CycleTransformer: Text-to-Image-to-Text Using Cycle Consistency

Hila Manor and **Matan Kleiner**
Faculty of Electrical and Computer Engineering, Technion - Israel Institute of Technology
{hila.manor, matankleiner}@campus.technion.ac.il

Abstract

Both Text-to-Image translation and Image-to-Text translation have been an active area of research in the recent past. Usually only one task is handled at a time, and the methods are tailored for extracting data from one domain and translating it to the other domain. Recently, some took inspiration from CycleGAN's use of duality for unpaired data, by leveraging the cycle consistency duality for paired data of different domains, such as text and images. Inspired by those papers and recent advancements in deep learning and NLP, we propose a novel architecture, CycleTransformer, to handle both Text-to-Image translation and Image-to-Text translation on paired data, using a unified architecture of transformers and CNNs and enforcing cycle consistency. Code is available at: https://github.com/HilaManor/CycleTransformer

1 Introduction

Generating captions of given images is useful for a wide variety of applications, from online shopping to driver-assistance system. In this task, known as the **Image-to-Text** or **Image Captioning** task, a given image is translated into a grammatically correct sentence that faithfully describes the content of the image. Generating an image from a given paragraph is also useful for a wide variety of applications, from bringing designs-ideas to life to creating more accurate simulations. Known as the **Text-to-Image** task, in this task an image is generated from a text prompt of a natural language, and should be a faithful visual representation of the given text. As both tasks are difficult and interesting problems to solve, they have been an active area of research in the recent past. Inspired by recent developments in deep learning and NLP, researchers have proposed some remarkable models such as GLIDE (Nichol et al., 2021) for image generation from text or Oscar (Li et al., 2020) for caption generating from images, among other capabilities. In this paper we introduce our model for both Image-to-Text and Text-to-Image tasks. Inspired by CycleGAN (Zhu et al., 2017), a model capable of learning mappings for image translation between two unpaired datasets, we propose CycleTransformer. CycleTransformer is a cycle-consistent model comprised of CNNs and transformers that can handle both the Image-to-Text task and the Text-to-Image task, using a multi-domain paired dataset.

2 Related Work

In this section we survey related research on the Text-to-Image and Image-to-Text tasks, and mutual translation between different domains using cycle consistency.

2.1 Text-to-Image

A variety of methods have been suggested over the years to tackle this very challenging task. One influential example is AttnGAN (Xu et al., 2018), which generates images from text using attention and adversarial training, and led to further exploration with other attention models (Li et al., 2019). Still an active area of research today, last year alone OpenAI introduced two new models for the Text-to-Image task: DALL-E (Ramesh et al., 2021) which combined transformers and discrete VAE, and GLIDE (Nichol et al., 2021) which combined diffusion models and guidance. Those models were trained on hundreds of millions of images and on clusters of GPUs to achieve their state of the art results.

2.2 Image-to-Text

In the recent past combined vision-language transformers models trained on millions of image-text pairs and then fine-tuned on down stream tasks (as image captioning) have achieved state of the art results in different vision-language tasks (Lu et al., 2019). Those models simply concatenate image region features and text features as input and resort

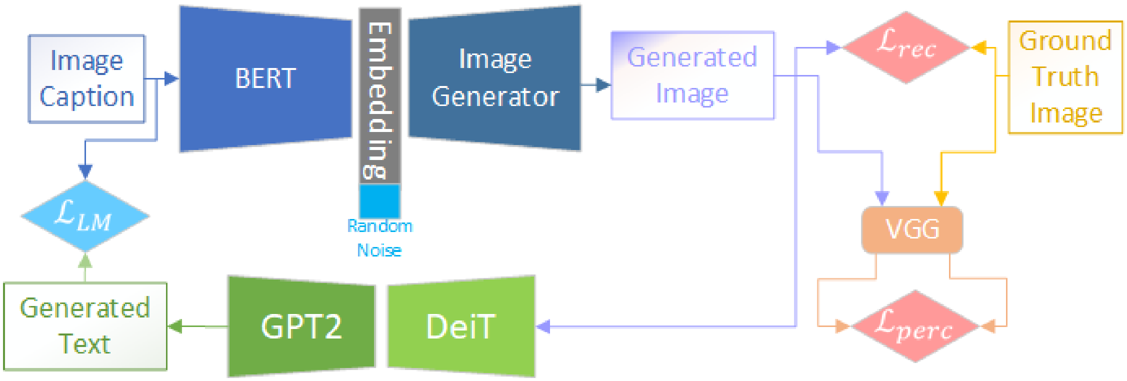


Figure 1: A diagram of CycleTransformer model

to the self-attention mechanism to learn semantic alignments between image and text. Lately, Li et al. (2020) suggested in Oscar to add to the input of the model tags detected in the images as anchor points.

2.3 Text-to-Image and Image-to-Text Using Cycle Consistency

As shown in CycleGAN, cycle consistency can help a model learn mutual transformations between images of different domains. Inspired by the idea of cycle consistency, some models were suggested in the past for Text-to-Image-to-Text using cycle consistency. Gorti and Ma (2018) used cycle consistency to improve a GAN's ability to generate images that better reflect the meaning of the sentence given to the model as input and embedded using an LSTM. Hagiwara et al. (2019) used cycle consistency to improve image captioning using GANs and LSTMs. Recently, TextCycleGAN (Alam et al., 2021) tried to create a Text-to-Image-to-Text model that performs well on both Text-to-Image generation and Image-to-Text translation using GANs, LSTMs, attention and cycle consistency.

3 Model

In this section, we present the CycleTransformer model with cycle-consistent training for Text-to-Image-to-Text translation. A diagram of our model is presented in Figure 1.

3.1 Text-to-Image

In order to generate an image from text we used a two step approach. First we use a pre-trained DistillBERT model (Sanh et al., 2019) for embedding the given caption as a matrix. We then concatenate a random noise vector sampled from the standard normal distribution to this embedding vector. The size of the noise vector is 20% of the size of the embedding vector. We feed the concatenated vector to an image generator model. The generator model is based on DCGAN's generator (Radford et al., 2015), and outputs an image of size 224×224 pixels. The DCGAN model is a well known GAN model for generating images with high perceptual quality across multiple images domain. A diagram of our image generator is presented in Figure 2.

3.2 Image-to-Text

In order to translate the generated image into text, we use an encoder decoder structure composed of a vision transformer model (Dosovitskiy et al., 2020) as an encoder and a transformer decoder as the decoder. Features are extracted from the generated image via the corresponding feature-extractor of the used vision transformer model, and passed to the encoder-decoder structure. Due to computational limitations our vision transformer is the small version of the distill-DeiT model (Touvron et al., 2021), and the transformer decoder is a GPT2 model (Radford et al., 2019). It's important to note that the images inputted to this part of the model at training is the generated image so far, which completes one side of the consistency cycle. In evaluation we use the original ground truth image, as expected for the original Image-to-Text task.

3.3 Objective Function

The model is trained with different objective functions for the Text-to-Image task and for the Image-to-Text task.
The Text-to-Image part of the model is trained with an L_2 reconstruction loss and a VGG-based perceptual loss, between the ground truth image x_i and the generated image $\hat{x}_i$. The perceptual loss encourages the image generator to generate more natural and perceptually pleasing results. Percep-

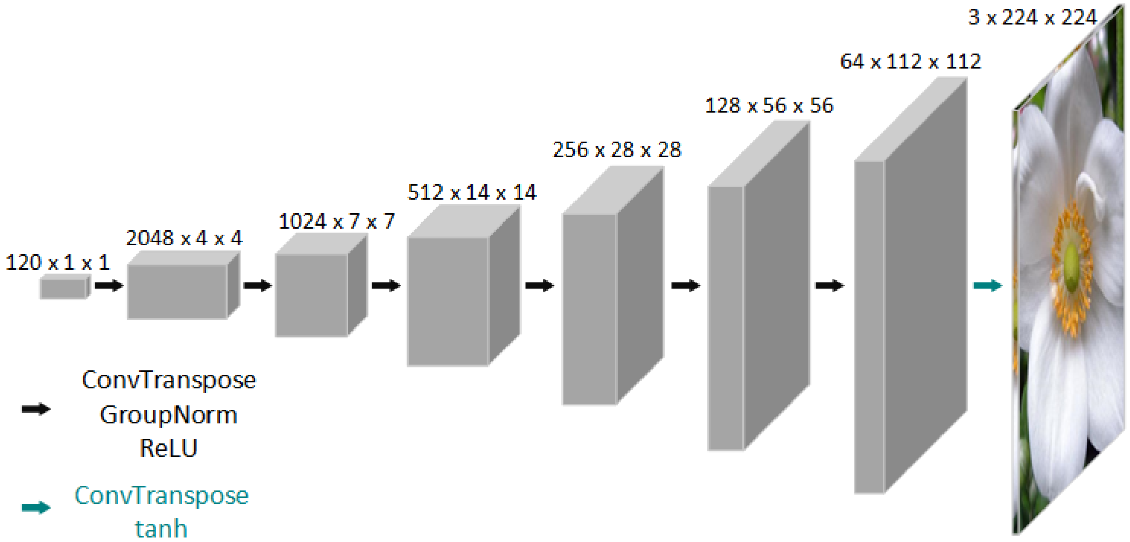


Figure 2: A diagram of the image generator in CycleTransformer. The generator is composed of 7 transpose convolution layers, seperated by a group normalization layer and a ReLU activation unit. The last layer is followed by a Tanh activation layer.

tual loss is based on the correlation between the human perception and features extracted from pre-trained classification networks. We follow Gatys et al. (2016), and use as perceptual loss a comparison between the deep-features of x_i and $\hat{x}_i$ extracted from VGG-19 (Simonyan and Zisserman, 2014). The overall Text-to-Image loss is thus:

$$\mathcal{L}_{txt2im}(x_i, \hat{x}_i) = \mathcal{L}_{rec}(x_i, \hat{x}_i) + \alpha \mathcal{L}_{perc}(x_i, \hat{x}_i) \quad (1)$$

Where α is a hyper-parameter that we set to 1×10^4. The Image-to-Text part of the model is trained using Language Modeling loss, which compares the generated caption $\hat{x}_t$ to the input caption x_t. This completes the second side of the consistency cycle. The overall Image-to-Text loss is thus:

$$\mathcal{L}_{im2txt}(x_t, \hat{x}_t) = \mathcal{L}_{LM}(x_t, \hat{x}_t) \quad (2)$$

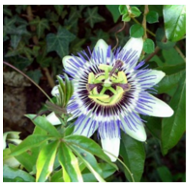
4 Data

We evaluated our model on the Oxford's 102 Flowers dataset (Nilsback and Zisserman, 2008), in which 8189 images of 102 different flowers species images are accompanied by 10 captions (Reed et al., 2016) per image. This makes the dataset a paired multi-domain dataset. This dataset is limited to the domain of flowers, and therefore it is often employed to study simple image generation from text and vice versa. This, combined with the fact that other works that have tried to enforce cycle consistency have also used this dataset lead us to choose this dataset. The images in the dataset vary greatly in scale, pose and lighting conditions, and some of the flower species display a great variety of specimens. In order to add captions to this dataset, Reed et al. (2016) used the Amazon Mechanical Turk (AMT) platform for data collection. Workers were instructed to describe only visual appearance, to avoid figures of speech, to avoid naming the species and to not describe the background. The added captions describe each image to varying degrees of success. A sample of the dataset is presented in Figure 3. 80% of the images and their corresponding captions were used for training, 10% for the validation split and 10% for the test split. Each image was paired with each of its 10 captions separately to create 10 (x_i, x_t) pairs, virtually making the dataset 10 times larger.

5 Experiments and Results

5.1 Training

We trained our model with batch of 12 images and sentences, truncating every sentence that is longer than 30 words (most of the sentences in the dataset are shorter). Both sizes were chosen to fit the GPU's memory capacity. The model was trained for 3 epochs, and for each epoch the Text-to-Image model was trained 9 iteration separately from the Image-to-Text model. This is done to give a head-start for the generator, as it is the only part of the model that is not pre-trained. We don't fine



1. the stamen are towering over the stigma which cannot be seen.
2. the flower is so and has disc of petals below the disc of stamens that are blue, white and violet
3. this flower is white and blue in color, with petals that are oval shaped.
4. this flower has blue petals as well as a green and purple pistil.
5. this flower has petals that are whgite and has purple stamen

Figure 3: A dataset sample. Only a subset of the 10 captions for this image are shown. The captions can contradict one another, can include spelling mistakes, and do not always describe core elements of the flower.

tune the distill BERT model at all, as it's used as basic word embedding, and fine-tuning leads to over-fitting and worse results. Both Text-to-Image model and Image-to-Text model were trained using Adam optimizer, with a learning rate of 5×10^{-3} and 5×10^{-6} respectively.

5.2 Baselines

We compare the cycle consistent model against two baselines, one for the Text-to-Image task, and one for the Image-to-Text task. The baselines share the original model's Text-to-Image part and Image-to-Text part architecture and hyper-parameters, respectively, and the only difference is the absence of the cycle consistency (i.e. the Image-to-Text part trains on the original ground truth image instead of the generated image, and there is no gradient flow backwards towards the Text-to-Image part).

5.3 Evaluation

The generated images are evaluated using Fréchet Inception Distance (FID) (Heusel et al., 2017), a common metric for image generation evaluation, which measures the deviation between the distribution of deep features of generated images and that of real images. For each image, the generated sentence is evaluated against all of the original image's captions using BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and ROUGE-L (Lin, 2004), common metrics for sentences similarity evaluation used also in previous cycle consistent works (2019, 2018, 2021). A quantitative comparison of the results is available in Table 1.

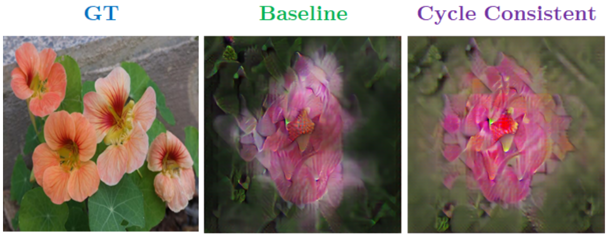
5.4 Results

The baseline model achieved better results than the cycle consistent model in all metrics, both for text and image quality, both in quantitative evaluations and in qualitative evaluations. The images created by both models present a big blob of colors (that usually corresponds to the color names present in the caption) on a green background. The static sharp artifacts are an expected direct result of using perceptual loss for image generation (Tej et al., 2020). The generated sentences by both models describe the input images quite well, however both models struggle with ending a sentence. Example of the results is presented in Figure 4. More examples can be found in Appendix A.

Model	B-4↑	R-L↑	M ↑	FID ↓
Baselines	0.4206	0.2092	0.088	114.8
Cycle Consistent	0.2246	0.1737	0.0724	143.8

Table 1: Quantitative comparison of the results between baseline and cycle consistent model.

One possible reason that the Image-to-Text baseline model achieved better results than the cycle consistent model is that it was trained with ground truth images and not with generated ones. Therefore, the features fed into GPT2 were more accurate. Another possible reason that the baseline model achieved better results is the small amount of training epochs used (due to time limitations). It's possible that for cycle consistency to reach its full potential it should be trained for a large amount of epochs, due to its difficult optimization process. We still believe that the cycle consistent model has potential to achieve better results than the baseline model, but the proper configuration is yet to be found. Finally, it seems that the proposed architecture (even without cycle consistency) also has great potential, since the baseline generations show varying shapes and intricate gradients of colors.



GT: this flower has round pale pink petal that are a deeper red color towards the ovule.
Baseline: this flower has petals that are pink with yellow shading and orange stamen. the petals are green and white. the stamen are green
Cycle Consistent: this flower has petals that are purple with shades of yellow and curled around a greed pedicel. the petals are bright red. the

Figure 4: Result comparison between the baseline models and the cycle consistent model, on the Text-to-Image task and the Image-to-Test task.

References

Mohammad R. Alam, Nicole A. Isoda, Mitch C. Manzanares, Anthony C. Delgado, and Antonius F. Panggabean. 2021. TextCycleGAN: cyclical-generative adversarial networks for image captioning. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III*, volume 11746, pages 213 – 220. International Society for Optics and Photonics, SPIE.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423.

Satya Krishna Gorti and Jeremy Ma. 2018. Text-to-image-to-text translation using cycle consistent adversarial networks. *arXiv preprint arXiv:1808.04538*.

Keisuke Hagiwara, Yusuke Mukuta, and Tatsuya Harada. 2019. End-to-end learning using cycle consistency for image-to-caption transformations. *arXiv preprint arXiv:1903.10118*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. 2019. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12174–12182.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–58.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Akella Ravi Tej, Shirsendu Sukanta Halder, Arunav Pratap Shandeelya, and Vinod Pankajakshan. 2020. Enhancing perceptual loss with adversarial feature matching for super-resolution. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

A Results

A.1 More Results from the Test Split

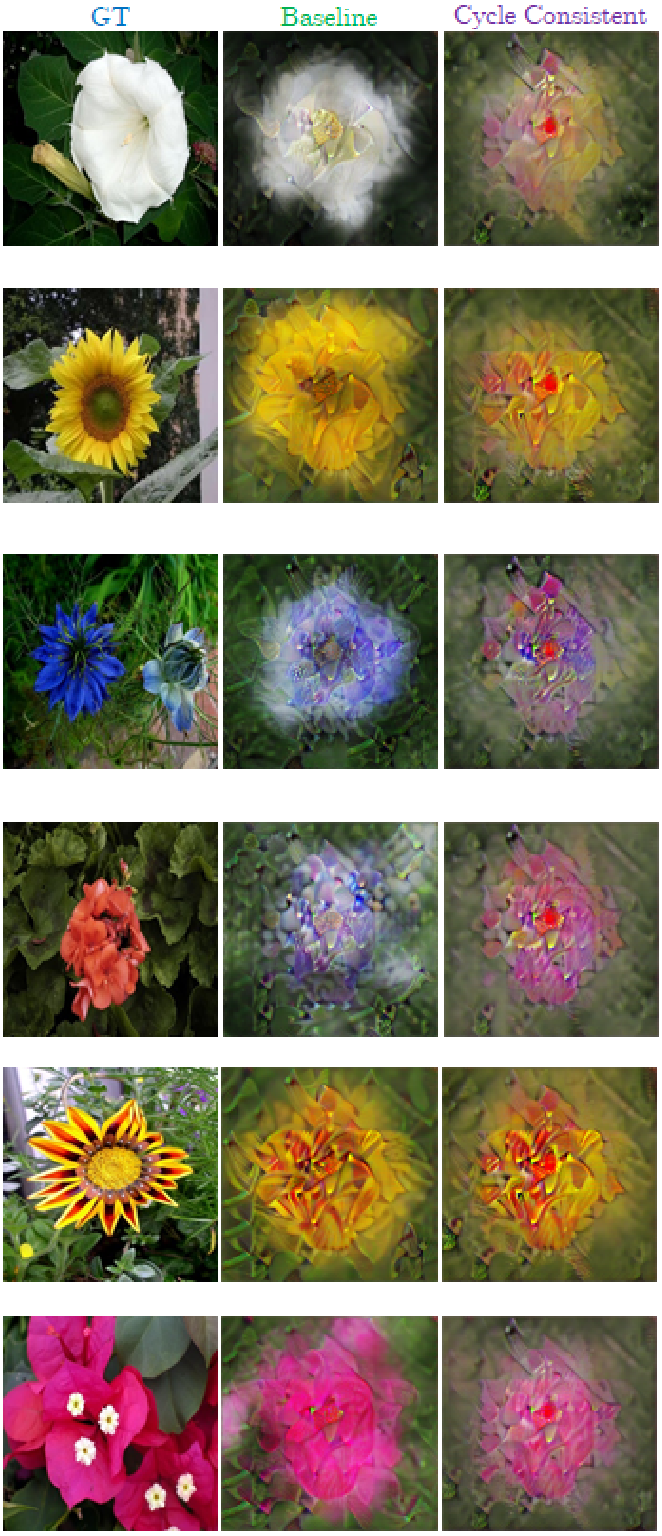
More results comparisons between the baseline models and the cycle consistent model, on the Text-to-Image task and then Image-to-Test task are available in Figure 5.

A.2 Custom Text Prompts for the Text-to-Image Task

We tried feeding into the Text-to-Image models new text prompts that we wrote. The text prompts are sentences describing flowers with different degrees of detail and complexity, and thus were categorized into 3 difficulties. The comparison between the results of the baseline Text-to-Image model and the cycle-consistent Text-to-Image model are available in Figure 6.

A.3 Custom Images for the Image-to-Text task

We also tried feeding into the Image-to-Text models new images of flowers found online (under creative commons license). We specifically chose flower species that do not appear in the original dataset and flower species that do appear. The comparison between the results of the baseline Image-to-Text model and the cycle-consistent Image-to-Text model are available in Figure 7.



GT: the petals are large and round and white and the stigma is delicate and yellow
Baseline: this flower has petals that are white and has yellow stamen. the stigma is green. the stamen is green. the petals are
Cycle Consistent: this flower is pink in color, with petals that are layered closely together. the stigma is white. the petals are light purple. the

GT: this flower has yellow petals and has green and orange stamen
Baseline: this flower has petals that are yellow and has black stamen. the stamen is yellow. the petals are arranged in a disc like
Cycle Consistent: the petals on this flower are yellow in color with many purple stamen. the stigma is green. the stigma is light purple. the stigma

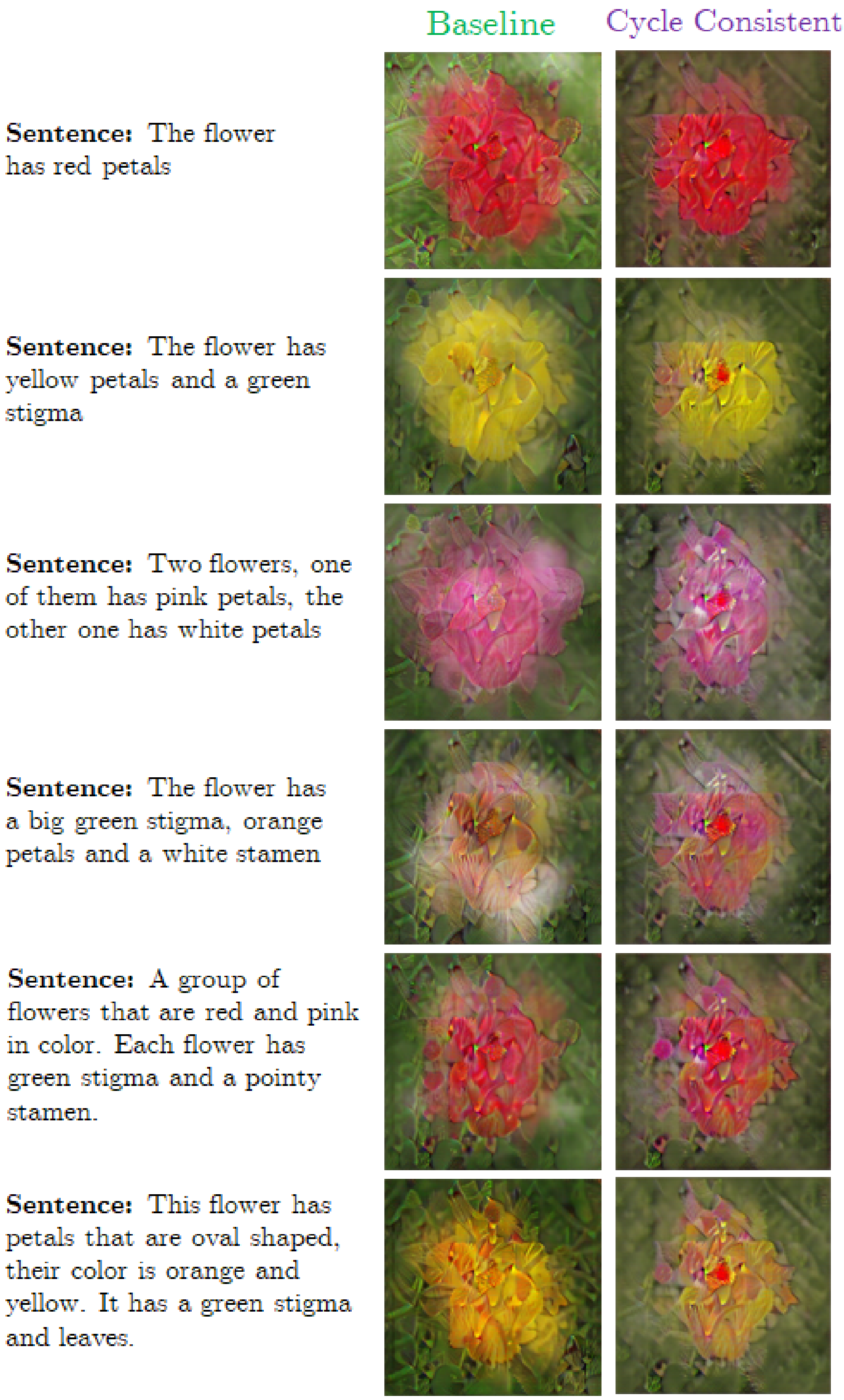
GT: this beautiful blow flower has several petals with green anther filaments in the center also the other flower is white with blue petals green ant
Baseline: this flower has petals that are blue and has green stamen. the petals are green and blue in color. the stamen is green
Cycle Consistent: the flower is purple with petals that are bright yellow and bunched together to form a ball. the filaments are green. the petals

GT: small flowers with vermilion colored petals are grouped in a cluster on the plant
Baseline: this flower has petals that are red and are bunched together. the stamen are green. the petals are small and are bunched
Cycle Consistent: the flower petals on this flower are a vibrant orange with shades of red to the pedicel. the petals are also a vibrant orange

GT: this flower is yellow and red in color, with petals that are striped red in the center.
Baseline: this flower has petals that are yellow and has red lines down the center. the petals are yellow and red in color. the stamen
Cycle Consistent: this flower has petals that are purple, with yellow stigmas. the leaves are light green. the petals are layered. the leaves

GT: this is a flower with large pink petals and small white flowers.
Baseline: this flower has petals that are pink and has flowery stigma. the stigma is white and has flowery stigma. the stigma is white and
Cycle Consistent: this flower has pink petals that are overlapping each other with dark purple stamen. the petals are bright purple. the petals are curled

Figure 5: Comparison between the baseline models and the cycle consistent model on more generation results, for the Text-to-Image task and the Image-to-Test task. The baseline models generations are more varied in shape and display more coherent colors. The generated sentences are also more diverse in the language used and the different elements of the flower are described with greater detail.



Sentence: The flower has red petals

Sentence: The flower has yellow petals and a green stigma

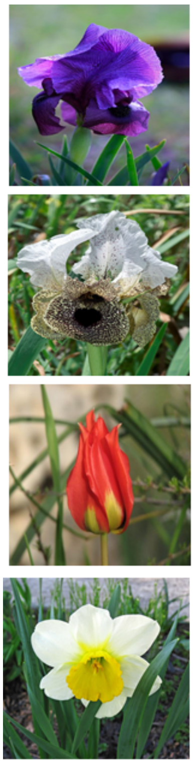
Sentence: Two flowers, one of them has pink petals, the other one has white petals

Sentence: The flower has a big green stigma, orange petals and a white stamen

Sentence: A group of flowers that are red and pink in color. Each flower has green stigma and a pointy stamen.

Sentence: This flower has petals that are oval shaped, their color is orange and yellow. It has a green stigma and leaves.

Figure 6: Comparison between the results of the baseline and the cycle-consistent Text-to-Image models, for custom text prompts on the Text-to-Image task. The first two sentences are simple, describing a single flower with one prominent color. In both cases, both models create a blob of the specified color in the center of the image, where the baseline model's blob's shape is a bit more flower-like. The next two sentences are more complicated, one of them describes more than one flower and the other describes 3 different-colored parts of the flower. The first sentence created a similar image response for both models. The seconds sentence caused the baseline model to generate a blob that merges two of the described colors whereas the cycle consistent model generated a colorful blob, without relating it to the mentioned colors. The last two sentences are the most complex ones. Both models generate similar results, while the the baseline model's results are a bit more pleasing to the eye.



Baseline: this flower has petals that are purple and has dark lines on them. the petals are purple near the ovary. the stamen are
Cycle Consistent: this flower has petals that are purple, smooth and arranged in a disc like manner around green sepals. the petals are light green with

Baseline: this flower has petals that are white and has purple dots on them. the stamen is green. the petals are white and has green
Cycle Consistent: this flower is purple in color, with petals that are drooping downward and long. the stamen are light green. the petals are

Baseline: this flower has petals that are red and has yellow stamen. the petals are layered and are yellow in color. the stamen is
Cycle Consistent: this flower is orange with petals that are bunched together and has dark green stamen. the petals are bright red. the flower is

Baseline: this flower has petals that are white and has yellow stamen. the petals are yellow near the ovary. the stamen is green
Cycle Consistent: this flower is yellow in color, with petals that are curled and purple. petals are green. petals are needle shaped. sepals

Figure 7: Comparison between the results of the baseline and the cycle-consistent Image-to-Text models, for custom images on the Image-to-Text task. All the image were found online and are under creative commons license. The first three flowers (Gilboa Iris, Nazareth Iris and Mountain Tulip, respectively) are not part of the original dataset flower species. The last flower (Daffodil) is a specie present in the original dataset. The caption created by the baseline model are more accurate and do not include colors that don't appear in the input image. The cycle consistent model also creates good captions but in the case of the Nazareth Iris and the Daffodil, it mentions that the flower's color is purple which is not the case.