Semantic Image Segmentation using Conditional Generative Adversarial Networks

Partik Adle
Department of Electronics
and Communication Engineering
Indian Institute of
Information Technology, Nagpur
Nagpur, India
pratikadle95@gmail.com

Aditya Kotasthane
Department of Electronics
and Communication Engineering
Indian Institute of
Information Technology, Nagpur
Wardha, India
kotasthaneaditya2377@gmail.com

Darpan Sarode
Department of Electronics
and Communication Engineering
Indian Institute of
Information Technology, Nagpur
Nagpur, India
darpansarode@gmail.com

Mayur Raul
Department of Electronics
and Communication Engineering
Indian Institute of
Information Technology, Nagpur
Nagpur, India
mayurraul67@gmail.com

Ambuj Yadav
Department of Electronics
and Communication Engineering
Indian Institute of
Information Technology, Nagpur
Prayagraj, India
yambuj151@gmail.com

Dr. Tapan Kumar Jain
Assistant Professor
Department of Electronics
and Communication Engineering
Indian Institute of
Information Technology, Nagpur
Nagpur, India
tapankumarjain@gmail.com

I. Abstract

Semantic image segmentation is a Computer Vision task in which each pixel in an image is assigned, one particular class. Semantic image segmentation models are very useful in various applications of image processing and computer vision domain and it is used in various domains such as medical area, autonomous vehicles, satellite image processing, etc. We have used conditional adversarial networks as a solution to the semantic segmentation problem. These conditional adversarial networks learn the mapping between the input image and the output image and at the same time also learn a loss function to train this mapping. In this paper, we propose a conditional adversarial training approach to train semantic segmentation models. We have trained a convolutional semantic segmentation network along with the conditional adversarial network that differentiates segmentation maps from the ground truth. The goal of our approach is that our model can predict the output image based on the available ground truth image and correct higher-order inconsistencies between ground truth segmentation maps.

II. Introduction

Semantic segmentation is a computer vision task in which each pixel in an image belongs to one particular class, i.e., it is the task of assigning a class to every pixel in an image. Currently, most of the methods [1] mainly rely on convolutional neural network (CNN) approaches as CNN has become the common workhorse for variety of image prediction problems. CNN's minimises a loss which is an objective that scores the quality of results. Although the learning process in CNN is automatic, a lot of manual effort is still required for designing effective losses. In other words, we still have to tell the CNN what we wish it to minimize which is an open problem and generally requires expert knowledge.

It would be highly desirable if we specify a high-level goal like if we make the output indistinguishable from reality and then automatically learn a loss function appropriate for satisfying this goal. This is done using Generative Adversarial Networks (GANs). These GANs learn a loss that tries to classify if the output image is real or fake and simultaneously trains a generative model to minimize this loss. Blurry images are not tolerated as they look fake. As these Generative Adversarial Networks (GANs) learn a loss that adapts to the data, they can be applied to various tasks that traditionally require different kinds of loss functions.

In this paper, we use conditional GANs. Just as GANs learn a generative model of data, conditional GANs (cGANs) learn a conditional generative model. This makes conditional GANs suitable for image-to-image translation tasks, where we put a condition on an input image and generate a corresponding output image. The adversarial term encourages the segmentation model to produce label maps that cannot be differentiated from ground truth maps by an adversarial binary classification model. Since the adversarial model can assess the joint configuration of many label variables, it can enforce forms of higher-order consistency that cannot be enforced using pair-wise terms, nor measured by a per-pixel cross-entropy loss. Our primary contribution is to demonstrate that on a wide variety of problems, conditional GANs (cGANs) produce reasonable results. Our second contribution is to present a framework that is sufficient to achieve good results.

III. Related work

Semantic segmentation has been widely investigated in the past years. Some of the existing methods aim at finding a graph structure over the image, by using Markov Random Field (MRF) or Conditional Random Field (CRF), to capture the context of an image and employ classifiers to label different pixels. These methods employ features for classification, and their performance on a variety of datasets is not that adequate.

A. Convolutional Neural Networks:

Convolutional Neural Networks (CNNs) is very popular recently in many computer vision applications including semantic segmentation. For instance, [2] and [3] leverage deep networks classify super-pixels and label the segments. More recent methods such as [4] apply classification on each pixel using a fully convolutional network. This is achieved by transforming fully-connected layers of CNN (VGG16) into convolutional layers and using the pre-trained ImageNet model to initialize the weights of the network. Using this kind of processing in Convolutional Neural Network method is expensive since for each image during the training phase, the iterative inference should be performed. This CNN method is based on supervised learning and relies on large data, which may often not be available.

B. Structured losses for image modeling:

Image to image translation problems is often formulated as a classification or regression problem on each pixel(e.g., [5]). These formulations treat the output space conditionally independent from all others given the input image. Structured losses penalize the output configuration. Conditional Generative Adversarial Networks (cGANs) instead learn a structured loss. In the conditional GAN, the loss is learned and in theory, it penalize any possible structure that differs between output and target.

C. Conditional Generative Adversarial Networks:

To overcome the limitations of current methods, we have used cGANs for semantic segmentation on available data and additional data to improve the fully supervised methods. These generative methods have been largely been used for visual classification tasks and very little has been done for semantic segmentation. This cGAN network aims at creating probability maps for each class for a given image, then the discriminator is used to distinguish between generated maps and ground truth image. Our method is different from this method as 1) we let the discriminator find the labels of pixels, 2) we use unlabeled data alongside generated data in an adversarial manner to compete in getting labels and 3) we use conditional GAN to enhance the quality of generated samples for better segmentation performance as well as to make cGAN training more stable.

Prior to this work, the conditioned GANs has been on discrete labels [6], text [7], and, indeed, images. Various other papers have also used GANs for the image to image mappings but only applied the GAN unconditionally, which relies on other terms such as L2 regression to force the output to be conditioned on the input. The changes were made to these methods for a specific application. Our framework differs from the other framework as in this nothing is application-specific. This makes our setup considerably simpler than most others.

This method also differs from the prior works [8], [9] in architectural choices for the generator and the discriminator. In this work, we use a "U-Net" architecture [10] for generator and we use a convolutional PatchGAN classifier network for discriminator, which only penalizes structure at the scale of image patches. In this work, we show that this approach is effective on a wider range of problems.

IV. Proposed Methodology

GANs are the generative models that learn a mapping from random noise vector z to output image $y, G : z \rightarrow y$. Contrast to the GANs, the conditional GANs learn a mapping from the observed image x and random noise vector z, to $y, G : x, z \rightarrow y$. The generator network G is trained to produce outputs that cannot be differentiated from the real images by an adversarially trained discriminator D which is trained to do and trained to detect the generator's fakes.

A. Objective

We can express the objective of the conditional GAN as:

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[logD(x, y)] + \mathbb{E}_{x,z}[log(1D(x, G(x, z))] \tag{1}$$

where, G tries to minimize this objective for an adversarial D which tries to maximize it, i.e. $G^* = arg\ min_G\ max_D\ \mathcal{L}_{cGAN}(G, D)$.

To test the importance of imposing a condition on the discriminator, we also compare to an unconditional variant of the discriminator in which the discriminator does not observe x:

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_{y}[logD(y)] + \mathbb{E}_{x,z}[log(1D(G(x, z))]. \tag{2}$$

It proves to be beneficial to mix the GAN objective with a traditional loss such as L2 distance. The job of discriminator does not change, but the generator's task is to not only fool the discriminator but also to be near the ground truth output in an L2 sense. We explore this option, using L1 distance rather than L2 as L1 encourages less blurring:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x, z)\|_1] \tag{3}$$

Our final objective is $G^* = arg\ min_G\ max_D\ \mathrm{L}_{cGAN}(G, D) + \lambda\mathcal{L}_{L1}(G)$.

Without z, the network still learns a mapping from x to y, but it produces deterministic outputs, and hence fails to match any distribution other than a delta function. Past conditional GANs has acknowledged this and has provided Gaussian noise z as an input to the generator, in addition to x. For the final model, we provide noise in the form of dropout which is applied on several layers of the generator at training as well as test time. We try to design the conditional GAN network that captures the full entropy of the conditional distributions.

B. Network Architecture

We adapt to generator and discriminator architectures from those in [11]. Both generator and discriminator use modules of the form convolution-BatchNorm-ReLu [12].

1) Generator with skips: A feature of the image to image translation problems is that they map a high-resolution input to a high-resolution output. Moreover, for the problems that we consider, the input and output differ in surface appearance, but both are the renderings of the same underlying structure. Therefore, structure in the input is roughly aligned with the structure in the output. We design the generator architecture using these considerations.

Many previous solutions [[13], [14]] to such problems in this area uses an encoder-decoder network. In such an encoder-decoder network, the input is passed through a series of layers that continuously downsample the layer until a bottleneck layer. After the bottleneck layer, the process is reversed. Such a network requires that all the information flow passes through all these layers, including the bottleneck layer. For many image translation problems, there is a great deal of low-level information sharing between the input and output layers. It is desirable to pass this information directly across the network.

To give the generator a way around the bottleneck for information we add skip connections in the network of the generator, following the general shape of a “U-Net” [10]. We add skip connections between each layer i and layer $n - i$, where n is the total number of layers. Each skip connection simply concatenates all channels at layer i with those at layer $n - i$.

2) Markovian discriminator: It is known that the L2 loss and L1 loss results in a blurry effect on image generation problems [15]. Although these losses fail to capture high frequencies, in many cases, they accurately capture the low frequencies. For such cases, we do not need an entirely new framework to enforce correctness at the low frequencies and the loss L1 will do that.

So, we restrict the cGAN discriminator to only model high-frequency structure, relying on an L1 term to force low-frequency correctness (Eqn. 4). Apart from that, in order to model high-frequencies, it is sufficient to restrict our structure in local image patches. Therefore, we design a discriminator architecture PatchGAN which only penalizes structure at the scale of patches. This discriminator tries to classify if each of $N \times N$ patch in an image is real or fake. We run this discriminator network convolutionally across the image, which takes an average of all responses to provide the required output of D.

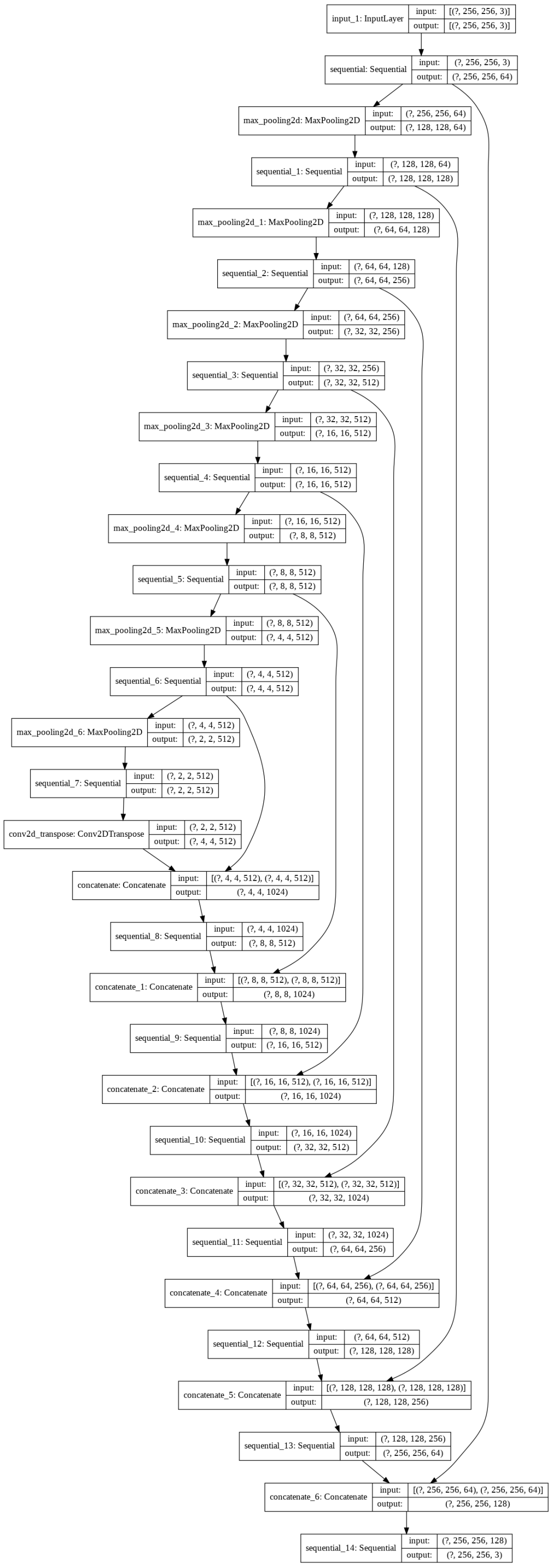


Fig. 1. Architecture of Generator

In this work, we demonstrate that N can be much smaller than the full size of the image and still produce high-quality results. This is the advantage because a smaller PatchGAN has

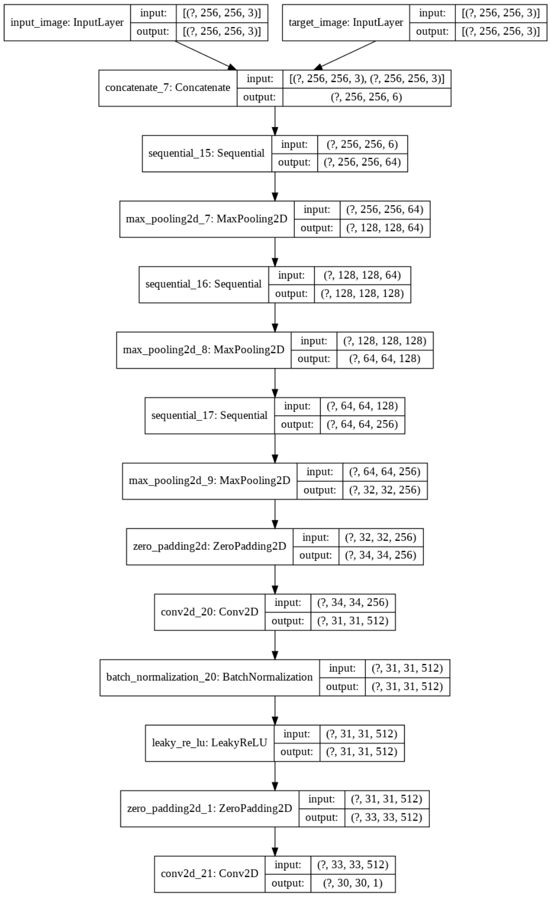


Fig. 2. Architecture of Discriminator

fewer parameters, runs faster and it can be applied to arbitrarily large images. Such a discriminator models the image as a Markov random field, assuming that there is independence between pixels separated by more than a patch diameter. This connection was explored in [16], and is also the common assumption in models of texture [17] and style [18].

C. Optimization and Inference

In order to optimize our networks, we follow the standard approach from [19]. We alternate between one gradient descent step on D, then one step on G. As suggested in the original GAN paper, rather than training G to minimize $log(1D(x, G(x, z))$, we instead train to maximize $logD(x, G(x, z))$ [19]. In addition to this, we divide the objective by 2 while optimizing D, which slows down the rate at which D learns relative to G. We use minibatch SGD and apply the Adam solver [20].

For inference, we run the generator network in the same manner as that during the training phase. This differs from the usual protocol in that we apply dropout at test time, and we apply batch normalization [12] using the statistics of the test batch, rather than aggregated statistics of the training batch. This approach to batch normalization, when the batch size is set to 1 is known as instance normalization and has been demonstrated to be effective at image generation tasks. In our experiments, we use batch sizes between 1 and 10.

V. Experimental Results

Dataset:

In this experiment, we used the Cityscapes Dataset [21] that focuses on semantic understanding of urban street scenes. Cityscapes Dataset consists of labelled videos taken from vehicles driven in Germany. We have used the processed subsample of this Dataset. This dataset that we have used has still images from the original videos, and the semantic segmentation labels are shown in images alongside the original image. This is one of the best datasets used for semantic segmentation tasks. This dataset has 2975 training images files and 500 validation image files. Each image file is 256x512 pixels, and each file is a composite with the original photo on the left half of the image and alongside it is the labelled image, i.e., output of semantic segmentation on the right half.

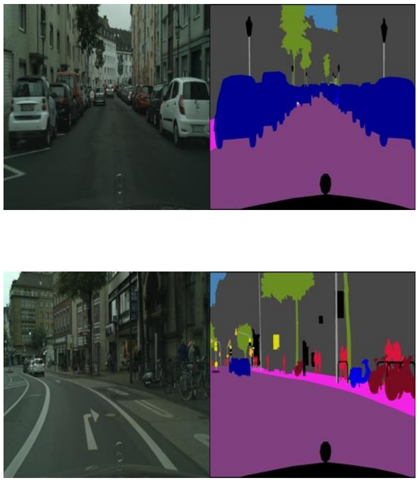
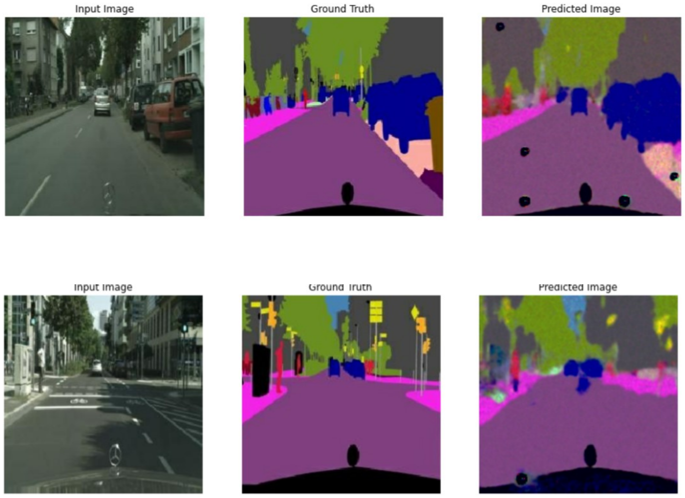


Fig. 3. Images from the Cityscapes Dataset



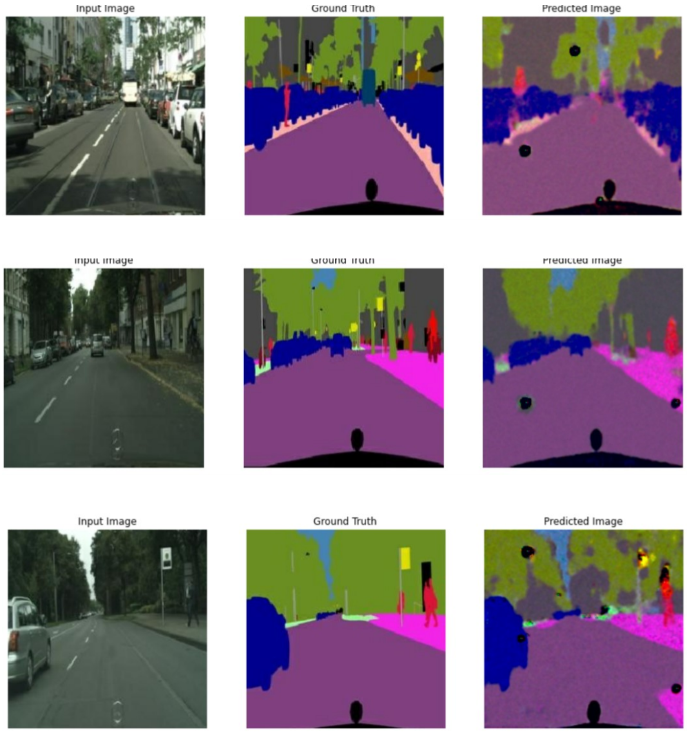


Fig. 4. Experimental Results

Semantic Segmentation:

Conditional GANs is effective on problems where the output is highly detailed, as is common in image processing and computer vision tasks. But in vision problems like semantic segmentation where the output is less complex than the input we train a cGAN (with/without L1 loss). Moreover, conditional GANs trained without L1 loss are able to solve this problem at a reasonable accuracy. Although conditional GANs achieve success, they are far from the best available method for solving this problem by simply using L1 regression gets better scores than using a cGAN. We were able to solve this vision problem, i.e., the goal of predicting output close to the ground truth in this work.

VI. Discussion

We have used the conditional adversarial approach to learn semantic segmentation models in this work. In the CNN based segmentation models, we use tractable computation of the exact multiclass cross-entropy loss. In this work, we have used the conditional adversarial network with adjustable parameters to regularize the segmentation model by enforcing high consistency in the prediction model.

The conditional GAN was used in this work during the training over the Cityscapes datasets. Our network was able to learn hidden structures that are then used to enhance the performance of our conditional GAN discriminator as they can be seen as additional pixel-level added data. Moreover, our conditional GAN framework was able to learn spatial object distributions, for example, roads are at the bottom of images, etc. Summarizing, the results achieved in this experiment indicate that the extra data provided through adversarial loss boosts the performance of semantic segmentation. The discriminator and the generator not only to generate images but it also amounts to learning more meaningful features for pixel classification.

VII. Conclusion

The results in this paper suggest that the conditional adversarial networks are a promising approach for many image-to-image translation tasks. These networks learn a loss adapted to the task and data which makes them applicable in a wide variety of settings. We conducted the experiments on the Cityscapes Dataset. The result shows that the adversarial training approach leads to improvements in semantic segmentation accuracy.

References

[1] A. v. d. H. G. Lin, C. Shen and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," *CVPR*, 2016.

[2] P. Y. M. Mostajabi and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 3376–3385, 2015.

[3] L. N. C. Farabet, C. Couprie and Y. LeCun, "Learning hierarchical features for scene labeling.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, p. 1915– 1929, 2013.

[4] E. S. J. Long and T. Darrell, "Fully convolutional networks for semantic segmentation," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 3431–3440, 2015.

[5] E. S. J. Long and T. Darrell, "Fully convolutional networks for semantic segmentation," *CVPR*, 2015.

[6] A. S. E. Denton, S. Chintala and R. Fergus, "Deep generative image models using a laplacian pyramid of adversarial networks," *NIPS*, 2015.

[7] X. Y. L. L. B. S. S. Reed, Z. Akata and H. Lee, "Generative adversarial text to image synthesis," *ICML*, 2016.

[8] T. Z. Phillip Isola, Jun-Yan Zhu and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv:1611.07004v3*, 2018.

[9] S. C. Pauline Luc, Camille Couprie, "Semantic segmentation using adversarial networks," *arXiv:1611.08408*, 2016.

[10] P. F. O. Ronneberger and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *MICCAI*, 2015.

[11] L. M. A. Radford and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *ICLR*, 2016.

[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *ICML*.

[13] J. D. T. D. D. Pathak, P. Krahenbuhl and A. A. Efros, "Context encoders: Feature learning by inpainting," *CVPR*, 2016.

[14] X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," *ECCV*, 2016.

[15] S. K. S. A. B. L. Larsen and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," *ICML*, 2016.

[16] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," *ECCV*, 2016.

[17] A. A. Efros and T. K. Leung, "Texture synthesis by nonparametric sampling," *ICCV*, 1999.

[18] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," *SIGGRAPH*, 2001.

[19] M. M. B. X. D. W.-F. S. O. A. C. I. Goodfellow, J. Pouget-Abadie and Y. Bengio, "Generative adversarial nets," *NIPS*, 2014.

[20] D. Kingma and J. B. Adam, "A method for stochastic optimization," *ICLR*, 2015.

[21] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.