Unsupervised and semi-supervised learning with Categorical Generative Adversarial Networks assisted by Wasserstein distance for dermoscopy image Classification

Xin Yi, Ekta Walia, Paul Babyn

***Abstract*—Melanoma is a curable aggressive skin cancer if detected early. Typically, the diagnosis involves initial screening with subsequent biopsy and histopathological examination if necessary. Computer aided diagnosis offers an objective score that is independent of clinical experience and the potential to lower the workload of a dermatologist. In the recent past, success of deep learning algorithms in the field of general computer vision has motivated successful application of supervised deep learning methods in computer aided melanoma recognition. However, large quantities of labeled images are required to make further improvements on the supervised method. A good annotation generally requires clinical and histological confirmation, which requires significant effort. In an attempt to alleviate this constraint, we propose to use categorical generative adversarial network to automatically learn the feature representation of dermoscopy images in an unsupervised and semi-supervised manner. Thorough experiments on ISIC 2016 skin lesion challenge demonstrate that the proposed feature learning method has achieved an average precision score of 0.424 with only 140 labeled images. Moreover, the proposed method is also capable of generating real-world like dermoscopy images.**

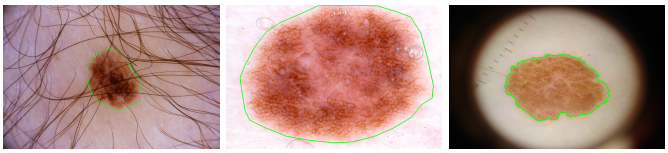
***Index Terms*—Dermoscopy, Categorical Generative Adversarial Networks, Unsupervised learning, Semi-supervised learning, Deep Learning , Melanoma classification**

I. Introduction

Skin cancer is the most prevalent cancer in Canada, with the incidence number almost equal to the four major cancers (lung, breast, colorectal, prostate) combined. Melanoma, as one of the two major skin cancer types, is the most deadly form with a 5-year survival rate of about 14% if detected late [22]. According to the Canadian Cancer Statistics 2014, the incidence rates of melanoma has increased by 2.05% per year for both sex combined between 1992 to 2013 [55]. An estimate of nearly 7300 people are expected to be diagnosed with melanoma and over 1200 are expected to die in 2017 [56]. This cancer arises when the pigment containing cells named melanocytes start to multiply without control and form malignant tumours. Despite its aggressiveness, this cancer is curable if detected early with a 5-year survival rate over 99% [22]. Therefore, a convenient and accurate method for early diagnosis of melanoma is of great practical impact.

Both X. Yi and P. Babyn are with the Department of Medical Imaging, University of Saskatchewan, Saskatoon, SK, S7N 0W8 Canada

E. Walia was with University of Saskatchewan at the time of inception of this work, and at present working with Philips Canada



Hair Air bubble Ruler

Fig. 1: Exemplar dermoscopy images with artifacts such as hair, air bubble and ruler.

Dermoscopy is a non-invasive skin imaging technique that has been widely adopted by dermatologists for the initial screening of skin disease. Utilizing a high magnification factor, it can minimize skin reflection interference, offering a better view of the skin lesion as compared to naked eyes [50]. However, the median wait time to see a dermatologist is over three months in Canada [4]. It would be of great value to develop an automated melanoma recognition system based on dermoscopy images to produce an accountable screening result in a more timely manner. The significance of automated image analysis and recognition for identification of melanoma has been highlighted in recent publications [23, 57, 40, 3].

There are a couple of challenges in the development of a dermoscopy image based automation system. First, there is no discernible boundary between the normal and lesion skin which makes the segmentation of skin lesions difficult, let alone the artifacts that can be seen in the image, e.g. hair, bubble and ruler. Sample images with these three artifacts can be seen in Figure 1. Another big problem is the large intra-class variation (among melanomas) and small inter-class variation (melanoma v.s. benign) as demonstrated in Figure 2. The skin lesions are segmented from the surrounding region to give a better illustration.

Recently, supervised learning with Convolutional Neural Networks (CNN) has provided excellent results for classification of skin cancer [22]. Dermotologist-level performance was claimed with 129,450 clinical images, including 3,374 labeled dermoscopy images used in the training. However, the authors stated that their method is constrained by data and the application can be extended to more visual conditions if sufficient labeled training examples exist. Similar claim has been made in [3] where Ali et al. stated that the method they presented is constrained by data and needs sufficient training examples to succeed in the classification of various

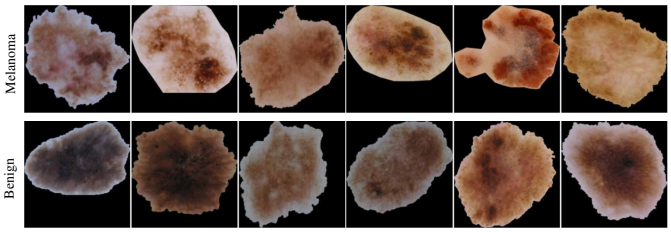


Fig. 2: Sample melanoma and benign images. Background is segmented to only show the skin lesion.

skin conditions.

Progressive performance seems to be straightforward by simply acquiring more labeled images. However, obtaining large volumes of labeled training data is time consuming and requires extensive expertise. In natural image field, we could rely on Amazon Turk to outsource the labeling task but this only suffices for general object categories. Fine-grained image labeling such as differentiating bird species would also require domain specific knowledge, not to mention dermoscopy images that require years of medical training and sometimes demand histopathology consultation. Moreover, privacy issues would further limit the number of qualified candidates.

Nonetheless, there is substantial amount of unlabeled images that have not been utilized, and their acquisition is relatively easier and inexpensive. Therefore, in this work, we explore unsupervised and semi-supervised learning techniques for dermoscopy image classification, in particular melanoma classification. Our proposed method combined categorical generative adversarial network (catGAN) [53] and Wasserstein distance (also known as earth mover's distance in computer science) [27] for its automatic feature learning. Evaluation on the ISIC skin lesion challenge 2016 dataset has shown promising results. We refer the proposed method as catWGAN and its detailed description can be found in Section III. Note that in this work, we mainly assess the feature learning capability of the network. As shape statistics of skin lesion plays an important role in the decision making of dermatologist, the skin lesions were segmented with ground truth segmentation maps to avoid interference of backgrounds and possible confusion that could be introduced by various segmentation algorithms.

II. Related works

High dimensional image data is usually assumed to lie on a lower dimensional manifold where the original data can be projected to get a more compact feature representation [13]. In automatic classification systems, this change of data representation can not only save computation, but also makes the system robust to potential transformations encountered in the real-world, such as illumination, rotation, scale change, and translation. Based on the strategies of how the features are designed, we broadly categorize the existing literature into two groups, one centred on use of hand-crafted features and the other focused on automatically learnt features.

A. Hand-crafted features

In the pre-deep learning era, hand-crafted features based on shape [24, 12], color [54, 14, 9] and texture [6, 39] of skin lesion played a key role in automatic melanoma classification systems. The features used are very broad and are generally adopted from the traditional computer vision literature. For example, Garnavi et al. used a wavelet decomposition based texture feature extraction method in conjunction with other geometric and border based features of the lesions [25]. Jafari et al. combined asymmetry, colour and border assessment features with a Support Vector Machine (SVM) classifier to automate melanoma detection [32]. Integration of local and global features is also exploited in some computer-aided diagnosis systems [10, 7]. Among various ways of fusing features from different sources, bag of visual word (BoW) is probably the most popular method to aggregate the low-level features into so-called mid-level features which are more robust to image variations. This has been adopted in many dermoscopy image classification works [8, 2, 52]. A more in-depth review of these traditional techniques can be found in [43].

B. Automatically learned features with deep neural networks

Recently, deep learning based algorithms have quickly dominated most vision based tasks. These advances have been quickly brought into the field of computer aided diagnosis, including retinopathy [28], breast cancer diagnosis [18], pulmonary nodules detection [15] and the focus of our work, melanoma image classification.

1) Supervised learning:: Previous deep learning work in melanoma classification has prevailingly used fully supervised learning for feature extraction and then attach a classifier such as random forest on top for classification. Constrained by the fact that not enough labeled samples are available to fully train the neural network, a large portion of related works use a transfer learning scheme by fine-tuning a pre-trained neural network. The underlying assumption is that the cascade level of features learned with natural images could also be beneficial for medical images especially those features learnt in the first few layers of the network which are mainly edges and some other simple image structures. The training could then focus on the deeper layers leaving the shallow layers untouched.

Codella et al. extracted features from a pre-trained CNN and further combined traditional sparse coding features for melanoma recognition [16]. Liao fine-tuned three different pre-trained CNNs (VGG15, VGG19, GoogleNet) for universal skin disease classification [36]. The effectiveness of transfer learning was also manifested in other similar works that adopts the same fine-tuning strategy [37, 59, 29].

2) Unsupervised and semi-supervised learning:: Unsupervised and semi-superviesd learning method is not commonly used in this field mostly due to its modest performance. The only work we have found, uses a stacked sparse autoencoder to learn hierarchical level of features for classification of skin lesion images [47]. However, we believe unsupervised and semi-supervised methods will play an important role in solving medical imaging problems due to the scarcity of labeled

medical datasets. Here we briefly review contemporary deep learning based semi-supervised learning methods that are not only scalable to large quantities of unlabeled images but are also capable of simultaneously performing the unsupervised and supervised learning task (opposed to unsupervised pre-training followed by supervised fine-tuning). For a more in-depth review of traditional semi-supervised learning methods, we refer the reader to these two works [49, 13].

Recent methods typically involve training of a feed-forward classifier together with some auxiliary unsupervised tasks, hoping the learnt features would generalize better. The most common unsupervised tasks includes minimizing reconstruction error of inputs [30, 34, 38] or learnt intermediate representations [46], encourage model invariance of input data perturbations [21, 48, 41], or some ways of data embedding [45, 58].

Generative modelling, which is a branch of unsupervised learning, has seen rapid progress during the last several years. Generative adversarial network (GAN) [26], in particular, has received attention due to its capability of generating synthetic real-world like samples. The extension of GAN into semi-supervised learning has achieved state-of-the-art results on CIFAR10 [20, 41], MNIST [53] and SVHN [19, 35]. In this work, we explored one type of generative model, named categorial GAN (catGAN) for unsupervised and semi-supervised learning. catGAN is a generalization of the GAN model to multiple classes where the adversarial loss used involves information maximization. Since the original catGAN training is not stable, we further adopted the Wasserstein distance for assistance and we refer the proposed approach as catWGAN. We have shown that with only 140 labeled samples, the learned feature outperforms simple hand-crafted features and baseline denoising autoencoder by a large margin. The model can also synthesize real-world like dermoscopy images (Figure 9) that could potentially be further used as a source of data augmentation and potentially for training of dermatologists. As far as we know, this is the first work that applied GAN based semi-supervised learning on melanoma classification.

III. Methodology

The traditional GAN is a generative model that implicitly estimates the sample distribution so that we can directly sample from the model. It consists of two types of networks: the generator G that generates synthetic samples from pure noise and the discriminator D_1 that differentiates between real and generated samples. G and D_1 update themselves alternatively during the training process with contradictory objective. catGAN adopts the general framework of GAN but modifies the objective of D_1 in a way that rather than classifying the input sample as "real" or "fake", it outputs confidence values of input belonging to each one of the underlying classes. Figure 3 following the orange line illustrates this process.

Let $G : \mathbb{R}^d \rightarrow \mathbb{R}^{m \times n}$ denote a mapping from random noise $z \sim p(z)$ to a generated image sample $\hat{x}$, and $D_1 : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^2$ denote a mapping from an input sample image to its predicted label distribution y. The input could be either real sample $x \sim p_r(x)$ or generated fake sample $\hat{x} \sim p_g(\hat{x})$. In unsupervised setting, the performance of D_1 was measured by the peakedness of the output label distribution using entropy. The overall objective of D_1 and G in catGAN formulation can be expressed mathematically as:

$$\mathcal{L}_{D_1}^{\text{catGAN}} = \max_{D_1} H_{x\sim p_r(x)}[p(y \mid D_1)] - \underbrace{\mathbb{E}_{x\sim p_r(x)}\big[H[p(y \mid x, D_1)]\big]}_{S_r} + \underbrace{\mathbb{E}_{z\sim p(z)}\big[H[p(y \mid G(z), D_1)]\big]}_{S_g}, \tag{1}$$

$$\mathcal{L}_{G}^{\text{catGAN}} = \min_{G} -H_G[p(y \mid D_1)] + \underbrace{\mathbb{E}_{z\sim p(z)}\big[H[p(y \mid G(z), D_1)]\big]}_{S_g}, \tag{2}$$

where $H_{x\sim p_r(x)}[p(y \mid D_1)]$ and $H_G[p(y \mid D_1)]$ is the entropy of the marginalized class distribution over real and generated samples respectively. These two entropies were maximized to ensure equal usage of samples from both classes. The second term in $\mathcal{L}_{D_1}^{\text{catGAN}}$ is the estimated entropy of the predicted class distribution over real samples, that D_1 tries to minimize. The S_g term as appearing in both $\mathcal{L}_{D_1}^{\text{catGAN}}$ and $\mathcal{L}_{G}^{\text{catGAN}}$ is the estimated entropy of the predicted class distribution over generated samples, over which D_1 tries to maximize and G tries to minimize. S_r and S_g are used to denote these two terms for future reference.

Similar to traditional GANs, stabilizing the training of catGAN so that neither G or D_1 is overpowered by the other is a significant issue, as already pointed out in the original catGAN paper. The generator would stop improving when the discriminator becomes too strong, where the loss of the discriminator gets saturated and leads to zero gradients for updating G. Although by adopting the DCGAN architecture [44], the likelihood of model collapse decreased a lot, we did observe this unstable phenomenon from time to time with different initializations. Therefore, to cope with this problem, we further employed a second discriminator D_2 with Wasserstein distance [5] for assistance as shown in Figure 3 following the red line.

Springenberg has compared the catGAN formulation to the regularized information maximization (RIM) framework, and found that the generator of catGAN can be thought of as an adaptively learned regularizer for its discriminator [53]. Under this perspective, the better the generated sample becomes, the more robust the D_1 becomes to the adversarial samples. This constitutes another part of the motivation of the integration of the second discriminator. When D_1 failed to supply enough gradients to update G, D_2 will still offer gradients to help G catch up.

The objective of D_2 is to differentiate between real and generated fake samples as in the traditional GAN without worrying about the composition of underlying classes. Traditional GANs minimize a f-divergence between the real data distribution and the generated data distribution [26, 42]. Since f-divergence is a function of the density ratio, it would either become zero or infinite when the support of the two distributions do not overlap. Using Wasserstein distance mitigates this problem by assuming a Lipschitz constraint on

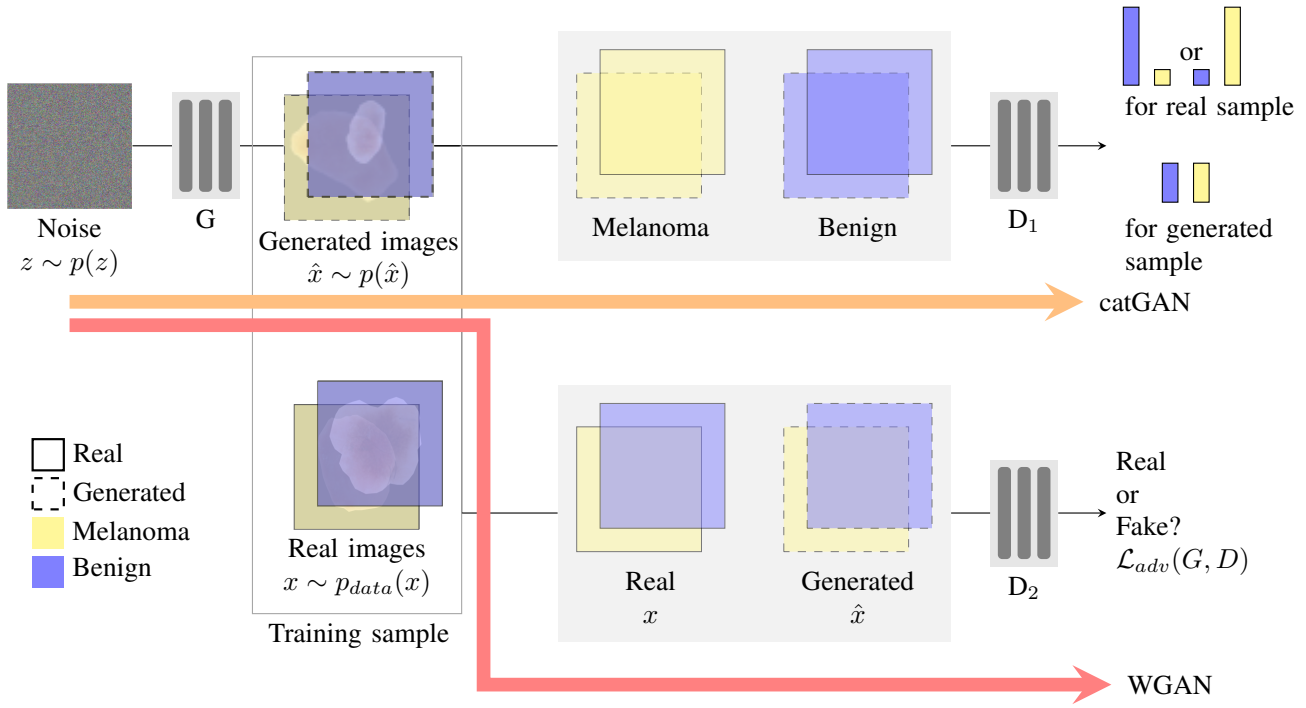


Fig. 3: Overview of catWGAN. G is the generator that is responsible for synthetic sample generation. The output D_1 are confidence values for the two classes we are interested in, melanoma and benign. The orange line is used to depict the catGAN architecture and the red line depicts the WGAN architecture used to assist the training of catGAN.

the discriminator[1]. Gradient penalty was employed here for the training of D_2, as it was shown to be beneficial for the training of WGAN with various architectures [27].

The loss can be expressed as follows in the WGAN formulation

$$\begin{aligned}\mathcal{L}_{D_2}^{\text{WGAN}} = \max_{D_2} \; & -\mathbb{E}_{x\sim p_r(x)}[(D_2(x)] + \mathbb{E}_{\hat{x}\sim p_g(\hat{x})}[D_2(\hat{x})] \\ & - \lambda \mathbb{E}_{\hat{x}\sim p_g(\hat{x})}[(||\nabla_{\hat{x}} D_2(\hat{x})||)^2],\end{aligned} \tag{3}$$

$$\mathcal{L}_G^{\text{WGAN}} = \min_G \; \mathbb{E}_{z\sim p(z)}[D_2(G(z))], \tag{4}$$

where λ is the weight of the gradient penalty. Combining equation (1) to (4) into the same framework, we have the full unsupervised objective in our case as:

$$\begin{aligned}\mathcal{L}_{\text{unsuper}}(G, D_1, D_2) = \min_G \max_{D_1, D_2} \; & \mathcal{L}_G^{\text{catGAN}} + \alpha \mathcal{L}_G^{\text{WGAN}} \\ & + \mathcal{L}_{D_1}^{\text{catGAN}} + \mathcal{L}_{D_2}^{\text{WGAN}},\end{aligned} \tag{5}$$

where α weights the influence of D_1 and D_2 to G. The negative of the first two terms of equation 3 is the Wasserstein distance between the generated distribution and the real sample distribution.

[1]In some publications, this is called critic because the output is no longer a confidence value of real or generated sample but a real number. We stick to the name of discriminator for the sake of consistency in this paper.

The extension to semi-supervised training is straightforward by incorporating the cross entropy (CE) loss for the labeled samples in equation 1, so that the loss for D_1 becomes:

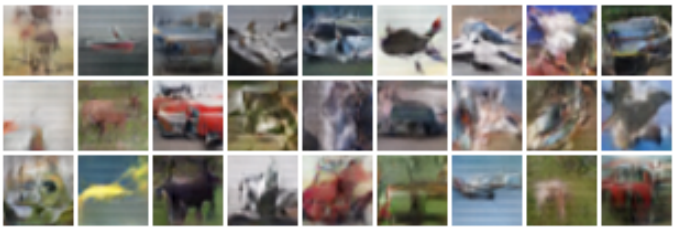
$$\begin{aligned}\mathcal{L}_{D_1}^{\text{catGAN}} = \max_{D_1} \; & H_{x\sim p_{(x)}}[p(y \mid D_1)] - \underbrace{\mathbb{E}_{x\sim p_r(x)}\big[H[p(y \mid x, D_1)]\big]}_{S_r} \\ & + \underbrace{\mathbb{E}_{z\sim p(z)}\big[H[p(y \mid G(z), D_1)]\big]}_{S_g} \\ & + \lambda \mathbb{E}_{(x,\hat{y})\sim \mathcal{X}^l}\big[CE\big[\hat{y}, p(y|x, D_1)\big]\big],\end{aligned} \tag{6}$$

where $\mathcal{X}^l$ is the set of labeled samples $\{(x^1, \hat{y}^1), (x^2, \hat{y}^2)\} \cdots (x^l, \hat{y}^l)\}$.

A. Network architectures

The architectures described below were used to generate images of size 64×64. For the catGAN part (G and D_1), we found the original architecture tends to produce low contrast images. Therefore, instead of using the original architecture from the catGAN paper, we customized with some key modifications. The detailed architecture can be found in Table I. Compared with catGAN's original architecture, the proposed architecture substituted all pooling layers with stride convolution. Compared with DCGAN's architecture, we employed leaky Relu instead of Relu for the generator to avoid dead gradients and batch normalization layer was inserted after every convolution layer. Also, the feature dimension of the discriminator was further compressed from $512 \times 4 \times 4$ to $512 \times 1 \times 1$ and a 2-way softmax layer was built on top to produce the confidence values of the two classes. Figure 4 shows the difference

(a)



(b)

Fig. 4: Generated images on CIFAR 10. (a) is the generated result from the original catGAN architecture (directly cropped from the original paper). (b) shows the generated results from the proposed catGAN architecture (no D_2 in this experiment).

between the image generated from the original and proposed catGAN architecture on CIFAR10 by just using the catGAN formulation without Wasserstein distance.

For D_2, we chose to restrict it to model high frequency structures as inspired by [31]. We found that by using D_1 alone, G could produce general structures but sometimes struggled in producing high frequency details, especially in generating high spatial resolution images. Therefore, it is reasonable to let D_2 focus on local image patches. The architecture of D_2 can be seen in Table I. It is a three layer fully convolutional network. The output of D_2 is an average over all responses. An advantage of this architecture is that it is suitable for arbitrary input sizes and has fewer parameters. The patch size in our case is 22×22.

B. Baseline methods

We chose another unsupervised learning algorithm named denoising autoencoder (DAE) as the baseline for comparison. A DAE is one type of neural network that learns to reconstruct a noise corrupted input. Similar to catGAN, DAE also consists of two networks, an encoder and a decoder. The encoder's job is to transform the input to a more compact feature representation that has smaller dimension than the input and the decoder reconstructs the input from this compressed representation. The resultant feature representation should preserve all necessary information needed for reconstruction.

The encoder adopts D_1's architecture with the last layer chopped off and the decoder reverses the encoder's operation accordingly to output a reconstructed image of size 64×64. In this manner, the learnt feature representation would have the same dimension as that of the proposed catWGAN for a fair comparison.

To demonstrate the effectiveness of the proposed method, we also compared two simple hand-crafted features that are commonly used in melanoma classification: edge histogram and color histogram [1, 17]

Generator (G)

Layer	Activation Size
Input noise	$z \in \mathbb{R}^{100}$
$512 \times 4 \times 4$ conv. +BN+lReLU	$512 \times 4 \times 4$
$256 \times 4 \times 4$ conv. stride $1/2$+BN+lReLU	$256 \times 8 \times 8$
$128 \times 4 \times 4$ conv. stride $1/2$+BN+lReLU	$128 \times 16 \times 16$
$64 \times 4 \times 4$ conv. stride $1/2$+BN+lReLU	$64 \times 32 \times 32$
4×4 conv. stride $1/2$+BN+lReLU	$3 \times 64 \times 64$
tanh	$3 \times 64 \times 64$

Discriminator 1 (D_1)

layer	Activation Size
Input image	$3 \times 64 \times 64$
$64 \times 4 \times 4$ conv. stride 2+BN+lReLU	$64 \times 32 \times 32$
$128 \times 4 \times 4$ conv. stride 2+BN+lReLU	$128 \times 16 \times 16$
$256 \times 4 \times 4$ conv. stride 2 +BN+lReLU	$256 \times 8 \times 8$
$512 \times 4 \times 4$ conv. stride 2+BN+lReLU	$512 \times 4 \times 4$
$512 \times 4 \times 4$ conv.+BN+lReLU	$512 \times 1 \times 1$
$2 \times 1 \times 1$ conv. +BN+lReLU	$2 \times 1 \times 1$
2-way softmax	$2 \times 1 \times 1$

Discriminator 2 (D_2)

layer	Activation Size
Input image	$3 \times 64 \times 64$
$64 \times 4 \times 4$ conv. stride 2	$64 \times 32 \times 32$
$128 \times 4 \times 4$ conv. stride 2	$128 \times 16 \times 16$
$256 \times 4 \times 4$ conv. stride 2	$256 \times 8 \times 8$
average	1

TABLE I: Architecture of the proposed catWGAN. The feature representation was extracted from the third to the last layer of D_1.

IV. Experiment setup

A. Datasets

Two datasets were used in this research. The first one is a fully annotated open access dataset from the International Symposium on Biomedical Imaging (ISBI) 2016 Skin Lesion challenge. This dataset is part of the International Skin Imaging Collaboration (ISIC) Archive, which is by far the largest publicly available dermoscopy image dataset. The 2016 challenge training dataset contains a total of 900 images with 173 melanomas and 727 benign cases. The test set is also released for analysis which consists of 75 melanomas and 304 benign cases. These are all 8-bit RGB colour images of varying spatial sizes. Segmentation masks were supplied by the organizer for the segmented lesion classification task.

We applied the ground-truth segmentation mask to extract the smallest square region that contains the lesion, and then resized it to 256×256 with bilinear interpolation. The images in the training set were further augmented with rotation (in

the range of [-180°, 180°]), horizontal and vertical flipping, and elastic transform [51] to further boost the dataset. The size of the dataset for unsupervised training is 20k (balanced 10k+10k). As for the semi-supervised training, 70 images were randomly selected from each class of the original training set and augmented to a size of 10k (5k+5k). Translation and scale changes (in the range of [0.3, 1.5]) were used along with the previously mentioned augmentation techniques in the semi-supervised case to alleviate overfitting. A python package named Augmentor [11] was used for the augmentation.

The second dataset is the PH2 [7]. It consists of a total of 200 dermoscopy images of melanocytic lesions, including 80 common nevi, 80 atypical nevi, and 40 melanomas. Since in this work, we only solve the problem of binary classification, for this dataset, melanoma images constitute a class depicting malignancy and remaining images of common nevi and atypical nevi constitute another class depicting benign cases. The skin lesion was also extracted using the method described above but without augmentation. The resultant images served as the validation dataset to select the best model during the training process.

B. Evaluation Metrics

Sensitivity (SE), specificity (SP), accuracy (AC), area under the receiver operating characteristic curve (AUC) and average precision (AP) are used for the evaluation as suggested by the ISBI 2016 challenge [29]. They are defined mathematically as follows:

$$\begin{aligned} \text{SE} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \\ \text{SP} &= \frac{\text{TN}}{\text{TN}+\text{FP}} \\ \text{AC} &= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \end{aligned} \tag{7}$$

where TP, TN, FP, FN represents the number of true positives, true negatives, false positives, false negatives respectively. AUC is defined as the integral of true positive rate (SE) with respect to the false positive rate (1-SP) under different thresholds. Similarly, AP is defined as the integral of precision with respect to recall under different thresholds. Precision and recall are expressed mathematically as:

$$\begin{aligned} \text{Precision} &= \frac{\text{T}P}{\text{TP}+\text{FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \end{aligned} \tag{8}$$

AP is used as the ranking metric for different methods because it is more sensitive to the change of TP. Note that there are different variants of the implementation of AP. To be consistent with the other works that also use the ISBI challenge dataset, we use the "scikit-learn" Python package for the computation of all the aforementioned metrics.

C. Implementation details

All the networks were trained on the Guillimin cluster of Calcul Québec. Adam optimizer [33] with $\beta_1 = 0.5$ and $\beta_2 = 0.9$ was used for all three networks with learning rate 0.0002. Batch size was chosen to be 200 to get a good estimate of the marginal entropy of the real and generated sample. D_2 was trained 5 times more often than G and D_1 to ensure the 1-Lipschitz assumption. λ and α were set to be 10 and 0.1. The implementation was based on the PyTorch framework. Training was stopped after 16K iterations.

D. Experiments

Three experiments were conducted to show the effectiveness of the proposed method. First, we monitored how the quality of our features evolves during the training of the catWGAN. We sampled the network every 50 iterations and extracted features from the third to last layer of D_1 and trained a linear SVM on top. PH2 dataset served as the validation dataset and 5-fold cross-validation was performed. Second, we performed horizontal comparison to the aforementioned baseline methods on the 2016 ISIC challenge test dataset. Lastly, we evaluated the image generated from the trained generator.

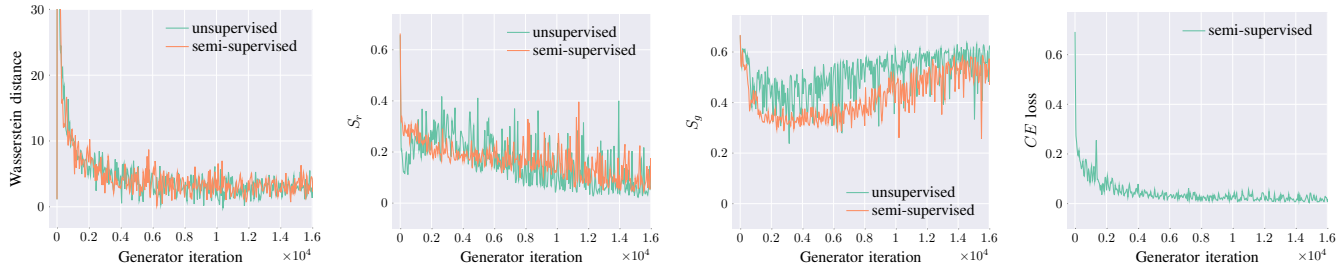
V. Results

Figure 5 shows the training loss statistics. (a) is the Wasserstein distance between the generated distribution and the real sample distribution and was shown to conform to the generated image quality [5] and we have found the same trend by examining the generated images at a series of checkpoints during the training as shown in Figure 6. (b) is the estimated entropy of the predicted class distribution over real samples (S_r) and (c) is the estimated entropy of the predicted class distribution over generated samples (S_g). The loss in (b) increases in the first 3000 iterations and then gradually decreases. The loss in (c) displays an opposite trend. (d) shows a gradually decreasing cross entropy loss for the labeled training samples in the semi-supervised learning.

A. Evolution of the feature representation during training

The effectiveness of the features learnt in unsupervised setting was evaluated using cross-validation on the PH2 dataset. Linear SVM was used and the results reported here are from stratified 5-fold cross-validation. We fixed c to be 1 in this experiment so that the iteration number was the only varying parameter to be validated. Furthermore, since the PH2 dataset is an unbalanced dataset with the amount of benign cases four times that of melanomas, the weights for melanoma samples were adjusted accordingly in the training.

As can be seen from Figure 7, for both unsupervised and semi-supervised training, the AC and AUC increases in the first 2000 iterations and then remains steady afterwards. The saturation of these two metrics is the result of large number of TN which makes small changes in TP less perceivable. AP on the other hand also exhibits similar trend in the very beginning but decreases and saturates to a lower value. On comparing Figure 7 (b) with Figure 5 (c), we can see that, the best AP was achieved when S_g is the smallest. Since the output of D_1 are two imaginary classes with no specific meaning, D_1 could have learnt something trivial to separate the input into two classes (e.g. symmetric v.s. asymmertric), which do not well align with the desired separation (melanoma v.s. benign).



(a) (b) (c) (d)

Fig. 5: Plot of the training loss during the unsupervised and semi-supervised training of catWGAN. (a) is negative of the first two terms of $\mathcal{L}_{D_2}^{\text{WGAN}}$ (Equation 3), representing the Wasserstein distance of the generated distribution and the real distribution. (b) is S_r and (c) is S_g as described in Section III (Equation 1). (d) is the CE loss for labeled samples.

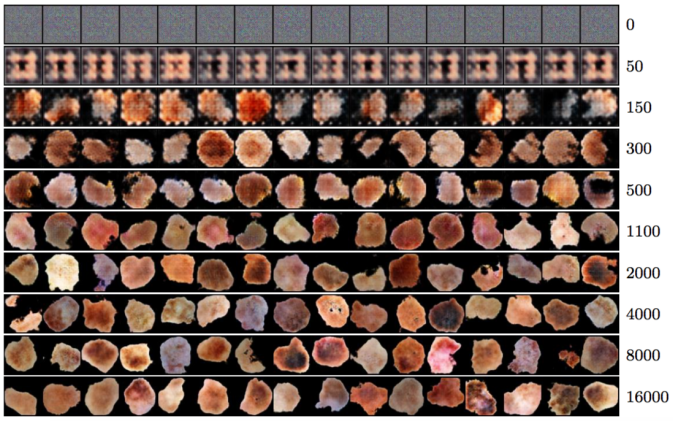


Fig. 6: Generated images from the generator during the unsupervised training of catWGAN. Numbers shown on the right are the generator iteration number. 16 images are shown at each checkpoint. Each image is of size 64×64. Semi-supervised catWAGAN's generated images (not shown here) exhibits similar trend.

Method	AP	AUC	AC
Edge Histogram	0.265	0.571	0.665
Color Histogram	0.36	0.626	0.789
DAE	0.329	0.634	0.794
catWGAN-unsuper (proposed)	0.351	0.613	0.812
catWGAN-semi (proposed)	0.424	0.690	0.81

TABLE II: Comparison to the baseline methods on the test set of the 2016 ISIC challenge dataset.

B. Comparison with baselines

With the validation results on PH2 dataset from Figure 7, we selected an ensemble of models that gives the best AP scores (around 2000 iterations) for the final testing on the 2016 ISIC test set. Results from different models are averaged to give the final result. Similar scheme was performed for DAE. Test results are shown in Table II. Unsupervised catWGAN performs slightly better than DAE in terms of AP and the performance of both methods is in between that of the edge and color histogram. Semi-supervised catWGAN achieves much better results than these four methods.

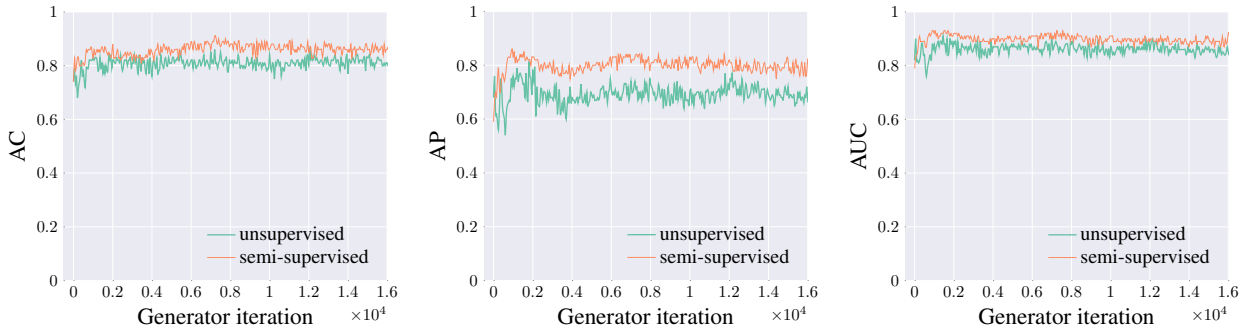
Figure 8 demonstrates the receiver operating characteristic (ROC) curve for both the unsupervised and semi-supervised catWGAN and the baseline DAE on the 2016 ISIC test set. We can see performance improvements at almost all operating points for the semi-supervised method over the unsupervised method.

VI. Discussion

From the unsupervised results of the first experiment, we can observe that even if the generator continues to improve the generated image, the features learnt from discriminator could oscillate instead of improving. It implicitly shows that without supervision, the network could not learn task dependent features, even if the objective of D_1 was to produce confidence values of two classes (expected to be melanoma and benign). Comparing the classification result of the unsupervised feature to the results of the edge and color histogram, we can see that both unsupervised catWGAN and DAE have their AP stuck in between the two simple hand-crafted features. This evidence further suggests the network might have learnt to separate the samples into two trivial classes, such as red vs orange, or symmetric vs asymmetric. Therefore, for images with variations in both color, shape and texture, unsupervised training might have limitations in capturing the desired task specific features. However, for images with distinct structural variations, like digits, unsupervised training will still be useful.

In this work, we have focused on the binary classification problem, but this method can be easily extended to multiple classes by replacing the last two layers of D_1 with the desired number of classes. Even in cases where the dataset contains images of only one semantic class, the output of D_1 can still be arbitrary number of imaginary classes, with each one corresponding to potential attributes.

There are some limitations of this work that we want to address. First, the generated images are only of a spatial size of 64×64 which makes the classification result (AP 0.424) not directly comparable to the state-of-the-art supervised melanoma classification methods (AP 0.624 [59]). We have applied our proposed method on larger images of size 128×128. Good looking images are generated as shown in Figure 9 but the features are not so robust. Effective architecture should be explored for D_1 in better learning of the feature representation in larger spatial resolution. Second,



(a) (b) (c)

Fig. 7: Performance of features learned by catWGAN in the process of training validated on PH2 dataset. From left to right shows the AC, AP, AUC respectively. Green line shows the result of unsupervised learning whereas orange line shows semi-supervised learning with 140 labeled images.

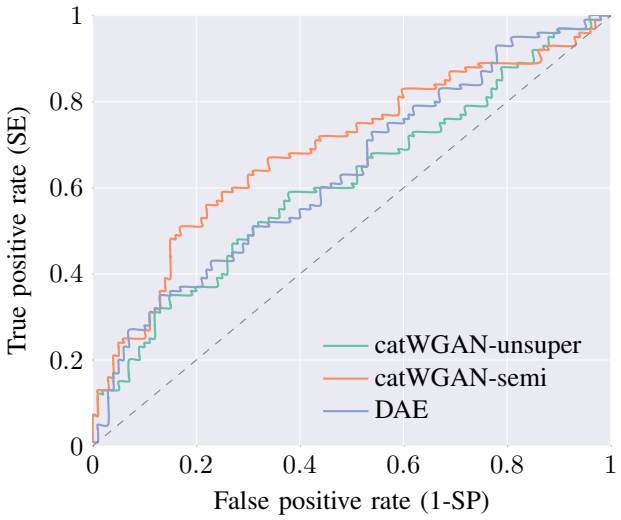


Fig. 8: ROC curves of the best proposed and baseline models on the 2016 ISIC challenge test dataset.

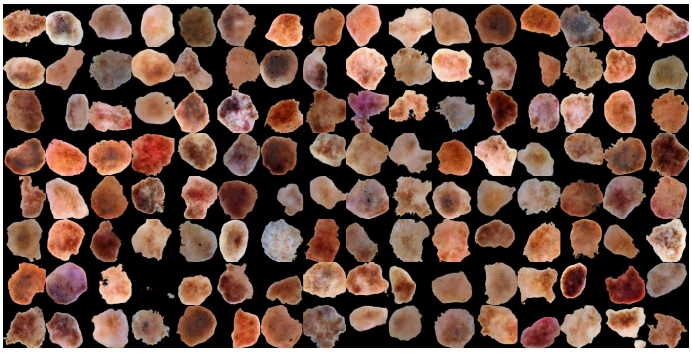


Fig. 9: Generated dermoscopy images of size 128×128 by using the proposed unsupervised method.

to make our results comparable with that of the other methods working on the 2016 ISIC challenge dataset, we restricted ourself on the 900 training images in the evaluation. Despite the fact that the size of unlabeled dataset used for training is augmented to 20k, the effective distinct number of images could limit the performance of the proposed method. In the future, we would like to use the full set 13,000 images of ISIC Archive as the unlabeled dataset and all the 900 labeled images in the semi-supervised setting to evaluate the full potential. Third, we believe explicit coordination between the supervised and unsupervised task should be explored in the future to ensure a more robust feature learning.

VII. Conclusion

In this work, we used categorial generative adversarial network assisted by Wasserstein distance for both unsupervised and semi-supervised learning. By just using 70 labeled samples from each class, the proposed model is able to learn a feature representation whose performance is much better than the denoising autoencoder and simple hand-crafted features. This demonstrates the efficacy of the proposed method and its application in many other medical image classification problems where feature representation is desired but limited labeled data is available. Another advantage of our proposed method is the ability to generate real-world like dermoscopy images with various shape, colour and surface texture.

VIII. Acknowledgement

This research was enabled in part by support provided by [2]Calcul Québec.

References

[1] Abedini M, Codella NC, Connell JH, Garnavi R, Merler M, Pankanti S, Smith JR, Syeda-Mahmood T (2015) A generalized framework for medical image classification and recognition. IBM Journal of Research and Development 59(2/3):1–1

[2] Alfed N, Khelifi F, Bouridane A (2016) Improving a bag of words approach for skin cancer detection in dermoscopic images. In: Control, Decision and Information Technologies (CoDIT), 2016 International Conference on, IEEE, pp 024–027

[2]http://www.calculquebec.ca/en/

[3] Ali ARA, Deserno TM (2012) A systematic review of automated melanoma detection in dermatoscopic images and its ground truth data. In: Proc. of SPIE Vol, vol 8318, pp 83,181I–1

[4] Alliance CSP (2012) Skin deep: a report card on access to dermatological care and treatment in canada

[5] Arjovsky M, Chintala S, Bottou L (2017) Wasserstein gan. arXiv preprint arXiv:170107875

[6] Ballerini L, Fisher RB, Aldridge B, Rees J (2013) A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions. In: Color Medical Image Analysis, Springer, pp 63–86

[7] Barata C, Ruela M, Francisco M, Mendonça T, Marques JS (2014) Two systems for the detection of melanomas in dermoscopy images using texture and color features. IEEE Systems Journal 8(3):965–979

[8] Barata C, Ruela M, Mendonça T, Marques JS (2014) A bag-of-features approach for the classification of melanomas in dermoscopy images: The role of color and texture descriptors. In: Computer vision techniques for the diagnosis of skin cancer, Springer, pp 49–69

[9] Barata C, Celebi ME, Marques JS (2015) Improving dermoscopy image classification using color constancy. IEEE journal of biomedical and health informatics 19(3):1146–1152

[10] Barata C, Celebi ME, Marques JS (2015) Melanoma detection algorithm based on feature fusion. In: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, IEEE, pp 2653–2656

[11] Bloice MD, Stocker C, Holzinger A (2017) Augmentor: An image augmentation library for machine learning. arXiv preprint arXiv:170804680

[12] Celebi ME, Iyatomi H, Schaefer G, Stoecker WV (2009) Lesion border detection in dermoscopy images. Computerized medical imaging and graphics 33(2):148–153

[13] Chapelle O, Scholkopf B, Zien A (2006) Semi-supervised learning. MIT Press

[14] Cheng YI, Swamisai R, Umbaugh SE, Moss RH, Stoecker WV, Teegala S, Srinivasan SK (2008) Skin lesion classification using relative color features. Skin Research and Technology 14(1):53–64

[15] Ciompi F, Chung K, Van Riel SJ, Setio AAA, Gerke PK, Jacobs C, Scholten ET, Schaefer-Prokop C, Wille MM, Marchianò A, et al (2017) Towards automatic pulmonary nodule management in lung cancer screening with deep learning. Scientific Reports 7

[16] Codella N, Cai J, Abedini M, Garnavi R, Halpern A, Smith JR (2015) Deep learning, sparse coding, and svm for melanoma recognition in dermoscopy images. In: International Workshop on Machine Learning in Medical Imaging, Springer, pp 118–126

[17] Codella N, Nguyen QB, Pankanti S, Gutman D, Helba B, Halpern A, Smith JR (2016) Deep learning ensembles for melanoma recognition in dermoscopy images. arXiv preprint arXiv:161004662

[18] Cruz-Roa A, Gilmore H, Basavanhally A, Feldman M, Ganesan S, Shih NN, Tomaszewski J, González FA, Madabhushi A (2017) Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. Scientific Reports 7:46,450

[19] Dai Z, Yang Z, Yang F, Cohen WW, Salakhutdinov R (2017) Good semi-supervised learning that requires a bad gan. arXiv preprint arXiv:170509783

[20] Donahue J, Krähenbühl P, Darrell T (2016) Adversarial feature learning. arXiv preprint arXiv:160509782

[21] Dosovitskiy A, Springenberg JT, Riedmiller M, Brox T (2014) Discriminative unsupervised feature learning with convolutional neural networks. In: Advances in Neural Information Processing Systems, pp 766–774

[22] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115–118

[23] Fornaciali M, Carvalho M, Bittencourt FV, Avila S, Valle E (2016) Towards automated melanoma screening: Proper computer vision and reliable results. arXiv preprint arXiv:160404024

[24] Friedman RJ, Rigel DS, Kopf AW (1985) Early detection of malignant melanoma: The role of physician examination and self-examination of the skin. CA: a cancer journal for clinicians 35(3):130–151

[25] Garnavi R, Aldeen M, Bailey J (2012) Computer-aided diagnosis of melanoma using border-and wavelet-based texture analysis. IEEE Transactions on Information Technology in Biomedicine 16(6):1239–1252

[26] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680

[27] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A (2017) Improved training of wasserstein gans. arXiv preprint arXiv:170400028

[28] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, et al (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama 316(22):2402–2410

[29] Gutman D, Codella NC, Celebi E, Helba B, Marchetti M, Mishra N, Halpern A (2016) Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:160501397

[30] Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. science 313(5786):504–507

[31] Isola P, Zhu JY, Zhou T, Efros AA (2016) Image-to-image translation with conditional adversarial networks. arXiv preprint arXiv:161107004

[32] Jafari MH, Samavi S, Karimi N, Soroushmehr SMR, Ward K, Najarian K (2016) Automatic detection of melanoma using broad extraction of features from digital images. In: Engineering in Medicine and Biology

Society (EMBC), 2016 IEEE 38th Annual International Conference of the, IEEE, pp 1357–1360
[33] Kingma D, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980
[34] Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint arXiv:13126114
[35] Li C, Xu K, Zhu J, Zhang B (2017) Triple generative adversarial nets. arXiv preprint arXiv:170302291
[36] Liao H (2016) A deep learning approach to universal skin disease classification. University of Rochester Department of Computer Science, CSC
[37] Lopez AR, Giro-i Nieto X, Burdick J, Marques O (2017) Skin lesion classification from dermoscopic images using deep learning techniques. In: Biomedical Engineering (BioMed), 2017 13th IASTED International Conference on, IEEE, pp 49–54
[38] Maaløe L, Sønderby CK, Sønderby SK, Winther O (2016) Auxiliary deep generative models. arXiv preprint arXiv:160205473
[39] Majtner T, Yildirim-Yayilgan S, Hardeberg JY (2016) Efficient melanoma detection using texture-based rsurf features. In: International Conference Image Analysis and Recognition, Springer, pp 30–37
[40] Mishra NK, Celebi ME (2016) An overview of melanoma detection in dermoscopy images using image processing and machine learning. arXiv preprint arXiv:160107843
[41] Miyato T, Maeda Si, Koyama M, Ishii S (2017) Virtual adversarial training: a regularization method for supervised and semi-supervised learning. arXiv preprint arXiv:170403976
[42] Nowozin S, Cseke B, Tomioka R (2016) f-gan: Training generative neural samplers using variational divergence minimization. In: Advances in Neural Information Processing Systems, pp 271–279
[43] Pathan S, Prabhu KG, Siddalingaswamy P (2018) Techniques and algorithms for computer aided diagnosis of pigmented skin lesions—a review. Biomedical Signal Processing and Control 39:237–262
[44] Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:151106434
[45] Ranzato M, Szummer M (2008) Semi-supervised learning of compact document representations with deep networks. In: Proceedings of the 25th international conference on Machine learning, ACM, pp 792–799
[46] Rasmus A, Berglund M, Honkala M, Valpola H, Raiko T (2015) Semi-supervised learning with ladder networks. In: Advances in Neural Information Processing Systems, pp 3546–3554
[47] Sabbaghi S, Aldeen M, Garnavi R (2016) A deep bag-of-features model for the classification of melanomas in dermoscopy images. In: Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the, IEEE, pp 1369–1372
[48] Sajjadi M, Javanmardi M, Tasdizen T (2016) Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In: Advances in Neural Information Processing Systems, pp 1163–1171
[49] Schoneveld L (2017) Semi-supervised learning with generative adversarial networks
[50] Silveira M, Nascimento JC, Marques JS, Marçal AR, Mendonça T, Yamauchi S, Maeda J, Rozeira J (2009) Comparison of segmentation methods for melanoma diagnosis in dermoscopy images. IEEE Journal of Selected Topics in Signal Processing 3(1):35–45
[51] Simard PY, Steinkraus D, Platt JC, et al (2003) Best practices for convolutional neural networks applied to visual document analysis. In: ICDAR, vol 3, pp 958–962
[52] Situ N, Yuan X, Chen J, Zouridakis G (2008) Malignant melanoma detection by bag-of-features classification. In: Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, IEEE, pp 3110–3113
[53] Springenberg JT (2015) Unsupervised and semi-supervised learning with categorical generative adversarial networks. arXiv preprint arXiv:151106390
[54] Stanley RJ, Stoecker WV, Moss RH (2007) A relative color approach to color discrimination for malignant melanoma detection in dermoscopy images. Skin Research and Technology 13(1):62–72
[55] Statistics C (2014) Canadian cancer society's advisory committee on cancer statistics. in. Canadian Cancer Society, Canadian Cancer Society
[56] Statistics C (2017) Canadian cancer society's advisory committee on cancer statistics. in. Canadian Cancer Society, Canadian Cancer Society
[57] Stoecker WV, Mishra N, LeAnder RW, Rader RK, Stanley RJ (2013) Automatic detection of skin cancer-current status, path for the future. In: VISAPP (1), pp 504–508
[58] Weston J, Ratle F, Mobahi H, Collobert R (2012) Deep learning via semi-supervised embedding. In: Neural Networks: Tricks of the Trade, Springer, pp 639–655
[59] Yu L, Chen H, Dou Q, Qin J, Heng PA (2017) Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE transactions on medical imaging 36(4):994–1004