IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 4, 2020, accepted February 8, 2020, date of publication February 11, 2020, date of current version February 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2973296

Attentively Conditioned Generative Adversarial Network for Semantic Segmentation

ARIYO OLUWASANMI[1], MUHAMMAD UMAR AFTAB[1], AKEEM SHOKANBI[2], JEHOIADA JACKSON[1], BULBULA KUMEDA[1], AND ZHIQUANG QIN[1]

[1]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
[2]Department of Computer Science and Technology, Southwest Jiaotong University, Chengdu 611756, China

Corresponding author: Zhiquang Qin (qinzg@uestc.edu.cn)

This work was supported in part by the NSFC-Guangdong Joint Fund under Grant U1401257, in part by the National Natural Science Foundation of China under Grant 61300090, Grant 61133016, and Grant 61272527, in part by the science and technology plan projects in Sichuan Province under Grant 2014JY0172, and in part by the opening project of Guangdong Provincial Key Laboratory of Electronic Information Products Reliability Technology under Grant 2013A061401003.

ABSTRACT Generative Adversarial Network has proven to produce state-of-the-art results by framing a generative modeling task into a supervised learning problem. In this paper, we propose Attentively Conditioned Generative Adversarial Network (ACGAN) for semantic segmentation by designing a segmentor model that generates probability maps from images and a discriminator model which discriminates the segmentor's output from the ground truth labels. Additionally, we conditioned the discriminator's dual inputs with extra information as a conditional adversarial model such that, an attention obtained probability distribution of the segmentor's feature maps is incorporated, and the ground truth is also accompanied by a vector of the class label. We demonstrate that our proposed model can provide better semantic segmentation results while stabilizing the discriminator to model long-range dependencies as a result of the supplementary inputs to the network. The attention network particularly provides more insights by extracting cues from the feature locations, and alongside the class label vector, gives the model an advantage to inform better spectral sensitivity. Experiments on the PASCAL VOC 2012 and the CamVid datasets show that our adversarial training technique yields improved accuracy.

INDEX TERMS Generative adversarial network, deep convolutional neural network, attention network, conditional gan, semantic segmentation, deep learning.

I. INTRODUCTION

Vision is a very crucial mechanism for our daily activity as humans rely heavily on visual information and cues which make up the largest percentage of our knowledge base. Objects interact with light through reflection and transmission which ultimately makes them visible to the human eyes. The transmitted and reflected light arrays are then translated into intuitive visual information. In the same manner, as the human eyes, images are made up of pixels which store the color information or intensity of the light photons at a particular point in an image [1]. With increasing technological advances, computers are being used to analyze digital data, for example, computer vision study attempts to use machine learning techniques to accomplish various advance image processing tasks such as object detection and classification, pattern recognition, object localization, landmark detection, captioning and image segmentation [2]. Computer vision generally aims at training computers to analyze and understand extracted information from digital forms like images, videos and scanners just like humans.

In this work, we focus on semantic segmentation which is a subset of computer vision. Image segmentation in itself involves the partitioning of an image into different components, and it differs from instance segmentation because instance segmentation not only partitions an image, it classifies the partitioned segments into a well-defined meaningful component [3]. This way, in semantic segmentation, each pixel in an image is identified to a specified class of objects such as cars, trees, dogs and so on [4]. Deep learning approaches have been used in recent years to accomplish the task of segmentation including the artificial neural networks (ANN) which uses artificial neurons that are designed to mimic the human biological neurons to computationally learn

The associate editor coordinating the review of this manuscript and approving it for publication was Byungcheol Song.

image mappings. However, convolutional neural network (CNN) has emerged as a better option for image analysis as it can extract image features and valuable attributes at different levels. CNN architecture uses weight sharing strategy to reduce parameter complexities and takes advantage of spatial coherence which is beneficial for training images [5].

In this paper, we investigate the problem of semantically segmenting objects in an image by proposing an Attentively Conditioned Generative Adversarial Network (ACGAN) which is based on the popular Generative Adversarial Network (GAN). The GAN model uses neural networks to learn the distribution of data such that it could generate the same distribution from a random sample space [6]. The novelty of our model compared to other GAN-Based segmentation techniques is presented in the structure and design of our conditioned discriminator model. Where n represents the total number of classes available in the dataset, our model design consists of a segmentation network (Segmentor) which generates n feature maps of the input image, an attention model which generates the probability distribution vector of the objects in the feature map, and a discriminator model which distinguishes the predicted feature maps from the ground truth label distribution.

Compared to other approaches, the two inputs to our discriminator are conditioned, first the probability features from the segmentor are attentively conditioned, and the second discriminator input which is the transformed multi-channel labels is also conditioned with a vector representing the constituent object classes. Obviously, convolutional-based GAN has been more successful in computer vision tasks because of its invariant property [7], but it was realized that the network encounters certain challenges in learning the mapping of datasets containing multiple classes, resulting in sample space mismatching [8]. Building on the idea of conditional GAN [9], we use an attention module to obtain the salient objects of the predicted probability maps and label distribution, which is then included as an additional input to the discriminator to inform the classes of the constituent objects in the image. In like manner, we accompany the ground truth images with a vector of the images' class label as extra information to the discriminator. Our adversarial training is designed such that the segmentation network outputs probability feature maps of the input image labels, and its loss is measured using cross-entropy while the discriminator maximizes the probability assigned to the output of the segmentor and the image labels [10].

II. RELATED WORK

A. GENERATIVE ADVERSARIAL MODELS

Generative Adversarial Network has accomplished incredible success in artificial intelligence, especially in the area of computer vision where it has been utilized for data generation [11], image-to-image translation [12], text-to-image synthesis [13], image super-resolution [14], face generation [15], object generation [16], and neural style transfer [17]. Taking samples from a fixed sample space distribution z, the first GAN was built using two multi-layer perceptron networks to learn and model such distribution into a copy of an existing distribution sample. Building on this, a convolutional GAN was introduced for advance image feature learning, then integrated with transposed convolution to upscale learned features to a higher resolution [18]. Reiteratively, adversarial models learn to discriminate generated samples from empirical data by approximating the distribution of the real training data through minimization of the discriminator's cross-entropy loss, while the generator, in turn, has a variational loss function which maximizes the log probability of the discriminant model. Supplementary information was considered for GAN input which is especially useful when considering tasks with single input mapping to an output of many mappings, such that class mismatching is avoided during training [19]. In most cases, the supplementary input is a modeled vector of the input data classes which acts as a conditional supervisory parameter.

Recently, several mechanisms have been devised to regulate GAN's susceptibility to factors such as sample space distribution, cost function, training stability, distribution disparity and flawed rivalry between discriminator and generator. The stability and quality of GANS were improved using a progressive growth approach rather than simultaneous learning, this explores learning large-scale features of an image at the initial stage before including finer features at later stages of training [20]. An Earth Mover (EM) distance was proposed to compute the distance between the real distribution and the predicted distribution of the GAN network, aiming at combating saturation in probability distribution convergence [21]. A variant method of calculating the distance of real data distribution and generated distribution over mini-batches was implemented using the Cramer Distance. This is done to match both distributions with an impartial function over the data batches in generating unbiased gradients [22].

B. SEMANTIC SEGMENTATION

Image segmentation, in general, relates to the computer vision task of labeling definitive regions on an image, however, semantic segmentation on the other hand specifically marks individual pixel in an image to a single segment or class. Recently, the encoder-decoder architecture has become the most famous semantic segmentation approach which mostly uses two different convolutional networks, whereby the encoder network extracts features from the input images and the decoder spatially expands the encoder's output to the input resolution. This is done to eliminate errors in matching the predicted results to the input images [23]. Over the years, the task of upsampling has been achieved via several methods such as Nearest Neighbor and Bed of Nails, but transpose convolutions which learn the upsampling procedures and simple bilinear interpolation which uses linear interpolating have been the most prevalent.

In most cases, the architectures are mostly fully convolutional networks, eliminating dense layers from the model

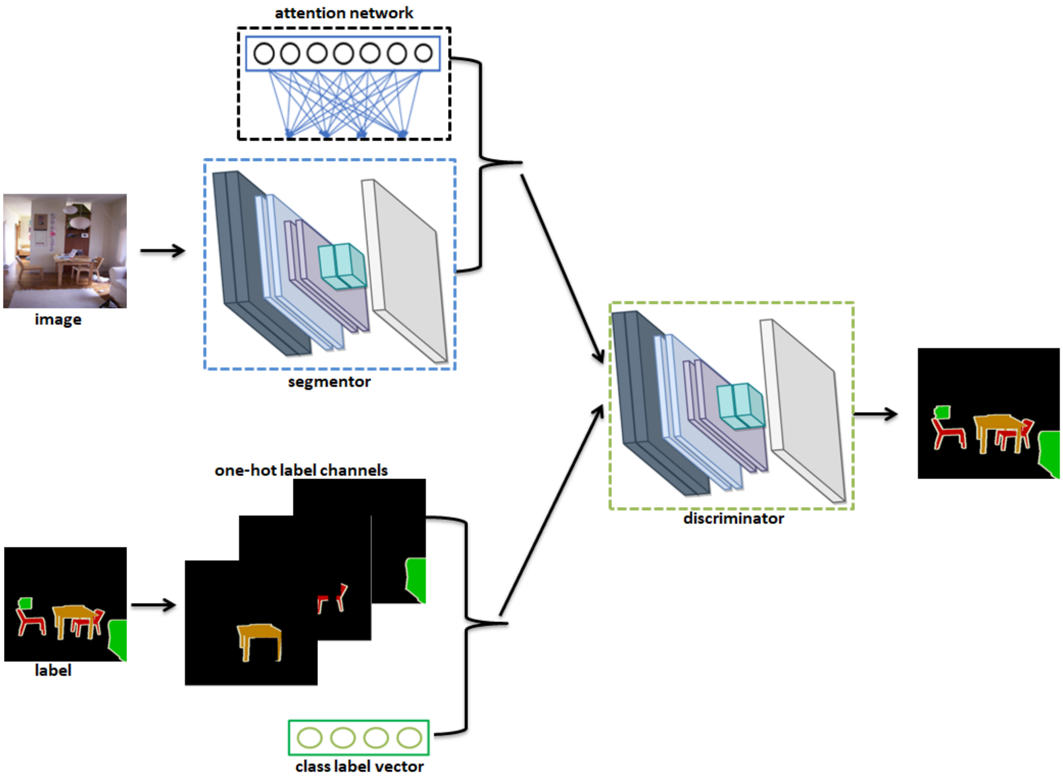

FIGURE 1. Proposed ACGAN framework. The segmentor's generated feature maps are conditioned on a weighted attention to the discriminator model. The ground truth label is also transformed into a one-hot C channel dimension and conditioned on the class label vector into the discriminator.

designs. Because of the boundary and shape inaccuracies as a result of the loss of information during processing, skip connections are introduced. Indices from previous layers on the downsample network which learns the higher-resolution features are transferred to layers on the upsample network to encode fine-grain details of the feature textures [24]. Likewise, contracting path and context modules have been researched to enable definite localization as well as multi-scale supervision [25]. Increased receptive field and field of view is achieved at the same time maintaining the spatial dimension of the features through dilated convolutions [26].

C. REFINEMENT NETWORKS

To improve the localization problem of segmentation as well as improve the accuracy of segmentation, many techniques have involved different approaches. The deeplab model connects the output of their segmentation to a Conditional Random Field (CRF) post-processing network [27]. The probability graphical model was able to achieve a pixel-level classification to enhance the localization of object detection. Subsequently, domain transform was included to ensure faster computation and improvement in object boundary classification [28].

Since semantic segmentation task involves establishing common grounds between object semantics and their localization, thus magnifying the invariance problem of convolutional neural network. Consequently, an end-to-end network comprising of a convolutional network and Markov Random Field (MRF) which embeds the label contexts and high-order relations was designed to both reduce post-processing model computation and increase segmentation accuracy [29]. A combination of long-range residual connection and residual pooling was employed to achieve multi-path fine-grained features for attaining efficient segmentation refinement. This allows increased flow of feature information flow along subsampling and upsampling layers, thereby increasing fine-grained resolution output [30]. In the same way, aggregated context prior of image global information containing all the different regions in the image was optimized using a specially designed pyramid pooling module [31].

D. ATTENTION MODELS

Convolutions generally extract features from images by learning high-level structures as the layer increases, but the task of segmentation is mostly affected by pooling operations which come along with increased layers. To eliminate the loss of spatial information, attention models have been included in different architecture to achieve an improved segmentation by modeling long-range dependencies [32]. Both global and local dependencies are combined to incorporate channel and spatial dimensions of features in an image by attributing a

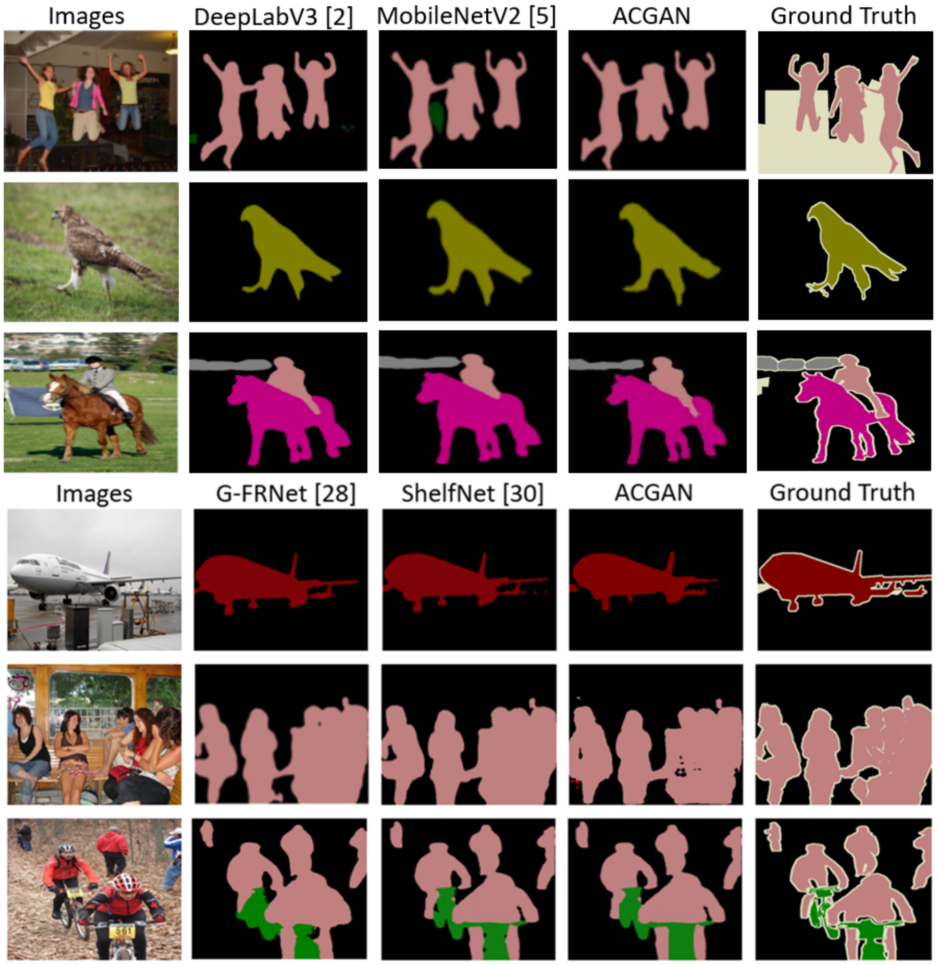

FIGURE 2. Qualitative results of segmentation depicting results of our model compared to the other models on the PASCAL VOC dataset.

weighted sum to all the regions in the image. This builds on the logic that identical objects would have similar features regardless of their distance apart [33].

The rate of object co-occurrence in an image was investigated by building an Aggregated Co-occurrence Feature Model which learns the mapping of invariant representations in a given feature. This gives a global elaboration about the interrelationship of objects in a scene [34]. The local neighborhood constraint of the convolution network was tackled by connecting all available points in an image, such that the prediction and the classification of a particular pixel are dependent on all other points in the image through the use of a point-wise spatial network of attention [35]. Multiscale resizing of input images has been used to demonstrate the effectiveness of different scales and positioning of objects [36].

III. ACGAN

We propose an adversarial segmentation model which is conditioned on an attended identification of the object class in an image, and the goal of the model is to semantically classify objects in an image into a definite range of classes. First, as displayed in Fig. 2, the input images are fed to the segmentor which outputs the feature probability maps of the input image such that the output image has channels equal to the number of predefined semantic classes c. Accordingly, a one-hot encoding technique is implemented to transform the images' ground truth labels into a map of C probability channel in the same scope as the output of the segmentor. This way, the discriminator model can then take as input the output of the segmentation model or the transformed ground truth label, after which it would generate a map, where each pixel p equals 1 or equals 0, which would be predicted as originating from the ground truth label or segmentation network. To incorporate more cues from the images feature locations, we apply an attention network to leverage responses from the feature positions which is added as additional information alongside the segmentor output. Also, the transformed ground truth label is fed into the discriminator with a vector representing the class label of the one-hot encoded ground truth. The attention network basically computes an attention vector of weighted sum of the feature map regions, thereby informing more intuition and cues for the discriminator's training. This efficiently helps to incorporate long-range dependencies over the internal states of the model and create a good balance of multi-class feature detection which clearly improves the segmentation accuracy.

A. ATTENTION NETWORK

The only two inputs to the discriminator are the probability feature maps of the segmentor and the transformed one-hot ground truth encoding, both considered as feature score maps f_n^c, where n represents all ranges of spatial positions and c is the label classes. Therefore, the weighted sum of the segmentor's feature score maps f_n^c of each inputs is represented as:

$$f_n^c = \sum_{n=1}^{N} w_n^c f_n^c \tag{1}$$

where the weight w_n^c is measured as:

$$w_n^c = \frac{exp(h_n^c)}{\sum_{t=1}^{c} exp(h_n^t)} \tag{2}$$

where h_n^c is the score map obtained from the attention model at time t. The attention network is designed with two layers of dense network having 64 nodes in the first layer and C nodes in the second layer.

B. LOSS FUNCTION

Denoting the input image as X, and the ground truth probability map as Y, the segmentator with the input image is represented as $S(X)$ while the discriminator alongside the segmentor's feature map is described as $D(S(X))$ and along the label map input is described as $D(Y)$. With this, we measure the losses of the network using the technique of [37], such that the cross entropy of the segmentor loss S_l prediction to the ground truth is given as:

$$S_l = -\sum_{h,w}\sum_{c=C} Y_n^{(h,w,c)} log(S(X_n)^{(h,w,c)}) \tag{3}$$

With h, w and c being the image height, width and number of label categories respectively, then the adversarial loss A_l of the model conditioned on the attended feature maps alongside the discriminator is measured as:

$$A_l = -\sum_{h,w} log(D((S(X_n)|Aatt)^{(h,w)}) \tag{4}$$

Such that the discriminator is trained by minimizing the loss stated as:

$$D_l = \sum_{h,w}(1-y_n)log(1-D(S(X_n)|Aatt)^{(h,w)}) + y_n log(D(Y_n|V)^{(h,w)}) \tag{5}$$

With $Aatt$ as the computed weight of the attention module and V as the class label vector, the objective function is optimized by training the discriminator to minimize the loss of the confidence map. To convert the image labels into probability channels, we ignore the scaling and product techniques [38] and implemented the basic approach as the one-hot probability channels have no effect when trained on fully convolution network [37], including the class vector as an additional input into the discriminator.

C. MODEL ARCHITECTURE

Our model is a conditioned adversarial network which uses two sub-models; segmentor and discriminator to semantically segment images in a supervised manner. Firstly, the segmentor which is a fully convolutional network is built on the ResNet-101 model with the last dense layer removed is fed with images of 3 dimensions which represents RGB, and outputs a C channel dimension, where C represents the number of object classes in the dataset. The generated feature maps from the segmentor is fed as input to the discriminator alongside the probability weights of the feature maps' salient objects. Also, the ground truth label is transformed into a one-hot encoding of C dimension. In the same way, the transformed one-hot ground truth is fed to the discriminator and conditioned with a vector representing each label's class. The discriminator which was designed as the method of [37] is also a fully convolutional network with five convolutional layers having 64, 128, 256, 512, 1 filters respectively with a stride of 2 and kernel size of $4 * 4$. The discriminator model however outputs a confidence map of one dimension, where each pixel belongs to a particular label class. Therefore, the discriminator is tasked with discriminating if a pixel in the confidence map belongs to the segmentor's feature maps or the ground truth distribution. This process eventually trains the segmentor to produce feature maps as similar as possible to the ground truth labels, as such, learning to accurately segment the training images. In both cases, the segmentor and discriminator output are up-sampled to the same size of the input images to ensure accurate comparison with ground truth label.

IV. EXPERIMENTAL RESULTS

A. DATASET

The PASCAL VOC 2012 Dataset [39] is arguably the most popular semantic segmentation data available today and beyond segmentation, it extends to four different tasks including classification, detection, action classification and person layout. For it segmentation task, the dataset is augmented with more annotations from [40] totaling 10,582 training images with 1, 456 and 1449 images for test and validation respectively. The PASCAL VOC 2012 dataset has 20 object class categories such as vehicle, animal and person.

The CamVid Dataset [41] is designed purposely for road scene understanding and semantic segmentation which is used for many autonomous driving learning. In total, it contains 701 annotated images of 367 training images, 233 testing images and 101 validation images respectively. The dataset has 11 semantic classes which are pixel-level annotated and is augmented during draining to increase learning capacity.

B. EVALUATION METRIC

The performance of the ACGAN-SEG on both the CamVid and PASCAL VOC 2012 Dataset is measured using the popular mean Intersection over Union (mIoU) technique proposed in [23] as:

$$mIoU = (1/n_c) \sum_i nii \Big/ \left(\sum_j nij + \sum_j nji - nii \right) \quad (6)$$

where n_c is the number of classes, nii is the number of pixels of class i predicted to be in class i and nij is the number of pixels of class i which are predicted to belong to class j. If the total pixels of a class i is represented as $t_i = \sum_j nij$, then the pixel accuracy $pacc$ and mean accuracy $macc$ are given as:

$$pacc : \sum_i nii / \sum_i t_i \quad (7)$$

$$macc : (1/n_c) \sum_i nii/t_i \quad (8)$$

C. ADVERSARIAL TRAINING

The model framework is optimized by the loss function stated in Eq. (5) by systematically framing the generative training as a supervised problem whereby the segmentation model which is also called segmentor produces predicted probability maps which is to be distinguished from ground truth distribution by the discriminator network. Firstly, the segmentor which is built on pre-trained ImageNet and MSCOCO dataset of the ResNet-101 model is modified by removing the classification layer such that the model is fully convolution and the reduced dimension is upsampled to the size of the input images, with each feature map representing each label class. The segmentor receives the training images as input and generates probability maps of their semantic labels.

Then the discriminator network which is also a fully convolutional network serves the adversarial learning purpose by differentiating the segmentor output from the ground truth labels. It accepts either of the segmentor network prediction or the ground truth label map as input. We complemented the segmentor's feature maps with an attended vector representing object classes to the discriminator while the ground truth is conditioned with a vector of the constituent class label. However, in preprocessing the image labels, we utilized the basic one-hot transformation technique, ignoring the product and scaling methods as there are no differences when running a fully convolutional network [37]. The final output of the discriminator is a single map where pixels are classified as binary, either from segmentor probability maps or ground truth labels.

The segmentor was trained using the Stochastic Gradient Descent (SGD) optimization method, thus applying a learning scheduler such that the initial learning rate is set as 0.0002, while the polynomial decay of 0.9, momentum of 0.9 and weight decay of 0.0001 is also same as the discriminator which has 0.0001 learning rate trained on Adam optimizer. The model was trained on the PASCAL VOC dataset for 20k iterations at a batch size of 10 and 40k iterations with a batch size of 2 on the CamVid dataset, running on an Nvidia Geforce 1080Ti Graphics Card. As the training continues, the evaluation loss is minimized and the model continues to converge.

TABLE 1. Comparison of results performance on the PASCAL VOC 2012 test dataset.

Method	mean IOU
WAILS [42]	55.9
PSP-CRF [43]	65.4
Semi-weakly [44]	65.8
FCN-8 [23]	67.2
DeepLab1 [26]	71.6
SmallFOV-light [38]	72.0
GCRF [45]	73.2
Dilation10 [25]	73.9
DPN [46]	74.1
Piecewise [47]	75.3
Baseline	74.9
Ours	75.6

D. EVALUATION

To analyze the efficiency of our ACGAN model, we compare the model's semantically segmented results to the image labels and their degree of correlation. Using our attentively conditioned GAN, we train the model to generate images corresponding to the real training data label. In evaluating the results, we compared the intersection over union of the labels and the generated images. We adequately juxtaposed the proposed model with other schemes to validate its quality alongside other state-of-the-art methods using adv-seg [37] without semi-supervised learning as the baseline.

1) RESULTS ON THE PASCAL VOC DATASET

As depicted in Table 1, the proposed model exceeds existing models for the task of segmentation, meaning that the predicted class label classification is more resembling of the ground truth label. From Table 1, it indicates that our GAN model is able to learn the distribution space of the classes and able to classifier each object or feature to the right cluster with a high assurance.

Compared to other state-of-the-art which did not have an adversarial training, our model shows better spatial consistency between label classes and obtain an improvement from 67.2% to 75.6% compared to FCN [23], and an improvement of 2.25% against Dilation10 [25]. In comparison with the saints that are trained on adversarial learning, our proposed model records an increase from 65.8% to 75.6% against Semi-weakly [42] and 2.25% against SmallFOV-light [38]. In essence, the shows the advantage of our adversarial model structure and the influence of the attended conditional parameter included in recognizing patterns.

In Fig. 2, an illustration of the proposed model is depicted compared to the labels and other models such as DeepLabV3 [2], MobileNet [5], G-FRNet [28] and ShelfNet [30], revealing the efficiency of the model. The

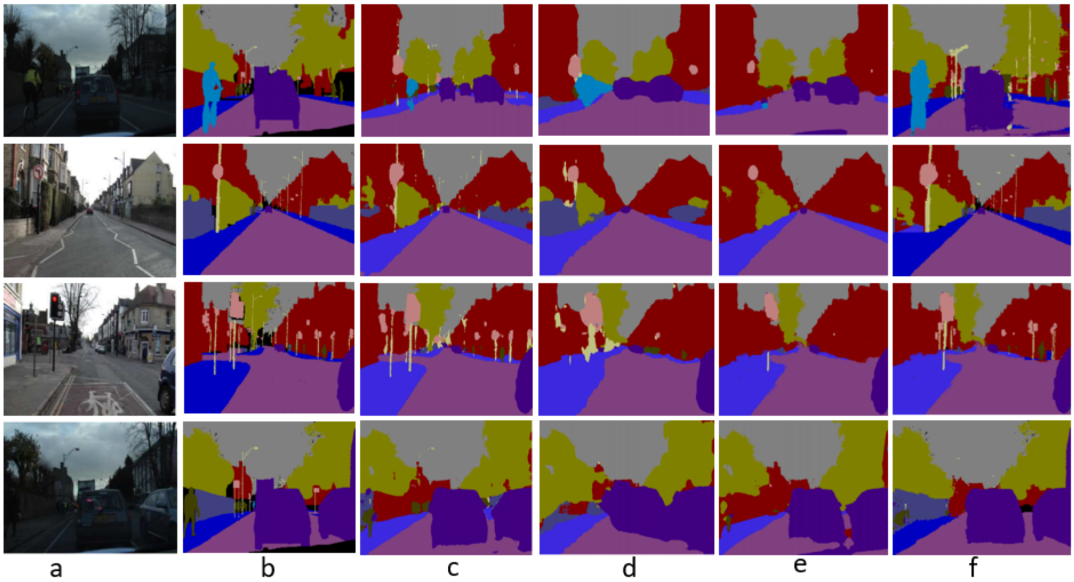

FIGURE 3. Qualitative results of segmentation on the CamVid dataset. Where a, b, c, d, e and f represents the input image, ground truth label, Segnet [24], FCN-Deconv [24], DeepLab-LargeFOV-denseCRF [24] and our ACGAN model result respectively.

TABLE 2. Comparison of results performance on the CamVid test dataset.

Method	mean IOU
ENet [48]	51.3
SegNet [24]	55.6
LinkNet [49]	55.8
FCN-8 [23]	57.0
ReSeg [50]	58.8
AttentionM [51]	60.1
LRN [52]	61.7
RTA-MC [53]	62.4
DeepLav-V2 [26]	65.2
Dilation10 [25]	65.3
Baseline	65.1
Ours	66.3

inclusion of the attention network is negligible in both time and computational cost, therefore estimating the adeptness of our model. Qualitatively, our ACGAN model depicted in Fig. 2 indicates that most of the objects are properly identified and segmented as close as possible to the labels. Meaning that pixels assignment is mostly accounted for in their right classes, reducing the false positives and false negatives of the model prediction compared to other approaches.

2) RESULTS ON THE CamVid DATASET

Table 2 shows the estimation of our proposed model on the CamVid dataset which achieves a performance improvement of 14% and 6.9% compared to FCN [23] and LRN [43] respectively. This shows the ability of our segmentor to generate images similar to the label truth, being trained via the adversarial objective function such that it maximizes its learning via the discriminator loss and converges its data distribution to the label distribution. Also, our model shows improved performance of 1.7% to DeepLabv2 [26] and 1.5% to Dilation10 [25]. In addition, Fig. 3 depicts clear vindication of our model in comparison with the result from other models.

The adversarial training significantly boosts the pixel classification, invariably learning the data sample distribution as close to the class distribution as possible and is able to decide which class each pixel belongs to, showing spatial consistency between the data curves and pattern. This, in turn, ensures the discriminator's task becomes more difficult and correspondingly improves the quality of the model result and strengthens the class. As displayed in Fig. 3, the model produces segmented images of the constituent objects compared to the ground truth.

This shows that the loss function of the model is well optimized to minimizing the loss of the segmentor such that its outputs are similar to the ground truth, making it difficult for the discriminator to distinguish between the predicted feature maps and the ground truth. It is fascinating to see that our model generates improved segmentation of object trough adversarial learning by learning the sample space of the data which would be resourceful for other transfer learning models in computer vision. Conclusively, it could be deduced that our ACGAN model benefited from the dual conditioned approach of the discriminator to enhance the learning process of the network compared to other methods.

E. ABLATION ANALYSIS

Without altering the model structure, parameters and hyperparameters, we remove the attention network of our work and the conditioned class label vector of the ground truth, meaning that the network becomes strictly adversarial training with no conditioned or additional input. The model then becomes the baseline which is closely similar to [37] without the semi-supervised training.

For the PASCAL VOC 2012 dataset, our model produced performance improvement of 0.9% from 74.9% to 75.6%. Evidently, the conditioned GAN which includes an attention generated vector as an additional input for both the segmentor network and the class vector for the ground truth label into the discriminator helps in identifying the object patterns and in its classification.

Also, on the CamVid dataset, the proposed model achieves a performance improvement of 1.8% from 65.1% to 66.3%, indicating that the segmentor can learn the ground truth sample space to look very similar to the labels, such that the discriminator cannot discriminate the labels from the segmentor's predictions.

Clearly, the extra inputs add to the information available to the discriminator which enhances feature identification and classification. This broadly improves the effectiveness and generalization of the discriminator's prediction, illustrating the importance of conditioned adversarial training in multiclass task which eliminates class mismatching and has improved GAN's application especially in image style transfer and transformation.

V. CONCLUSION

In this paper, we propose an Attentively Conditioned Generative Adversarial Network (ACGAN), which cleverly builds on the logic of the Generative Adversarial Network (GAN) to achieve a supervised learning problem for the task of semantic segmentation using two sub-models; the segmentor (segmentation model) and discriminator. Our proposed network learns via a conditional adversarial network such that, on the one hand, an additional input A_{att}, which represents attended feature probability of the segmentor's feature maps X is incorporated as input $p(X|A_{att})$ to the discriminator, and on the other hand, the second input to the discriminator which is the ground truth Y is conditioned on a vector V of the class label as $p(Y|V)$.

By experimenting on the PASCAL VOC 2012 and the CamVid dataset, ACGAN demonstrates its efficiency and effectiveness by generating plausible segmented images and shows an improve segmentation accuracy, as well as stabilizing the discriminator in modeling long-range dependencies due to extra information provided to the network.

REFERENCES

[1] X. Zhang, D. Zhang, T. Li, and S. Yu, "Application of heuristic search algorithm in sub-pixel image segmentation," in *Proc. 5th Int. Conf. Syst. Inform. (ICSAI)*, Nov. 2018, pp. 710–716.

[2] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[3] R. Meng, S. G. Rice, J. Wang, and X. Sun, "A fusion steganographic algorithm based on faster R-CNN," *CMC, Comput., Mater. Continua*, vol. 55, no. 1, pp. 1–16, 2018.

[4] J. Zhang, X. Jin, J. Sun, J. Wang, and A. K. Sangaiah, "Spatial and semantic convolutional features for robust visual object tracking," *Multimedia Tools Appl.*, pp. 1–21, Aug. 2018, doi: 10.1007/s11042-018-6562-8.

[5] M. Sandler, A. Howard, M, Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," Jan. 2018, *arXiv:1801.04381*. [Online]. Available: https://arxiv.org/abs/1801.04381

[6] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," 2018, *arXiv:1809.07294*. [Online]. Available: http://arxiv.org/abs/1809.07294

[7] Y. F. Tu, Y. Lin, J. Wang, and J.-U. Kim, "Semi-supervised learning with generative adversarial networks on digital signal modulation classification," *Comput., Mater. Continua*, vol. 55, no. 2, pp. 243–254, 2018.

[8] J. Liu, Y. Li, X. Tian, A. Sangaiah, and J. Wang, "Towards semantic sensor data: An ontology approach," *Sensors*, vol. 19, no. 5, p. 1193, Mar. 2019.

[9] J. Liu, C. Gu, J. Wang, G. Youn, and J.-U. Kim, "Multi-scale multi-class conditional generative adversarial network for handwritten character generation," *J. Supercomput.*, vol. 75, no. 4, pp. 1922–1940, Apr. 2019.

[10] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. NIPS*, 2017, pp. 5767–5777.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, p. 2672–2680.

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[13] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. ICML*, 2016, pp. 1060–1069.

[14] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[15] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[16] J. Wu, C. Zhang, T. Xue, W. Freeman, and J. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling," in *Proc. NIPS*, 2016, pp. 82–90.

[17] L. Gatys, A. Ecker, and M. Bethge, "A neural algorithm of artistic style," *Nature Commun.*, vol. abs/1508.06576, 2015. [Online]. Available: http://arxiv.org/abs/1508.06576 and https://dblp.org/rec/journals/corr/GatysEB15a.bib

[18] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434

[19] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[20] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*. [Online]. Available: http://arxiv.org/abs/1710.10196

[21] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: https://arxiv.org/abs/1701.07875

[22] T. Salimans, H. Zhang, A. Radford, and D. Metaxas, "Improving GANs using optimal transport," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–13.

[23] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, p. 3431–3440.

[24] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[25] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122

[26] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*. [Online]. Available: https://arxiv.org/abs/1412.7062

[27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[28] M. A. Islam, M. Rochan, N. D. B. Bruce, and Y. Wang, "Gated feedback refinement network for dense image labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4877–4885.

[29] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1377–1385.

[30] J. Zhuang, J. Yang, L. Gu, and N. Dvornek, "ShelfNet for fast semantic segmentation," Sep. 2019, *arXiv:1811.11254*. [Online]. Available: https://arxiv.org/abs/1811.11254

[31] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[32] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[33] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," 2018, *arXiv:1809.02983*. [Online]. Available: http://arxiv.org/abs/1809.02983

[34] H. Zhang, H. Zhang, C. Wang, and J. Xie, "Co-occurrent features in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 548–557.

[35] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. Loy, D. Lin, and J. Jia, "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 267–283.

[36] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.

[37] W. Hung, Y. Tsai, Y. Liou, Y. Lin, and M. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *Proc. BMVC*, 2018, pp. 1–17.

[38] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," 2016, *arXiv:1611.08408*. [Online]. Available: http://arxiv.org/abs/1611.08408

[39] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[40] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 991–998.

[41] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Marseille, France, Oct. 2008, pp. 44–57.

[42] H. Zhou, K. Song, X. Zhang, W. Gui, and Q. Qian, "WAILS: Watershed algorithm with image-level supervision for weakly supervised semantic segmentation," *IEEE Access*, vol. 7, pp. 42745–42756, 2019.

[43] L. Zhang, H. Li, P. Shen, G. Zhu, J. Song, S. A. A. Shah, M. Bennamoun, and L. Zhang, "Improving semantic image segmentation with a probabilistic superpixel-based dense conditional random field," *IEEE Access*, vol. 6, pp. 15297–15310, 2018.

[44] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5689–5697.

[45] C. Han, Y. Duan, X. Tao, and J. Lu, "Dense convolutional networks for semantic segmentation," *IEEE Access*, vol. 7, pp. 43369–43382, 2019.

[46] G. Lin, C. Shen, A. V. D. Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3194–3203.

[47] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*. [Online]. Available: http://arxiv.org/abs/1606.02147

[48] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, p. 1–4.

[49] F. Visin, A. Romero, K. Cho, M. Matteucci, M. Ciccone, K. Kastner, Y. Bengio, and A. Courville, "ReSeg: A recurrent neural network-based model for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 41–48.

[50] L. Fan, W.-C. Wang, F. Zha, and J. Yan, "Exploring new backbone and attention module for semantic segmentation in street scenes," *IEEE Access*, vol. 6, pp. 71566–71580, 2018.

[51] M. A. Islam, S. Naha, M. Rochan, N. Bruce, and Y. Wang, "Label refinement network for coarse-to-fine semantic segmentation," 2017, *arXiv:1703.00551*. [Online]. Available: http://arxiv.org/abs/1703.00551

[52] P.-Y. Huang, W.-T. Hsu, C.-Y. Chiu, T.-F. Wu, and M. Sun, "Efficient uncertainty estimation for semantic segmentation in videos," in *Proc. ECCV*, 2018, pp. 520–535.