# Learning Segmentation Masks with the Independence Prior

**Songmin Dai, Xiaoqiang Li,**[*] **Lu Wang, Pin Wu, Weiqin Tong, Yimin Chen**
School of Computer Engineering and Science, Shanghai University, China
Shanghai Institute for Advanced Communication and Data Science, Shanghai University, China
{laodar, xqli, luwang, wupin, wqtong, ymchen}@shu.edu.cn

## Abstract

An instance with a bad mask might make a composite image that uses it look fake. This encourages us to learn segmentation by generating realistic composite images. To achieve this, we propose a novel framework that exploits a new proposed prior called the independence prior based on Generative Adversarial Networks (GANs). The generator produces an image with multiple category-specific instance providers, a layout module and a composition module. Firstly, each provider independently outputs a category-specific instance image with a soft mask. Then the provided instances' poses are corrected by the layout module. Lastly, the composition module combines these instances into a final image. Training with adversarial loss and penalty for mask area, each provider learns a mask that is as small as possible but enough to cover a complete category-specific instance. Weakly supervised semantic segmentation methods widely use grouping cues modeling the association between image parts, which are either artificially designed or learned with costly segmentation labels or only modeled on local pairs. Unlike them, our method automatically models the dependence between any parts and learns instance segmentation. We apply our framework in two cases: (1) Foreground segmentation on category-specific images with box-level annotation. (2) Unsupervised learning of instance appearances and masks with only one image of homogeneous object cluster (HOC). We get appealing results in both tasks, which shows the independence prior is useful for instance segmentation and it is possible to unsupervisedly learn instance masks with only one image.

## Introduction

Deep Convolutional Neural Networks (DCNNs) have been widely used and have achieved remarkable success in supervised semantic segmentation (Long, Shelhamer, and Darrell 2015; Chen et al. 2014; Ronneberger, Fischer, and Brox 2015; Güçlü et al. 2017). But their training needs a large amount of images with pixel-level annotation. Such annotation is much more time consuming than box-level or image-level annotation, especially for the images of homogeneous object cluster (HOC) (Wu et al. 2018) because of densely distributed objects with various degrees of occlusion. Is it
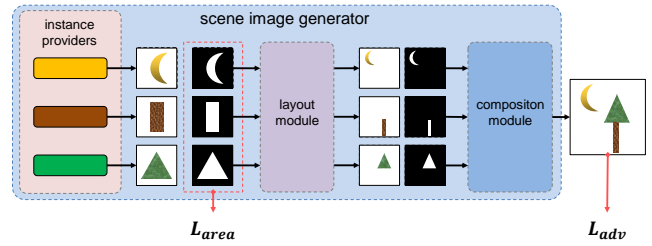
Figure 1: Our GAN based framework generates scene images by compositing instances with independent appearances and specific categories, which is trained with adversarial loss and mask area penalty.

possible to learn image segmentation with low-cost annotation or even just one image of HOC without any annotation? The answer suggested by our work is yes.

A natural image consists of multiple separated coherent objects. And objects often appear, move and change independently. It seems that the image segmentation task is just to group the coherent pixels into independent components, which can be treated as a independent component analysis (ICA) or blind source separation problem. But thoughts like grammar of images (Zhu, Mumford, and others 2007), scene graph (Newell and Deng 2017; Johnson, Gupta, and Fei-Fei 2018) and geometric relationship (Hu et al. 2018; Lin et al. 2018) imply that objects seeming disentangled are actually weakly interrelated by their poses (position, scale and orientation) at larger spatial scales. To formalize and better understand these intuitions, a graphical model for scene images is developed firstly. Then we point out that the real independent factors among objects are their appearances. Actually the appearances of objects are conditionally independent given the categories of objects. We call it as the independence prior.

We argue that the independence prior is important for building a strong connection between the quality of a composite image and the quality of the instance masks it uses. To exploit it for unsupervised instance segmentation, we propose a framework based on generative adversarial networks (GANs) (Goodfellow et al. 2014), which learns instance segmentation by optimizing the composite images from instances with specific categories, independent appear-

ances, the smallest possible masks and suitable layout. Figure 1 shows an overview of the proposed framework. We apply our framework on two practical visual tasks. In the category-specific foreground segmentation task, there are two different category-specific instance providers. One is a trainable foreground instance provider adapted from a segmentation network, and the other is an untrainable random sampler on prepared backgrounds collected from natural images. Fake images are generated by pasting carved foregrounds from input images into various backgrounds. The foreground instance provider learns instance masks by struggling to find minimal necessary regions that keep foreground instances unbroken and realistic. We use data distillation (Radosavovic et al. 2018) for the raw result to improve performance, which is useful when the training data is very limited. In the task of unsupervised learning of appearances and masks with only one image of HOC, there are many enough trainable instance providers with shared parameters to generate instances for dense homogeneous objects. Fake images are generated by compositing them with structure-free layouts. With reducing the artifacts and mask area penalty, learned appearances gradually get improved and each mask finally just covers a region that is as small as possible but enough for exactly one object.

Our main contributions are as follows:

- We formalize intuitions about natural images with a graphical model, which captures the dependence of appearance on category, the conditional independence of appearances and the geometric co-occurrence of objects.

- We propose a novel framework that learns instance appearances and segmentation masks with the independence prior.

- Our first application provides a new feasible solution for foreground segmentation on category-specific images with box-level annotation.

- Our second application shows the possibility of learning instance appearances and segmentation masks from just one image of HOC without any annotations.

## Related Work

**Localization cues based on the dependence between image part and instance category.** Methods like (Zhou et al. 2016; Selvaraju et al. 2017) use learned classification networks to find out the image parts with category-specific patterns. But such localization cues only help us to find sparse and local regions with discriminative features, which rely on the dependence between image part and instance category. (Wei et al. 2017) tried to gradually erase as much as possible discriminative region contributing for classification through a progressive trained classifier. But region without discriminative features may still be missing without their postprocessing.

**Grouping cues based on the dependence between image parts.** Intra-object parts have a high affinity with each other, but it is much lower for the affinity between inter-object parts. We think such prior about the dependence between image parts is important for bottom up grouping and

instance segmentation. Recent weakly supervised semantic segmentation methods widely use grouping cues that are built in CRF, MCG, GrabCut, ect. techniques. (Roy and Todorovic 2017; Güçlü et al. 2017; Khoreva et al. 2017; Zhou et al. 2018). But untrainable CRF, MCG and GrabCut are human designed grouping prior based on color, texture etc. appearance, which is unlearnable. Trainable CRF is possible to learn the dependence of image parts without human designed potential functions, but costly segmentation labels are necessary (Güçlü et al. 2017). (Ahn and Kwak 2018) tried to learn affinities between image parts with image-level supervision. But they modeled such affinities only on local pairs. (Pathak et al. 2017) models the dependence of image parts through the motion association of image parts from videos. By using motion cue to correct the grouping based on low-level appearance, their experiments show that this indeed improves the segmentation performance.

**Layered generative models for images.** Some generative models use image layers to composite a scene image that contains multiple objects, which can get some benefits from modeling each object separately. In the works of (Johnson, Gupta, and Fei-Fei 2018; Yang et al. 2017; Vondrick, Pirsiavash, and Torralba 2016), generators can be trained to produce a coarse soft mask for each instance without pixel-level annotation. However, these generators didn't exploit the independence prior, image layers may communicate with and thus compensate each other to reduce the artifacts of composite results, which hinders the optimization of soft masks. In (Eslami et al. 2016; Huang and Murphy 2015), it was assumed that both the appearances and poses of objects are factorized, VAE (Kingma and Welling 2013) was used to inference each layer and reconstruct the image by compositing the inferenced layer. In Tagger (Greff et al. 2016), a RNN model was designed to progressively group coherent pixels and separate the independent objects, which was supervised via a denoise task. It was assumed that removing the interference between different objects can lead to a better denoise performance. But what are the real independent factors was still not pointed out, and the dependence of these factors are not minimized directly.

**Minimizing the dependence between factors with generative models.** Our framework forces generator to take the independence prior by compositing instances with independent appearances. In the foreground segmentation tasks, we paste the carved foreground into a different background sampled independently, which is inspired by the resampling operation proposed in (Brakel and Bengio 2017). They minimized mutual information by using GAN to minimize the f-divergence (Nowozin, Cseke, and Tomioka 2016) between joint distribution and product of its marginals. A more formally expression can be found in Mutual Information Neural Estimator (MINE) (Belghazi et al. 2018).

## The Independence Prior

In this paper, we hierarchically decompose an image into three level representations: part (pixels, feature, patch, etc. visual primitives), instance (single object) and image. And

we use a corresponding description for an image with hierarchical latent factors: *appearance* (both appearance and surface mask) for the pixel level, object category and pose for the instance level and scene for the image level. Without loss of generality, we assume that a natural scene image is the result of compositing $N$ instances according to the scene-specific layout and their visible order. Images with $M < N$ instances can be treated as a special case with $N - M$ invisible instances. More formally, a natural image $x$ of scene $S$ is generated from $N$ instances with categories $c_1, c_2, ..., c_N$, category-specific *appearances* $a_1, a_2, ..., a_N$ and scene-specific poses $r_1, r_2, ..., r_N$ according to their stacking orders $1, 2, ..., N$:

$$p(x) = \int_z p(x|z)p(z)\,dz\,, \quad (1)$$

where $z$ denotes all of the above latent factors,

$$p(z) = p(c_1, a_1, r_1, c_2, a_2, r_2, ..., c_N, a_N, r_N, S)\,. \quad (2)$$

Based on our hierarchical decomposition, $p(z)$ can be further factorized into:

$$p(S)p(c_1, r_1, c_2, r_2, ..., c_N, r_N|S)\prod_{i=1}^{N} p(a_i|c_i)\,, \quad (3)$$

where $p(c_1, r_1, c_2, r_2, ..., c_N, r_N|S)$ term models the scene-specific geometric co-occurrence association of instances and captures the dependence of object categories and their poses. The *appearance* of each instance depends only on its category but not any other. The factorization term $\prod_{i=1}^{N} p(a_i|c_i)$ shows the conditional independence of instance *appearances*. See Figure 2 for a graphical model depiction.
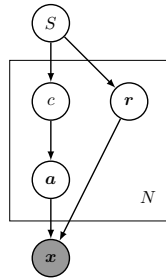


Figure 2: A graphical model for scene images that contain $N$ instances.

We can summarize the following three priors about the dependence between these factors:

- **The dependence between image part and instance category** The appearance of an object part is dependent on its object category.

- **The dependence between image parts** The appearances of inter-object parts are independent of each other given object categories, but intra-object parts are still related to each other. Intuitively, there is a much higher affinity between intra-object parts than inter-object parts.

- **The dependence between instance poses** Objects of a specific scene are related to each other by the geometric co-occurrence association, such as the coexistence of eye glasses and human face, fish and water, car and road, etc.

These properties lead to following factors that mainly affect the realistic of a composite image:

- **Appearance** The appearance of image parts should obey the statistical association corresponding to its category so that we can't find any artifacts from it.

- **Segmentation mask** The mask of an instance for image composition shouldn't break down the integrity of its surface so that the dependence between image parts won't be violated when the carved surface is combined with new surfaces with independent appearances.

- **Geometric relationship** Instances should be placed with suitable poses according to the geometric co-occurrence association between instances.



Figure 3: Images with artifacts caused by different factors. Left: Artifacts comes from the appearance. Middle: Artifacts comes from segmentation mask of the face. Right: Artifacts comes from the unsuitable placement of the eye glasses on that face.

They imply that it is possible to train generative models to jointly learn object appearances, segmentation masks and geometric relationships unsupervisedly. But jointly learning to place multiple objects and inference segmentation masks is still hard for high resolution images as far as we know, actually ST-GAN (Lin et al. 2018) shows that problems may exist even if only two objects with perfect appearances and masks need to be arranged. Our framework only focuses on the special case where layout $r_1, r_2, ..., r_N$ is easy to design by hand and object categories $c_1, c_2, ..., c_N$ are known and fixed. In such cases, artifacts caused by unsuitable layout can be avoided, and object appearances are independent of each other because the $a_1, a_2, ..., a_N$ are d-separated by $c_1, c_2, ..., c_N$. With this independence prior, we propose a framework that learns object appearances and segmentation masks by reducing artifacts like the first two examples showed in Figure 3.

## Our Framework

To learn object appearances and segmentation masks by generating realistic composite images of specific scene with GANs. We need to avoid the image layers communicating with and compensating each other so that the artifacts of each instance can be effectively showed in the final composite image. And the discriminator shouldn't win easily by
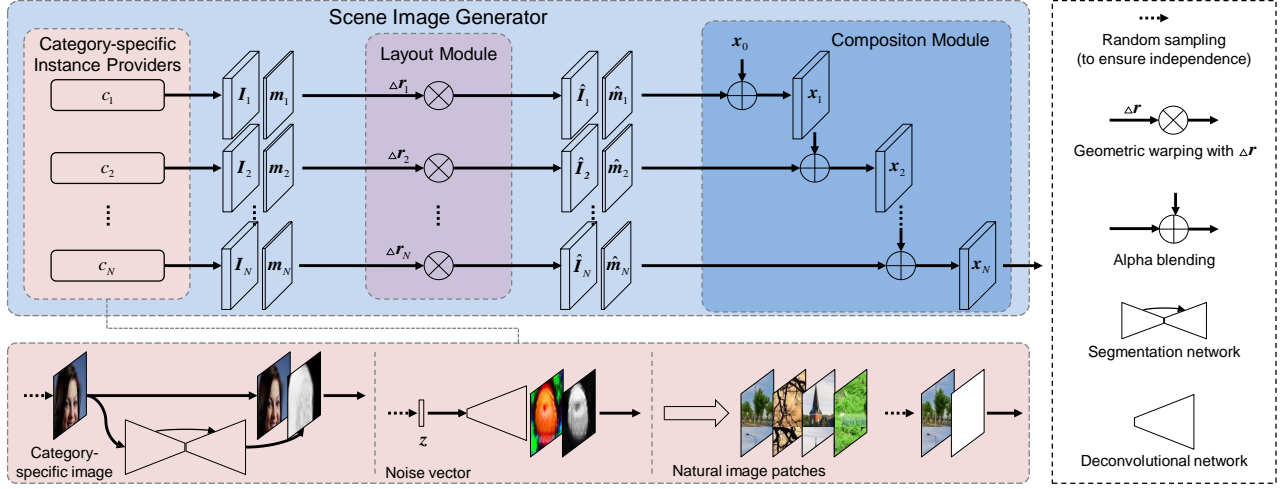
Figure 4: Generator architecture in the proposed framework. The bottom shows details of three implementation cases of instance providers which are used in our applications. The first case produces instances from randomly sampled category-specific image with a segmentation network, the second case produces instances from randomly sampled noise vector with a deconvolutional network and the last case produces instances from randomly cropped natural image patches.

finding out the artifacts caused by layout, which discourages the discriminator from finding more artifacts caused by instance masks. As showed in Figure 4, our model generates a scene-specific image by compositing multiple instances of categories $c_1, c_2, ..., c_N$ with a suitable layout. To ensure that object appearances are independent of each other, the instance images for compositing are sampled independently from category-specific instance providers. To place objects in a suitable geometric relationship, geometric warpings are used for instance images to correct their poses appearing in the final composite images according to scene-specific layout priors. Such priors could be summarized based on human observations or adapted from bounding box annotations of scene images like (Johnson, Gupta, and Fei-Fei 2018).

## Generator for Scene Images

**Category-specific Instance Providers** Category-specific instance providers are designed to independently output images of four channels for categories $c_1, c_2, ..., c_N$ according to specific tasks, the first three channels RGB capture the color appearance, the last channel Alpha represents the soft mask. Any architecture that produces category-specific instances can be implemented according to our requirements. Figure 4 shows some implementation cases of instance providers, which are used and described in our applications. More details can be found in the application section.

**Layout Module** Because the instances are outputted independently, they are also forced to own independent poses $r_1, r_2, ..., r_N$. So geometric warpings are needed to correct them into a suitable layout. The geometric warping parameters $\Delta r_1, \Delta r_2, ..., \Delta r_N$ are shared among RGBA channels:

$$
\begin{aligned}
\hat{I}_i &= ST(I_i, \Delta r_i) \\
\hat{m}_i &= ST(m_i, \Delta r_i),
\end{aligned}
\tag{4}
$$

where $\hat{I}_i$ is the warping result of $i$th layer instance's appearance $I_i$(the RGB channels), $\hat{m}_i$ is the warping result of the $i$th layer instance's soft mask $m_i$ (the Alpha channel), $ST$ is the geometric warping operator.

**Composition Module** We composite a complex image that contains many instances by recursively applying Alpha blending according to their stacking orders:

$$
x_i = \hat{I}_i \odot \hat{m}_i + x_{i-1} \odot (1 - \hat{m}_i),
\tag{5}
$$

where the $x_i$ is the composite image of the first $i$ layer instances, $x_0$ is set to be a pure black image.

## Adversarial Training with Mask Area Penalty

The key for instance providers is to provide instance images independently and thus break down the statistical association among them. This encourages instance providers not to separate high affinity parts and output complete surfaces so that artifacts will not be shown when composite with irrelevant neighbor pixels. We optimize the generator with adversarial training, but there are trivial solutions if only a adversarial loss is used. The instance outputted by a provider maybe contain completed surfaces of multiple objects even a complete scene image in order to avoid breaking down any surfaces and showing any artifacts. So a mask area penalty is used for encouraging instance providers to output a minimal necessary surface region that covers a single complete instance. Intuitively, instance providers struggle to remove parts with low affinities to necessary discriminative parts and group together the ones with high affinities. We train the discriminator $D$ and the generator $G$ by following loss functions:

$$
L_D = \mathbb{E}_{\boldsymbol{x}}[\log(D(\boldsymbol{x}))] + \mathbb{E}_{\boldsymbol{x}_g}[\log(1 - D(\boldsymbol{x}_g))]
\tag{6}
$$

$$
L_G = \mathbb{E}_{\boldsymbol{x}_g}[\log(D(\boldsymbol{x}_g))] + \lambda L_{area},
\tag{7}
$$

where $L_{area} = \mathbb{E}_{\boldsymbol{m}}[\frac{1}{N}\sum_i^N(\frac{\|\boldsymbol{m}_i\|_1}{A} - a)^2]$, $\boldsymbol{x}$ and $\boldsymbol{x}_g$ are drawn from real data and fake composite distributions, $\boldsymbol{m}_i$ is the soft mask outputted from the $i$th layer instance provider, $A$ is the image area in pixels, $a$ is an empirical constant for the smallest mask area.

## Two Applications

### Foreground Segmentation on Category-specific Images

Firstly we apply our framework on category-specific foreground segmentation tasks with box-level annotation. In this task, all images are preprocessed by cropping with box annotations. Every image contains at least one foreground of the target category. Actually there are only two instances with known and fixed categories and stacking orders. We use two instance providers to independently provide the foreground instances and background instances.

**Foreground Provider**   The foreground provider is implemented with a segmentation network. As shown in the first implementation case of Fig. 4, a foreground instance is constructed by concatenating the input RGB image and the corresponding inferred foreground mask, where we assume that foreground objects are almost always showed with little occlusion.

**Background Provider**   Generating or inferencing high-resolution backgrounds (by image inpainting) may be hard due to the variety of appearance and the non-negligible occlusion. But it is much easier to select good enough image as background, like crawl context related images from web or sample enough patches without any foreground region from the original given images. So we construct background instances by sampling patches from scene-specific natural images, and their masks just cover all the pixels of such patches, as shown in the third implementation case of Figure 4.



Figure 5: Some images of HOC. The right column is the special case, images of DSF-HOC.

**Layout Prior**   It is assumed that the dependence between foregrounds and backgrounds can be reduced to weak enough after cropping by bounding boxes, which means the instance poses are also independent of each other. So no geometric corrections are needed, and $\Delta\boldsymbol{r}_i = \boldsymbol{0}$.

### Learning Instance Appearances and Masks From One Image of HOC

Sometimes we have images of HOC that contains a large number of foreground objects with the same category and diversity appearances. Figure 5 shows some of them. Theoretically, one image of HOC contains a sufficient amount of information about the appearance of foreground objects. So it is possible to learn the appearances, segmentation masks of foreground objects from just one image of HOC even if many of them are overlapped. To efficiently demonstrate this possibility, we apply our framework on the special cases of HOCs, images of dense structure-free homogeneous object cluster (DSF-HOC), because of the simplicity of instance layout and the invisibility of background, as showed in the right column of Figure 5.

We generate small image patches but not a whole image, main reasons are: 1. We can collect much more training samples rather than only a single one image; 2. We can greatly reduce the size of image layers $N$; 3. Smaller input size is required for maintaining the visual quality of instances.

**Foreground Providers**   To learn the appearance and object mask of foreground objects, we provide foreground instances by a trainable generator that maps a random noise vector $\boldsymbol{z}$ to a RGBA image as shown in the second implementation case of Figure 4. All of $N$ foreground providers share the same generator because of the category uniformity, but sample $\boldsymbol{z}$ independently to guarantee the independence between appearances.

**Background Provider**   We construct background images for DSF-HOC patches by just sampling from pure black images or real patches with missing regions. The foreground providers are forced to generate complete foreground objects to cover them.

**Layout Prior**   Because of the structure-free property of DSF-HOCs, instances appear randomly and independently with allowed poses that range from $\boldsymbol{r}_{min}$ to $\boldsymbol{r}_{max}$. It is expected that the learned instances own canonical poses, so $\Delta\boldsymbol{r}_i$ is sampled from uniform distribution $U(\boldsymbol{r}_{min}, \boldsymbol{r}_{max})$.

## Experiments and Results

### Experimental Settings

In all experiments, we use TensorFlow (Abadi et al. 2016) library to build and train our models. Discriminators in our experiments are designed based on the SN-GAN(Miyato et al. 2018). All the other neural networks use batch normalization in most of layers. The segmentation network is adopted from UNet (Ronneberger, Fischer, and Brox 2015). Generators in the DSF-HOC tasks are designed based on deep deconvolution neural networks. All models are trained with the ADAM (Kingma and Ba 2014) optmizer with $\beta_1 = 0.5$ and

Figure 6: Foreground segmentation results on CelebA dataset. Soft masks are showed. The first and fourth rows are real images, the second and fifth rows are the corresponding images of the raw results, the third and six rows are the corresponding images of the results with data distillation. Ground truth is unavailable.
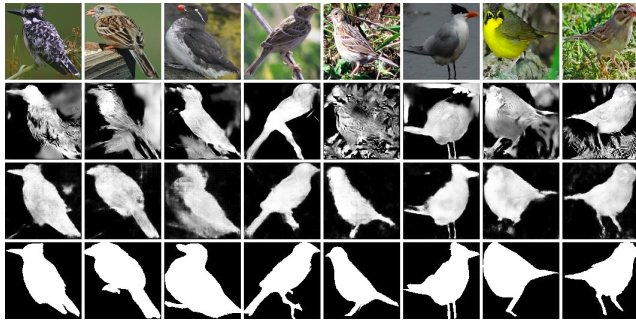


Figure 7: Foreground segmentation results on Caltech-200 bird dataset. Soft masks are showed. From top to bottom, the image rows are real images, raw results, results with data distillation and the ground truth.

$\beta_2 = 0.999$. The learning rate is fixed as 0.0002 for discriminators and 0.0004 for generators. See Appendix for further details.

## Foreground Segmentation on Category-specific Images

We evaluate our foreground segmentation method on the CelebA (Liu et al. 2015) and Caltech-200 bird (Wah et al. 2011) datasets. The CelebA dataset contains 202,599 face images with box annotation. Caltech-200 bird dataset contains 11,788 images with both box-level and pixel-level annotation.

**Data Preparation.** Firstly we crop images using provided bounding boxes. Then cropped images are scaled to size $128 \times 128$ as category-specific images. We construct the reference background images for CelebA by sampling patches with the same size from INRIAPerson (Dalal and

Triggs 2005) dataset's negative images but from images crawled from the web with keywords 'landscape' or 'tree' for Caltech-200 bird because of more tree and landscape background appearing in bird images.

**Raw Results.** The results by adversarial training are shown in the first rows under input images of Figure 6 and Figure 7. The segmentation model with adversarial training tries to infer a minimal region that covers a integral foreground object. There is no try to search architectures for segmentation networks and discriminators, but satisfactory segmentation masks are obtained on CelebA dataset. There are weird noises especially for the Caltech-200 bird dataset. These noises seem to be nothing to do with the texture or with explainable shapes. This may be caused by: 1. Intrinsic problems of adversarial training; 2. Training samples are not enough for showing the foreground and background independent effectively, enough cases with similar foregrounds appear in various backgrounds are needed; 3. Training samples may be not enough to show the generative appearance. We believe that intrinsic problems of adversarial training can be avoided with better adversarial training methods in the future. Problems caused by dataset size could be solved by finding more bird images from the web and cropping with a bird detector.

**Results with Data Distillation.** We show that the raw result can be easily improved by data distillation in the second rows under input images Figure 6 and Figure 7. With data distillation, we get appealing results on both dataset. We compute the mIOU score on Caltech-200 bird dataset using the ground-truth, and the final result gets performance of 0.78. Qualitative evaluation on CelebA dataset is hard because pixel-level annotation is missed and costly. But a visually accuracy for semantic boundary can be observed clearly. The segmentation results on Caltech-200 bird dataset show the powerful unsupervised grouping ability, even though
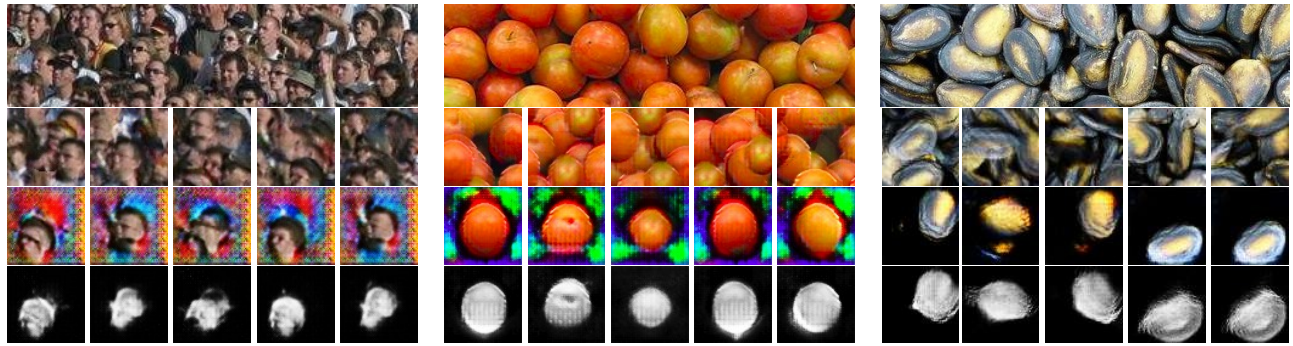
Figure 8: Instance appearances and masks learning results on three selected images of DSF-HOC. From top to bottom, the image rows are real images (only a small portion is showed), the generated composite images by our model, the learned instance appearances by our model and the learned soft mask by our model. See Appendix for more detailed results.

there is highly similar texture between the background and foreground, our method can segment out a bird shape mask which is very close to the ground truth. We argue that it is hard for discriminative methods because of lacking ability to model generative detail, part-part dependency and shape structure. And the soft mask generated by our method will be more suitable for image synthesis because of the direct optimization of the composite quality.

## Learning Instance Appearances and Masks From One Image of HOC

We evaluate our method for appearances and masks learning from one image of HOC on three selected images of DSF-HOC. They are an image of the crowd, an image of fruit cluster and an image of seed cluster. They are showed in the first row of Figure 8.

**Data Preparation.** To collect as many patches as possible and reduce the number of instance in each patch, we firstly cut out image patches with a small enough window size. We find that a window size of roughly 1.5 time of the width of the biggest instance works nicely. All patches are scaled into $64 \times 64$. Both horizontal flip and vertical flip are used for the image of fruit cluster and the image of seed cluster so we can get more patches by taking advantage of isotropy. Patches collected in this way are used as real images for training. The background images are constructed by placing a black disk with radius of roughly 20 pixels in the center of real images, which work better than pure black images.

**Layout Details.** To layout generated instances in a structure-free way, we firstly pad generated images and then randomly crop them into $64 \times 64$. The padding sizes for the crowd image, fruit cluster image and seed cluster images are set to 30 pixels, 32 pixels and 26 pixels.

**Results.** The results are shown in Figure 8. Even if there are densely distributed instances in the original images, the generators output only one single instance for each RGBA image. The object surface of every inferred instance is completed, unobscured and realistic even though the most of instances in the original image are more or less occluded.

And the generated object-like masks show highly agreement to the corresponding appearance. We believe this demonstrates great potential to extend our method to unsupervisedly learning of instance generation, segmentation, detection and counting with only a single image of DSF-HOC. In addition, we find setting padding size to 32 pixels (the fruit cluster case) leads to best alignment and centering for instances in generated images. and there are much more position variance when padding size is 26 pixels (the seed case). Intuitively, the generators may try to compensate for the insufficient poses provided by layout module.

## Conclusion and Future Works

We proposed a graphical model for natural images capturing dependencies of object appearances, categories and poses. And we pointed out that the conditional independence of object appearances is an important prior for image segmentation. Moreover, a framework was proposed for learning instance segmentation masks with low cost. Our first application provides a new low-cost solution for category-specific foreground segmentation. The generated soft masks are more suitable for image composition than the masks from discriminative methods because of our direct optimization of the composite quality. More impressively, our second application learns object appearances and soft masks with only one image of DSF-HOC without any annotation. It shows a possibility of extremely low-cost unsupervised learning of segmentation, detection, counting, etc. by transferring the knowledge from the instance generators. Our framework also provides a new perspective on universal perceptual grouping that it can be treated as separating the inputs into components with independent factors. In the future, we plan to develop a trainable multiple objects layout module by advancing the ST-GAN (Lin et al. 2018) and design an effective way to jointly train the instance providers and layout module end-to-end.

## Acknowledgments

# References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. 2016. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, 265–283.

Ahn, J., and Kwak, S. 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*.

Belghazi, I.; Rajeswar, S.; Baratin, A.; Hjelm, R. D.; and Courville, A. 2018. Mine: mutual information neural estimation. In *ICLR*.

Brakel, P., and Bengio, Y. 2017. Learning independent features with adversarial nets for non-linear ica. *arXiv preprint arXiv:1710.05050*.

Chen, L. C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *Computer Science* (4):357–361.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, 886–893.

Eslami, S. A.; Heess, N.; Weber, T.; Tassa, Y.; Szepesvari, D.; Hinton, G. E.; et al. 2016. Attend, infer, repeat: Fast scene understanding with generative models. In *NIPS*, 3225–3233.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*, 2672–2680.

Greff, K.; Rasmus, A.; Berglund, M.; Hao, T.; Valpola, H.; and Schmidhuber, J. 2016. Tagger: Deep unsupervised perceptual grouping. In *NIPS*, 4484–4492.

Güçlü, U.; Güçlütürk, Y.; Madadi, M.; Escalera, S.; Baró, X.; González, J.; van Lier, R.; and van Gerven, M. A. 2017. End-to-end semantic face segmentation with conditional random fields as convolutional, recurrent and adversarial networks. *arXiv preprint arXiv:1703.03305*.

Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; and Wei, Y. 2018. Relation networks for object detection. In *CVPR*, volume 2.

Huang, J., and Murphy, K. 2015. Efficient inference in occlusion-aware generative models of images. *arXiv preprint arXiv:1511.06362*.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image generation from scene graphs. In *CVPR*.

Khoreva, A.; Benenson, R.; Hosang, J. H.; Hein, M.; and Schiele, B. 2017. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, volume 1, 3.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. In *ICLR*.

Lin, C.-H.; Yumer, E.; Wang, O.; Shechtman, E.; and Lucey, S. 2018. St-gan: Spatial transformer generative adversarial networks for image compositing. In *CVPR*, 9455–9464.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*, 3431–3440.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral normalization for generative adversarial networks. In *ICLR*.

Newell, A., and Deng, J. 2017. Pixels to graphs by associative embedding. In *NIPS*, 2171–2180.

Nowozin, S.; Cseke, B.; and Tomioka, R. 2016. f-gan: Training generative neural samplers using variational divergence minimization. In *NIPS*, 271–279.

Pathak, D.; Girshick, R. B.; Dollár, P.; Darrell, T.; and Hariharan, B. 2017. Learning features by watching objects move. In *CVPR*, volume 1, 7.

Radosavovic, I.; Dollár, P.; Girshick, R.; Gkioxari, G.; and He, K. 2018. Data distillation: Towards omni-supervised learning. In *CVPR*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241.

Roy, A., and Todorovic, S. 2017. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *CVPR*, 7282–7291.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 618–626.

Vondrick, C.; Pirsiavash, H.; and Torralba, A. 2016. Generating videos with scene dynamics. In *NIPS*, 613–621.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wei, Y.; Feng, J.; Liang, X.; Cheng, M.-M.; Zhao, Y.; and Yan, S. 2017. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*, volume 1, 3.

Wu, Z.; Chang, R.; Ma, J.; Lu, C.; and Tang, C. K. 2018. Annotation-free and one-shot learning for instance segmentation of homogeneous object clusters. In *IJCAI*, 1036–1042.

Yang, J.; Kannan, A.; Batra, D.; and Parikh, D. 2017. Lr-gan: Layered recursive generative adversarial networks for image generation. In *ICLR*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*, 2921–2929.

Zhou, Y.; Zhu, Y.; Ye, Q.; Qiu, Q.; and Jiao, J. 2018. Weakly supervised instance segmentation using class peak response. In *CVPR*.

Zhu, S.-C.; Mumford, D.; et al. 2007. A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision* 2(4):259–362.