Regularized Adversarial Training for Single-shot Virtual Try-On

Kotaro Kikuchi
Waseda University
Tokyo, Japan
kotaro@pcl.cs.waseda.ac.jp

Kota Yamaguchi
CyberAgent, Inc.
Tokyo, Japan
yamaguchi_kota@cyberagent.co.jp

Edgar Simo-Serra
Waseda University
Tokyo, Japan
ess@waseda.jp

Tetsunori Kobayashi
Waseda University
Tokyo, Japan
koba@waseda.jp

Abstract

Spatially placing an object onto a background is an essential operation in graphic design and facilitates many different applications such as virtual try-on. The placing operation is formulated as a geometric inference problem for given foreground and background images, and has been approached by spatial transformer architecture. In this paper, we propose a simple yet effective regularization technique to guide the geometric parameters based on user-defined trust regions. Our approach stabilizes the training process of spatial transformer networks and achieves a high-quality prediction with single-shot inference. Our proposed method is independent of initial parameters, and can easily incorporate various priors to prevent different types of trivial solutions. Empirical evaluation with the Abstract Scenes and CelebA datasets shows that our approach achieves favorable results compared to baselines.

1. Introduction

In this paper, we consider the problem of naturally placing a new object onto a background image, such that the resulting composition looks realistic, enabling applications such as virtual try-on. Geometrically placing an object is a basic operation in graphic design. The difficulty in object placement is that subtle misalignment of the object to the background can severely hurts the design quality. Designers spend a lot of efforts in completing visually pleasing graphics layout because of this quality requirement. Automating this operation using machine learning techniques can help designers' productivity and open the door for intelligent tools for creating magazine covers [8], posters [10], banners [9], and virtual try-on [4].

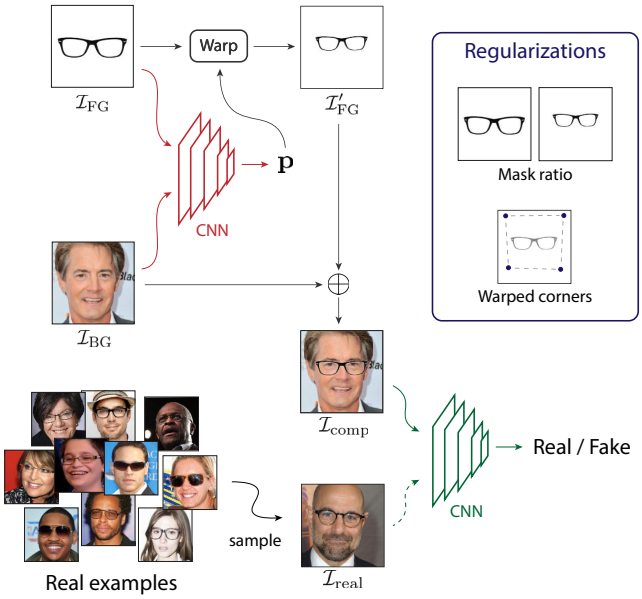

Figure 1. *Overview of our image compositing pipeline.* Given a foreground and background image, our single-shot placement model produces plausible image warping parameters to place the foreground image on the background. We propose a simple yet effective regularization to stabilize the adversarial training of the spatial transformer network.

Object placement can be formulated as the task of finding appropriate geometric warping given foreground and background images. One of the promising approaches is Spatial Transformer Generative Adversarial Networks (ST-GAN) [6], which comprises of multiple spatial transformer networks and adversarial training. The main idea is geometrically correcting the location of the foreground object given the initial warping parameters. For this purpose, ST-GAN iteratively updates the initial warping parameters to

refine resulting composite image. The drawbacks of this iterative approach is an increase in computation time due to costly multiple forward computations, initial parameters highly affecting the result, it is unclear how many iterations should be applied to a given input, and most importantly, the adversarial training process becomes unstable. We observed that the ST-GAN model often suffers from falling into trivial solutions such as excessive scaling and framing out , because these solutions can easily fool the discriminator during training.

In this paper, we propose a simple yet effective regularization technique to guide the parameters based on user-defined trust regions. Our approach effectively stabilizes the training process of spatial transformer networks, and enables an accurate single-shot inference (Figure 1), unlike the iterative approach in ST-GAN [6]. Furthermore, our approach does not require any initial warping parameters, and can easily incorporate different types of user-defined priors (*e.g.*, avoiding excessive skewing) for further guidance. Our experiments with the Abstract Scenes and CelebA datasets shows that our proposed approach shows favorable results compared to the state-of-the-art approaches.

We summarize our main contributions in the following:

- We propose a novel regularization technique to guide warping parameters during the adversarial training of spatial transformer networks, enabling high-quality single-shot inference of object placement for virtual try-on.
- We show that our approach achieves favorable results on the Abstract Scenes and CelebA datasets in comparison to existing approaches.

2. Related Work

Virtual try-on based on 2D image composition has recently started to attract the attention from the research community. Most recent approaches are based on pixel-level transformations of the image [4], however, it is challenging in this setting to generate realistic high-resolution compositions without modifying the content of the image. Additionally, large amounts of training data limit the applicability. In this paper, we do not assume the target is wearing a similar garment to that being tried-on and learn to predict a transformation of the original image, allowing high-resolution image composition.

Spatial Transformer Networks (STNs) [5] introduce learnable image warping module within a deep learning approach, allowing overlaying a masked foreground image onto a background image. The main components of STNs are a neural network to predict a set of warp parameters and a differentiable warping function. We build upon STNs to implement our single-shot inference.

Generative Adversarial Networks (GAN) [2] are generative models that learn a generator network $\mathcal{G}$ and a discriminator network $\mathcal{D}$. In GAN framework, a well-trained generator network can reproduce a generative distribution that matches the empirical distribution of a given data collection. One advantage of GAN is that the loss function is defined by the discriminator network, and therefore does not require labeled datasets. Unsupervised training by GAN framework only requires data collections representing the desired domain distribution.

Recently proposed ST-GAN [6] introduces the GAN paradigm into STNs. ST-GAN generates the distribution of possible updates to the current warping parameters. Since the generator produces updates to warping, the overall model iteratively applies updates to the initial warping parameters to solve for the final warping. Although ST-GAN nicely fits our purpose of object placement, there are several drawbacks arising from unstable training of an iterative model. In this paper, we propose a single-shot inference approach to object placement that overcomes the instability in STNs training by proper regularization.

3. Spatial Transformer Generative Adversarial Networks (ST-GAN)

We briefly introduce ST-GAN [6] in the following section. Given a background image $\mathcal{I}_{\text{BG}}$ and a foreground image $\mathcal{I}_{\text{FG}}$ with a corresponding alpha mask $\mathcal{M}_{\text{FG}}$, the process of image compositing is expressed by:

$$\mathcal{I}_{\text{comp}} = \mathcal{I}_{\text{FG}} \odot \mathcal{M}_{\text{FG}} + \mathcal{I}_{\text{BG}} \odot (1 - \mathcal{M}_{\text{FG}}) \,. \tag{1}$$

A realistic looking composition is then obtained by warping the foreground image with

$$\begin{aligned} \mathcal{I}'_{\text{FG}} &= \text{warp}\big(\mathcal{I}_{\text{FG}}, \mathbf{p}\big) \\ \mathcal{M}'_{\text{FG}} &= \text{warp}\big(\mathcal{M}_{\text{FG}}, \mathbf{p}\big), \end{aligned} \tag{2}$$

where warp$(.)$ is a differentiable warping function [5], usually comprised of a homography transformation and bilinear interpolation, and p are the warping parameters.

Original ST-GAN [6] iteratively applies Spatial Transformer Networks (STN) to predict a series of warping updates. At the i-th iteration, given the input images and the previous warping parameters $\mathbf{p}_{i-1}$, the warping update $\Delta\mathbf{p}_i$ and the new warping parameters $\mathbf{p}_i$ can be written by:

$$\begin{aligned} \Delta\mathbf{p}_i &= \mathcal{G}_i\big(\text{warp}\big(\mathcal{I}_{\text{FG}}, \mathbf{p}_{i-1}\big), \mathcal{I}_{\text{BG}}\big) \\ \mathbf{p}_i &= \mathbf{p}_i + \Delta\mathbf{p}_i, \end{aligned} \tag{3}$$

where $\mathcal{G}_i$ is the i-th geometric prediction network.

ST-GAN learns the model parameters for the geometric prediction networks and the discriminator with Wasserstein GAN [1] objective with a gradient penalty [3] to force the discriminator to be a 1-Lipschitz function. The warping update $\Delta\mathbf{p}_i$ is constrained to lie within a trust region by introducing an additional penalty $\mathcal{L}_{\text{update}} = \|\Delta\mathbf{p}_i\|_2^2$ [6], which

avoids trivial solutions, *e.g.*, removing the foreground and leaving only the background image. The final loss function is written by:

$$\mathcal{L}_{\mathcal{D}} = \mathbb{E}\big[\mathcal{D}\big(\mathcal{I}_{\text{comp}}(\mathbf{p}_i)\big)\big] - \mathbb{E}\big[\mathcal{D}(\mathcal{I}_{\text{real}})\big] + \lambda_{\text{grad}} \cdot \mathcal{L}_{\text{grad}} \quad (4)$$

$$\mathcal{L}_{\mathcal{G}_i} = -\mathbb{E}\big[\mathcal{D}\big(\mathcal{I}_{\text{comp}}(\mathbf{p}_i)\big)\big] + \lambda_{\text{update}} \cdot \mathcal{L}_{\text{update}}, \quad (5)$$

where $\mathcal{I}_{\text{comp}}(\mathbf{p}_i)$ denotes the composite image using $\mathcal{I}_{\text{BG}}$ and $\mathcal{I}'_{\text{FG}}$ warped by $\mathbf{p}_i$, $\mathcal{I}_{\text{real}}$ is a real example sampled from training data collections, $\mathcal{L}_{\text{grad}}$ is a gradient penalty term [3], λ_{grad} and λ_{update} are hyper-parameters to adjust the weights for the gradient penalty term and the warping penalty term respectively. For more detail, refer to [6].

4. Proposed Approach

Our approach aims at penalizing the warping parameters $\mathbf{p}$ falling into undesirable regime. The main idea is to introduce plausible warping regions as a prior. To do this, we introduce the regularization function f to enforce target parameters x to lie within the range of given minimum value x_{min} to given maximum value x_{max}. The regularizer f is defined by a rectifier function:

$$f(x, x_{min}, x_{max}) = \text{ReLU}(x_{min} - x) + \text{ReLU}(x - x_{max}). \quad (6)$$

Let us consider two common issues in spatial transformer networks: excessive scaling and framing out . For preventing excessive scaling, we can set a determinant of the (inverse) affine matrix as a representative parameter, assuming we are parameterizing the warp by an affine transformation. We use the ratio between the sum of the original mask $\mathcal{M}_{\text{FG}}$ and the sum of the transformed mask $\mathcal{M}'_{\text{FG}}$, where $\mathbf{p}$ is applied, to approximate the scaling factor:

$$r = \frac{\sum_{i,j} \mathcal{M}'_{\text{FG}}(i,j)}{\sum_{i,j} \mathcal{M}_{\text{FG}}(i,j)}. \quad (7)$$

Then, our scaling regularizer is defined by:

$$\mathcal{L}_{\text{mask}} = f(r, r_{min}, r_{max}). \quad (8)$$

For preventing framing out, we can apply regularization to the coordinates of the warped corners $\mathbb{C}$, where $\mathbf{p}$ is applied, of the foreground object:

$$\mathcal{L}_{\text{coord}} = \sum_{c \in \mathbb{C}} f(c^x, c^x_{min}, c^x_{max}) + f(c^y, c^y_{min}, c^y_{max}) \quad (9)$$

Our final loss functions become:

$$\mathcal{L}_{\mathcal{D}} = \mathbb{E}\big[\mathcal{D}\big(\mathcal{I}_{\text{comp}}(\mathbf{p})\big)\big] - \mathbb{E}\big[\mathcal{D}(\mathcal{I}_{\text{real}})\big] + \lambda_{\text{grad}} \cdot \mathcal{L}_{\text{grad}} \quad (10)$$

$$\mathcal{L}_{\mathcal{G}} = \mathbb{E}\big[\mathcal{D}(\mathcal{I}_{\text{real}})\big] - \mathbb{E}\big[\mathcal{D}\big(\mathcal{I}_{\text{comp}}(\mathbf{p})\big)\big] + \lambda_{\text{mask}} \cdot \mathcal{L}_{\text{mask}} + \lambda_{\text{coord}} \cdot \mathcal{L}_{\text{coord}}, \quad (11)$$

where λ_{mask} and λ_{coord} are hyper-parameters to adjust the weights for respective regularization terms. We modify the generator loss for more stable hyper-parameter tuning since the size of Wasserstein GAN objective for a generator changes as training proceeds. Note that the first regularization term $\mathcal{L}_{\text{mask}}$ may suppress framing out as well, however, we found that using both $\mathcal{L}_{\text{mask}}$ and $\mathcal{L}_{\text{coord}}$ is stable. It is also straight-forward to apply our regularization technique to penalize any summary statistics from the warping parameters $\mathbf{p}$, such as skewing, etc.

Thanks to the stable adversarial learning by our loss (Eq. (11)), we find thati, in contrast to ST-GAN's iterative updates, a single-shot inference model can produce high-quality prediction of object placement, significantly lowering the computational cost. Prediction is then simplified to the following:

$$\mathbf{p} = \mathcal{G}\big(\mathcal{I}_{\text{FG}}, \mathcal{I}_{\text{BG}}\big). \quad (12)$$

We emphasize that this single-shot inference model does not converge without our regularization.

5. Experiments

We evaluate our approach quantitatively with the Abstract Scenes dataset [11], and qualitatively with the CelebA datasets [7].

5.1. Abstract Scenes Evaluation

Dataset. We use the Abstract Scenes dataset [11] to evaluate our approach in terms of the reproducibility of the ground-truth placement. The dataset contains 11,000 clip-art scenes of children playing. Here, we consider a task of placing glasses and hats in the scene. We split the dataset into a training, validation and test set of 8,775, 1,111 and 1,109 scenes respectively. We generate background images by placing all the objects under the target objects (glasses or a hat), and real images by rendering all of them including targets. Background images are all resized to 144×144. We create foreground images by placing and resizing target objects onto the center of 144×144 transparent pixels.

Warping parameters. We estimate three warping parameters: scaling, horizontal translation and vertical translation. We can directly regularize warping parameters with our regularizing function f (Eq. (6)), but we use $\mathcal{L}_{\text{mask}}$ (Eq. (8)) and $\mathcal{L}_{\text{coord}}$ (Eq. (9)) for emphasizing the generality of the choice of these parameters.

Evaluation metrics. We regard a frame of foreground image as a bounding box and compute the accuracy with Intersection over Union (IoU) under different thresholds. We denote this metric as IoU@θ where θ indicates the IoU threshold. We show results for $\theta = .25, .5, .75$.

Results. We summarize the IoU evaluation on the test set in Table 1. ST-GAN (initial) is the evaluation at the initial parameters, and ST-GAN (warp 5) is the result at the final

Figure 2. *Virtual try-on results on the CelebA dataset.*

Table 1. *Image compositing evaluation in Abstract Scenes.* For ST-GAN (warp 5) we show the average value of 10 trials.

	IoU@θ		
Method	0.25	0.50	0.75
ST-GAN (initial)	0.13	0.03	0.00
ST-GAN (warp 5)	0.41	0.36	0.25
Ours	**0.47**	**0.43**	**0.32**

warping parameters. Our method achieves higher scores than the maximum scores in ST-GAN while being much more computationally efficient.

5.2. CelebA Evaluation

CelebA is a large dataset of facial images [7]. Here, we evaluate our approach in the virtual try-on task of placing eyeglasses onto faces. We only conduct qualitative evaluation since the dataset does not contain ground truth annotation. Following procedures in [6], we create an evaluation dataset which contains 152,249 training and 18,673 test images without glasses, and 10,521 training images with glasses. We use images of 10 glasses provided by [6] as foreground image. Following [6], we use a homography transformation for warping glasses. Results are shown in Figure 2. We find that our method is able to produce compelling more compelling results than ST-GAN.

6. Conclusions

We proposed an effective regularization technique to guide the warping parameters of Spatial Transformer. Experiments demonstrate that our approach achieves favorable results compared to ST-GAN baseline. In the future, we wish to evaluate our approach in a more realistic virtual try-on scenario, and extend our approach to enable simultaneous placement of multiple objects.

Acknowledgement

This work was partially supported by Waseda University Leading Graduate Program for Embodiment Informatics, and JST PRESTO (Simo-Serra, Grant Number: JPMJPR1756).

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein Generative Adversarial Networks. In *ICML*, volume 70, pages 214–223, 2017.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *NIPS*, pages 2672–2680. 2014.

[3] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved Training of Wasserstein GANs. In *NIPS*, pages 5767–5777. 2017.

[4] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. Viton: An image-based virtual try-on network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial Transformer Networks. In *NIPS*, pages 2017–2025. 2015.

[6] C.-H. Lin, E. Yumer, O. Wang, E. Shechtman, and S. Lucey. ST-GAN: Spatial Transformer Generative Adversarial Networks for Image Compositing. In *CVPR*, 2018.

[7] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep Learning Face Attributes in the Wild. In *ICCV*, 2015.

[8] X. Yang, T. Mei, Y.-Q. Xu, Y. Rui, and S. Li. Automatic Generation of Visual-Textual Presentation Layout. *ACM Trans. Multimedia Comput. Commun. Appl.*, 12(2):33:1–33:22, 2016.

[9] Y. Zhang, K. Hu, P. Ren, C. Yang, W. Xu, and X.-S. Hua. Layout Style Modeling for Automating Banner Design. In *ACM Multimedia Thematic Workshops*, Thematic Workshops '17, pages 451–459, New York, NY, USA, 2017. ACM.

[10] N. Zhao, Y. Cao, and R. W. H. Lau. What Characterizes Personalities of Graphic Designs? *SIGGRAPH*, 37(4), 2018.

[11] C. L. Zitnick and D. Parikh. Bringing Semantics into Focus Using Visual Abstraction. In *CVPR*, pages 3009–3016, 2013.