Object Transfiguration with CycleGAN: Resolving background distortions in images and videos

Divya Kothandaraman
Indian Institute of Technology, Madras
ramandivya27@yahoo.in

Abstract

Object transfiguration is a problem that has been widely explored in the past. Significant progress was made by CycleGAN which trains on a large number of unpaired examples to generate a mapping from one class to another. However, the method faces many problems like background distortion and hue changes. In images, objects can be localized by object detection. Foreground detection can do the same to moving objects in videos. In our proposed method, detection bounding boxes along with cycleGAN transformed images are used to generate the final results. Our method performs better than vanilla cycleGAN for images. For videos, the final transformation depends heavily on the robustness of the background subtraction algorithm.

1. Introduction

Image generation is an important problem in computer vision. Variational Autoencoders [5] use a probabilistic approach to generate images while autoregressive methods such as pixelRNNs [8] condition on previous pixels to generate new pixels. The most popular generative modelling approaches use an adversarial loss function in a type of neural network called the Generative Adversarial Networks [4]. Here, the network jointly learns a generator for synthesizing images and a discriminator classifying images as real or fake.

Many problems require us to transform a given image from one domain to another. For example, even semantic segmentation can be considered as a domain transfer problem where the task is to transform the image to a segmentation map. In [15], the authors propose an unsupervised method for style transfer. They present a network that learns to capture special characteristics of one image collection and translates these into the other image collection in the absence of any paired training examples. Their method performs well in a variety of scenarios like transforming a photograph to monet and vangogh style paintings, a zebra to a horse, an apple to an orange, day pictures to night pictures and vice versa. However, the method encounters failure in many object transfiguration cases. For example, when the image of a man riding a horse is transformed to a zebra, the generator paints the stripes of the zebra on not just the horse, but also the rider. In object transfiguration, we would ideally want nothing other than the object in concern to get transformed.

In the past, many researchers have worked on the problem of resolving background disfigurement. The key idea behind most such methods is to focus on the object in particular and then selectively transform. For image-to-image translation, we would want semantically similar regions to get mapped to each other followed by appropriate transformation. This idea is explored in [1] wherein the authors propose a method called 'SemGAN'. Here, cycleGAN is applied on semantic segmentation masks of an image rather than the image directly.

In this paper, we propose a simple method that applies object detection bounding boxes(for images)/background subtraction bounding boxes(fovideos) on cycleGAN transformed images to refine the outputs. The method doesn't require us to modify or re-train the cycleGAN network. Given a test image, the object in question is detected and the patch containing the object is transformed, leaving the background as such. Pre trained models like Faster RCNN, YOLO, SSD can be used for object detection. In case of videos, this needs to be done for every frame in the video. We consider the where the object that needs

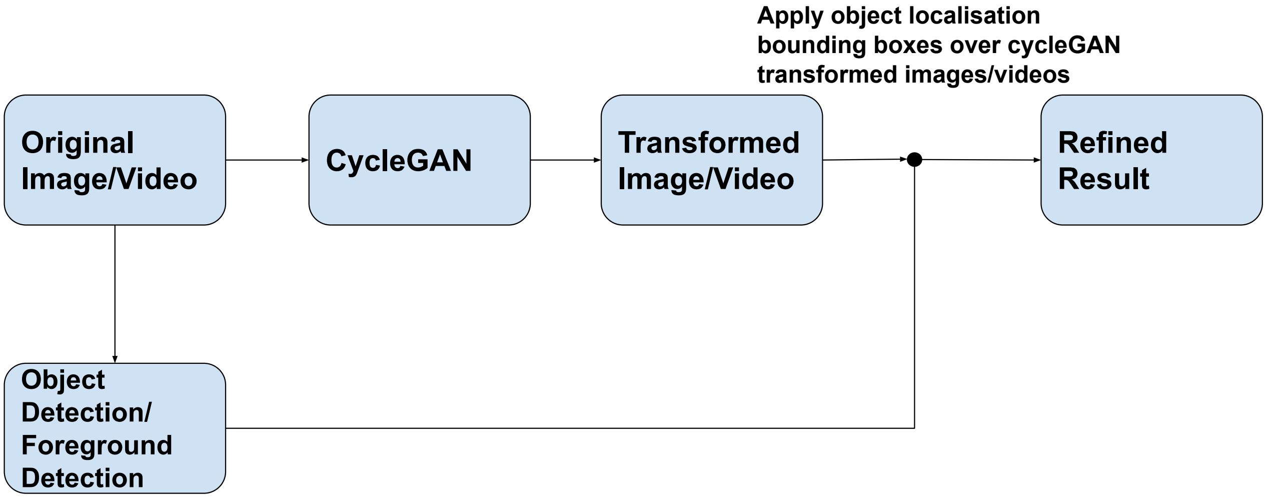

Apply object localisation bounding boxes over cycleGAN transformed images/videos
Original Image/Video
CycleGAN
Transformed Image/Video
Refined Result
Object Detection/ Foreground Detection

Figure 1. Our pipeline

to be transformed is the only moving entity in the video. This lets us segment out the object using a background subtraction algorithm. The main contributions of this paper are as follows:
(i) Object detection + cycleGAN pipeline to selectively transform objects from one domain to another in an image
(ii) Background subtraction + cycleGAN pipeline to selectively transform a moving object in a video

2. Related work

2.1. Style Transfer

[3] describes an artistic style transfer method to independently process and manipulate the content and the style of natural images. The addition of a photorealism regularization term [6] constrains the reconstructed image to be represented by locally affine color transformations of the input thus preventing distortions. Further, incorporating semantic labellings into the transfer procedure ensures that the transfer happens between semantically equivalent sub-regions. [15] describes an approach for image-to-image translation when paired training data is not available. The goal is to learn a mapping from one domain to another such that the distribution of images obtained by the transformation of images from the first domain to the second domain is indistinguishable from the distribution of images in the second domain. An adversarial loss with two discriminator losses are used. Forward cycle consistency and backward cycle consistency losses are applied.

2.2. Background Subtraction

Background subtraction/Foreground detection is a technique which allows a moving object in a video to be segmented out from its background. [11] uses Gaussian Mixture Models to solve this problem wherein each pixel is modelled as a weighted sum of Gaussians. [2] describes a probabilistic approach to video segmentation using Gaussian Mixture Models. Conditional Random Fields (CRF) have also been explored intensely for video segmentation. [10] proposes a discriminative model wherein appearance, shape and context information are incorporated efficiently to gives a classifier which can be applied to a large number of classes. [9] uses a heirarchial CRF model for the same problem. [13] deals with spatio temporal segmentation. By generating hypotheses of motion, coherent motion regions are identified iteratively. This classifies each location of the image to one of the hypotheses. The algorithm in efficient algorithm that [12] considers video segmentation and optical flow estimation simultaneously.

2.3. SemGAN and Attention methods

[1] describes an end to end neural network training mechanism for semantic segmentation followed by style transfer. Since invertibility does not necessarily enforce semantic correctness, a semantic loss term is applied. Pre-trained weights are used to initialize the semantic segmentation network. The paper performs better than cycleGAN in multi modal scenarios. Failure cases (incorrect mapping, for example, tree being mapped to the sky) are addressed better than cycleGAN but results

of other object transfiguration cases(unimodal cases) are not as good as cycleGAN results. [7], [14] propose architectures that have an attention layer in their model. Thus, the object in question is localized with the help of attention after which transformation is done.

3. Our Method

3.1. Images

We propose a pipeline wherein the first step is to compute the bounding boxes of the object in question. This can be done using an efficient object detection algorithm such as faster RCNN or YOLO. After this, one approach would be to crop off the object and create a new dataset containing the cropped images which can be used for training. While testing, the transformed cropped object needs to be re-sized and stitched back with the background to yield the entire transformed image. However, given the vastness of training set, we find it redundant to re-train cycleGAN on cropped images. We instead apply the trained cycleGAN model(trained on the original dataset) on the test images, perform object detection to detect the object in question and replace the pixels outside the bounding box with the corresponding pixels in the original image.

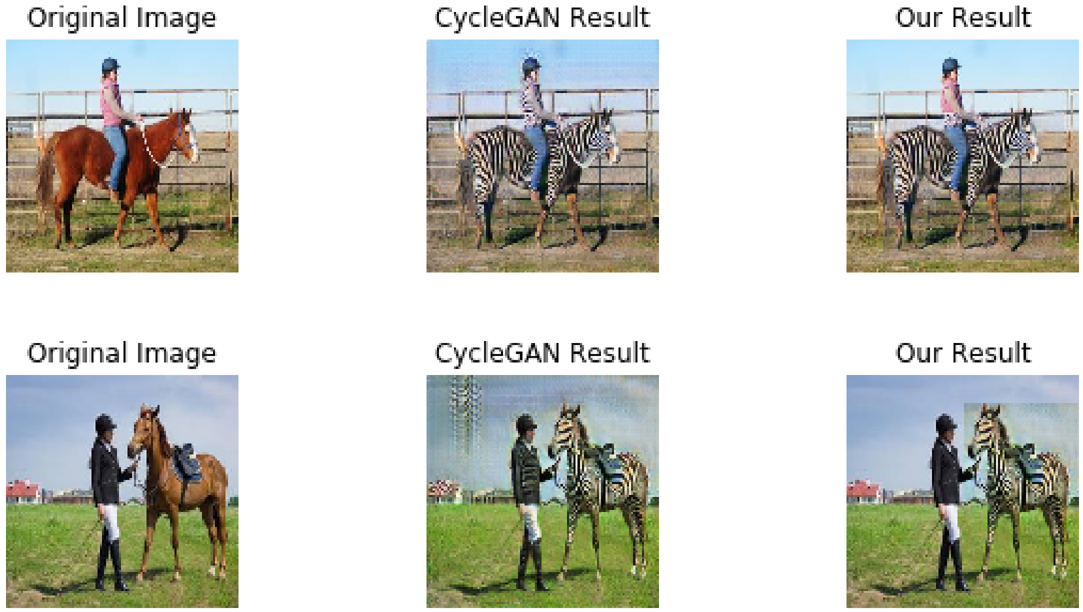
3.2. Videos

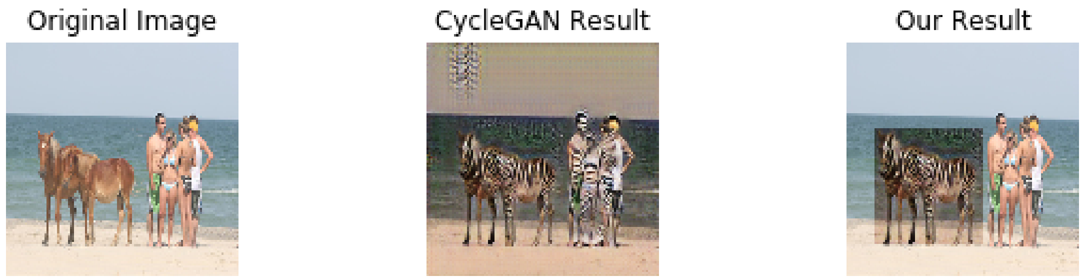
In case of videos, the object in question needs to be detected in every frame. Training a neural network, even for object detection, is a costly process. The number of frames per video can be really large and the number of distinct objects per video can also vary. We propose a method based on video segmentation for object style transfer in videos with diverse backgrounds.For this, we require the object in question to be the only moving entity in the video. Moving objects can be detected by background subtraction. The trained cycleGAN model is first applied on every frame of the video. We then replace the background pixels with the corresponding pixels in the original video. This gives us a video wherein only the moving object is transformed.

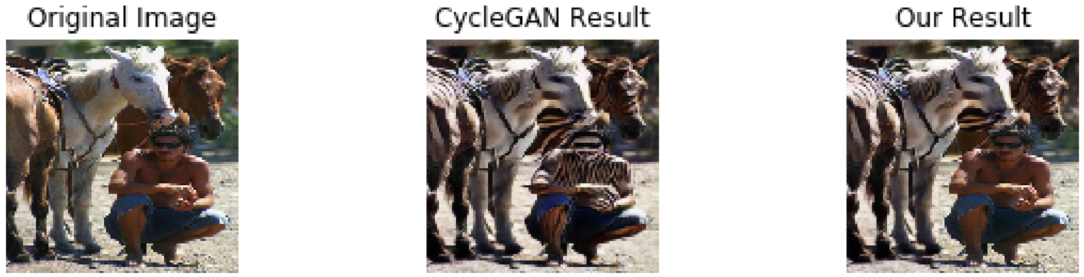
4. Results

4.1. Images

Results on a few images are shown below.

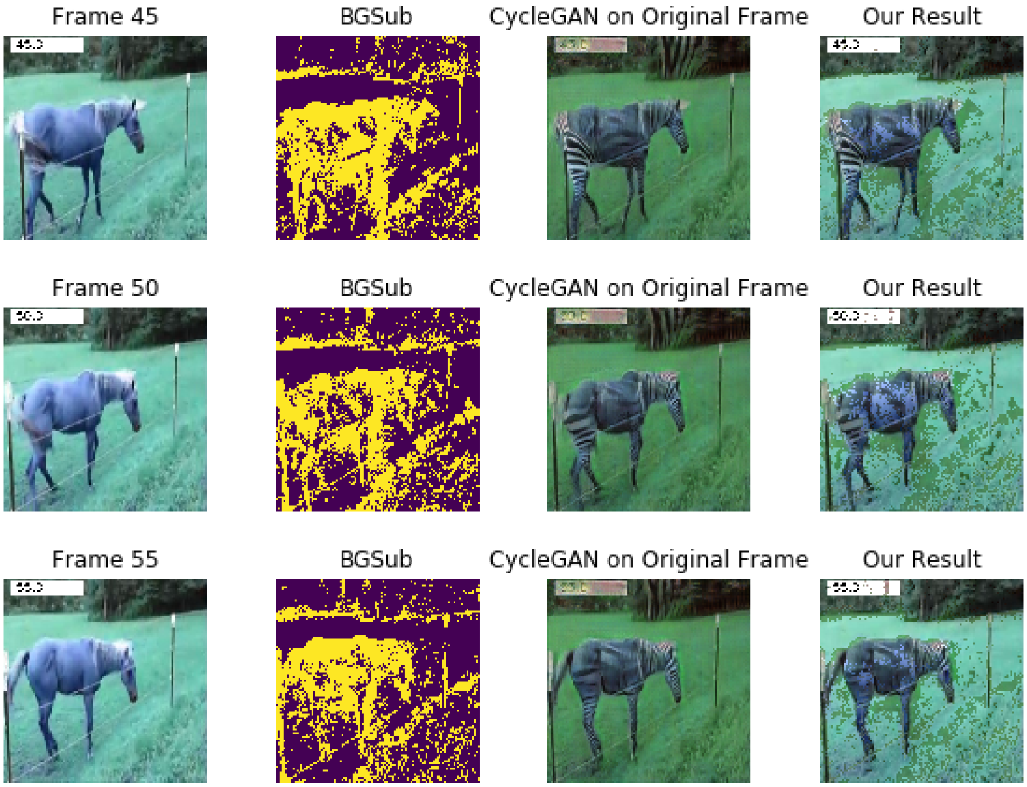


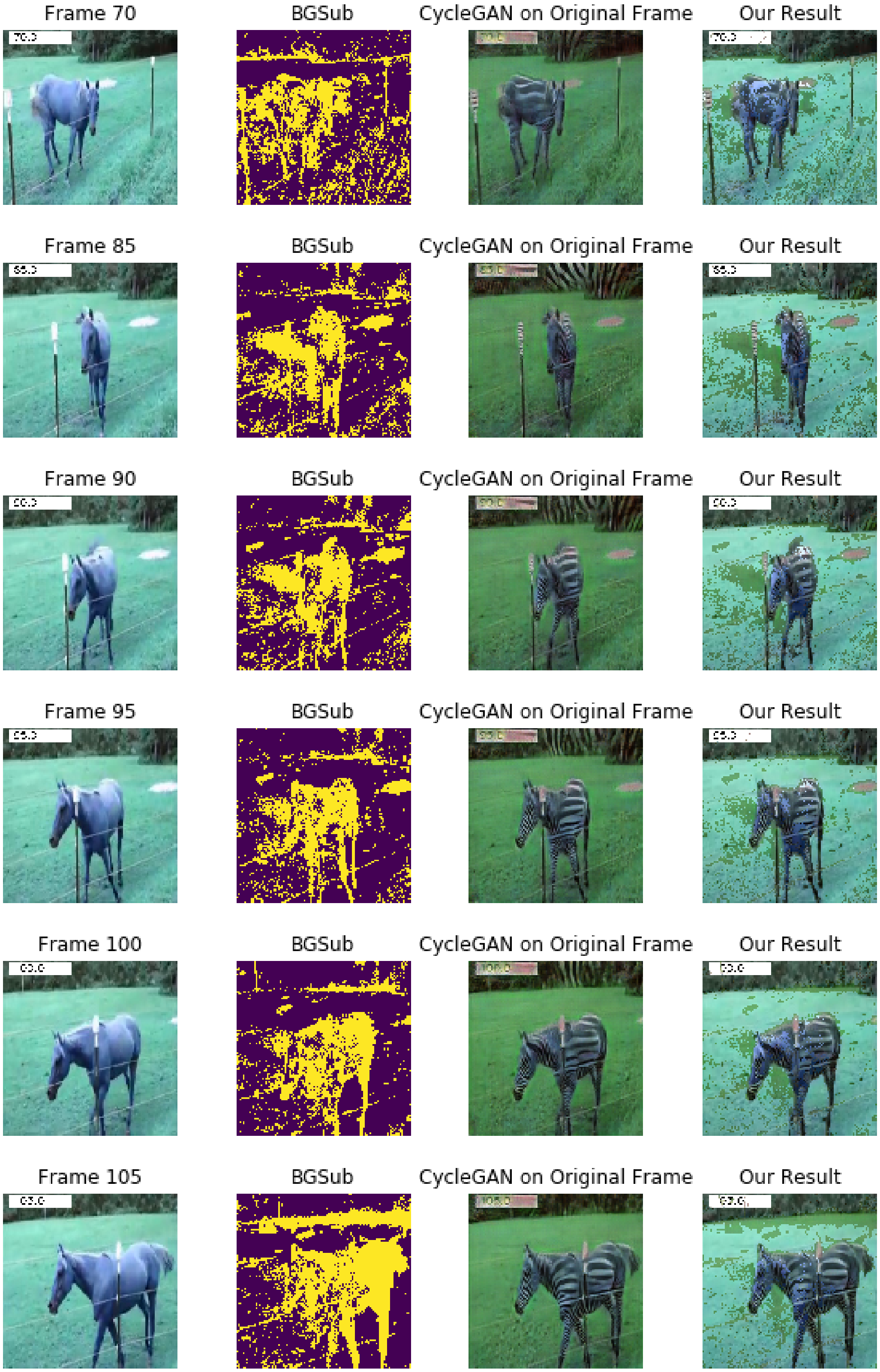


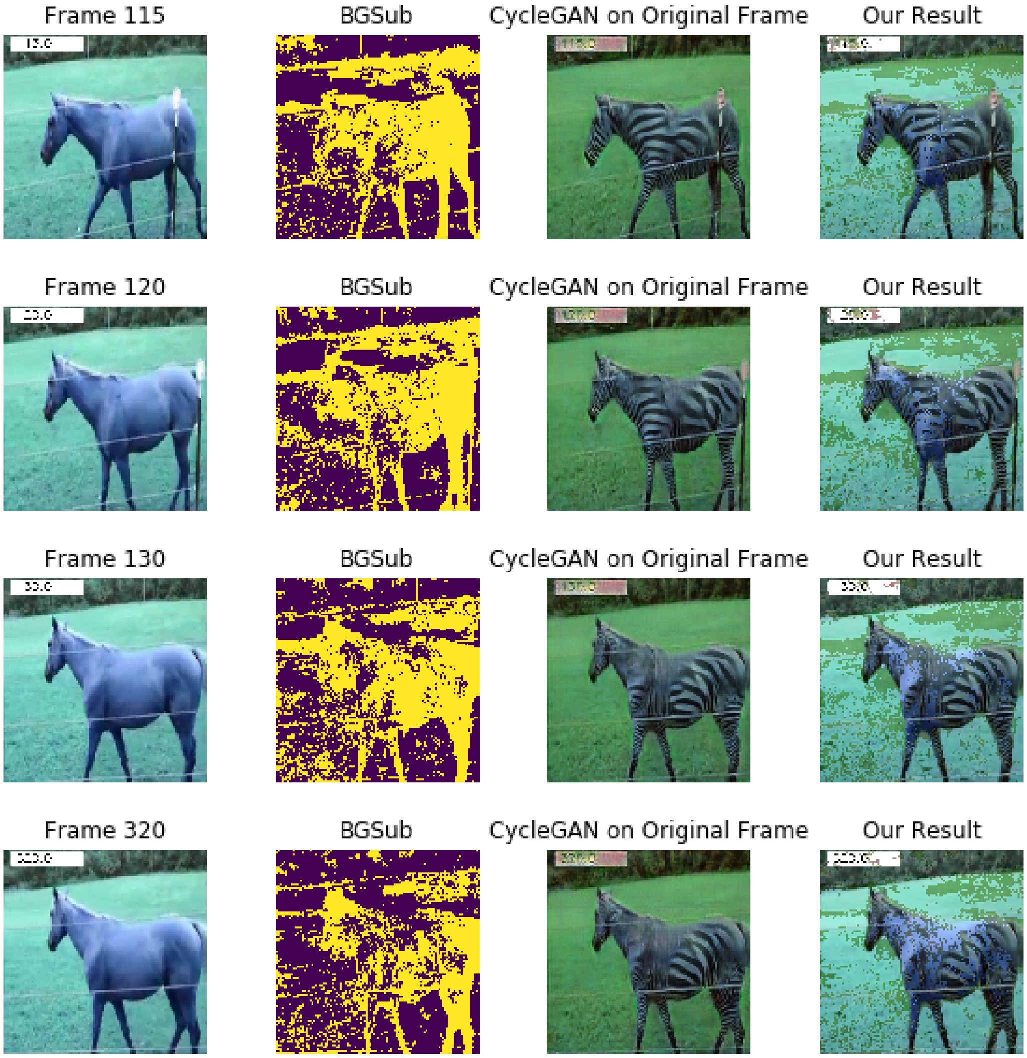


4.2. Video

The algorithm has been applied on a horse to zebra transformation video. For convenience, only a few frames are displayed.


Frame 45
BGSub
CycleGAN on Original Frame
Our Result
Frame 50
BGSub
CycleGAN on Original Frame
Our Result
Frame 55
BGSub
CycleGAN on Original Frame
Our Result


Frame 70
BGSub
CycleGAN on Original Frame
Our Result
Frame 85
BGSub
CycleGAN on Original Frame
Our Result
Frame 90
BGSub
CycleGAN on Original Frame
Our Result
Frame 95
BGSub
CycleGAN on Original Frame
Our Result
Frame 100
BGSub
CycleGAN on Original Frame
Our Result
Frame 105
BGSub
CycleGAN on Original Frame
Our Result


Frame 115
BGSub
CycleGAN on Original Frame
Our Result
Frame 120
BGSub
CycleGAN on Original Frame
Our Result
Frame 130
BGSub
CycleGAN on Original Frame
Our Result
Frame 320
BGSub
CycleGAN on Original Frame
Our Result

5. Conclusion

The main goal of this paper was to overcome background distortions that occur when style transfer methods like cycleGAN are applied directly on the entire image. We developed a pipeline that uses object localization results along with cycleGAN transformed images to generate final results where only the object of interest is transformed. Our method for images produces really good results. However, for videos, the results depend heavily on the robustness of the foreground detection algorithm.

References

[1] A. Cherian and A. Sullivan. Sem-gan: Semantically-consistent image-to-image translation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1797–1806. IEEE, 2019. 1, 2

[2] N. Friedman and S. Russell. Image segmentation in video sequences: A probabilistic approach. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pages 175–181. Morgan Kaufmann Publishers Inc., 1997. 2

[3] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1

[5] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[6] F. Luan, S. Paris, E. Shechtman, and K. Bala. Deep photo style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4990–4998, 2017. 2

[7] Y. A. Mejjati, C. Richardt, J. Tompkin, D. Cosker, and K. I. Kim. Unsupervised attention-guided image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 3693–3703, 2018. 3

[8] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016. 1

[9] C. Russell, P. Kohli, P. H. Torr, et al. Associative hierarchical crfs for object class image segmentation. In *2009 IEEE 12th International Conference on Computer Vision*, pages 739–746. IEEE, 2009. 2

[10] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European conference on computer vision*, pages 1–15. Springer, 2006. 2

[11] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 2, pages 246–252. IEEE, 1999. 2

[12] Y.-H. Tsai, M.-H. Yang, and M. J. Black. Video segmentation via object flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3899–3908, 2016. 2

[13] J. Wang and E. H. Adelson. Spatio-temporal segmentation of video data. In *Image and Video Processing II*, volume 2182, pages 120–132. International Society for Optics and Photonics, 1994. 2

[14] C. Yang, T. Kim, R. Wang, H. Peng, and C.-C. J. Kuo. Show, attend and translate: Unsupervised image translation with self-regularization and attention. *IEEE Transactions on Image Processing*, 2019. 3

[15] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1, 2