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 $I \colon \Omega \mapsto \mathbb{R}$ 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 $\chi_R \colon \Omega \mapsto \{0, 1\}$ that partitions the image domain Ω as follows:

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

where $R \subset \Omega$ denotes a region of interest. We introduce an embedding function $\phi \colon \Omega \mapsto (0, 1)$ for a relaxed form of the characteristic function χ_D 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 $u \colon \Omega \mapsto \mathbb{R}$ that is desired to be robust to bias field leading to a subsequent optimization with respect to ϕ in maximizing the following probability:

$$P(\phi, u \mid I) = P(\phi \mid u, I) P(u \mid I), \tag{2}$$

where the conditional joint probability for segmenting function ϕ 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:

$$P(\phi \mid u, I)P(u \mid I) \propto P(\phi \mid u, I)P(I \mid u)P(u), \tag{3}$$

where we have the following by the chain rule:

$$P(\phi \mid u, I)P(I \mid u) = P(\phi, I \mid u). \tag{4}$$

Thus, we have:

$$\begin{aligned} P(\phi \mid u, I)P(u \mid I) &\propto P(\phi, I \mid u)P(u), && (5) \\ &\propto P(\phi \mid u)P(I \mid u)P(u), && (6) \\ &\propto P(\phi \mid u)P(u \mid I), && (7) \end{aligned}$$

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

$$P(\phi, u \mid I) \propto P(\phi \mid u)P(u \mid I). \tag{8}$$

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

$$I = \nu(u + \eta), \tag{9}$$

where the noise process η is assumed to follow a normal distribution with mean 0 and the bias field ν 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 $P(u|I)$ where we introduce an auxiliary bias field function ν as follows:

$$\begin{aligned} P(u, \nu \mid I) &\propto P(I \mid u, \nu)P(u \mid \nu)P(\nu), && (10) \\ &\propto P(I \mid u, \nu)P(u)P(\nu), && (11) \end{aligned}$$

where u and ν 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:

$$P(I \mid u, \nu) \propto \exp\left(-\|\frac{I}{\nu} - u\|_2^2\right), \tag{12}$$

and the prior probabilities for u and ν are given by:

$$\begin{aligned} P(u) &\propto \exp\left(-\|\nabla u\|\right), && (13) \\ P(\nu) &\propto \exp\left(-\|\nabla \nu\|_2^2\right) \exp\left(-\|\nu - 1\|_2^2\right), && (14) \end{aligned}$$

where the gradients of u and ν 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 ν is assumed to have a smoothly varying intensity field leading to the L_2^2 regularization. The optimal solutions of u and ν can be given by the joint minimization of the following objective functional $\mathcal{L}_1$ derived by taking the negative $\log$ of the posterior probability:

$$\begin{aligned} \mathcal{L}_1(u, \nu; I) = \|\frac{I}{\nu} - u\|_2^2 + \lambda\|\nabla u\| \\ + \alpha\|\nabla \nu\|_2^2 + \beta\|\nu - 1\|_2^2, \quad (15) \end{aligned}$$

where λ, α, β are control parameters given by positive constants. The intrinsic image u and its associated undesirable bias field ν are represented by the outputs of a neural network where the model parameters are optimized by minimizing the objective function $\mathcal{L}_1$ 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 ϕ for partitioning region of interest with a Gaussian noise process η as follows:

$$u(x) = a(x) \cdot \phi(x) + b(x) \cdot (1 - \phi(x)) + \eta(x), \tag{16}$$

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

$$P(\phi \mid u) \propto P(u \mid \phi)P(\phi), \tag{17}$$

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

$$\begin{aligned}\mathcal{L}_2(\phi, a, b; u) = \gamma_1 \|\nabla \phi(x)\| + \gamma_2 \|\nabla a(x)\| + \gamma_2 \|\nabla b(x)\| \\ + \int_\Omega |u(x) - a(x)|^2 \phi(x) \, \mathrm{d}x \\ + \int_\Omega |u(x) - b(x)|^2 (1 - \phi(x)) \, \mathrm{d}x, \quad (18)\end{aligned}$$

where γ_1 and γ_2 are positive control parameters for the regularization of total variation on ϕ and a, b, respectively. The partitioning function ϕ 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 ϕ follows a Laplace distribution leading to an implicit regularization term as follows:

$$P(\phi) \propto \exp\left(-\|\nabla \phi\|\right), \quad (19)$$

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

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 ϕ to the joint prior probability $P(\phi, \psi)$ with an additional variable ψ as follows:

$$P(\phi, \psi) = P(\phi \mid \psi) P(\psi), \quad (20)$$

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

$$\psi_i(x) = \chi_S \circ t_i(x), \quad (21)$$

where $t_i \colon \Omega \mapsto \Omega$ and we omit the symbol S in the notation ψ_i for ease of presentation. We construct an empirical distribution of the prior probability $P(\psi)$ in Eq. (20) by the equivalence class $\mathcal{S} = \{\psi_i = \chi_S \circ t_i \mid t_i \in \mathcal{T}\}$ 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 ϕ by a conditional probability $P(\phi \mid \psi)$ in Eq. (20). We denotes empirical distribution of the probability density function $Q(\phi)$ of partitioning function ϕ by $\mathcal{R} = \{\phi_j\}$ where ϕ_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:

$$-\log P(\phi \mid \psi) \propto D_{JS}(Q(\phi) \parallel P(\psi)). \quad (22)$$

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 $\mathcal{S}$ from its non-equivalence class generated by the partitioning function ϕ 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:

$$\begin{aligned}\min_g \max_h \big(\mathbb{E}_{\psi \sim P(\psi)}[\log(h(\psi))] \\ + \mathbb{E}_{\phi \sim Q(\phi)}[\log(1 - h(\phi))]\big). \quad (23)\end{aligned}$$

Due to the limitation of the objective function such as vanishing gradient and model collapse in Eq. (23) using Kullback-Leivler 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:

$$\begin{aligned}\mathcal{L}_3(\rho, \theta; \mathcal{S}, \mathcal{R}) = \mathbb{E}_{\psi \sim P(\psi)}[\log(h(\psi))] \\ - \mathbb{E}_{\phi \sim Q(\phi)}[\log(h(\phi))] - \frac{\kappa}{2} \mathbb{E}_{\psi \sim P(\psi)}[\|\nabla h(\psi)\|_2^2], \quad (24)\end{aligned}$$

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 ρ and θ 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 discriminator in Eq. (24). As shown in [40], the objective function defined in Eq. (24) is known to achieve better convergence property than the Wasserstein GAN [3]. Note that $\phi \colon \Omega \mapsto (0, 1)$ is a smooth function whereas $\psi \colon \Omega \mapsto \{0, 1\}$ 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 ϕ can be achieved instead by the back-propagation from the objective function in Eq. (24) due to the binary representation of ψ.

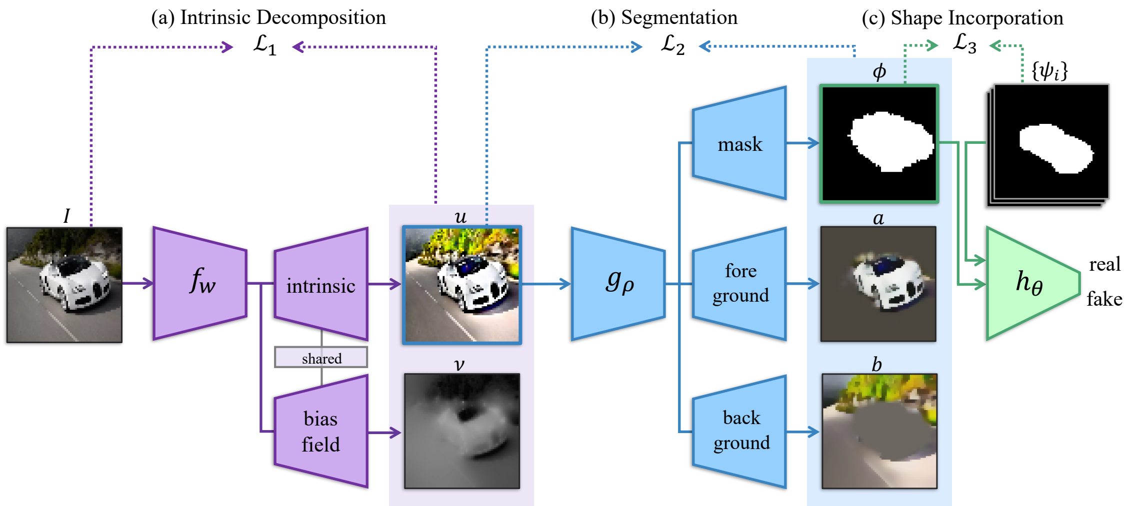

Figure 1: Schematic illustration of the proposed neural network architecture. The problems of interest consist of three constituent parts: (a) obtaining an intrinsic image representation u that is robust to a multiplicative bias field ν for a given image I, (b) deriving a partitioning function ϕ that 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 ϕ using a given set of prior shapes $\{\psi_i\}$. The intrinsic decomposition auto-encoder f is optimized by minimizing $\mathcal{L}_1$. The obtained optimal u is fed into the segmentation auto-encoder g that is optimized by minimizing $\mathcal{L}_2 + \mathcal{L}_3$. To impose the geometric constraint on ϕ, the discriminator h classifies ϕ and ψ by minimizing $\mathcal{L}_3$.

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 $(u, \nu) = f(I; w)$ be an auto-encoder parameterized by w for the reconstruction of intrinsic image u and the multiplicative bias field ν given input I. Let $(\phi, a, b) = g(u; v)$ be an auto-encoder parameterized by ρ for segmenting function ϕ and its associated estimates a and b given u. Let $h(\cdot; \theta)$ be a classifier parameterized by θ discriminating real shape ψ from segmenting shape ϕ. The optimal model parameter w is obtained by minimizing $\mathcal{L}_1$ in Eq. (15). Similarly, the optimal model parameters ρ and θ are obtained by minimizing $\mathcal{L}_2 + \mathcal{L}_3$ in Eq. (18), Eq. (24) and $\mathcal{L}_3$ in Eq. (24), respectively. The generative adversarial training scheme between g and h driven by $\mathcal{L}_3$ in Eq. (24) imposes the geometric properties of real shape ψ on the resulting segmenting function ϕ. 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 64×64 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

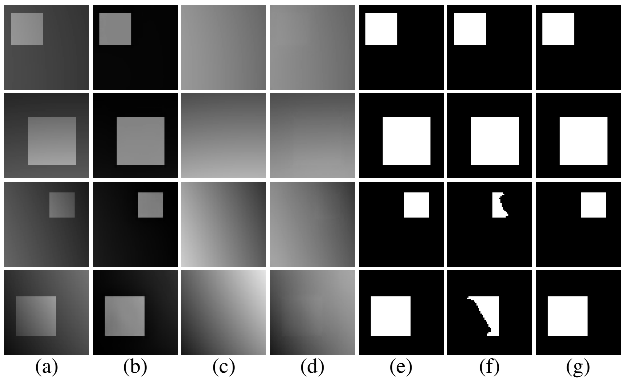

Figure 2: Segmentation results without shape prior on the synthetic square images multiplied by bias fields with different standard deviations 0.1, 0.2, 0.3 and 0.4 from (top) row to (bottom). (a) original image. (b) obtained intrinsic image. (c) ground truth of bias field. (d) obtained bias field. (e) ground truth of the shape. (f) obtained segmentation on the original input. (g) obtained segmentation on the intrinsic (full model).

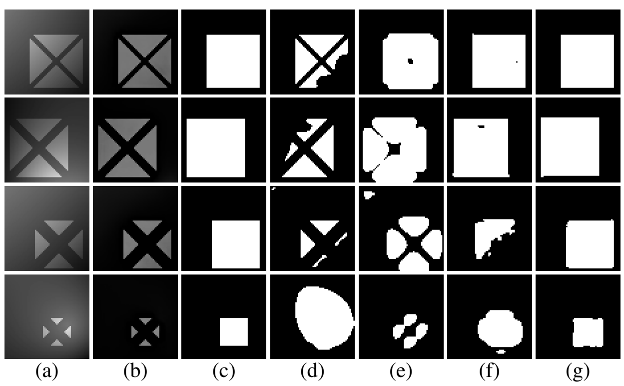

Figure 3: Segmentation results on the synthetic square images with occlusions at varying degrees 20%, 40%, 60% and 80% from (top) row to (bottom) in addition to bias fields with std 0.4. (a) original image. (b) obtained intrinsic image. (c) ground truth of the shape. (d) obtained segmentation on the 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 λ, α, β as 1e-02, 1.5, 1e-04. For the parameters in Eq. (18), we set γ_1, γ_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.

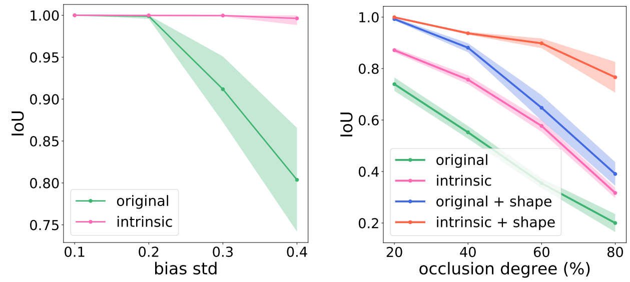

Figure 4: Results on the ablation study for the segmentation with different methods on the square images with (left) bias fields at varying std and (right) occlusions at varying degrees in addition to the bias fields with std 0.4. x-axis represents the degree of degrading factors (left) bias field and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

method	occlusion (%)	bias (std) 0.1	0.2	0.3	0.4
original	20	0.8652	0.7808	0.7776	0.7394
	40	0.7405	0.5890	0.5884	0.5524
	60	0.5880	0.3961	0.3871	0.3544
	80	0.3164	0.2358	0.2151	0.1999
intrinsic	20	0.8866	0.8971	0.8757	0.8714
	40	0.7671	0.7790	0.7555	0.7569
	60	0.6332	0.5576	0.5781	0.5770
	80	0.6319	0.4183	0.3451	0.3168
original + shape	20	0.9985	0.9962	0.9951	0.9923
	40	0.9918	0.9970	0.9256	0.8810
	60	0.9007	0.6497	0.6201	0.6475
	80	0.6131	0.4053	0.3795	0.3911
intrinsic + shape **(full model)**	20	**0.9990**	**0.9983**	**0.9983**	**0.9991**
	40	**0.9987**	**0.9977**	**0.9492**	**0.9365**
	60	**0.9555**	**0.9570**	**0.9278**	**0.9490**
	80	**0.8958**	**0.7839**	**0.7458**	**0.7649**

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 obtained from the original images in green and the intrinsic images in pink. We give ablation results on the square images including occlusions with varying degrees from low (top) to

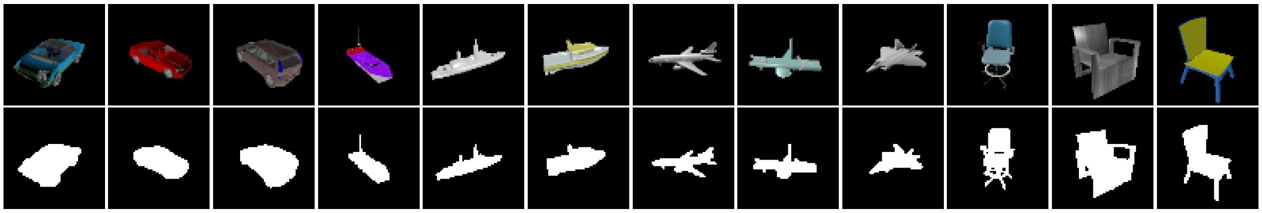

Figure 5: Examples of the rendered object images (top) and their shape images (bottom) generated from ShapeNet.

	standard (g)	skip (g)	residual (g)
standard (h)	**0.6303**	0.6190	0.5876
residual (h)	0.6013	0.6013	0.5793

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.

method	car	boat	airplane	chair
original	0.3126	0.2002	0.2131	0.3543
intrinsic	0.3250	0.2348	0.2307	0.3683
original + shape	0.6303	0.4756	0.4544	0.4824
intrinsic + shape	**0.6340**	**0.4901**	**0.4714**	0.4776
PerturbedGAN	0.5026	0.3122	0.3049	0.3902
ReDO	0.4637	0.3618	0.4110	0.4181
GrabCut	0.5122	0.3325	0.4026	**0.5127**

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

high (bottom) in addition to a fixed degree of bias fields (0.4) in Fig. 3. The segmentation results without shape prior are shown to suffer from the occlusions as shown in (d) and (e). Similarly, the results on the original images with bias fields yield partial failure due to the unsatisfied assumption on the image model as shown in (d) and (f) whereas it is shown that the intrinsic images alleviate the degrading effects in (e) and (g). Our full model is shown to be robust to both occlusions and uneven biases as shown in (g). The ablation results with different methods on the shape images with occlusions and bias fields are provided in Fig. 4 (right) where x-axis indicates the occlusion degrees and y-axis indicates the IoU score. The average IoU scores with the different methods using the square images with varying degrees of degrading factors that are occlusions and bias fields are presented in Tab. 1 where our full model (intrinsic + shape) yields the best results and the performance gap increases over the degrees of degradation factors.

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 64×64. 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 λ, α and β as 1e-02, 15 and 1e-04. For the parameters in Eq. (18), we set γ_1 and γ_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-

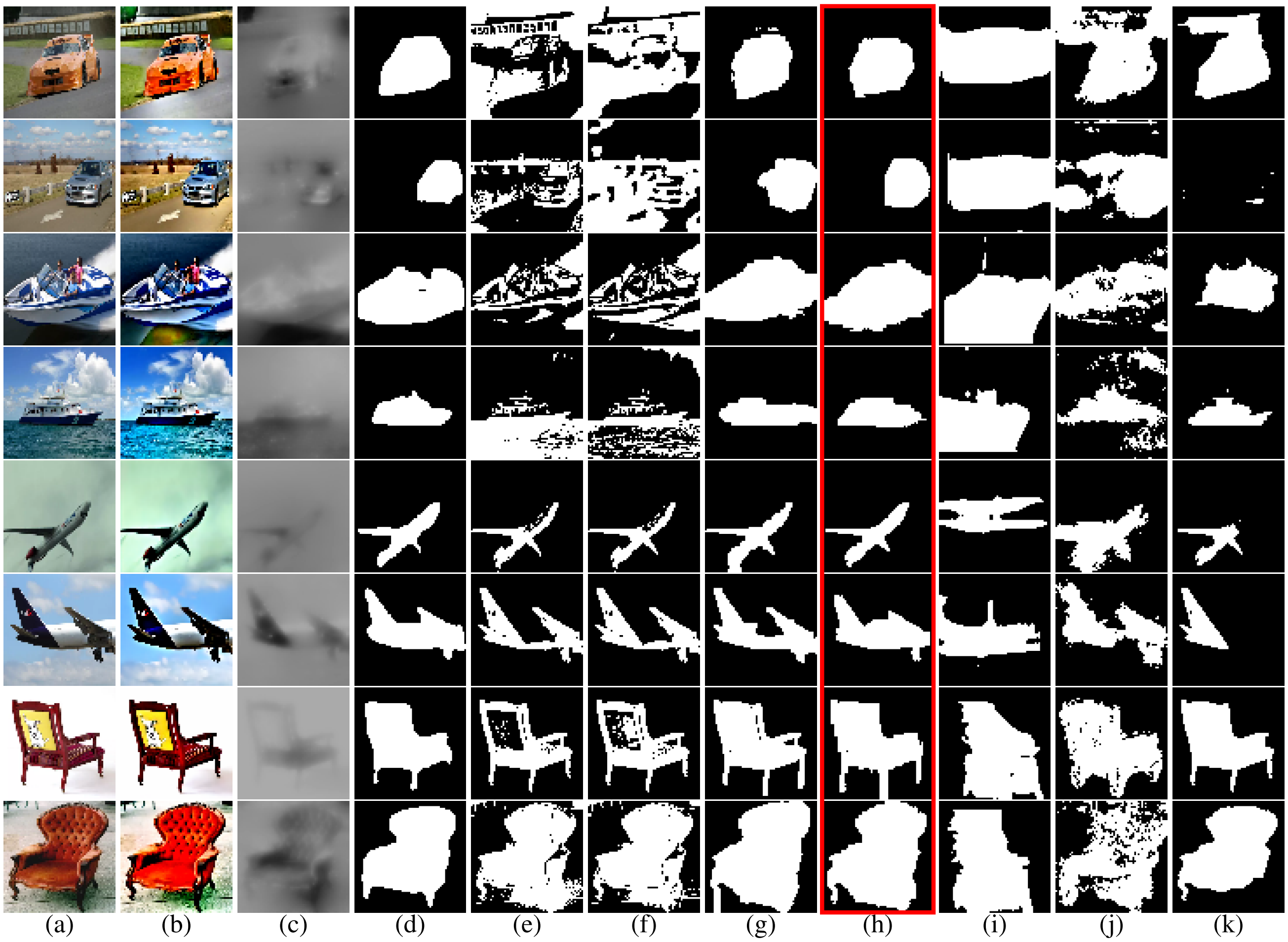

Figure 6: Segmentation results on LSUN dataset. (a) original. (b) obtained intrinsic. (c) obtained bias field. (d) pseudo-label by Mask R-CNN. (e) our result on the original without shape prior. (f) ours on the intrinsic without shape prior. (g) ours on the original with shape prior. (h) ours on the 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 PerturbedGAN, (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, indicating that simplifying the distribution to be learned by GAN leads to more robust performance. The quantitative evaluation 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 algorithm developed in a deep learning framework where a shape prior is incorporated by generative adversarial networks. In addition, we have developed an unsupervised deep learning 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 Intelligence Graduate School (CAU).

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. i

[2] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 328–335, 2014. i

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 214–223, 2017. i, iv

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. i, ii

[5] Anil S Baslamisli, Thomas T Groenestege, Partha Das, Hoang-An Le, Sezer Karaoglu, and Theo Gevers. Joint learning of intrinsic images and semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–302, 2018. ii

[6] Yaniv Benny and Lior Wolf. Onegan: Simultaneous unsupervised learning of conditional image generation, foreground segmentation, and fine-grained clustering. *arXiv preprint arXiv:1912.13471*, 2019. i

[7] Adam Bielski and Paolo Favaro. Emergence of object segmentation in perturbed generative models. In *Advances in Neural Information Processing Systems*, pages 7254–7264, 2019. i, ii, vii

[8] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989. i

[9] Xavier Bresson, Selim Esedoḡlu, Pierre Vandergheynst, Jean-Philippe Thiran, and Stanley Osher. Fast global minimization of the active contour/snake model. *Journal of Mathematical Imaging and vision*, 28(2):151–167, 2007. i, ii

[10] Vicent Caselles, Ron Kimmel, and Guillermo Sapiro. Geodesic active contours. *International journal of computer vision*, 22(1):61–79, 1997. i, iv

[11] Antonin Chambolle. Finite-differences discretizations of the mumford-shah functional. *ESAIM: Mathematical Modelling and Numerical Analysis*, 33(2):261–288, 1999. i, ii

[12] Tony Chan and Wei Zhu. Level set based shape prior segmentation. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 1164–1170. IEEE, 2005. ii

[13] Tony F Chan and Luminita A Vese. Active contours without edges. *IEEE Transactions on image processing*, 10(2):266–277, 2001. i, iii

[14] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. vii

[15] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2018. ii

[16] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. i

[17] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. ii

[18] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016. ii

[19] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. i

[20] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. In *Advances in Neural Information Processing Systems*, pages 12705–12716, 2019. i, ii, vii

[21] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2061–2069, 2019. ii

[22] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799, 1995. i

[23] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619, 2002. i

[24] Timothee Cour, Florence Benezit, and Jianbo Shi. Spectral segmentation with multiscale graph decomposition. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 1124–1131. IEEE, 2005. i

[25] Daniel Cremers, Nir Sochen, and Christoph Schnörr. Towards recognition-based variational segmentation using shape priors and dynamic labeling. In *International Conference on Scale-Space Theories in Computer Vision*, pages 388–400. Springer, 2003. ii

[26] Ioana Croitoru, Simion-Vlad Bogolin, and Marius Leordeanu. Unsupervised learning of foreground object detection. *arXiv preprint arXiv:1808.04593*, 2018. i

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

Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. i, ii, iv
[28] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. ii, vii
[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. v
[30] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4233–4241, 2018. ii
[31] Qin Huang, Chunyang Xia, Chihao Wu, Siyang Li, Ye Wang, Yuhang Song, and C-C Jay Kuo. Semantic segmentation with reverse attention. *arXiv preprint arXiv:1707.06426*, 2017. ii
[32] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018. ii
[33] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. i
[34] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. ii
[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. vii
[36] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015. ii
[37] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. i, ii
[38] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016. ii
[39] Wei-Chiu Ma, Hang Chu, Bolei Zhou, Raquel Urtasun, and Antonio Torralba. Single image intrinsic decomposition without a single intrinsic image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–217, 2018. ii
[40] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International Conference on Machine Learning*, pages 3481–3490, 2018. iv
[41] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. i
[42] David Mumford and Jayant Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 42(5):577–685, 1989. i, ii, iii
[43] Stanley Osher and James A Sethian. Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations. *Journal of computational physics*, 79(1):12–49, 1988. i
[44] Dileep Kumar Panjwani and Glenn Healey. Markov random field models for unsupervised segmentation of textured color images. *IEEE Transactions on pattern analysis and machine intelligence*, 17(10):939–954, 1995. i
[45] Thomas Pock, Antonin Chambolle, Daniel Cremers, and Horst Bischof. A convex relaxation approach for computing minimal partitions. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 810–817. IEEE, 2009. i
[46] Thomas Pock, Daniel Cremers, Horst Bischof, and Antonin Chambolle. An algorithm for minimizing the mumford-shah functional. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1133–1140. IEEE, 2009. i, ii
[47] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. i, iv
[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. ii
[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. i, ii, v
[50] Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *Advances in neural information processing systems*, pages 2018–2028, 2017. i, iv
[51] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004. vii
[52] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000. i
[53] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5688–5696, 2017. ii
[54] Remez Tal, Jonathan Huang, and Matthew Brown. Learning to segment via cut-and-paste. In *Proceedings of the European Conference on Computer Vision*, pages 37–52, 2018. ii
[55] Andy Tsai, Anthony Yezzi, and Alan S Willsky. Curve evolution implementation of the mumford-shah functional for image segmentation, denoising, interpolation, and magnification. *IEEE transactions on Image Processing*, 10(8):1169–1186, 2001. i, ii, iii

[56] Luminita A Vese and Tony F Chan. A multiphase level set framework for image segmentation using the mumford and shah model. *International journal of computer vision*, 50(3):271–293, 2002. ii

[57] Bo Wahlberg, Stephen Boyd, Mariette Annergren, and Yang Wang. An admm algorithm for a class of total variation regularized estimation problems. *IFAC Proceedings Volumes*, 45(16):83–88, 2012. iv

[58] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 879–888, 2019. i

[59] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. v, vii

[60] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001. i

[61] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. i