM Siggraph Comput. Graph. **12**, 12–19 (1978).

53 C. A. Poynton, *A Technical Introduction to Digital Video* (John Wiley & Sons, Inc., 1996), p. 175.

54 Z. Papakipos and J. Bitton, "AugLy: Data augmentations for adversarial robustness," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2022), pp. 156–163.

55 I. Farup, C. Gatta, and A. Rizzi, "A multiscale framework for spatial gamut mapping," IEEE Trans. Image Process. **16**, 2423–2435 (2007).

56 A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. Kalinin, "Albumentations: fast and flexible image augmentations," Information **11**, 125 (2020).