CENG 796 - Peer-review form

Reviewed project ID: Group **07**

Reviewed project's title (title of the paper): StyleSwin: Transformer-based GAN for High-resolution Image Generation
Reviewer name(s): Özgür Aslan, Burak Bolat

Instructions:

- Answer = *Yes*, *No* or *Partial*.
- You may expand sections as necessary.
- For most questions, you do not need to add comments, unless the instructions tell you otherwise.
- "Notebook" refers to "Jupyter Notebook" file that is expected to be named as main.ipynb

Question	**Answer**	**Comments**
Contains a jupyter notebook file	Yes	
Notebook is located at <project_root>/main.ipynb	Yes	
Notebook's first section contains paper information (paper title, paper authors, and project group members' name & contact information) Some good examples: see group03, group10, group11 (and a couple of other groups).	Parital	The paper authors' name is not included.
Notebook contains a section for hyper-parameters of the model.	No	
Notebook contains a section for training & saving the model.	No	
Notebook contains a section (or a few sections) for loading a pre-trained model & computing qualitative samples/outputs.	No	
Notebook contains reproduced plots and/or tables, as declared.	No	
Notebook contains pre-computed outputs.	No	

Data is included and/or a proper download script is provided.	No	
Notebook contains a section describing the difficulties encountered.	No	*Explain anything that looks ambiguous, hard-to-understand, etc. in this section.*
The paper has achieved its goals and/or explained what is missing.	No	Does paper mean main notebook? If it is NoN
The notebook contains a section that reproduces the figure(s) and table(s) declared in the goals.	No	
The notebook also reports the original values of the targeted quantitative results, for comparison.	No	
MIT License is included.	Yes	
As the reviewer(s), you have read the paper & understood it.	Yes	This paper uses transformer based feature extraction instead of convolution based methods in StyleGAN. They exploit the power of self attention of vision transformers. To clarify, they removed the conv layer between AdaIN operations in a Style Block and used Attention instead. However, using global attention requires quadratic complexity (one of the N patches attends N patches = N^2). Instead of using that mechanism, authors used Swin Transformer which attends M patches in a window (thus, for N patches each attends to M patches, the complexity is MN) Swin Transformer has the problem of losing long range dependencies as they limit the attention area. They provide this property by shifting the windows at each layer. Also, to improve the receptive field of the attention layers, they introduce double attention. Which divides the attention heads to two. One half is used in regular attention and the other is used for slide windowed attention. Due to using Swin Transformer, which is a local attention mechanism, in high resolution the generator produces block artifacts. To overcome this issue, they use the wavelet discriminator, since these artifacts can be easily found in the fourier space.
Implementation of the model seems correct.	Partially	** If you have not been able to find any errors , give a list of things that you have been able to match between the paper and the code. (eg. "I have located Eq. 3, 5, 7 and they seem to be corrected implemented.")*

		** Also denote any part that looks possibly problematic. You may use "additional comments" section below for your detailed comments.* **Generator:** They used **Pixel Norm** which is not mentioned in the paper. But added for normalizing values before values get out of control. I understand the purpose but I do not miss it, the group is used for noise the vector. Noise vector has already been drawn from Normal Distribution (in the train.py). I wonder about the effects of this. What I expect is applying Pixel Norm to the output of some layers of Generator or something not from Normal Dist. **SPE:** Sin-pos embed, they have borrowed this module but have not decided how to init (or simply use). They mentioned in the code. **tRGB:** Correct **Upsample:** Correct **Z → W:** They implemented conceptually, layers will be determined later. See train.py, initialization for networks are not handled for this version. :) Funny wording at generator.py for class NoiseMassage at line 576 (also, massage_mlp) **Discriminator:** Authors stated they have used StyleGAN's discriminator and the StyleGAN paper states (in Appendix C) that ProgressiveGAN's discriminator is used. In the Appendix A of the ProgGAN (https://arxiv.org/pdf/1710.10196.pdf), architecture is given at Table 2. I understand that they used two 3x3 conv without changing the input sizes (simply by using padding) and then downsampled it with avg pooling. However, the provided code discriminator.py does downsampling with stride of 2. I am not sure that stride 1 3x3 convolution and avg pooling of 2x2 is equal to stride 2 3x3 convolution. The group implemented spectral normalization. Also, I could not see any normalisation for discriminator of StyleGAN. I refer to it since the author stated they used StyleGAN's disc. About normalization: *"We do not use batch normalization [29], spectral normalization [45], attention mechanisms [63], dropout [59], or pixelwise feature vector normalization [30] in our networks." (from StyleGAN)* About the StyleGAN assumption, the group stated that they have not dealt with artifacts (artifacts mentioned in the paper) yet. Thus, I assumed the sentence from the paper: "*Since the discriminator severely affects the stability of adversarial training, we opt to use a Conv-based discriminator directly from [31] (StyleGAN)."* If the assumption is wrong, I still feel uncomfortable that downsampling is done with stride. From Wavelet paper: *"In order to downscale the image between blocks, the wavelet coefficients are combined back into a full image via IWT,* ***the image is bilinearly downsampled****, and the lower-resolution image is decomposed back into the wavelet coefficients, which are fed into the next block."*

Notebook looks professional (in terms of notation, readability, etc.)	Partial	The notebook gives enough information about the paper but it does not include any other information/code about the other sections. Therefore, we can say it is between No - Partial.
Source code looks professional (in terms of coding style, comments, etc.)	Yes	

Additional comments:

Please write any suggestions that can content-wise and/or aesthetically improve the notebook or the source code. *You may also add your lengthy comments (eg. mathematical problems that you have found in the implementation) here, and, refer to this text in your comments above.*