Toward Realistic Image Compositing with Adversarial Learning

Bor-Chun Chen*
University of Maryland
College Park
sirius@umd.edu

Andrew Kae
Verizon Media Group
andrewkae@verizonmedia.com

Abstract

Compositing a realistic image is a challenging task and usually requires considerable human supervision using professional image editing software. In this work we propose a generative adversarial network (GAN) architecture for automatic image compositing. The proposed model consists of four sub-networks: a transformation network that improves the geometric and color consistency of the composite image, a refinement network that polishes the boundary of the composite image, and a pair of discriminator network and a segmentation network for adversarial learning. Experimental results on both synthesized images and real images show that our model, Geometrically and Color Consistent GANs (GCC-GANs), can automatically generate realistic composite images compared to several state-of-the-art methods, and does not require any manual effort.

1. Introduction

Image compositing aims to create a realistic-looking image by taking the foreground object of one image and combining it with the background from another image (see Figure 1). In order to make the composite image look realistic, many factors need to be considered, such as scene geometry, object appearance, and semantic layout. It is a challenging task and usually requires a human expert carefully adjusting details including geometry and color using professional image editing software such as PhotoShop [1] to create a single composition.

Many previous works [4, 27, 18, 14, 32, 19, 35, 30] try to alleviate this manual burden by creating algorithms that can automatically adjust the appearance of the foreground image and make it fit into the background naturally. While this may work in some cases, many of these approaches still require human supervision to help with tasks such as determining the appropriate location and size of the foreground object or capturing the lighting conditions of the scene.

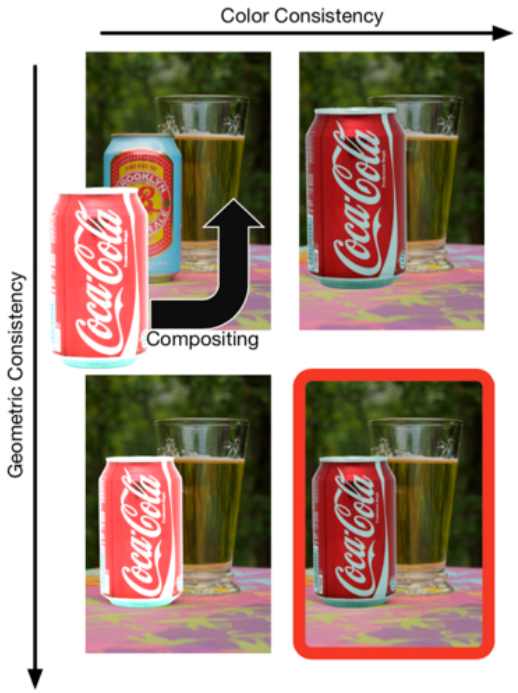

Figure 1. The goal of image composition is to create a realistic image by combining a foreground object with a background image. The x-axis corresponds to increasing color consistency in the composite image, while the y-axis corresponds to increasing geometric consistency. However, the composite image only looks realistic when both geometry and color consistency are considered (i.e. image in the red box). (Best viewed in color)

Recently generative adversarial networks (GANs) have been shown to have the ability to generate realistic looking images [7, 11, 3, 33, 31, 36, 5, 28] by learning to deceive an adversarially trained discriminator network. However, image composition is a different task from image generation because the composite image must maintain details from the input images and apply only slight changes to improve the realism of the composition. Recent work [21] modified the GAN framework by restricting the range of the generator to a geometric manifold using a spatial transformer net-

* Work done during internship at Verizon Media Group.

work [12] in order to generate realistic composite images that are geometrically consistent. However, such a model only works if the foreground appearance is already consistent with the background image. If the domain of the foreground and the background images are different, geometric transformation alone does not have the ability to generate a natural-looking composite image. As shown in Figure 1, for a composite image to be realistic, the model needs to account for both geometric and color consistency. However, it is not trivial to combine previous works to automatically adjust both color and geometry since these two properties are interdependent: geometric correction relies on color consistency while color correction also relies on geometric consistency.

To address the above issue, we propose a novel GAN architecture called Geometrically and Color Consistent GAN (GCC-GAN) for image compositing that simultaneously learns both geometric and color correction with adversarial learning. GCC-GAN contains four sub-networks: a transformation network, a refinement network, a discriminator network, and a segmentation network. The transformation and refinement networks act together as the generative compositing model, which aim to generate a realistic composite image while considering geometric, color, and boundary consistency. At the same time, the discriminator and segmentation networks help to increase the realism of the composite image through adversarial learning. In particular, the discriminator network learns to separate composite images from real images while the segmentation network learns to separate the foreground object from the background in the composite images. GCC-GANs are trained end-to-end with a geometric loss, an appearance loss, an adversarial loss, and an adversarial segmentation loss. Unlike previous works that restrict the generator to geometric transformations, our model can apply both geometric and color correction as well as boundary refinement to generate a composite image. Experimental results show that our model can generate geometrically and color consistent images in both synthetic and real-world datasets.

The contributions of this paper include: (1) demonstrating the need for both geometric and color consistency for the image compositing task, (2) proposing a novel end-to-end model that creates realistic composite images based on the generative adversarial network framework, and (3) extensive evaluations including human perception experiments showing the ability of the proposed model to generate realistic composite images compared to different state-of-the-art methods.

2. Related Work

Image Compositing models combine a foreground image with a background image seamlessly. Many prior works focus on how to modify the appearance of the foreground image to better fit into the background based on color gradients [4, 27] or color statistics [18, 14, 32]. Agarwala *et al*. [2] provide a system to combine multiple source images taken in the same scene with the help of user inputs. Lalonde *et al*. [19] develop an interactive system to create composite images by selecting foreground objects from a large database. With the advancement of deep learning research in computer vision, various deep learning models [35, 30, 21, 29] were also introduced for image compositing. Similar to our approach, Zhu *et al*. [35] use a discriminative model to estimate the realism of composite images. However, their discriminative model is fixed during the image compositing process and cannot be improved for better composition. Tsai *et al*. [30] introduced an end-to-end encoder-decoder network for image harmonization. Although these methods can generate realistic compositions, they still rely on a human for tasks such as deciding the location and size of the foreground objects. Most recently, Tan *et al*. [29] propose to use deep neural networks to learn the location and size of the foreground object for compositing human into background image; Lin *et al*. [21] use generative adversarial networks (GANs) with a spatial transformer network [12] to learn the correct geometry transformation of the foreground object. These works consider geometric consistency in image compositing, but they can only work when the domain of the foreground and background images are similar. Our work extends previous works by providing a unified end-to-end framework that learns to adjust both the geometry and appearance consistently, which allows our model to automatically composite images from different sources.

3D Synthesis There have been many works that combine synthetic 3D objects with images [6, 15, 16, 9, 8]. However, these methods require explicitly reconstructing the scene geometry and environment illumination in order to render the 3D object. On the other hand, our model can directly take the rendered object as input for composition.

Generative adversarial networks [7] have been utilized for many different image generation tasks [23, 3, 26, 11, 36, 10, 33, 20]. Conditional GANs [23] provide a way to generate images from different classes given different inputs. Isola *et al*. [11] provide a framework that translates an image from one domain to another, given pairs of training images. Zhu *et al*. [36] further extend the framework to work over unpaired training images using cycle consistency. However, these frameworks cannot be directly applied to the image composition task since the composed images need to retain the fine details of both the foreground and background images in a consistent manner. Instead of direct image generation, our model utilizes the adversarial learning process to learn geometric and color corrections for realistic composition.

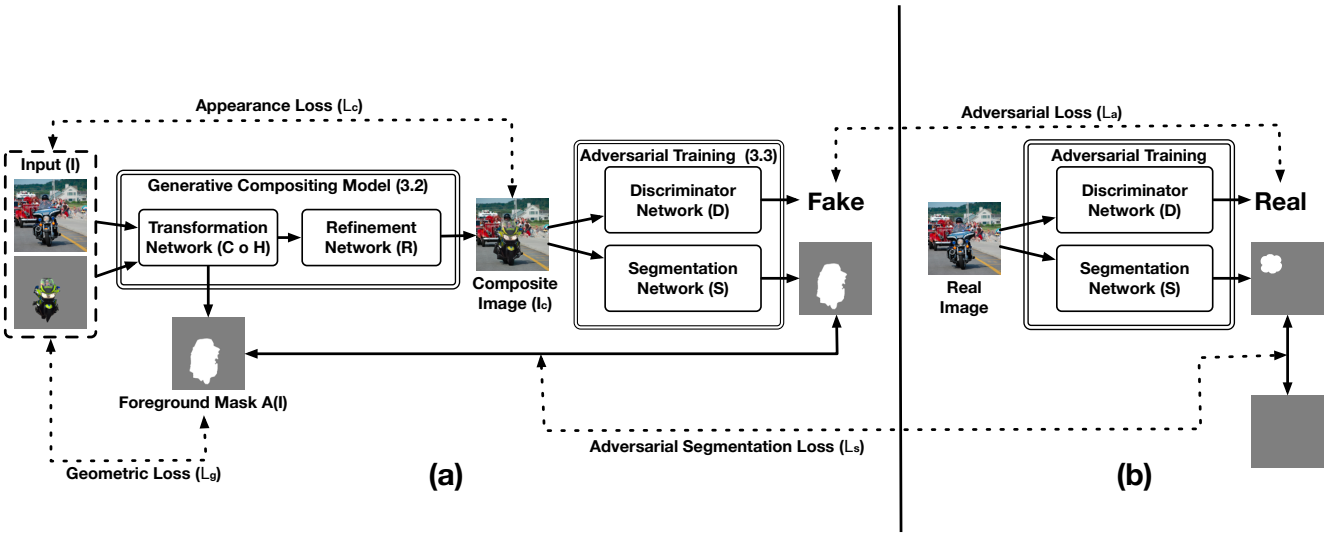

Figure 2. **Overview of the proposed network architecture.** (a) Given an input triplet consisting of a foreground object, a foreground mask, and a background image, the generative compositing model (consisting of transformation and refinement networks) learns to create a realistic composite image, in order to fool both the discriminator network and the segmentation network. (b) Given an image, the discriminator network learns to predict real vs fake while the segmentation network learns to segment the foreground from background.

3. Proposed Method

3.1. System Overview

Figure 2 shows an overview of the proposed network architecture. The model consists of four sub-networks: a transformation network, a refinement network, a discriminator network, and a segmentation network. The transformation network and refinement network act together as the generative compositing model and is described in Section 3.2. The discriminator network and the segmentation network improve the generative model through adversarial learning and is described in Section 3.3. Given an input triplet consisting of a background image, a foreground image, and an object mask, the compositing model learns to composite realistic images while the discriminator network learns to distinguish composite images from real images. In addition, the segmentation network tries to separate the foreground object from the background in the composite image. The model is trained to optimize the min-max objective function described in section 3.4.

3.2. Generative Compositing Model

Given a foreground image with N pixels $I_f \in [0,1]^{N\times 3}$ with a foreground mask $\alpha \in \{0,1\}^N$ and a background image $I_b \in [0,1]^{N\times 3}$ as inputs $I = \{I_f, I_b, \alpha\}$, the process of image compositing can be formulated as follows:

$$\begin{aligned} I_c &= G(I;\theta_G) \\ &= A(I) \circ F(I) + (1 - A(I)) \circ I_b, \end{aligned} \tag{1}$$

where $\circ$ is the Hadamard product, G is the compositing model which combines the foreground region of I_f indicated by the mask α and the background image I_b; θ_G is the model parameters. $F(I) \in [0,1]^{N\times 3}$ is the transformed foreground and $A(I) \in [0,1]^N$ is the alpha mask. Under this formulation, a simple alpha composition model can then be described as identity functions: $A(I) = \alpha; F(I) = I_f$.

If only the geometric correction is considered as in [21], the model becomes:

$$A(I) = H(\alpha, T_h(I;\theta_G)) \tag{2}$$

$$F(I) = H(I_f, T_h(I;\theta_G)), \tag{3}$$

where $H(\cdot)$ is the geometric transformation function, such as homography, affine or similarity transform, and $T_h(\cdot)$ the transformation matrix. We use the spatial transformer network [12] to predict the transformation parameters.

On the other hand, if we assume foreground/background geometry is consistent and only consider the color correction, $F(I)$ becomes a color transformation function $F(I) = C(I_f, T_c(I;\theta_G))$ which adjusts the appearance of the foreground image. We use a linear brightness and contrast model as in [35]:

$$C(I_f, T_c(I;\theta_G)) = \begin{bmatrix} I_f & \mathbf{1} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \\ \beta_1 & \beta_2 & \beta_3 \end{bmatrix}, \tag{4}$$

where $T_c(I;\theta_G) = (\lambda_1, \lambda_2, \lambda_3, \beta_1, \beta_2, \beta_3)$ is a transformation network that predicts the contrast and brightness parameters.

To apply both geometric and color correction to the composite image, we can then combine Equations 3 and 4:

$$F(I) = C\big(H(I_f, T_h(I;\theta_G)), T_c(I;\theta_G)\big). \tag{5}$$

Note that we can use a single network to predict both color and geometric transformation parameters at the same time, so that $T(I;\theta_G) = [T_h(I;\theta_G); T_c(I;\theta_G)]$ and simplify Equation 5 as:

$$F(I) = (C \circ H)(I_f, T(I;\theta_G)). \tag{6}$$

Equations 2 and 6 together describe our compositing model $I_c = G(I;\theta_G)$. However, the composite image might still contain some boundary artifacts. To address this issue, we introduce a refinement network R with an encoder-decoder architecture that further refines the composite image. So the final composition model can be described as:

$$\begin{aligned} I_c &= G(I;\theta_G) \\ &= R(A(I) \circ F(I) + (1 - A(I)) \circ I_b). \end{aligned} \tag{7}$$

3.3. Adversarial Learning

Equation 7 describes our compositing model $I_c = G(I;\theta_G)$ incorporating a transformation network and refinement network. We adopt a similar procedure described as in [7] to train a discriminator network $D(x;\theta_D)$ with adversarial learning. Adversarial learning maximizes the following adversarial loss $\mathcal{L}_a$ to distinguish natural image I_b from the composite image I_c:

$$\mathcal{L}_a(D,G) = \mathbb{E}_{I_b}\Big[logD(I_b)\Big] + \mathbb{E}_{I_c}\Big[log(1 - D(I_c))\Big]. \tag{8}$$

We use a basic three-layer convolutional network for the discriminator network and adopt spectral normalization [24] to stabilize the training process. To reduce the discrepancy between foreground and background in the composite image, we propose to train an additional segmentation network S that learns to separate the foreground object from the background in the composite image. This network is trained with adversarial segmentation loss:

$$\begin{aligned} \mathcal{L}_s(S,G) = &\sum_{s \in fg} \mathbb{E}_{I_c}\Big[log(1 - D_s(I_c))\Big] \\ &+ \sum_{s \in bg} \mathbb{E}_{I_c}\Big[log(D_s(I_c)\Big], \end{aligned} \tag{9}$$

where $s \in \{fg \cup bg\}$ indicate different spatial locations, and fg, bg are sets of foreground and background spatial locations in the composite image. The segmentation network S detects the foreground region by generating foreground/background probabilities for each spatial location.

3.4. Geometric and Color Consistent GAN (GCC-GAN)

Following [7], we optimize the composition model described in Equation 7 by minimizing a min-max objective:

$$\min_{\theta_G} \max_{\theta_D, \theta_S} \mathcal{L}_a(D,G) + \lambda \mathcal{L}_s(S,G). \tag{10}$$

Additional constraints are needed since directly minimizing the above objective will usually lead to the trivial solution where the compositing model simply removes the foreground in the composite image using geometric transformations. Therefore, we add a geometric loss term to our objective function:

$$\mathcal{L}_g = \mathbb{E}_I\Big[\|T(I;\theta_G)\|_2^2 + \lambda_{mask} e^{-k\frac{\|A(I)\|_1}{N}}\Big] \tag{11}$$

The first term in Equation 11 penalizes large transformations, similar to the update loss in [21]; the second term is an exponential loss that directly penalizes the size of the foreground mask if it is too small. For data with ground-truth geometric transformation parameters, we directly use mean square error between the predicted parameters and the ground-truth parameters as our geometric loss.

Finally, we use a pixel-wise L1 loss $\mathcal{L}_c$ to anchor the transformed foreground image to the original foreground image:

$$\mathcal{L}_c = \mathbb{E}_I\left[\frac{\|\big(H(I_f, T(I;\theta)) - F(I)\big) \circ A(I)\|_1}{\|A(I)\|_1}\right]. \tag{12}$$

Combining the above three loss terms, the final loss function for our GCC-GAN becomes:

$$\min_{\theta_G} \max_{\theta_D} \lambda_a \mathcal{L}_a + \lambda_s \mathcal{L}_s + \lambda_g \mathcal{L}_g + \lambda_c \mathcal{L}_c, \tag{13}$$

where λ_a, λ_s, λ_g, and λ_c are hyper-parameters that control the weights between different loss terms.

3.5. Implementation Details

We implemented GCC-GAN with PyTorch [25] and trained on the Nvidia GTX 1080TI GPUs. The input is resized to 128×128 for the experiments on synthesized data and 256×256 for the experiments on COCO. We use the Adam [17] optimizer with an initial learning rate of 0.0002 and (λ_a, λ_s, λ_g, and λ_c) are set to (0.01, 0.01, 1, 1) to empirically balancing the loss terms. We use a batch size of 1 for both experiments, and train the model for 200 epochs for the synthesized dataset and 5 epochs for the COCO experiment. Lastly, we use affine transformations as our geometric transformation function, and we adopt the architecture in [13] for both the refinement network and the segmentation network.

4. Experiments

4.1. Image Compositing with Synthesized Objects

We first validate our model in a simplified artificial setting with a synthesized dataset. We use the Panda3D game engine[1] to render images containing a table and a soda can.

[1] https://www.panda3d.org/

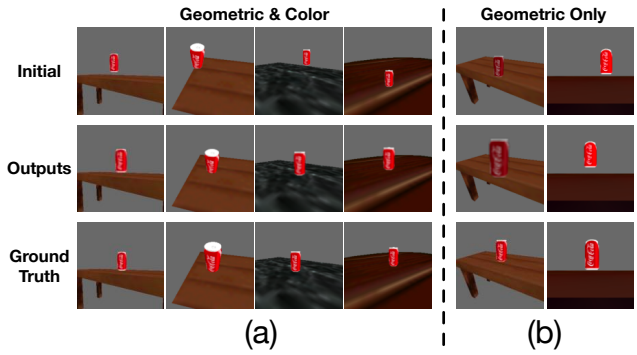

Figure 3. **Experiments on synthesized data.** (a) Through geometric and color transformations, our model learns the relationship between the soda can and the table, and successfully generates composite images with the soda can placed on the table. (b) Without color transformation, the model cannot learn the correct transformation because geometric transformations alone cannot move the composite image on to the manifold of the training data.

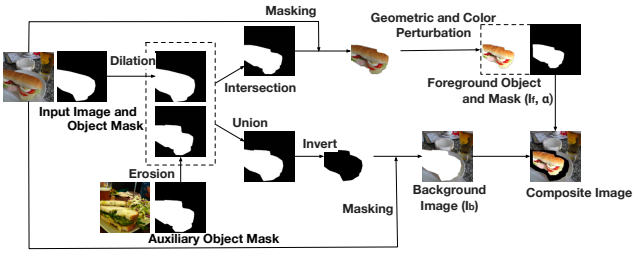

Figure 4. **Training data generation process.** For a given image and its object mask, we first select an auxiliary object mask from a different image in the dataset with the same semantic category. We use morphological operations to remove the boundary in the foreground object and background image. We then combine the object mask with the auxiliary ones to simulate the boundary mismatch during testing. Finally, we apply geometric and color perturbation to simulate the inconsistency during testing.

We render three images for each 3D configuration, including a foreground image with a soda can, a background image with a table, and a ground-truth composite image with a soda can on the table. We then apply random geometric and color perturbations to the foreground and learn a model to composite the perturbed foreground into the background image. Since the synthesized images have a perfect segmentation mask, there will be no boundary artifact in the composite image. As a result, we omit the refinement network and segmentation network in our model for the experiment. We train our model on 15,000 synthesized training triplets with 200 epochs. Figure 3 (a) shows some example results where the first row is the initial composition with foreground perturbation, the second row is the output of our model, and the third row is the ground truth composite image. Our model is able to correct the geometry and color of the foreground object and generate a plausible composite image.

Importance of Color Consistency. To demonstrate the importance of color consistency in the composite image, we also train a model with only the geometric transformation network similar to [21]. Figure 3 (b) shows the result of the model applying only geometric corrections. The model fails to generate plausible composite images because geometric transformation alone cannot move the composite image on to the manifold of the training data.

4.2. Image Compositing with Common Objects

We use the Common Object in Context (COCO) [22] dataset for our compositing experiments. COCO consists of 330K images with segmentation masks of 80 common object categories.

Training Data Generation. Our goal is to generate a composite image by inserting an object from a foreground image into a new background image. However, we do not have training data with realistic composite images, which requires intensive human annotation with professional image-editing software. Instead, we automatically generate training data by perturbing the input images. Figure 4 shows the process of training data generation. For each input image with corresponding object mask, we first select an auxiliary object mask from another image in the dataset with same object category. We then use morphological operations and combine the object mask with auxiliary mask to remove the boundaries from the image, simulating the boundary mismatch during testing. Finally, we apply geometric and color distortion to the foreground to simulate the geometric and color mismatch during testing. For each input image I, we generate a background image I_b, a foreground image I_f and an object mask α as input to our model. Our model then tries to composite the foreground object into the background and generate a realistic composite image. We select object segments that occupy between 5% to 50% of the whole image for our experiments. For each segment, we select 5 auxiliary object masks with the largest intersection over union with the original object mask, resulting in 516,070 training triplets. During testing, we simply remove an object from the background image, and composite another foreground object with the background. Note that our goal is to evaluate image compositing algorithms, therefore we use the ground-truth object mask to segment the objects, however, we can also use semantic segmentation to segment objects for image compositing.

Compared Baselines. We compare our model with the following baselines:

- **Alpha Composition**: a linear combination of the foreground and background using the alpha mask.
- **Poisson Blending [27]**: a gradient based method that minimizes gradient changes in the composite image.
- **Deep Harmonization [30]**: an end-to-end encoder-decoder network with semantic segmentation.

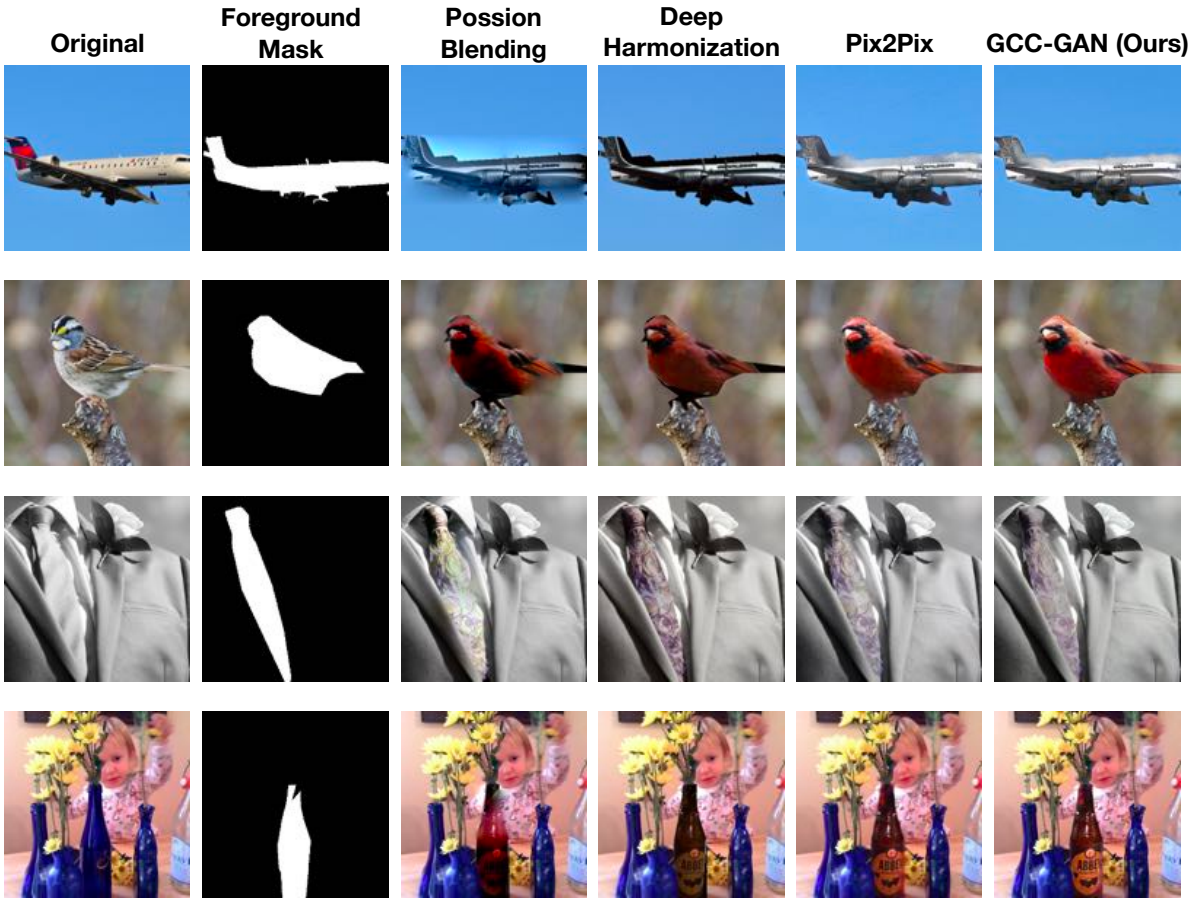

Figure 5. **Qualitative results of different algorithms.** The first column is the original image and the second column is the foreground object mask. The remaining columns show the outputs of different algorithms. Note that since the baseline methods do not account for geometric consistency, for fair comparison, we select foreground objects best matching the background to ensure geometric consistency.

- **Pix2Pix [11]**: an image-to-image translation network with adversarial loss.

Figure 5 shows some qualitative results of the proposed method compared to baselines. Note that since the baselines do not account for geometric consistency, for fair comparison, we select a foreground object that best matches the background, and adjust the geometry to match the foreground and background mask before input to the baselines. Even without geometric mismatch, our model can automatically generate competitive or more realistic composite image compared to all baseline methods. Pix2Pix can generate images of similar quality compared to the proposed method, however, in the following section, we show that when there is geometric inconsistency between foreground and background, Pix2Pix fails to generate plausible composite image since their model does not incorporate geometric losses (c.f. Figure 7).

Importance of Geometric consistency. Figure 6 shows the process of geometric correction of the proposed model with some examples. The first column is the background image and the second column is the foreground object with mask. The third column shows initial alpha composition with a simple copy-paste operation. Notice that the foreground and background in this initial composition is geometrically inconsistent. In the fourth column, our model first transforms the foreground to make the composite image geometrically consistent using the spatial transformer network. Finally, the last column shows the result of the refinement network which makes the boundary more realistic and achieves more realistic image compositing. Figure 7 shows comparison between our model and the Pix2Pix model which does not incorporate geometric correction. Compared to the composite image generated by Pix2Pix, our model is able to perform geometric transformation to the foreground image and thereby generate a plausible composite image.

Human Perceptual Experiments. We also conducted human perceptual experiments to quantitatively evaluate our model. In the first experiment, we want to verify how well our composite image can fool a human subject under close examination compared to baseline method. We randomly select ten images from each of the 80 categories in the COCO dataset with a total of 800 images. For each image,

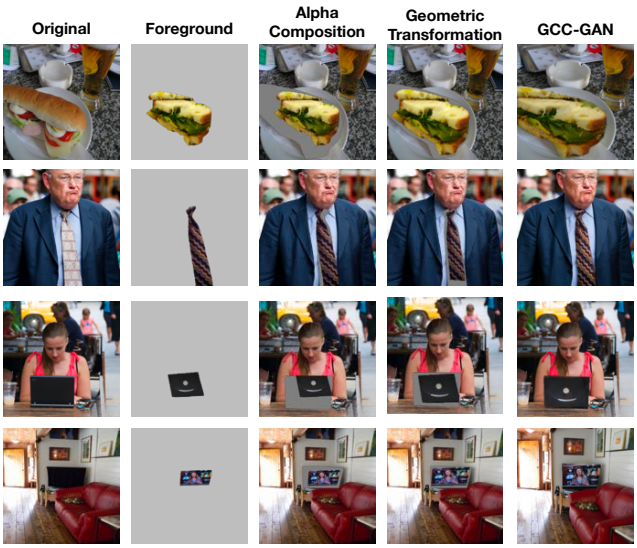

Figure 6. Geometric correction of GCC-GAN. The first and second column show the original image and foreground object. The third column shows the composite image using alpha composition, where the geometry is inconsistent between foreground and background. The fourth column shows the composite image after geometric transformation, and the last column shows the output of GCC-GAN with the final refinement network.

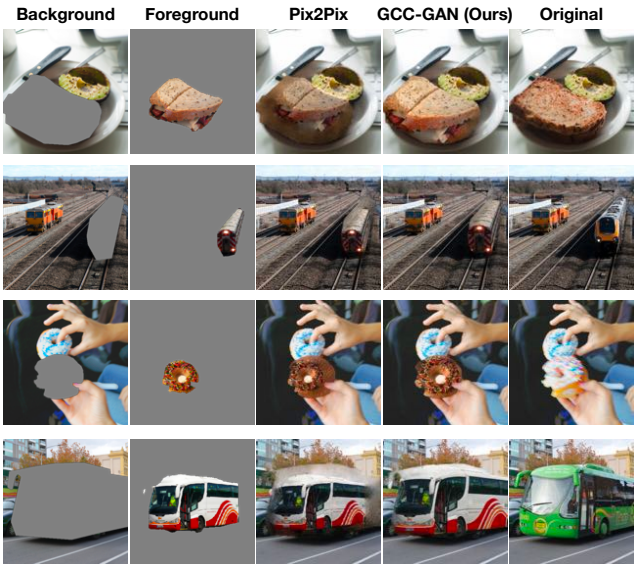

Figure 7. Comparison between Pix2Pix and GCC-GAN when the input geometry is inconsistent. GCC-GAN is able to correct the geometric error and generate more plausible composite images compared to Pix2Pix.

we generate five composite images using the baseline algorithms mentioned earlier. We show the composite image as well as original real image to the annotator with random order and ask them to check if there is any unusual artifact in the image and obtained a total of 4,800 annotations. Table 1 shows the results of the experiment. Even though the input image does not require any geometric correction,

Method	% Real
Alpha composition	4.1%
Poisson blending [27]	10.0%
Deep harmonization [30]	8.6%
Pix2Pix [11]	10.2%
GCC-GAN (Ours)	**11.0%**
Real image	73.8%

Table 1. **Human perceptual experiment with a single image.** We ask the annotator to check if there is any unusual artifact in the image. GCC-GAN can fool the annotator 11% of the time compared to baselines. Note that for fair comparison, we ensure geometric consistency by selecting foreground object best matching the background.

Method	GCC-GAN Performs Better
Alpha composition	82.5%
Poisson Blending [27]	67.3%
Deep Harmonization [30]	71.4%
Pix2pix [11]	56.7%

Table 2. **Human perceptual experiment with pairs of images.** Given two images, we ask the annotator to select the more realistic image from the pair. The output of GCC-GAN is selected more than half of the time compare to all other baselines.

our model still outperforms all baselines in term of human perception, which demonstrates the effectiveness of the adversarial learning process with segmentation network. Note that 26.2% of real images were actually annotated as fake, which shows the annotator is very strict and inspect image meticulously.

In the second experiment, we want to directly compare our algorithm with baselines. We randomly collect five images from each category from COCO, for a total of 400 images. We show the annotator two composite images. One image is generated by our model while the other is generated with one of the baseline methods. To ensure fair comparison, both images are generated with the same foreground and background, with the matching object mask to ensure the composite image is geometrically consistent, and is shown to the annotator in no particular order. Table 2 shows the results of the experiment and again, even without geometric correction, our model can outperform all baseline method and generates better composite image.

Qualitative results and failure cases. Figure 8 shows composite images generated by our model along with the original image for different object categories. GCC-GAN is able to generate realistic composition. Figure 9 show some failure cases our model. In the first example, our model does not have any pose information and was not able to consider semantic layout of the street scene. Therefore, the model generates a composite of the car with an inconsistent pose. In the second example, the foreground segmentation mask is imperfect (i.e. the wheel of the bike), so the model gen-

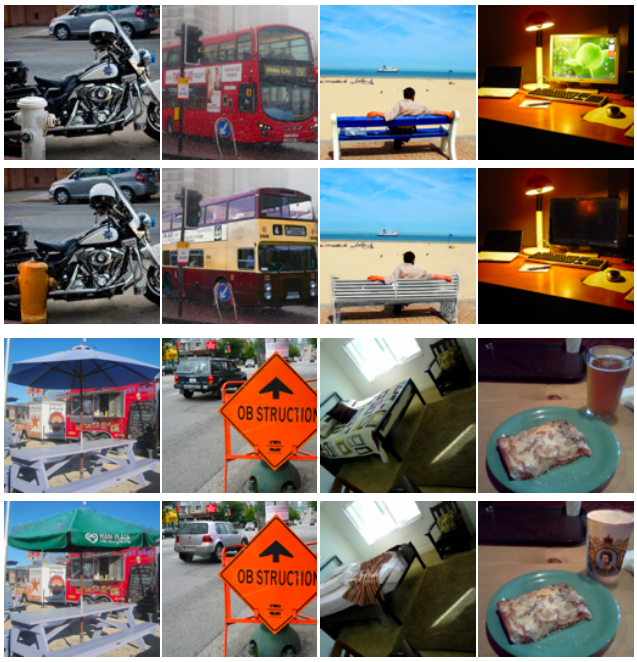

Figure 8. **Qualitative results.** The first and the third rows show the original images and the second and fourth rows shows the output of GCC-GAN.

erate a composite image with inconsistent appearance. In third example, we insert a color train into a black and white background. Since most of our training data consists of color images, the model did not learn to change the appearance of foreground into black and white. In the last example, we show the failure case of genearting a composite image with an animal. Our model works better with rigid objects, and has difficulty modeling animals with diverse poses.

Image Manipulation Detection. In this experiment, we want to see how well the composite image generated by our model can fool an image manipulation detection algorithm. To this end, we utilize a well-trained state-of-the-art image manipulation detection model, RGB-N [34], which uses a two-stream faster-rcnn network to detect different types of image manipulation. We randomly selected 50 images output by each of the baseline algorithms and pass them to the RGB-N model to generate manipulation scores. Table 3 shows the average manipulation scores of different compositing algorithms. Our model obtains the lowest RGB-N score, which indicates that the RGB-N model considers composite images generated by our GCC-GAN model are more realistic compared to baselines.

5. Conclusion

We proposed GCC-GAN for image compositing which considers geometric, color, and boundary consistency. Based on experiments with synthesized data as well as real-world data, we show that both geometric and color consistency are crucial for generating realistic-looking composite images. We also show that GCC-GAN yields better results compared to several state-of-the-art baselines for experiments involving human perception and image manipulation detection. Despite the promising results, we also show the limitations of GCC-GAN, such as failing to deal with objects with diverse poses. Future work includes incorporating pose information into our image compositing framework and using GCC-GAN to improve image manipulation detection algorithms.

Method	Average RGB-N Score
Alpha composition	75.4%
Poisson blending [27]	75.8%
Deep harmonization [30]	77.0%
Pix2Pix [11]	69.1%
GCC-GAN (Ours)	**63.7%**
Real image	57.8%

Table 3. **Average manipulation score for different compositing algorithms.** The score is generated by a state-of-the-art manipulation detection algorithm [34] where a higher score indicates a higher possibility that the image is manipulated. GCC-GAN is able to generate more realistic images that fool the manipulation detection algorithm. Note that Poisson blending and deep harmonization perform worse than alpha composition probably because the compositing process introduces additional artifacts that are captured by the manipulation detection algorithm.

Figure 9. **Failure cases.** (1) GCC-GAN does not incorporate pose information and does not learn the semantic layout of the street. Therefore, the composite image contains a car with unrealistic pose. (2) GCC-GAN generates an unrealistic image due to segmentation error and motion blur, which is not accounted for. (3) Since most of our training data are color images, GCC-GAN composites a color train into a black and white background. (4) GCC-GAN performs better with rigid objects and has difficulty compositing object with diverse poses such as animals.

Acknowledgements. We would like to thank the CAKE (Content Analysis & Knowledge Engineering) team at Verizon Media Group for their help with image annotation. We would also like to thank Guy Dassa for helpful feedback and discussion.

References

[1] Photoshop. https://www.adobe.com/products/photoshop.html. 1

[2] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive digital photomontage. *ACM Transactions on Graphics (ToG)*, 23(3):294–302, 2004. 2

[3] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017. 1, 2

[4] P. J. Burt and E. H. Adelson. A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics*, 2(4):217–236, Oct. 1983. 1, 2

[5] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1511–1520, 2017. 1

[6] P. Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *ACM SIGGRAPH 2008*, page 32. ACM, 2008. 2

[7] 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, 2, 4

[8] Y. Hold-Geoffroy, K. Sunkavalli, J. Eisenmann, M. Fisher, E. Gambaretto, S. Hadap, and J.-F. Lalonde. A perceptual measure for deep single image camera calibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2354–2363, 2018. 2

[9] Y. Hold-Geoffroy, K. Sunkavalli, S. Hadap, E. Gambaretto, and J.-F. Lalonde. Deep outdoor illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7312–7321, 2017. 2

[10] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36(4):1–14, July 2017. 2

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1, 2, 6, 7, 8

[12] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 2, 3

[13] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 4

[14] M. K. Johnson, K. Dale, S. Avidan, H. Pfister, W. T. Freeman, and W. Matusik. Cg2real: Improving the realism of computer generated images using a large collection of photographs. *IEEE Transactions on Visualization and Computer Graphics*, 17(9):1273–1285, 2011. 1, 2

[15] K. Karsch, V. Hedau, D. Forsyth, and D. Hoiem. Rendering synthetic objects into legacy photographs. In *ACM Transactions on Graphics (TOG)*, volume 30, page 157. ACM, 2011. 2

[16] K. Karsch, K. Sunkavalli, S. Hadap, N. Carr, H. Jin, R. Fonte, M. Sittig, and D. Forsyth. Automatic Scene Inference for 3d Object Compositing. *ACM Transactions on Graphics*, 33(3):1–15, June 2014. 2

[17] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[18] J.-F. Lalonde and A. A. Efros. Using color compatibility for assessing image realism. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 1, 2

[19] J.-F. Lalonde, D. Hoiem, A. A. Efros, C. Rother, J. Winn, and A. Criminisi. Photo clip art. *ACM transactions on graphics (TOG)*, 26(3):3, 2007. 1, 2

[20] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2

[21] C.-H. Lin, E. Yumer, O. Wang, E. Shechtman, and S. Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9455–9464, 2018. 1, 2, 3, 4, 5

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[23] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[24] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 4

[25] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. 4

[26] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[27] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Transactions on graphics (TOG)*, 22(3):313–318, 2003. 1, 2, 5, 7, 8

[28] X. Qi, Q. Chen, J. Jia, and V. Koltun. Semi-parametric image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8808–8816, 2018. 1

[29] F. Tan, C. Bernier, B. Cohen, V. Ordonez, and C. Barnes. Where and who? automatic semantic-aware person composition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1519–1528. IEEE, 2018. 2

[30] Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, and M.-H. Yang. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3789–3797, 2017. 1, 2, 5, 7, 8

[31] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 1

[32] S. Xue, A. Agarwala, J. Dorsey, and H. Rushmeier. Understanding and improving the realism of image composites. *ACM Transactions on Graphics (TOG)*, 31(4):84, 2012. 1, 2

[33] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018. 1, 2

[34] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1053–1061, 2018. 8

[35] 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. 1, 2, 3

[36] 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