Unsupervised Segmentation incorporating Shape Prior via Generative Adversarial Networks

Dahye Kim and Byung-Woo Hong
Chung-Ang University, Seoul, Korea
dahye@image.cau.ac.kr, hong@cau.ac.kr

Abstract

We present an image segmentation algorithm that is developed in an unsupervised deep learning framework. The delineation of object boundaries often fails due to the nuisance factors such as illumination changes and occlusions. Thus, we initially propose an unsupervised image decomposition algorithm to obtain an intrinsic representation that is robust with respect to undesirable bias fields based on a multiplicative image model. The obtained intrinsic image is subsequently provided to an unsupervised segmentation procedure that is developed based on a piecewise smooth model. The segmentation model is further designed to incorporate a geometric constraint imposed in the generative adversarial network framework where the discrepancy between the distribution of partitioning functions and the distribution of prior shapes is minimized. We demonstrate the effectiveness and robustness of the proposed algorithm in particular with bias fields and occlusions using simple yet illustrative synthetic examples and a benchmark dataset for image segmentation.

1. Introduction

The image segmentation problem plays a significant role in providing both the appearance (such as texture or brightness) and geometry of objects by partitioning the domain of image into mutually disjoint regions. It is often considered as a basis for a higher level of visual understanding of image contents. Various classical image segmentation algorithms have been developed based on the variational framework [10, 42, 11, 55, 13, 9, 45, 46] where an objective functional that defines a discrepancy between model and observation is optimized in a solution space of partitioning function. The variation of observation from the defined model is typically computed based on a single measurement leading to an unsupervised algorithm. Albeit a number of successful unsupervised variational algorithms have been developed using normalized cuts in graph representations [24, 52], markov random field models [44, 60], density estimations in a feature space [22, 23], level set embedding functions [43, 13] and hierarchical methods in multi-scale representations [24, 2], their associated limitations that stem from the complexity of statistical properties in characterizing regions of interest naturally lead to the development of supervised algorithms using a large number of training images. The development of supervised image segmentation algorithms based on the resurgent neural networks in particular with locally characteristic convolutional kernels has been making a significant improvement over the classical unsupervised approaches [16, 49, 37, 41, 61, 4, 19] where convolutional neural networks predict the probability of indication for region of interest. However, the supervised algorithms generally require extensive manual annotations that are rarely available and often result in coarse-grained. It is also often insufficient to generalize an effective segmentation model with respect to both appearance and geometry albeit data-driven supervision due to the inherited complexity from the variations in lighting conditions and physical properties of objects. The difficulties in coping with high dimensional distributions with huge variations lead to the development of segmentation algorithms by unsupervised learning schemes using abundant training examples with partial or crude labels [33, 26, 6, 20, 7, 1, 58]. In particular, the successful application of generative adversarial networks (GAN) [27, 47, 50, 3] has been extended to an image segmentation problem [20, 7, 6] where the distribution of composite images formed by the foreground of the object of interest and its realistic background is desired to be learned. However, the distribution for both appearance and geometry of object turns out difficult to be learned due to its enormous dimensionality and variations despite a relatively large number of coarse-grained labels. Thus, it is desired to improve the learnability [8] of a characteristic distribution for segmentation in a generative learning scheme, which motivates to simplify a generative model to learn. In this work, we present an unsupervised segmentation algorithm that learns an embedding function for a bipartitioning model based on the statistical homogeneity of appearance and incorporates...
a shape prior that is imposed on the segmentation model in a GAN framework. Our proposed algorithm considers a generative learning model only for the geometric property excluding the appearance (intensity) property of an object so that such simpler distribution is easier to learn and turns to be more effective. It is often feasible to create a three dimensional model for the shape of object and generate a large collection of projected images from arbitrary viewing directions. Thus, we propose to learn an unsupervised segmentation model based on the intensity of an object and impose its geometrical constraint using its shape images of the same category in the GAN framework. We also propose to learn an intrinsic image representation that is robust with respect to undesirable bias fields in an unsupervised way, so that the proposed unsupervised segmentation model can be less sensitive to the inhomogeneity of object appearance. Our unified framework combines the intrinsic image representation model and the segmentation model incorporating a shape constraint that is learned by the GAN algorithm.

2. Related Works

The image segmentation problem has been typically considered as an optimization problem minimizing an energy functional that is designed to measure the discrepancy between an observation and a model in a variational framework [42, 11]. A number of image models have been developed based on edge [55], region [56] and convex optimization [9, 46]. The image model based on the statistical homogeneity of intensity has been extended to incorporate shape information as a prior knowledge [25, 12] where an alternative optimization is performed to minimize a partitioning energy and a distance between a partitioning function and an embedding function for a desired shape. Meanwhile, there have been a number of works proposing intrinsic representations robust with respect to imaging conditions [5, 39]. With the increasing popularity of machine learning techniques using deep neural network architectures, a fully convolutional network has been developed for semantic segmentation in a supervised framework where local [37], global [36] and their combined [17] approaches are proposed using a set of manually annotated images. Another popular supervised deep models for image segmentation have been developed based on the convolutional encoder-decoder architecture [49, 4] where characteristic features are encoded and its symmetric decoding leads to localization. To overcome the limitation of available fine-grained annotations, weakly supervised methods have been proposed using bounding boxes [54, 30], regional convolutional network [48, 28], direction features [15], dense sliding windows [21] and attention networks [18, 31]. In contrast to the discriminative models that predict the probability of segmentation labels, generative segmentation models have been developed due to the introduction of effective generative algorithms [34, 27]. An adversarial training approach for segmentation has been proposed in [38] where a discriminator is learned to distinguish between the ground truth segmentation maps and the ones yielded by a generator. In order to cope with the lack of manual annotations, semi-supervised learning algorithms based on GANs have been developed in [53, 32] where fully convolutional discriminators are learned to differentiate the ground truth labels from the probability maps obtained by generators in combination with the adversarial loss on unlabeled data. Another GAN-based segmentation methods that are most closely related to our approach include [7, 20] where an adversarial learning is applied to generate a realistic composite image that consists of layers for the parts of foreground images and the natural background images. It is assumed that the composition of image parts obtained from the foreground image under perturbations and the natural background is shown to be realistic when the image parts correspond to the desired segmentation. The distribution to be learned by the proposed GAN methods in [7, 20] is aimed to characterize both appearance and geometry of object, thus leading to a complex and high dimensional discriminator. In contrast, the desired distribution to be learned by our generative model only considers the geometrical property of object as a constraint to an unsupervised segmentation model based on the object appearance. Thus, our method employs different learning schemes depending on the characteristic properties, namely appearance and geometry.

3. Segmentation with Shape Prior via GAN

Let \( I: \Omega \rightarrow \mathbb{R} \) be an image that is assumed to be a scalar function for ease of mathematical presentation, yet it can be extended to a vector-valued function for images with multiple channels. The objective of an image segmentation task is to obtain a characteristic function \( \chi_R: \Omega \rightarrow \{0, 1\} \) that partitions the image domain \( \Omega \) as follows:

\[
\chi_R(x) = \begin{cases} 
1 & : x \in R \\
0 & : x \notin R, 
\end{cases} 
\]

(1)

where \( R \subset \Omega \) denotes a region of interest. We introduce an embedding function \( \phi: \Omega \rightarrow (0, 1) \) for a relaxed form of the characteristic function \( \chi_R \) for computational convenience [9] as defined by \( R = \{ x| \phi(x) > \xi \} \) where \( \xi \in (0, 1) \) denotes a threshold that is typically given by 0.5. In our segmentation model, we consider an intrinsic image representation \( u: \Omega \rightarrow \mathbb{R} \) that is desired to be robust to bias field leading to a subsequent optimization with respect to \( \phi \) in maximizing the following probability:

\[
P(\phi, u | I) = P(\phi | u, I)P(u | I),
\]

(2)

where the conditional joint probability for segmentation function \( \phi \) and intrinsic image \( u \) given image \( I \) is computed by
the product of the marginal probability $P(\phi \mid u, I)$ and the conditional probability $P(u \mid I)$. Then, the Bayes theorem leads to the following:

$$P(\phi \mid u, I)P(u \mid I) \propto P(\phi \mid u, I)P(I \mid u)P(u),$$

(3)

where we have the following by the chain rule:

$$P(\phi \mid u, I)P(I \mid u) = P(\phi, I \mid u).$$

(4)

Thus, we have:

$$P(\phi \mid u, I)P(u \mid I) \propto P(\phi, I \mid u)P(u),$$

(5)

$$\propto P(\phi \mid u)P(I \mid u)P(u),$$

(6)

$$\propto P(\phi \mid u)P(u \mid I),$$

(7)

where we assume that $\phi$ and $I$ are conditionally independent given $u$. Thus, we have:

$$P(\phi, u \mid I) \propto P(\phi \mid u)P(u \mid I).$$

(8)

The problem of interest is to obtain the sequential estimation of optimal intrinsic image $u$ and segmenting function $\phi$ by maximizing $P(\phi \mid u)P(u \mid I)$ where $P(\phi \mid u)$ is a conditional probability for an optimal segmenting function $\phi$ given $u$ and $P(u \mid I)$ is a posterior probability defined for an optimal intrinsic image $u$ given $I$. We develop an unsupervised learning algorithm for estimating $u$ and $\phi$ in a deep learning framework where $u$ and $\phi$ are represented by parameterized functions constructed by nested composition of linear and nonlinear functions. We also incorporate a shape prior into the estimation of $\phi$ using a generative adversarial network.

### 3.1. Intrinsic Image Representation

We propose to obtain a robust representation of image with respect to undesirable bias field so that the homogeneity of appearance statistics is better characterized resulting in more accurate segmentation. We consider an image formation model using an additive noise with a multiplicative bias field as follows:

$$I = \nu(u + \eta),$$

(9)

where the noise process $\eta$ is assumed to follow a normal distribution with mean 0 and the bias field $\nu$ is assumed to follow a log-normal distribution with mean 0 imposing a positive constraint $\nu > 0$. The computation of an optimal intrinsic representation $u$ from observation $I$ can be obtained by maximizing the posterior probability $P(u \mid I)$ where we introduce an auxiliary bias field function $\nu$ as follows:

$$P(u, \nu \mid I) \propto P(I \mid u, \nu)P(u \mid \nu)P(\nu),$$

(10)

$$\propto P(I \mid u, \nu)P(u)P(\nu),$$

(11)

where $u$ and $\nu$ are assumed to be independent so that we have $P(u|\nu) = P(u)$. We have the likelihood probability based on the Gaussian noise assumption as follows:

$$P(I \mid u, \nu) \propto \exp \left( -\frac{1}{\nu} - u \right),$$

(12)

and the prior probabilities for $u$ and $\nu$ are given by:

$$P(u) \propto \exp \left( -\|\nabla u\| \right),$$

(13)

$$P(\nu) \propto \exp \left( -\|\nabla \nu\|^{2} + (\nu - 1)_{2}^{2} \right),$$

(14)

where the gradients of $u$ and $\nu$ are assumed to follow a Laplace and a Normal distribution, respectively. In addition, $\log \nu$ is assumed to follow a Normal distribution with mean 0. It is desired to preserve significant geometric features in the reconstruction of $u$, thus we use a total variation regularization for $u$, whereas bias field $\nu$ is assumed to have a smoothly varying intensity field leading to the $L_{2}^{2}$ regularization. The optimal solutions of $u$ and $\nu$ can be given by the joint minimization of the following objective functional $L_{1}$ derived by taking the negative log of the posterior probability:

$$L_{1}(u, \nu; I) = \| \frac{I}{\nu} - u \|^{2} + \lambda \|\nabla u\|$$

$$+ \alpha \|\nabla \nu\|^{2} + \beta \|\nu - 1\|_{2}^{2},$$

(15)

where $\lambda, \alpha, \beta$ are control parameters given by positive constants. The intrinsic image $u$ and its associated undesirable bias field $\nu$ are represented by the outputs of a neural network where the model parameters are optimized by minimizing the objective function $L_{1}$ in an unsupervised manner. The optimal intrinsic representation $u$ for image $I$ is used as an input for segmentation, as will be discussed in the following section.

### 3.2. Segmentation Model

We use the obtained intrinsic image $u$ instead of the original observation $I$ for segmentation where a piecewise smooth Mumford-Shah model [42, 13, 55] is applied based on an embedding function $\phi$ for partitioning region of interest with a Gaussian noise process $\eta$ as follows:

$$u(x) = a(x) \cdot \phi(x) + b(x) \cdot (1 - \phi(x)) + \eta(x),$$

(16)

where $a : \Omega \rightarrow \mathbb{R}$ and $b : \Omega \rightarrow \mathbb{R}$ are continuous functions that respectively estimate the interior and exterior of a segmenting region that is characterized by function $\phi$. An optimal partitioning function $\phi$ given $u$ can be obtained by maximizing the posterior probability $P(\phi \mid u)$ in Eq. (8):

$$P(\phi \mid u) \propto P(u \mid \phi)P(\phi),$$

(17)
where we have the likelihood probability $P(u \mid \phi)$ based on the Gaussian noise assumption leading to the following objective functional:

\[
\mathcal{L}_2(\phi, a, b; u) = \gamma_1 \Vert \nabla \phi(x) \Vert^2 + \gamma_2 \Vert \nabla a(x) \Vert^2 + \gamma_2 \Vert \nabla b(x) \Vert^2 + \int_{\Omega} \left[ u(x) - a(x) \right]^2 \phi(x) \, dx + \int_{\Omega} \left[ u(x) - b(x) \right]^2 (1 - \phi(x)) \, dx, \tag{18}
\]

where $\gamma_1$ and $\gamma_2$ are positive control parameters for the regularization of total variation on $\phi$ and $a$, $b$, respectively. The partitioning function $\phi$ and its associated estimates $a$ and $b$ for the foreground and the background of segmenting region are represented by the separate outputs of a neural network. Note that the optimal estimates $a$ and $b$ can be directly obtained by applying alternating direction method of multipliers algorithms [57], but we instead learn the associated parameters with $a$ and $b$ in an unsupervised manner. The prior probability $P(\phi)$ in (17) is generally given by the assumption that the gradient of $\phi$ follows a Laplace distribution leading to an implicit regularization term as follows:

\[
P(\phi) \propto \exp \left( -\Vert \nabla \phi \Vert \right), \tag{19}
\]

which penalizes the length of partitioning boundary [10]. Whereas we rather establish the joint prior probability $P(\phi, \psi)$ with an additional variable $\psi$ that represents a prior shape in the construction of the prior probability $P(\phi)$ imposed on the segmenting function $\phi$.

We propose to incorporate shape information about a region of interest into its segmentation exploiting a prior knowledge of segmenting function using a generative adversarial network (GAN) [27]. We extend the prior probability $P(\phi)$ in Eq. (17) leading to an implicit regularization imposed on the segmenting function $\phi$ to the joint prior probability $P(\phi, \psi)$ with an additional variable $\psi$ as follows:

\[
P(\phi, \psi) = P(\phi \mid \psi) P(\psi), \tag{20}
\]

where $\psi$ represents an explicit shape. Let $\mathcal{S} \subset \Omega$ be a shape and $\chi_S$ be its characteristic function. Let $\mathcal{T}$ be a transformation group acting on the domain $\Omega$. We denote by $\psi_t$ a deformed shape from $\chi_S$ by an element $t \in \mathcal{T}$ as follows:

\[
\psi_t(x) = \chi_S \circ t(x), \tag{21}
\]

where $t : \Omega \rightarrow \Omega$ and we omit the symbol $S$ in the notation $\psi_t$ for ease of presentation. We construct an empirical distribution of the prior probability $P(\psi)$ in Eq. (20) by the equivalence class $\mathcal{S} = \{ \psi_t = \chi_S \circ t \mid t \in \mathcal{T} \}$ of shape $S$ under the action of the transformation group $\mathcal{T}$. Shape is represented in the form of binary image, and its statistics are explicitly formed by a variety of shapes within the same category. Given a prior probability $P(\psi)$ on shape $S$, its geometric property leads to a constraint in the determination of partitioning function $\phi$ by a conditional probability $P(\phi \mid \psi)$ in Eq. (20). We denote empirical distribution of the probability density function $Q(\phi)$ of partitioning function $\phi$ by $\mathcal{R} = \{ \phi_j \}$ where $\phi_j$ is associated with an input image $I_j$, equivalently, with its intrinsic representation $u_j$. We construct a conditional probability $P(\phi \mid \psi)$ using Jensen–Shannon divergence $D_{JS}$ as a discrepancy measure between probability distributions of partitioning function $Q(\phi)$ and prior shape $P(\psi)$ as follows:

\[
- \log P(\phi \mid \psi) \propto D_{JS}(Q(\phi) \parallel P(\psi)). \tag{22}
\]

The optimization procedure is suited in the GAN framework [27, 47, 3]. Let $h$ be a discriminator for the classification of shapes and $g$ be a generator for the determination of partitioning function. The classifier $h$ aims to discriminate the equivalence class of shape $S$ from its non-equivalence class generated by the partitioning function $\phi$ induced by $g$. Then, the objective function that aims to obtain optimal sets of model parameters for the discriminator network $h$ and the segmentation network $g$, respectively, is defined by:

\[
\min_g \max_h \left( \mathbb{E}_{\psi \sim P(\psi)} \left[ \log(h(\psi)) \right] + \mathbb{E}_{\phi \sim Q(\phi)} \left[ \log(1 - h(\phi)) \right] \right). \tag{23}
\]

Due to the limitation of the objective function such as vanishing gradient and model collapse in Eq. (23) using Kullback–Leibler divergence, we apply a non-saturating loss for the generator and add a regularization that is designed to penalize the gradients of the discriminator [50] as follows:

\[
\mathcal{L}_3(\rho, \theta, \mathcal{S}, \mathcal{R}) = \mathbb{E}_{\psi \sim P(\psi)} \left[ \log(h(\psi)) \right] - \mathbb{E}_{\phi \sim Q(\phi)} \left[ \log(h(\phi)) \right] - \frac{\kappa}{2} \mathbb{E}_{\psi \sim P(\psi)} \left[ \Vert \nabla h(\psi) \Vert^2 \right], \tag{24}
\]

where $\kappa > 0$ is a control parameter for the regularization, and $\mathcal{S}$ and $\mathcal{R}$ represent the equivalence class of shapes $\{ \psi_i \}$ and a set of partitioning functions $\{ \phi_j \}$, respectively, and $\rho$ and $\theta$ are sets of model parameters associated with the segmentation network $g$ and the classifier network $h$, respectively. The latent space is induced by a set of intrinsic images and the generator is driven by the segmentation loss in Eq. (18) whose solution space is constrained by a shape prior via the Wasserstein GAN [3]. Note that $\phi : \Omega \rightarrow (0, 1)$ is a smooth function whereas $\psi : \Omega \rightarrow \{ 0, 1 \}$ is a characteristic function. It is generally required to impose a sparsity constraint $\nabla \phi$ following the assumption in Eq. (19) in order to obtain a binary representation for partitioning boundary. However, the sparsity constraint on the function $\phi$ can be achieved instead by the back-propagation from the objective function in Eq. (24) due to the binary representation of $\psi$. 

7327
Figure 1: Schematic illustration of the proposed neural network architecture. The problems of interest consist of three constituent parts: (a) obtaining an intrinsic image representation \( u \) that is robust to a multiplicative bias field \( v \) for a given image \( I \), (b) deriving a partitioning function \( \phi \) that determines a region of interest based on the intrinsic representation \( u \) with its associated foreground and background estimates \( a \) and \( b \), respectively, and (c) imposing a geometric constraint to the partitioning function \( \phi \) using a given set of prior shapes \( \{ \psi_i \} \). The intrinsic decomposition auto-encoder \( f \) is optimized by minimizing \( L_1 \). The obtained optimal \( u \) is fed into the segmentation auto-encoder \( g \) that is optimized by minimizing \( L_2 + L_3 \). To impose the geometric constraint on \( \phi \), the discriminator \( h \) classifies \( \phi \) and \( \psi \) by minimizing \( L_3 \).

4. Neural Network Architectures

The schematic illustration of the neural network architectures for each component of the proposed algorithm is presented in Fig. 1. Let \( (u, v) = f(I; w) \) be an auto-encoder parameterized by \( w \) for the reconstruction of intrinsic image \( u \) and the multiplicative bias field \( v \) given input \( I \). Let \( (\phi, a, b) = g(u; v) \) be an auto-encoder parameterized by \( \rho \) for segmenting function \( \phi \) and its associated estimates \( a \) and \( b \) given \( u \). Let \( h(\cdot; \theta) \) be a classifier parameterized by \( \theta \) discriminating real shape \( \psi \) from segmenting shape \( \phi \). The optimal model parameter \( w \) is obtained by minimizing \( L_1 \) in Eq. (15). Similarly, the optimal model parameters \( \rho \) and \( \theta \) are obtained by minimizing \( L_2 + L_3 \) in Eq. (18), Eq. (24) and \( L_3 \) in Eq. (24), respectively. The generative adversarial training scheme between \( g \) and \( h \) driven by \( L_3 \) in Eq. (24) imposes the geometric properties of real shape \( \psi \) on the resulting segmenting function \( \phi \). For the selection of neural networks for the auto-encoder \( g \) and the discriminator \( h \), we consider a standard convolutional neural network architecture and its variants with skip connections [49] or residual blocks [29]. The standard structures are adopted for both \( g \) and \( h \) based on the results shown in Tab. 2 that compares the performance of different combinations of \( g \) and \( h \).

5. Experiments

We demonstrate the robustness and effectiveness of each component of our proposed algorithm. We perform quantitative and qualitative analysis of the performance in the reconstruction of intrinsic images and the segmentation of the object of interest. We use a set of simple yet illustrative synthetic images and LSUN dataset [59] in the evaluation.

5.1. Results on Synthetic Dataset

Dataset. We randomly generate binary images representing square shapes with varying sizes and locations as shown in Fig. 2 (e). For the demonstration of the reconstruction of intrinsic images, we randomly generate a bias field with intensity gradation within a given variation from an arbitrary viewing direction as shown in Fig. 2 (c) where the standard deviations of gradation are set to be 0.1, 0.2, 0.3 and 0.4 from top row to bottom. As shown in Fig. 2, we apply randomly generated bias fields in (c) to binary square images in (e) to construct composite images in (a) using the multiplicative model. In order to show the effectiveness of our shape prior model, we apply occlusions along the diagonal lines in addition to the bias fields to the binary square images as shown in Fig. 3 (a) where the occlusion degrees are set to be 20%, 40%, 60% and 80% with respect to the regions of interest from top row to bottom. For the evaluation, we generate 60k images of size 64 x 64 for each configuration of experiment and use 50k for training, 5k for validation and 5k for testing.

Hyper-parameters. We apply a dynamic scheduling of learning rate following a sigmoid function with the initial value 5e-05 and the final value 1e-06 for \( f \), but we use the
Figure 2: Segmentation results without shape prior on the synthetic square images multiplied by bias fields with different standard deviations 0.1, 0.2, 0.3 and 0.4 from (top) row to (bottom). (a) original image. (b) obtained intrinsic image. (c) ground truth of bias field. (d) obtained bias field. (e) ground truth of the shape. (f) obtained segmentation on the original input. (g) obtained segmentation on the intrinsic (full model).

Figure 3: Segmentation results on the synthetic square images with occlusions at varying degrees 20%, 40%, 60% and 80% from (top) row to (bottom) in addition to bias fields with std 0.4. (a) original image. (b) obtained intrinsic image. (c) ground truth of the shape. (d) obtained segmentation on the original without shape prior. (e) obtained segmentation on the intrinsic without shape prior. (f) obtained segmentation on the original with shape prior. (g) obtained segmentation on the intrinsic with shape prior (full model).

fixed values 1e-05 and 1e-04 for $g$ and $h$. We use mini-batch sizes of 120 for $f$ and 128 for $g$ and $h$. For the parameters in Eq. (15), we set $\lambda, \alpha, \beta$ as 1e-02, 1.5, 1e-04. For the parameters in Eq. (18), we set $\gamma_1, \gamma_2$ as 1e-05 and 0.1.

Evaluation. We provide visual illustrations of qualitative results on the binary shape images with bias fields at varying variations for the reconstruction of intrinsic images and the unsupervised segmentation without a shape prior in Fig. 2.

Figure 4: Results on the ablation study for the segmentation with different methods on the square images with (left) bias fields at varying std and (right) occlusions at varying degrees in addition to the bias fields with std 0.4. x-axis represents the degree of degrading factors (left) bias field and (right) occlusion and y-axis represents IoU score.

<table>
<thead>
<tr>
<th>method</th>
<th>occlusion (%)</th>
<th>bias (std)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>original</td>
<td>20</td>
<td>0.8652</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.7405</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.5880</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.3164</td>
</tr>
<tr>
<td>intrinsic</td>
<td>20</td>
<td>0.8866</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.7671</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.6332</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.6319</td>
</tr>
<tr>
<td>original + shape</td>
<td>20</td>
<td>0.9985</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.9918</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.9007</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.6131</td>
</tr>
<tr>
<td>intrinsic + shape (full model)</td>
<td>20</td>
<td>0.9990</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.9987</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.9555</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.8958</td>
</tr>
</tbody>
</table>

Table 1: Segmentation results by the ablation study with different configuration of the methods. The average IoU values are presented for the square images with varying degrees of occlusions and bias field variations.

It is clearly demonstrated that the segmentation results obtained from the intrinsic images are better across all the variations in the bias fields whereas the segmentation quality on the original ones deteriorates as the standard deviation of the bias field increases. Their quantitative comparisons based on the intersection of union (IoU) are provided in Fig. 4 (left) where x-axis indicates the standard deviation of gradation and y-axis indicates the IoU score. We give ablation results on the square images including occlusions with varying degrees from low (top) to
Table 2: Comparison of the segmentation IoU on LSUN car with different network architectures of our model incorporating shape prior. Each column represents different auto-encoder network \( g \) and each row indicates different discriminator network \( h \).

Table 3: Segmentation IoU on LSUN dataset with different methods. (intrinsic + shape) denotes our full model.

5.2. Results on LSUN Dataset

Dataset. In the evaluation of our algorithm for real images, we consider 4 categories including airplane, boat, car and chair in LSUN dataset [59] where images are color and of the size \( 64 \times 64 \). Since the ground truth for the object segmentation in LSUN dataset is not available, we employ a Mask R-CNN model [28] that has been trained using COCO dataset [35] to obtain pseudo-labels for the object segmentation. In this experiment, we only consider images with a single object whose size is between 5% and 95% with respect to the image size. Examples of object images and their pseudo-labels are shown in Fig. 6 (a) and (d), respectively. For the 4 different categories airplane, boat, car and chair, the numbers of images used are 71,590, 49,642, 75,973 and 60,606 for training, 7,954, 5,516, 8,441 and 6,734 for validation, and 8,726, 6,196, 9,407 and 7,271 for testing. In the construction of a shape prior model for each category, we generate binary shape images by random projections from 3 dimensional object models in ShapeNet [14]. We apply morphological operations to the obtained projection images in order to have simple shapes without holes and the numbers of generated images are 97,080, 46,536, 179,904 and 162,672 for airplane, boat, car and chair, respectively. In Fig. 5, examples of the rendered projection images and their binary shapes are shown at top row and bottom, respectively.

Hyper-parameters. We apply the same learning rate scheduling to \( f \) as done in Sec. 5.1. We use 1e-03 for the fixed learning rate of \( g \) and \( h \). We use the same mini-batch size as done in Sec. 5.1. For the parameters in Eq. (15), we set \( \lambda, \alpha \) and \( \beta \) as 1e-02, 15 and 1e-04. For the parameters in Eq. (18), we set \( \gamma_1 \) and \( \gamma_2 \) as 1e-02 and 0.1.

Evaluation. We perform an ablation study and a comparative analysis based on LSUN dataset. In our comparison, we consider the state-of-the-art techniques including perturbedGAN [7], redrawing of objects (ReDO) [20], GrabCut [51]. For the implementation of the algorithms under comparison, we use their official codes with the recommended parameters. Since the network that maps input images to the generator is unavailable in the perturbedGAN work, we add an encoder to the publicly available codes. For the initial condition of GrabCut, we employ a generic condition using central squares. Examples of the qualitative comparisons are provided in Fig. 6 where (a) original im-

![Figure 5: Examples of the rendered object images (top) and their shape images (bottom) generated from ShapeNet.](image-url)
Figure 6: Segmentation results on LSUN dataset. (a) original. (b) obtained intrinsic. (c) obtained bias field. (d) pseudo-label by Mask R-CNN. (e) our result on the original without shape prior. (f) ours on the intrinsic without shape prior. (g) ours on the original with shape prior. (h) ours on the intrinsic with shape prior (full model). (i) PerturbedGAN. (j) ReDO. (k) GrabCut.

age. (b) obtained intrinsic image, (c) obtained bias field, (d)
pseudo-label obtained by Mask R-CNN, and segmentation
result by (e) our model without shape prior on the original
image, (f) our model without shape prior on intrinsic image,
(g) our model with shape prior on original image, (h) our
model with shape prior on intrinsic image, (i) result by Per-
turbedGAN, (j) result by ReDO and (k) result by GrabCut
are shown. It is visually demonstrated that our full model
(intrinsic + shape) outperforms the other algorithms under
comparison. In particular, our model provides more accurate
results compared to (i) and (j) where the GAN framework
considers both appearance and geometric properties, indicat-
ing that simplifying the distribution to be learned by GAN
leads to more robust performance. The quantitative evalua-
tion is presented based on IoU in Tab. 3. Our ablation studies
show that using the intrinsic representation and shape priors
significantly improves the quality of the segmentation.

6. Conclusions

We have presented an unsupervised segmentation algo-
rithm developed in a deep learning framework where a shape
prior is incorporated by generative adversarial networks. In
addition, we have developed an unsupervised deep learn-
ing technique to obtain an intrinsic representation that is
robust to undesired bias fields. We have demonstrated the
effectiveness of our algorithm to biases and occlusions using
synthetic images. The comparative analysis with the recent
benchmark works on LSUN dataset indicates the potential
of our method to real applications.

Acknowledgements

This work was supported by Korea government: NRF-
2017R1A2B4006023 and IITP-2021-0-01341, Artificial In-
telligence Graduate School (CAU).
References


[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and


