# DeepLabV3+ and SegFormer robustness analysis

[1]Mohamed El Amine Boudjoghra,[2] Anees Ur Rehman Hashmi,[3] Muhammad Huzaifa

{[1]mohamed.boudjoghra, [2]anees.hashmi, [3]muhammad.huzaifa}@mbzuai.ac.ae

## Abstract

*In this project, we conducted a comparative analysis of Segformer, a ViT-based semantic segmentation model, and DeeplabV3+, a CNN-based semantic segmentation model, using various image perturbations (e.g., image patch shuffling and removal) and noise addition (e.g., salt and pepper noise). Segformer has been proven to be very efficient due to its hierarchical encoder, and the spatial pyramid pooling in DeeplabV3+ allows it to compete with SOTA semantic segmentation models. Previous studies have compared these models' performances on the segmentation of natural images without any noise; however, the performance of these models on perturbed and noisy images was unknown. Both models were trained on the ADE20k dataset and tested on segmentation of the person class, which alleviates the effect of Segformer pre-training on ImageNet. Results suggest that Segformer can tolerate more perturbation than DeeplabV3+ and can perform well on noisy images; however, both models' performance drops significantly as we increase the noise.*

## 1. Introduction

Segmentation is a fundamental task in computer vision where pixel-level class is predicted instead of image-level class prediction as in classification. The semantic segmentation refers to the classification of pixels in an image into semantic classes, and it does not produce separate masks for each instance of a class. A typical semantic segmentation model first extracts the features thereby shrinking the size of the input image and then up-samples it to reproduce the segmentation mask. The feature extraction part is called as encoder, while the decoder up-samples the output of the encoder.

After the introduction of fully convolutional networks (FCNs) [7] for image segmentation, there have been a number of successful improvements to the architecture of the segmentation models. Many state-of-the-art image classification architectures have been shown to improve the performance of segmentation as well, but this improvement has been mainly made to the encoder (backbone) part of the model and considerably fewer changes have been observed in the decoder [11]. After the great success of ViTs in image classification [4], the first vision transformer-based semantic segmentation model (SETR) was proposed by [13]. SETR contains a ViT encoder and decoders consisting of several CNN. After this, other transformer-based architectures including Pyramid Vision transformer (PVT) [10] and Twins [3] were proposed, improving the backbone of the models.

One of the major drawbacks that CNN-based semantic segmentation techniques have, is the inability to analyze global information in images because of the local nature of convolutional filters [9]. To solve this problem of locality, spatial pyramid pooling and aggregated features with dilated convolutions were used in deeplabV3 and deeplabV3+ [1, 2] to extract multiscale features. A novel Vision transformer-based approach, inspired from NLP transformers, uses attention between patches extracted from images to capture the global context of an image. Nevertheless, Vision transformers, abbreviated ViTs, are data hungrier than CNNs, thus one can outperform the other depending on the size of the training dataset. Furthermore, ViTs are commonly known to be associated with a quadratic computational cost, because of the between-patches attention computation.

We will compare the mean Intersection over Union of the Segformer [11], which was trained on ImageNet and finetuned on ADE20K, and Deeplabv3+ [1] which was trained on ADE20k. The chosen person dataset, contains the person class only to assure that the pretraining of the SegFormer on ImageNet, which doesn't include a person class, will not contribute to the segmentation results on the person dataset and will serve as an initialization to the training on the ADE20K only.

### 1.1. DeeplabV3+

The semantic segmentation process in the DeeplabV3+ relies mainly on the encoder where a spatial pyramid constructed with dilated separable convolution, is used to capture more spatial information. The encoded information of

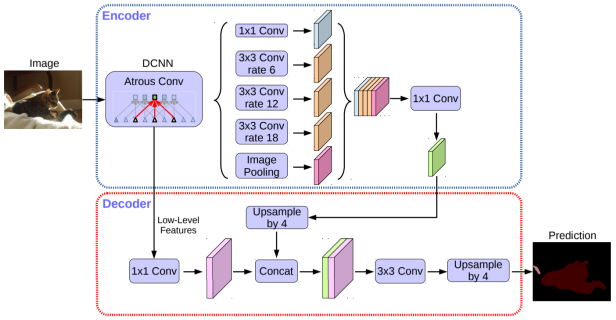


Figure 1: DeeplabV3+ architecture

the input image is obtain using a 1 × 1 convolution on the outputs of the spatial pyramid. Since the latter guarantees a very powerful encoder, the decoder can be simplified and still can give very good results. In the decoder part, the encoded feature tensor is up-sampled by a factor of 4, and then concatenated with the feature tensor with the same dimension from the second layer of the encoder, then convolved with a 3 × 3 kernel and up-sampled by 4 to give the predicted mask. The visual demonstration of the architecture is depicted in Figure 1, and taken from the original paper [2].

### 1.2. SegFormer

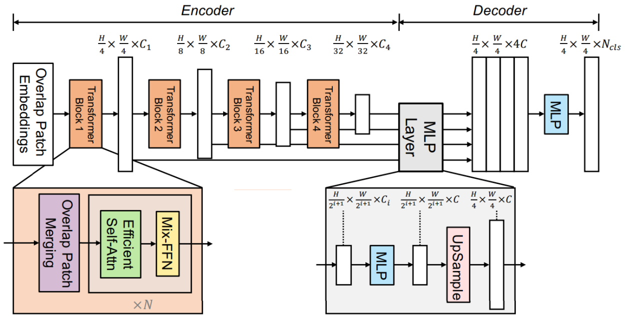

Encoder
Decoder
$\frac{H}{4}\times\frac{W}{4}\times C_1$
$\frac{H}{8}\times\frac{W}{8}\times C_2$
$\frac{H}{16}\times\frac{W}{16}\times C_3$
$\frac{H}{32}\times\frac{W}{32}\times C_4$
$\frac{H}{4}\times\frac{W}{4}\times 4C$
$\frac{H}{4}\times\frac{W}{4}\times N_{cls}$
Overlap Patch Embeddings
Transformer Block 1
Transformer Block 2
Transformer Block 3
Transformer Block 4
MLP Layer
MLP
Overlap Patch Merging
Efficient Self-Attn
Mix-FFN
×N
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i$
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C$
$\frac{H}{4}\times\frac{W}{4}\times C$
MLP
UpSample

Figure 2: SegFormer architecture

Similarly to the DeeplabV3+, the SegFormer architecture is articulated on the high performance of the encoder, where hierarchical transformers are incorporated to generate different spatial features at different scales. The building block of the Segformer is called *Transformer block*, where it takes features tensor and returns a tensor of embeddings. The first component of this block is for overlapped patch merging; It takes a features tensor and applies a convolution with a kernel size equal to the size of the patch, and a stride less than the kernel size to have overlapped patches. The second component is aimed to replace the positional embedding and is called Mix-FNN; It uses depth wise convolution with a kernel size of 3 × 3 to leak positional information from the neighboring embeddings. The last component is the Efficient self attention, where it reduces the dimension of the input tensor of features with a ratio $r$, then it returns the attention matrix of this new tensor using the conventional multihead-attention mechanism explained in Figure 3 taken from [6]. Figure 2 explains the architecture of the SegFormer; It was initially taken from the original paper [11], and then modified to match original code provided with the paper.

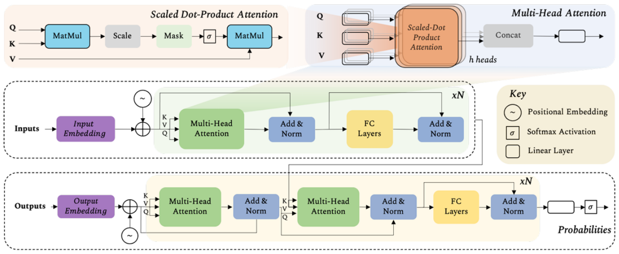


Figure 3: Multihead attention mechanism

## 2. Contribution

This analysis enabled us to provide a more solid proof of the superiority of SegFormer on DeeplabV3+. Furthermore, this work can be a footing stone for a more thorough comparison between CNN and ViT based architectures to generalize the hypothesis that ViT based semantic segmentation is way more robust than CNN based ones, in most cases.

## 3. Related work

After the success of vision transformers [4] over CNNs in classification, different transformer-based architectures have been proposed for semantic segmentation [12]. [11] introduced SegFormer, an efficient, but yet a very powerful semantic segmentation framework that unifies Transformers with lightweight multi-layer perceptron (MLP) decoders. A survey [6] summarizes and compares the vision transformer techniques in the computer vision domain. From CNN's segmentation, DeepLabV3+ [2] is considered to be one of the most promising ones to date.
An extensive study of the ViTs' properties was conducted in [8], to show their robustness in classification when introducing nuisance(e.g. occlusions and distributional shifts) in the input. Clearly, this study showed that ViTs outperformed CNNs, where ViTs were able to retain up to 60% top-1 accuracy when dropping up to 80% of information in the image, while CNN models showed 0% top-1 accuracy; the aforementioned study uses the encoders of different architectures for classification tasks; Hence the segmentation

results remain unknown. Even though new methods had been suggested to increase the robustness of CNNs to noise in [5], they still focused only on corruptions, perturbations, and adversarial perturbation.

# 4. Methodology

## 4.1. Dataset

The person dataset is a publicaly available dataset for semantic segmentation task in a kaggle competition. It has 5678 labeled images, and was specifically chosen to be completely new and easy to segment for both models. To test the robustness of DeeplabV3+ and SegFormer, we applied different data perturbations with different degrees and tested the pre-trained models on all of them to perform our analysis.

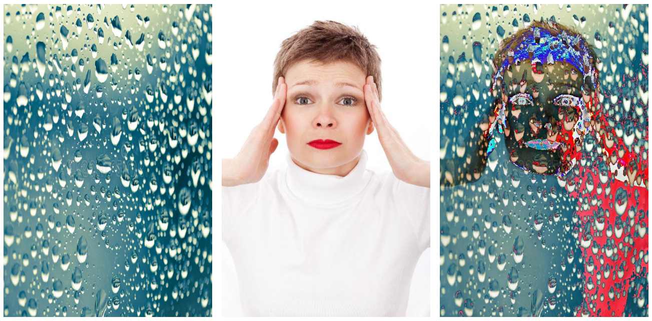
## 4.2. Data perturbation

### 4.2.1 Contextual information disturbance

Given an image $Img_o$ from the original dataset, we added a different contextless image denoted $Img_d$ with redundant information across all patches.The addition operation is the following

$$Img_N = Img_o + a \times Img_d$$

multiplied imgd with a weight a to controlwe multiplied $img_d$ with a weight $a$ to control the amount of information about the person present in the generated image.



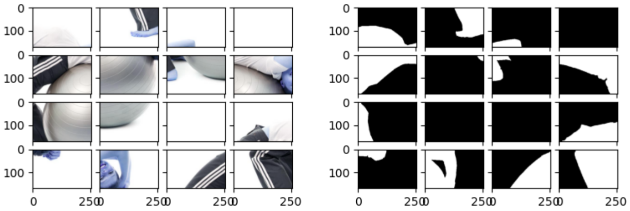
((a)) $Img_d$ ((b)) $Img_o$ ((c)) $Img_N$ with $a = 1$

Figure 4: Context loss

Using the new generated data with $a \in (0.005, 0.01, 0.05, 0.1, 0.5, 1)$, we can observe the behaviour of the two models when the global context of the image is not completely meaningful. Since vision transformers rely on self attention between patches to predict the mask, we aimed to see how it is going to perform when most of them give similar embeddings (i.e. have the same context, see Figure 9(b)). Furthermore, the DeeplabV3+ is expected to give better results in this case, as it relies on local features which are still present in the generated image, like edges and blobs, to recognize and then segment the person.

### 4.2.2 Random patch shuffling

This test was first performed in [8], where the ViT proved to be robust to patch shuffling in image classification. In order to see if it can retain the spatial information in addition to the class information, we did the same analysis in the segmentation task. First, we divided the original images and masks into $gridsize \times gridsize$ patches, and then randomly shuffled them, while assuring that a pair (image, mask) are shuffled in the same way for a correct IoU measurment.
For the experiments, we choose $gridsize \in (2, 4, 8, 16, 32)$.



((a)) shuffled image ((b)) corresponding mask

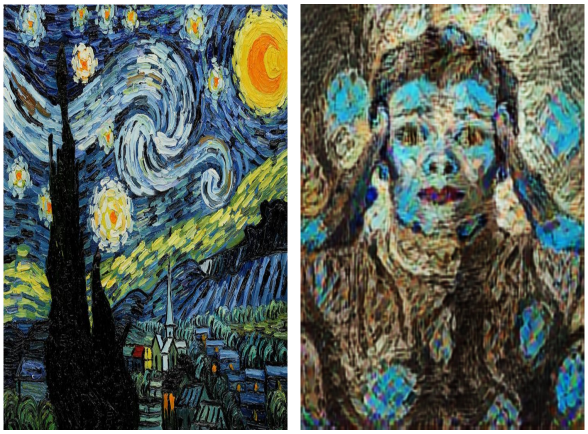
Figure 5: Image shuffling $gridsize = 4$

### 4.2.3 Stylization

The stylized images are generated by feeding a Picasso drawing and the original dataset to an arbitrary image stylization model available on tensorflow_hub. The new images will loose most of the textures that define the person class, in addition to the meaning in the patches which will make both models struggle during segmentation. In this case, both attention and convolutions must fail, and the aim of the analysis is to see which one can capture more spatial and contextual information about the person.

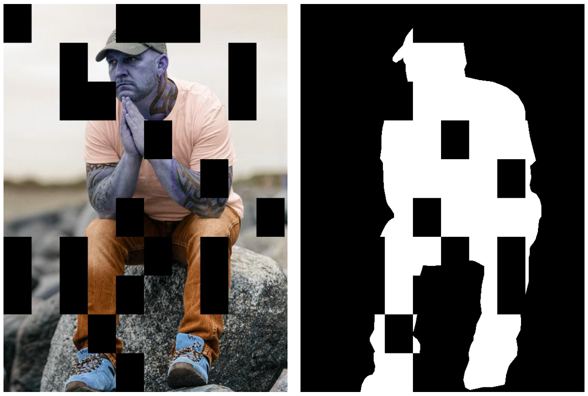
### 4.2.4 Random patch dropping

The original image was divided into a total of 100 patches (10x10 grid). The size of each patch varies depending on the size of the original image. Then a random subset of N patches is dropped, where N represents the % of dropped patches, which is taken to be 30%, 50% and 70%. This results in three different copies of the original dataset with 30%, 50% and 70% patches randomly dropped, respectively.



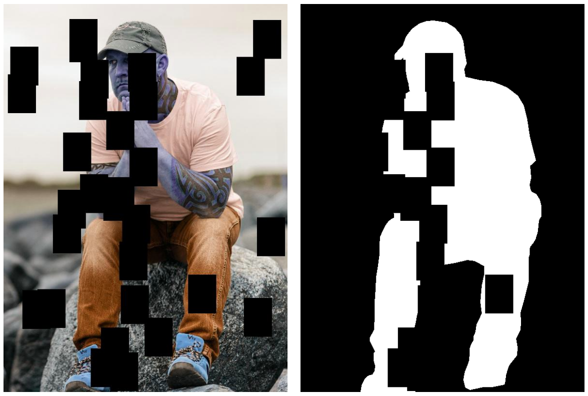
((a)) Stylization image ((b)) Stylized image

Figure 6: Image stylization



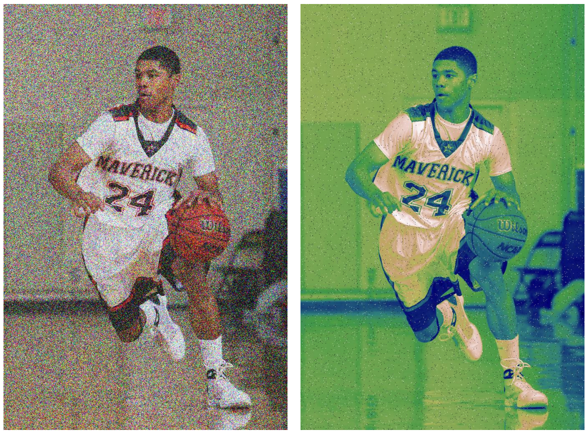
((a)) Generated image ((b)) Generated mask

Figure 7: Random patch dropping



((a)) Generated image ((b)) Generated mask

Figure 8: Random cropping



((a)) Gaussian noise ((b)) Salt and pepper noise

Figure 9: Noise effects

### 4.2.5 Random cropping

Similar to the random patch dropping, we divide the original image into 100 patches to calculate the patch size for each image. Then, we randomly crop the image N times with a crop window size equal to the calculated patch size, where is the number of patch-sized windows to be cropped. Here, we used N to be 30, 50 and 70, and tested our models on each of these modified copies of the dataset.

### 4.2.6 Gaussian Noise

One interesting thing to test is the behavior of these models when gaussian noise is applied. Gaussian noise is different from salt and pepper noise because the noise values are randomly selected which could range from any value between 0 - 255, unlike salt and pepper where only maximum or minimum pixel value noise is added. The equation of gaussian noise is given by

$$GN \sim \mathcal{N}(\mu = 0, \sigma^2)\,.$$

$$NoisyImage = OriginalImage + GN$$

Here standard deviation is variable. Hence we tested with different standard deviation value of 1.0, 2.5, and 5.0.

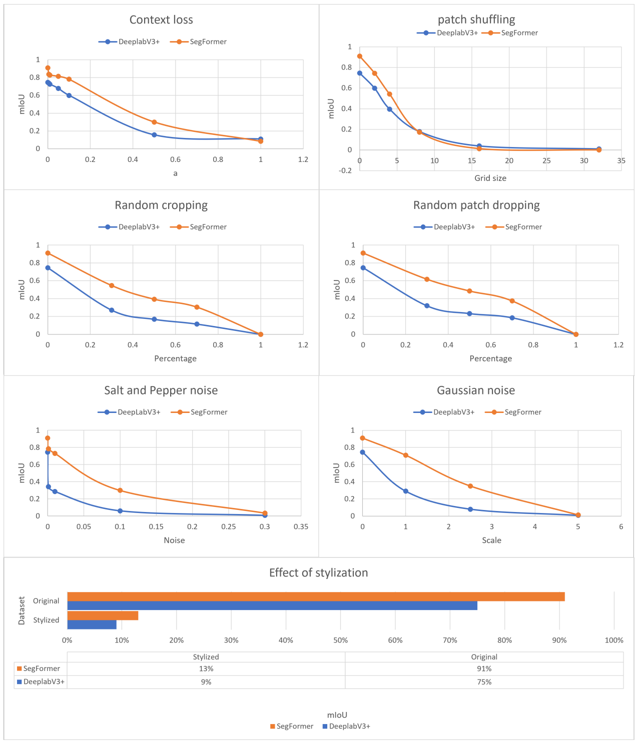
### 4.2.7 Salt and pepper noise

We generated and tested salt and pepper noise dataset with different noise intensity values. Salt and pepper noise is also known as Impulse noise. It is the addition of pixel spikes(either 0 or 255-pixel value) on the original image. We randomly add noise with a given probability. Here we tested with 0.001, 0.01, 0.1, and 0.3 probability values.

### 4.3. Experimentation

The experimentation was conducted in the HPC machines for both models, where each test took approximately 20mins for the segformer, and 20 mins for the deeplabv3+ on each transformed dataset.After we ran the tests, we plotted the interpolation of the obtained mIoUs, see Figure 10.

## 5. Results & discussion



| | Stylized | Original |
|---|---|---|
| SegFormer | 13% | 91% |
| DeeplabV3+ | 9% | 75% |

Figure 10: Experimental results

Both models were evaluated on the *person* dataset and applied to each of the perturbed image datasets. Segformer outperformed DeeplabV3+ in all the experiments performed and gave better segmentation masks. Figure 11 and 12 depict the sample predicted masks of both models that are discussed below.

On the dataset containing images with Gaussian noise, we increased the noise with a constant factor to check the extent of robustness in both models. The gaussian noise performance graph in Figure 10 shows that the performance of DeeplabV3+ drops sharply; however, Segformer gives a significant intersection over union (mIoU) score even for a scale of 2.5, suggesting the models' high capacity to deal with serious levels of Gaussian noise.

Both models fail when tested on the salt and pepper noise addition, with Segformer maintaining a higher mIoU than its competitor model. This can be due to the fact that the salt and pepper noise induces both dark and bright values in the image pixels, thereby introducing sharp changes and gradients and thereby fooling the models easily.

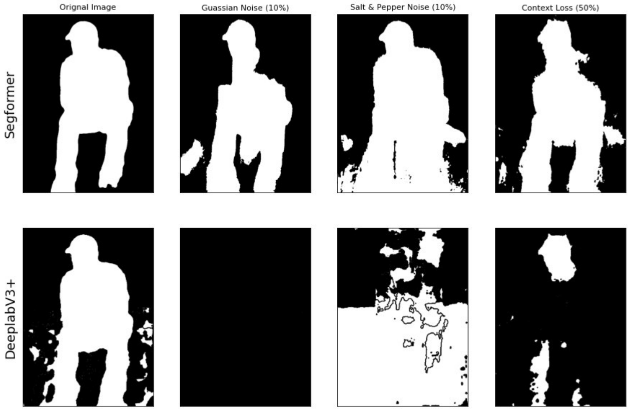


Figure 11: Predicted masks

For the patch shuffling task, tested models show a similar trend, and their performance drops below 20% as we divide images into an 8x8 grid. This is the only experiment where DeeplabV3+ and Segformer maintain a very narrow performance gap. This also demonstrates that Segformer does not provide significantly better results on such images (see patch shuffling graph in Figure 10).

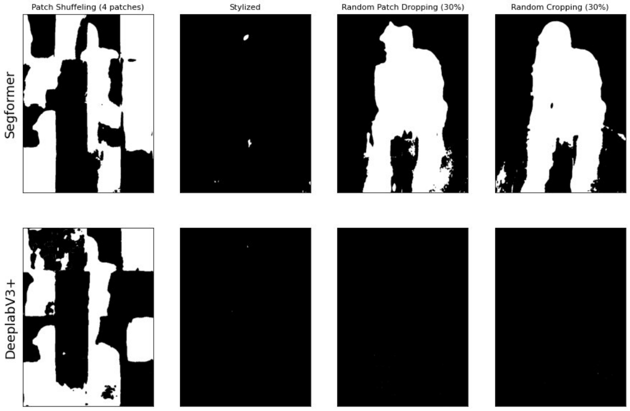


Figure 12: Predicted masks

We also applied stylization to the original images to see the effect of domain change on performance of these models. The stylization graph in Figure 10 shows that both models do not perform well on this dataset. DeeplabV3+ and Segformer achieved 0.9 and 0.13 mIoU, respectively, which is significantly lower than 0.75 and 0.91 mIoU on the original dataset. Here, it is clear that the relative performance of DeeplabV3+ drops more than that of Segformer. Further experiments can be helpful to find out how this affects other similar stylization changes. Surprisingly, neither model fails immediately on images with lost context.

These models gradually lose performance as the proportion of stray image mixture is increased in the original image.

Furthermore, both models exhibit a gradual decline in performance in images with dropped and randomly cropped patches. The mIoU results in both experiments show a similar trend (fig 10) however, both models perform relatively poor on randomly cropped images compared to patch dropping.

Finally, our experiments show that Segformer gives better mIoU in all the cases, which strongly suggests its capacity to segment images with significant levels of diverse perturbations and noise. However, further experiments to test these models on other image perturbations can provide more in-depth information.

## 6. Conclusion

Segformer provides better segmentation results compared to DeepLabV3+ because of the attention mechanism that looks at the global context of the image. Its robustness indicates its practicality in applications where noise is present in the data, as segformer is less influenced by the noise than DeepLabV3+ (e.g. medical imaging, autonomous driving). Analyzing the robustness of ViT and CNN based segmentation models on different weather conditions would be an interesting future research topic to explore for more in-depth analysis of these architectures.

## References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1, 2

[3] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34:9355–9366, 2021. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[5] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 3

[6] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 2

[7] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[8] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021. 2, 3

[9] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021. 1

[10] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 1

[11] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1, 2

[12] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511, 2019. 2

[13] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence

perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 1