Reinforcement Learning based Spatial Transformer GAN for Realisitc Multi-Object Image Compositing

Gaurav Mittal
Robotics Institute
Carnegie Mellon University
gauravm@andrew.cmu.edu

Tanya Marwah *
Robotics Institute
Carnegie Mellon University
tmarwah@andrew.cmu.edu

Abstract

Generating realistic image composites by placing geometrically correct instance of the foreground onto a background image is an area of active research. Though it is possible to do so for a single foreground object [7], doing it sequentially for multiple objects is not trivial since every addition of object drastically changes the image which can cause the model to fail due to radical change in the underlying distribution. To this end, we propose a generative adversarial approach to multi-object compositing based on [7] where we introduce multiple discriminators to handle the different distributions. We use reinforcement learning to train an agent which treats the discriminators as bandits and learns to choose the right discriminator to train the generator. The generator learns to iteratively predict the parameters of a spatial transformer to warp the foreground object onto the background. This ability allows the generator to generalize over different stages of multi-object compositing. We demonstrate our approach on CLEVR dataset and show that our approach is giving promising results and has the potential to generalize over more complicated visual scenes.

1. Introduction

Image compositing, in general, refers to fusing two images together into a single image. In this report, we consider the task of image compositing as overlaying a masked foreground object onto a background image after appropriately warping the foreground object such that it matches the geometry of the overall visual scene. Image compositing is important in the computer vision paradigm especially because it can be used to tackle the shortcomings of direct image generation. Images are very high-dimensional data entities and direct image generation methods such as Generative Adversarial Networks (GANs) [3] are limited by finite network capacity making them work only on restricted domains. Image compositing can be leveraged by allowing direct generation methods to continue working on these restricted domains while compositing can fuse them together to generate the overall realistically-looking visual scene.

Recently, [7] proposed a GAN based approach called ST-GAN which leverages Spatial Transformer Networks (STNs) [5] by having the generator predict the parameters to realistically warp the foreground object onto the background. Their method is the state-of-the-art to address the problem of realistic image generation through geometric transformations in a GAN framework. One of the limitations of their approach that it cannot be trivially extended to sequentially composite multiple objects in the visual scene. This is because with every addition of object to the scene results in a drastic change in the data distribution. Moreover, since the discriminator works by distinguishing between geometrically correct/incorrect images, once an object is placed with correct geometry over the background, the image will start matching the real distribution with no incentive left for the model to correctly place the next object.

To deal with this limitation, we introduce a novel architecture where we have multiple discriminators to handle the different data distributions arising from having different number of objects in the scene. We also employ Reinforcement Learning to train an agent that treats the multiple discriminators as bandits and choose the appropriate discriminator to train the generator. To the best of our knowledge, this is the first attempt to model the distributions separately and use RL in conjunction with GAN for multi-object compositing. The idea behind doing this is that when the generator generates an image composite with a certain number of objects in the image, *the extent to which the discriminators rate the image as fake is relative*. This relative fakeness can be used as the reward to determine the appropriate discriminator (bandit) to train the generator.

*contributed to the project but not enrolled in the course 16-824

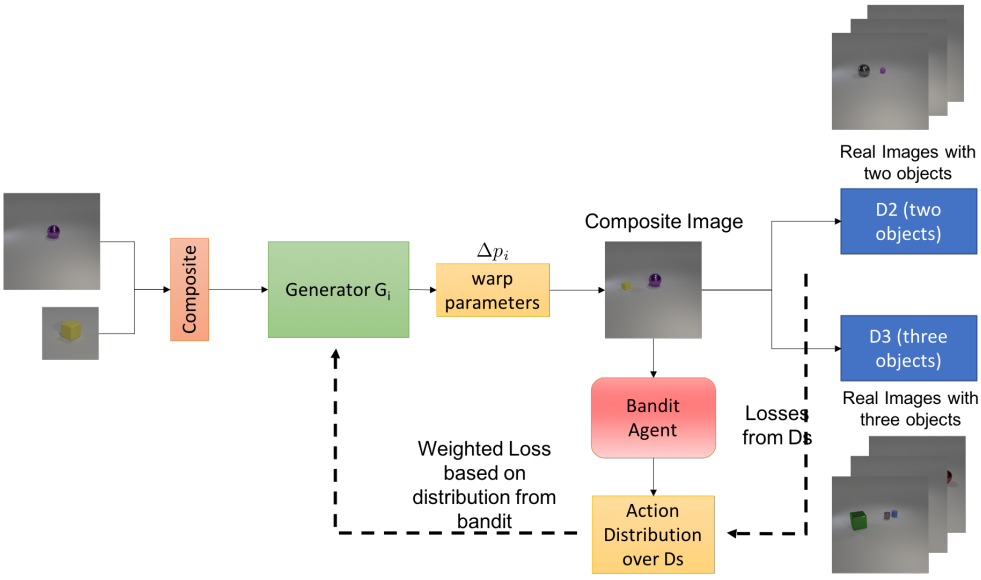

Figure 1. High-level model architecture of RL based ST-GAN. Here we are trying to place two objects on the background image (leading to two discriminators).

2. Related Work

Previously image compositing has been attempted through Possion blending [8] and deep learning approaches [10, 13]. Further, Spatial Transformer Networks (STNs) [5] are one way to incorporate learnable image warping withing a deep learning framework.

Generative Adversarial Networks (GANs) [3] are a class of generative models that are learned by playing a minimax optimization game between a generator network G and a discriminator network D. GANs are utilized for data generation in various domains, including images [9], videos [11] and 3D voxel data [12]. ST-GAN [7] combines GANs with STNs to generate realistic composites of single object.

3. Methodology

Let K be the number of objects to be composited onto an image. At some point k, given a background image I_{BG_k} and foreground object I_{FG_k} with a corresponding mask M_{FG_k}, the compositing process can be expressed as,

$$\begin{aligned} I_{comp_k} &= I_{FG_k} \odot M_{FG_k} + I_{BG_k} \odot (1 - M_{FG_k}) \\ &= I_{FG_k} \oplus I_{BG_k} \end{aligned}$$

Given the composite parameters p_{0_k} (as the initial warp state), we can rewrite the above as,

$$I_{comp_k}(p_{0_k}) = I_{FG_k}(p_{0_k}) \oplus I_{BG_k}$$

In our work, we restrict the geometric warp function to homography transformations making the assumption that the perspective of the foreground object is close to the correct perspective. Additionally, predict large displacement parameters directly from image pixels is extremely challenging [4, 2]. So the geometric transformation is predicted iteratively using iterative STNs to predict a series of warp updates. At the i^{th} iteration of compositing an object k, given image I and previous warp parameters p_{i-1_k}, the correcting warp update Δp_{i_k} and the new warp update p_i can be written as,

$$\begin{aligned} \Delta p_{i_k} &= G_{i_k}(I_{FG_k}(p_{i-1_k}), I_{BG_k}) \\ p_{i_k} &= p_{i-1_k} \circ \Delta p_{i_k} \end{aligned}$$

This preserves the original image from loss of information due to multiple warping operations. Figure 1 shows the high level overview of the entire architecture of RL-based ST-GAN.

Sequential Generators: The STNs are integrated into the GAN framework to allow them to learn geometric warps that map images closer to the natural image manifold. In ST-GAN, the generator G generates a set of low-dimension warp parameter updates (instead of images) and the discriminator D gets as input the warped foreground image composited with the background. To learn gradual geometric improvements towards the natural image manifold, a sequential adversarial training strategy is adopted where a stack of generators is trained G_i one after the other by fixing the weights of all previous generators trained so far.

Multiple Discriminators: The key limitation of ST-GAN is that once an object is placed with correct geometry in the

scene, the image begins matching the real distribution with no incetive left for the model to correctly place the next object. To overcome this restriction, we introduce multiple discriminators such that each discriminator is trained to model is different real image data distribution. This way each discriminator is trained in a one-versus-all fashion where all composite images are considered fake while only real images with a certain number of objects are considered real. The intuition behind this setup is that whenever the discriminators are fed a composite image from the generator, the discriminators will rate the image as fake relatively. *For instance, if the composite image has 4 objects, the discriminator corresponding to real images will 4 objects is least likely to rate the image as fake in comparison to all other discriminators.*

Multi-bandit agent: Each discriminator models a disjoint portion of the real distribution. Therefore, given the discriminators, our model needs to decide which one to choose to train the generator for any given composite image. We model this problem as a Reinforcement Learning task where we treat the discriminators as bandits, each giving some reward for a given composite image. We train a bandit agent whose task is choose (turn) the appropriate discriminator (bandit) that is least likely to rate the composite as fake. In other words, the bandit agent is choosing the region of the image manifold subjective to the composite image where it could be best matched to.

Objective: We follow a modular approach towards training our model. We first train the discriminators, then use them to train the bandit agent and then train the generators with them while fine-tuning the discriminators and bandit agent end-to-end. To pre-train a discriminator, we feed it real images y_k with only a certain number of objects and fake images as the composites $x(p_{i_k})$ generated by an untrained generator. The loss function is defined as,

$$L_D = -\mathbb{E}\left[\log(D(y_k) + \log(1 - D(x(p_{i_k}))\right]$$

Using the trained discriminators, the agent is trained by receiving a $+1$ reward if the predicted action corresponds to the discriminator with the least likelihood of fakeness for the composite image and receive -1 otherwise. The loss function for the agent is,

$$L_{agent} = -\mathbb{E}[\log(W_{action\_taken}).Reward]$$

After the discriminators and agent have been trained, the entire model is trained by optimizing the Wasserstein GAN (WGAN) [1] objective. Based on the agent's action, we choose the appropriate discriminator D and update the i^{th} generator G_i with D alternating the respective loss functions,

$$L_D = \mathbb{E}_{x,p_i}\left[D(x(p_i))\right] - \mathbb{E}_y\left[D(y)\right] + \lambda_{grad}.L_{grad}$$
$$L_{G_i} = -\mathbb{E}_{x,p_i}\left[D(x(p_i))\right] + \lambda_{update}.L_{update}$$

where we add penalty to gradients and warp updates for regularization as in [7].

4. Experiments and Results

To demonstrate the soundness of our approach, we experiment with compositing 2 objects onto an image. Therefore, our model has two discriminators.

Model Architecture: The architecture of the generator G is C(32)-C(64)-C(128)-C(256)-C(512)-L(1024)-L(6) where output is 6-dimensional for affine (8-dimensional for homography). The generator receives an input of 7 channels: RGBA for foreground and RGB for background. The architecture of each discriminator D is C(32)-C(64)-C(128)-C(256)-C(512)-L(1024)-L(1) with LeakyReLU as non-linear activation. The agent A has architecture similar to discriminator C(32)-C(64)-C(128)-C(256)-C(512)-L(1024)-L(2) where the output is 2-dimensional representing the weights for taking the corresponding action (that is choosing the discriminator).

Dataset: We wish to use a dataset that gave us flexibility in the number of objects that we can have in the scene and at the same time, follow some geometry making it look like a natural 3D scene. So we chose CLEVR dataset [6] for this purpose. The dataset consists of photo-realistic rendered images with objects of different shapes, colors, materials, size and possible occlusions. We used Blendr to generate scenes with iteratively increasing number of objects. Along with each generated scene, we also generated the corresponding foreground object and its mask that was added to the scene. We, in all, experimented with around 10,000 such images equally distributed between two object and three object images.

Task: The task we demonstrate our model on is to place two foreground objects on a given background image. So the unified task of the generator Gi's is to generate the geometrically-correct warp parameters for both the objects. The task of discriminator D_2 is to discriminate between real images having two objects from the composite images generated based on warp predicted. Similarly the task of discriminator D_3 will be for real images having three objects. The agent needs to choose the right discriminator given the composite image which can have either two or three objects.

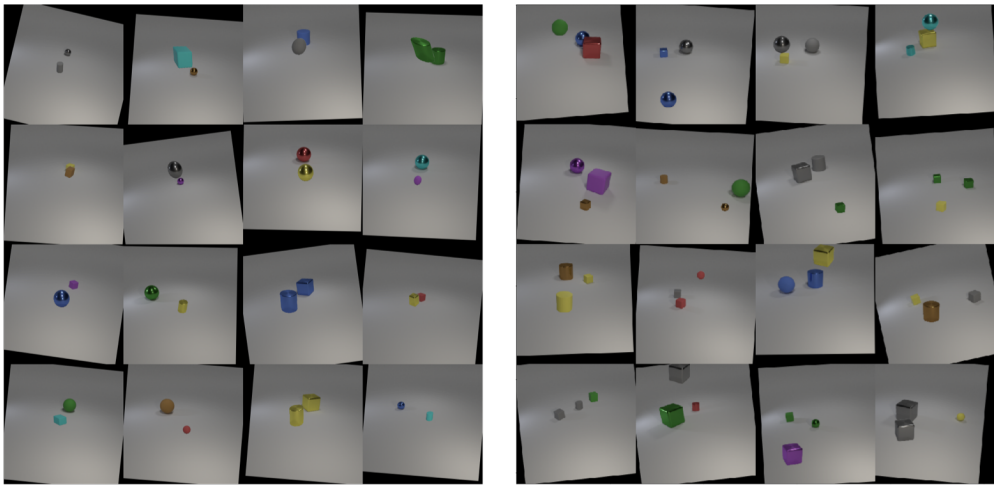

Figure 2. Single-object compositing on different types of background images. (a) shows compositing of a single object on a background with only one object. (b) shows compositing of a single object on a background with two objects.

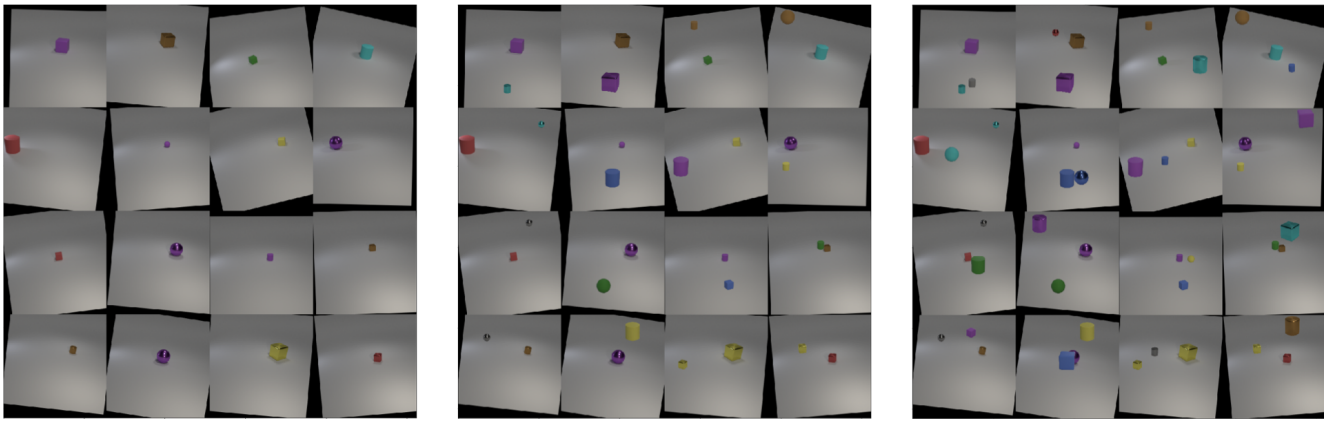

Figure 3. Multi-object Compositing using RL-based ST-GAN. (a) shows the original background image. (b) shows background image composited with one object. (c) shows background image (composited with one object) composited with a second object.

Image Type	D_2	D_3
Two Objects (Real)	0.68 ± 0.18	0.22 ± 0.23
Three Objects (Real)	0.10 ± 0.17	0.90 ± 0.09
Two Objects (Comp)	0.53 ± 0.23	0.24 ± 0.29
Three Objects (Comp)	0.11 ± 0.19	0.83 ± 0.18

Table 1. Probability Statistics of classifying an image as real by the two discriminators (D_2 and D_3). Here Comp refers to composite image (fake).

Discriminator: Table 1 shows the statistics for the probability scores returned by the two discriminators D_2 and D_3 for different kinds of images fed into them. We can clearly observe that there is a significant difference in the probability scores predicted by the discriminators. For instance, for real images with three objects, D_2 that is trained to model the distribution of images with two objects gives a very low probablity on an average while the probability score is very high for D_3 that has been modeled for images with three objects. The interesting thing to watch out here is the scores for composite images. We can see that even in case of fake images, one discriminator considers one type of composite images less fake than the other image. This clearly proves our hypothesis that given the discriminators are modeled for different image distributions, there exists a *relative measure of fakeness* where one of the discriminators is least likely to call a composite image fake amongst all other discriminators.

Bandit Agent Given that the *relative measure of fakeness* of the discriminators is established, the agent is trained

Image Type	Average Reward [-1, 1]
Two Objects (Real)	0.52
Three Objects (Real)	0.64
Two Objects (Comp)	0.41
Three Objects (Comp)	0.49

Table 2. Average reward accumulated by the bandit agent for different types for images

using the probability scores from the discriminators. The agent receives as reward of $+1$ if the action (or the discriminator) it choose gives a better probability score and receive -1 otherwise. Table 2 summarizes the average reward accumulated by the bandit agent for different types of images. We can observe that the reward is considerably high for all kinds of images. The reward is relatively low for composite images compared to real images which was expected since due to arbitrary warping, it could lead to occlusion and other defects in the image confusing the discriminators/action to give incorrect scores/actions.

Generator As the last stage of the training, we train the generators G_i's with the pre-trained discriminators and agent to predict the warp parameters for image compositing. Due to constraints of time and compute resources, we couldn't allow our model to run for a long time. So we instead posed restrictions on the warp parameters and perturbation in the background (in comparison to ST-GAN). Figure 2 shows compositing of a single object on background images consisting of one object and two objects. We can notice the composited foreground objects as those for which the shadow and reflection effects are absent. Since here we are only dealing with geometric corrections, those effects will be omitted from the foreground object. Figure 3 shows multi-object compositing where on the same background object, multiple foreground objects are composited one after the other. We can clearly observe that our model is able to generate decently realistic looking image composites for a highly varying background (as we go from one object composite to another).

5. Discussion and Conclusion

In this report, we introduce a novel approach that integrated spatial transformer based generative adversarial networks with Reinforcement Leraning to learn to generate realistic multi-object image composites. We established that if we have a generative adversarial network with multiple discriminators such that each discriminator models a different portion of the data distribution, there will exist a relative measure of fakeness assigned by the discriminators to a given fake (generated) image. We leveraged this difference in fakeness to train an RL bandit agent which chooses an appropriate discriminator (which is least likely to rate the image as fake) to train the generator. Through various experiments we conducted, we found the approach considerably efficient in generating multi-object image composites. We believe that our approach holds potential to be extended to an even greater number of discriminators and even to scenarios where image composites are generated using real-world objects.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision*, 56(3):221–255, 2004.

[3] 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.

[4] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[5] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[6] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1988–1997. IEEE, 2017.

[7] C.-H. Lin, E. Yumer, O. Wang, E. Shechtman, and S. Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. *arXiv preprint arXiv:1803.01837*, 2018.

[8] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Transactions on graphics (TOG)*, 22(3):313–318, 2003.

[9] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[10] Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, and M.-H. Yang. Deep image harmonization.

[11] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621, 2016.

[12] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, pages 82–90, 2016.

[13] J.-Y. Zhu, P. Krahenbuhl, E. Shechtman, and A. A. Efros. Learning a discriminative model for the perception of realism in composite images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3943–3951, 2015.