ADVENT: A DiVersE Neural styles Transfer framework

XUANBIAO ZHU [ZHUXB@SEAS]
JIAHANG SHA [JHSHA@SAS]
XINMENG HUANG [XINMENGH@SAS]

ABSTRACT. The most popular style transfer technique, generative adversarial network (GAN), is only capable of transferring an input image to one style. We propose ADVENT (**A** **D**i**V**ers**E** **N**eural styles **T**ransfer framework) to translate multiple styles for different objects of an images simultaneously. The core idea is to apply generative methods over detected objects with pre-learned masks. Our framework is general that embraces all semantic segmentation models like SegFormer and image style transfer methods like CycleGAN. We fine-tune two semantic segmentation models: SegFormer and FCHarDNet and train a CycleGAN model from scratch with a novel L_1 regularization to preserve content. Experiments show that our suggested method delivers pleasant results.

1. INTRODUCTION

How Vincent Van Gogh or Picasso would paint it with his brush when he saw the same thing as you? The goal of style transfer is to use Neural Network to mimic the artists, like Van Gogh and Picasso, to draw something they have never seen before. More specifically, the task of transferring a style from a given image to a target image while keeping its original details is known as style transfer. As an example, if we want to have the appearance of a painting, like the Starry Night, while remaining the fundamental things of our original image, like an apple, we can use Style transfer to realize that.

Although existing methods can actually realize the function as what we say before, they can only be applied on the whole image. But if we want our Van Gogh only draw just an apple but remain our background as the same, the current methods will fail. A more intelligent algorithm is to be selective and interactive, the user can choose which part of the content image should to transfer to the style of Van Gogh, which part should be the style of Monet and which part should remain the same style as original.

In order to solve the problem we listed before, we combine object detection with generative adversarial network(GAN) to realize diverse neural style transfer. In details, object detection helps recognize the class of objects and GAN helps generate different styles. User can determine which object in the image should be transferred to which style.

1.1. **Contributions.** Our contributions are as follows:

- We propose ADVENT framework to achieve diverse neural styles transfer over different objects within a single image. Our framework embraces all object segmentation models and style transfer methods.
- We fine-tune two models: SegFormer and FCHarDNet in the objection segmentation module.
- We implement and train a CycleGAN model from scratch. In our training, a novel L_1 penalty term is imposed to preserve image's content under generators' transformation.
- Our experiments show a remarkably visible effect of diverse neural styles, which sheds light on potential commercial value of augmented reality.

2. BACKGROUND

2.1. **Harmonic Densely Connected Network(HarDNet).** ResNet, MobileNet, and DenseNet are examples of cutting-edge neural network designs that obtains outstanding accuracy when they have low MACs and small model size. However, when it comes to real-time recognition, the inference time might be as important as accuracy. Although the above network can significantly reduce their parameters through compression, the DRAM traffic problem haven't been solved by these networks since their feature maps are usually take up a large share in DRAM. To solve that problem, HarDNet use the different sparsification from LogDenseNet. The layer k in each block connects to layer $k - 2^n$ if 2^n divides k where n is an non-negative integer and $k - 2^n \geq 0$. When the network uses this scheme, Layers 1 through $2^n - 1$ could be flushed from memory once layer 2^n has been processed. Obviously, Layers divide by a larger power of

two have more influence on the proposed network than those divided by a smaller power of two. Therefore, the network also add weights to balance that.

2.2. **SegFormer.** When solving the task of semantic segmentation, researchers typically use traditional Fully convolution network(FCN). However, SegFormer is another branch, using transformer to handle the same problem. The encoder of SegFormer doesn't use interpolating positional codes when doing inference on image which resolution is different from the training image, which means the model can adapt test image with arbitrary resolution without affecting its performance. In terms of the bottleneck of the encoders, the self-attention layers computes as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_{\text{head}}}}\right)V \tag{1}$$

In SegFormer, it uses the sequence reduction process, with a reduction ratio R to reduce the length of sequence:

$$\hat{K} = \text{Reshape}\left(N/R, C \cdot R\right)(K)$$
$$K = \text{Linear}(C \cdot R, C)(\hat{K})$$

K is the sequence to be reduced, the Reshape function reshape K to the shape $N/R \times (C \cdot R)$, the $\text{Linear}(C_{\text{in}}, C_{\text{out}})(\cdot)$ function is just a linear layer transforms the dimension from C_{in} to C_{out}. Besides that, SegFormer also uses MiX-FFN to replace normal positional encoding(PE) to introduce location information. The process of Mis-FFN is:

$$x_{\text{out}} = \text{MLP}(\text{GELU}(\text{Conv}_{3\times3}(\text{MLP}(x_{\text{in}})))) + x_{\text{in}} \tag{2}$$

As we can see, Mix-FNN combines 3×3 convolution with normal PE to realize a data-driven positional encoding.

As for the decoder, SegFormer uses a lightweight MLP decoder, called All-MLP to make full use of features generated from transformer. The whole process contains four steps:

$$\hat{F_i} = \text{Linear}(C_i, C)(F_i), \forall i$$
$$\hat{F_i} = \text{Upsample}\left(\frac{H}{4} \times \frac{W}{4}\right)(\hat{F_i}), \forall i$$
$$F = \text{Linear}(4C, C)(\text{Concat}(\hat{F_i})), \forall i$$
$$M = \text{Linear}(C, N_{cls})(F)$$

where M is the predicted mask and N_{cls} is the number of classes.

2.3. **CycleGAN.** Naive GAN targets to realize generation of a single data domain where a generator G tries to generate data from distribution as close to the real data domain as possible, while another discriminator D tries to distinguish the real data from the generated data as much as possible. In order to achieve image-to-image translation, CycleGAN uses two discriminators D_A, D_B for each domain. Each discriminator determines whether the data in the domain is real. As for the generators G_{AB} and G_{BA}, they act a bit like an Autoencoder structure, except that the decoder outputs the counterpart in the target domain instead of the original one. To make full use of both discriminators, the two generators should play roles like inverse, which is regularized by ℓ_1 loss: $\mathcal{L}_{\text{CYC}}(G_{BA}, G_{AB})$. The final objective is like (3).

$$\min_{G_{BA}, D_A, GAB, D_B} \mathcal{L}_{\text{GAN}}(G_{BA}, D_A) + \mathcal{L}_{\text{GAN}}(G_{AB}, D_B) + \lambda\mathcal{L}_{\text{CYC}}(G_{AB}, G_{BA}) \tag{3}$$

3. Related Work

3.1. **Semantic Segmentation.** Semantic segmentation can be thought of as a pixel-level extension of image classification. FCN [Shelhamer et al., 2017a], a fully convolutional network that conducts pixel-to-pixel classification in an end-to-end manner, is the foundational work of semantic segmentation in the deep learning age. Following that, researchers concentrated on expanding the receptive field [Zhao et al., 2017, Yang et al., 2018], refining the contextual information [Yuan and Wang, 2018, Yu et al., 2020], introducing boundary information [Ding et al., 2019, Bertasius et al., 2016], constructing various attention modules [Fu et al., 2019]. These methods considerably increase semantic segmentation performance at the cost of incorporating a large number of empirical modules, resulting in a computationally intensive framework. Transformer-based architectures have been shown to be useful for semantic segmentation in more recent methods [Zheng et al., 2021, Xie et al., 2021]. These approaches, however, are still computationally intensive.

3.2. **Style Transfer.** Image-to-image translation has been proposed in Image Analogies by [Hertzmann et al., 2001], which uses a non-parametric texture model [Efros and Leung, 1999] on a single input-output training image pair. More recent approaches use CNNs and GANs to learn a parametric translation function from a dataset of input-output samples (e.g., [Shelhamer et al., 2017b, Zhu et al., 2017]). Our method is based on the "'CycleGAN" framework of [Zhu et al., 2017], which learns a mapping from input images to a target domain using a conditional generative generation [25]. However, unlike previous research, we do not learn the mapping using paired training examples. [Gatys et al., 2016b, Johnson et al., 2016, Ulyanov et al., 2016, Gatys et al., 2016a] is another way of image-to-image translation, in which the content of one image is combined with the style of another image based on matching the Gram matrix statistics of pre-trained deep features to create a novel image. Instead of learning the mapping between two distinct photos, we are primarily interested in learning the mapping from one image collection to other many styles by attempting to capture correspondences between higher-level appearance patterns. As a result, our method can be used for various tasks where single sample transfer methods fail, such as painting photos, object transformation, and so on.

4. Approach

4.1. **Framework.** Our proposed ADVENT framework mainly consists of two components: 1. an object semantic segmentation algorithm to get masks of the selected objects; 2. one or many (can be designated but usually small) style transfer algorithms to obtain the stylized version of the content image. With the help of the masks, the styled images are then blended with the original content image to produce the result of migrating the styles of a reference image to the specified objects. The pipeline of the proposed method is depicted in Figure 1. Note that the choice of style and object in an image can be designated by users in an interactive way.

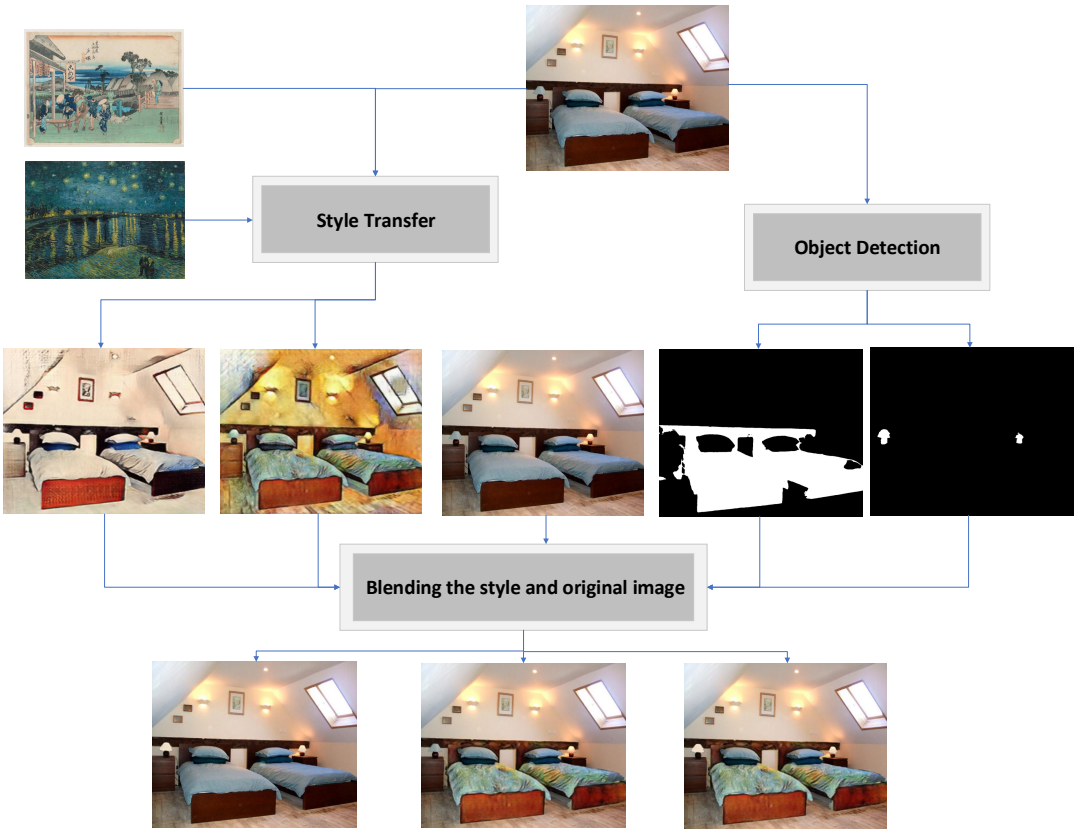

Figure 1. An overview of ADVENT. It consists of three major components: (i) the Style Transfer Module, (ii) the Object Segmentation Module, and (iii) the Style Blending Module.

4.2. **Blending.** The original content image X_{Content}, the stylized images $X_{\text{Style}_1}, \ldots, X_{\text{Style}_k}$ from the style transfer module, and the disjoint binary masks $M_1, \ldots, M_k$ of the segmented objects $O_1, \ldots, O_k$ from the object segmentation module are all accepted by the blending module. This module employs a simple masking technique based on the following formula:

$$X_{\text{Transferred}} = (\mathbb{1} - \sum_{i=1}^{k} M_i) X_{\text{Content}} + \sum_{i=1}^{k} M_i X_{\text{Style}_i}.$$

Based on the mask of the segmented object $M_1, \ldots, M_k \in \{0, 1\}$, every pixel in the generated picture $X_{\text{Transferred}}$ would be equivalent to the RGB values of either the unaltered content image X_{Content} or the styled images $\{X_{\text{Style}_i}\}_{i=1}^{k}$.

5. Experimental Results

The following pipeline is used: First, the images are all resized: for FCHarDNet with HarDNet-70 backbone, this would be 1024×1024; for SegFormer with MiT-B3 backbone, this would be 512×512. Then, the images are passed through the semantic segmentation modules to detect the regions of interest and their corresponding masks. Next, neural style transfer modules take in the regions of interest and output the images with styles transferred on each part as directed. At the final stage, all regions of interest with transferred styles are to be assembled into the same image.

5.1. **Data sets.** We experiment with three publicly available dataset for different purposes: ADE20K, Cityscapes, and COCO-Stuff. Cityscapes is a dataset with street images taken while driving for semantic segmentation that includes 5000 fine-annotated high-resolution photos divided into 19 categories which we use for FCHarDNet with HarDNet-70. ADE20K is a scene parsing dataset published for challenging with images from 150 fine-grained semantic categories that we use for SegFormer with MiT-B3 backbone. COCO-Stuff is a scene parsing dataset published for challenging with images from 182 categories from which we randomly choose a few images to verify the final composite effects of our model.

5.2. **Object Segmentation.** For SegFormer with MiT-B3 backbone, we resize all the input images to 512×512. For FCHarDNet with HarDNet-70, we resize the input images to 1024×1024. For both models, we fine-tune the pretrained model with data augmentation through random resize and cropping on both images and labels. For the hyper-parameter for fine tuning SegFormer, we chose all as close as possible to those used in the pretraining process: AdamW as optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$, 6e-5 as the learning rate with a Polynomial LR scheduler imported from repository, 0.01 as the weight decay, and 30 epochs. For the hyper-parameter for fine tuning FCHarDNet, we choose SGD as optimizer with Nesterov momentum, 5e-2 as the learning rate with cosine learning rate decay, 6e-5 as the weight decay, and 100 epochs.

5.3. **Style Transfer.** For the style transfer module in the project, we implement the famous CycleGAN model from scratch. We first resize images to 256×256 and transform images to have pixel-wise 0.5 mean and 0.5 standard deviation. In our model, the discriminators D_A and D_B are naive CNNs that are made up of only 6 convolutional layers and leaky-RELU activation. D_B promotes and categorizes the transition from domain A to domain B, e.g. making normal photos into Monet styles. Our two generators $G_{B\to A}$ and $G_{A\to B}$ are the two generators consist of three components: 1. encoder (compressing the image into a smaller feature representation); 2. residual blocks (connecting the output of one layer to the input of a previous layer); 3. decoder (reconstructing images from feature representation). Beyond standard loss defined in (3), we also add an additional term $\mathbb{E}\left[\|X - G_{A\to B}(X)\|\right] + \mathbb{E}\left[\|Y - G_{B\to A}(Y)\|\right]$ in the loss to preserve the content. We use Adam optimizer with hyper-parameter $\text{lr} = 0.0002$, $\beta_1 = 0.5$, and $\beta_2 = 0.99$ in our training.

5.4. **Results.** The figure 2 shows four of result images among $\binom{4}{2}$ combinations of styles for bed and rug. The result images show that the effect of our proposed method is desired: the style transfer on bed and that on the rug are both obvious while disjoint. The comforter wrinkles are also visible in the result images, which indicates that the content in the original image is preserved to a satisfactory extent during style transfer. Different combinations of styles on individual items result in dramatic visual effects, which implies a potential use in augmented reality.

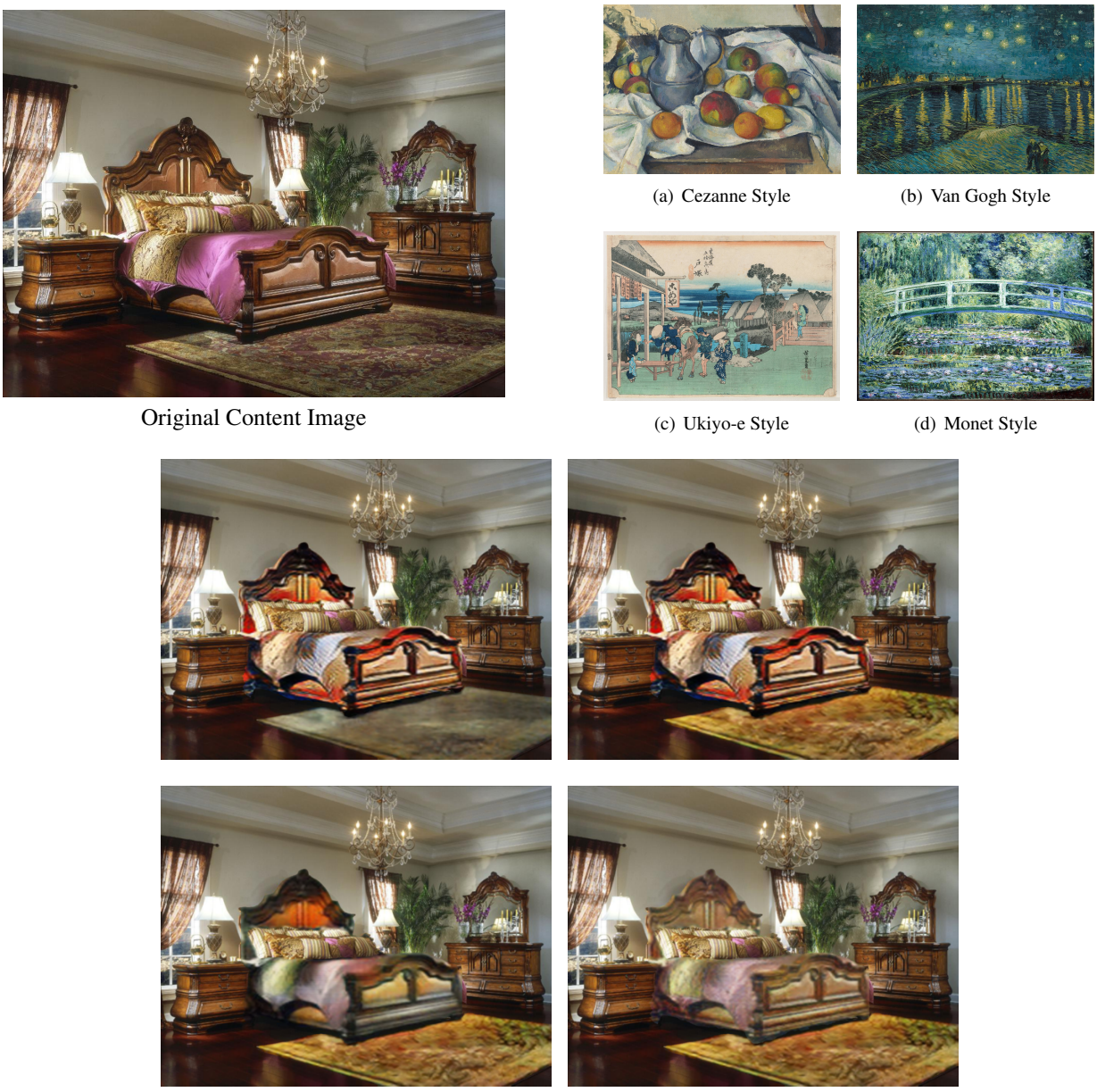

FIGURE 2. Four mix-n-match combinations of diverse transferred style effects of the bed and the rug

6. DISCUSSION

In this project, we propose a framework named ADVENT for diverse neural styles transfer. Due to time/resource limitation, our semantic segmentation module does not show highly satisfactory performance. The boundaries of style transferred objects also often blur. Therefore, the style transfer module and the object detection module could be combined into more methods based on state-of-the-art deep network backbones in future work for this study. Further research of additional filters to maintain the varied properties of the content image or style image is another prospective future development, as this would make the style transfer more controlled in terms of how it combines the styles into the content.

REFERENCES

[Bertasius et al., 2016] Bertasius, G., Shi, J., and Torresani, L. (2016). Semantic segmentation with boundary neural fields. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3602–3610.

[Ding et al., 2019] Ding, H., Jiang, X., Liu, A. Q., Magnenat-Thalmann, N., and Wang, G. (2019). Boundary-aware feature propagation for scene segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6818–6828.

[Efros and Leung, 1999] Efros, A. A. and Leung, T. K. (1999). Texture synthesis by non-parametric sampling. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 2:1033–1038 vol.2.

[Fu et al., 2019] Fu, J., Liu, J., Tian, H., Fang, Z., and Lu, H. (2019). Dual attention network for scene segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3141–3149.

[Gatys et al., 2016a] Gatys, L. A., Bethge, M., Hertzmann, A., and Shechtman, E. (2016a). Preserving color in neural artistic style transfer. *ArXiv*, abs/1606.05897.

[Gatys et al., 2016b] Gatys, L. A., Ecker, A. S., and Bethge, M. (2016b). Image style transfer using convolutional neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423.

[Hertzmann et al., 2001] Hertzmann, A., Jacobs, C. E., Oliver, N., Curless, B., and Salesin, D. (2001). Image analogies. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*.

[Johnson et al., 2016] Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *ECCV*.

[Shelhamer et al., 2017a] Shelhamer, E., Long, J., and Darrell, T. (2017a). Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:640–651.

[Shelhamer et al., 2017b] Shelhamer, E., Long, J., and Darrell, T. (2017b). Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:640–651.

[Ulyanov et al., 2016] Ulyanov, D., Lebedev, V., Vedaldi, A., and Lempitsky, V. S. (2016). Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*.

[Xie et al., 2021] Xie, E., Wang, W., Wang, W., Sun, P., Xu, H., Liang, D., and Luo, P. (2021). Segmenting transparent objects in the wild with transformer. In *IJCAI*.

[Yang et al., 2018] Yang, M., Yu, K., Zhang, C., Li, Z., and Yang, K. (2018). Denseaspp for semantic segmentation in street scenes. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3684–3692.

[Yu et al., 2020] Yu, C., Wang, J., Gao, C., Yu, G., Shen, C., and Sang, N. (2020). Context prior for scene segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12413–12422.

[Yuan and Wang, 2018] Yuan, Y. and Wang, J. (2018). Ocnet: Object context network for scene parsing. *ArXiv*, abs/1809.00916.

[Zhao et al., 2017] Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239.

[Zheng et al., 2021] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H. S., and Zhang, L. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6877–6886.

[Zhu et al., 2017] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251.