

Learning High-level Prior with Convolutional Neural Networks for Semantic Segmentation

Haitian Zheng, Feng Wu, Lu Fang,
 University of Science and Technology of China
 Hefei, China
 {zhenght, fengwu, fanglu}@mail.ustc.edu.cn

Yebin Liu
 Tsinghua University
 Beijing, China
 liuyebin@mail.tsinghua.edu.cn

Mengqi Ji
 The Hong Kong University of Science and Technology
 HongKong, China
 mji@ust.hk

Abstract

This paper proposes a convolutional neural network that can fuse high-level prior for semantic image segmentation. Motivated by humans' vision recognition system, our key design is a three-layer generative structure consisting of high-level coding, middle-level segmentation and low-level image to introduce global prior for semantic segmentation. Based on this structure, we proposed a generative model called conditional variational auto-encoder (CVAE) that can build up the links behind these three layers. These important links include an image encoder that extracts high level info from image, a segmentation encoder that extracts high level info from segmentation, and a hybrid decoder that outputs semantic segmentation from the high level prior and input image. We theoretically derive the semantic segmentation as an optimization problem parameterized by these links. Finally, the optimization problem enables us to take advantage of state-of-the-art fully convolutional network structure for the implementation of the above encoders and decoder. Experimental results on several representative datasets demonstrate our supreme performance for semantic segmentation.

1. Introduction

Recent years have witnessed a great success in using supervised Convolutional Neural Networks (CNN) for vision recognition problems such as image classification and object detection [18, 31, 32]. Taking advantage from the extremely powerful feature learning capability of CNN, Fully Convolutional Networks (FCN) [24] adapts the CNN structures for dense pixel-wise prediction and significantly

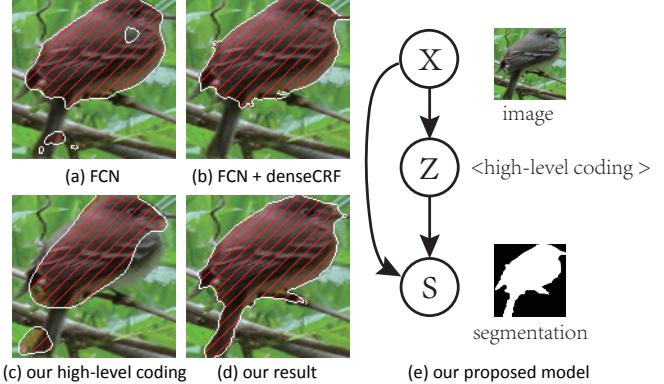


Figure 1. Our image generative model (e): a high-level coding \mathbf{z} is first sampled. After that, a mid-level semantic segmentation \mathbf{s} is sampled conditioned on \mathbf{z} . Finally image \mathbf{x} is sampled conditioned on \mathbf{s} . Comparison of results using our proposed method (d) and results using FCN segmentation (a), results using FCN + dense CRF segmentation [17] (b), and results using our high-level coding only (c).

boosts the accuracy of semantic segmentation [24]. However, disadvantage of FCN remains clear: its intrinsic local receptive field makes dense prediction locally and sometimes inconsistent with the global structure of an object. To mitigate the predication inconsistency between pixels with similar appearance, recent works introduce Conditional Random Field (CRF) into the Neural Network framework [39, 30, 22, 28, 21].

Although the integration of CRF is able to refine the poor local prediction of FCN, either the FCN or the CRF stage can still lead to problematic segmentation prediction. First, CRF is essentially a post-processing based on the FCN re-

sults, and heavily relies on the adjacent prediction. When the local receptive field of FCN causes the FCN produce largely mistaken results, CRF is incapable to recover the mislabeling region occurred at the FCN stage. Second, CRF is basically a low-level vision technique without utilizing such high-level image information as the global shape of segmentation, causing ambiguity among pixels with similar low-level features. As illustrated in Fig. 1(a) and Fig. 1(b), when the tail of the bird shares similar low-level features with background, either FCN or CRF post-processing cannot distinguish the tail from background.

In contrast to the local-oriented FCN-CRF modeling, segmentation process of the human visual system usually starts with the high-level “global scene” recognition before doing fine segmentation in local region. As an example, in Fig. 1(e), the general shape and the topic of “a bird with long tail is facing right” quickly come into human mind before the fine segmentation is carefully obtained. Such high-level semantic information is particularly helpful to avoid local ambiguity, for example, the confusion between the tail and the branch in Fig. 1.

To take advantage of the high-level semantic information, we propose in this paper a deep neural networks that can integrate high-level prior for high quality semantic image segmentation. However, because normally neural networks is not designed to perform information integration and abstraction from target signal, available supervised neural networks are infeasible for learning global semantic-level features directly from the pixel-wise annotation of segmentation. Therefore, in contrast, we model the natural image and the semantic segmentation in a generative perspective, and build a three-layer generative model called Conditional Variational Auto-encoder (CVAE) where the semantic segmentation generated from natural image as well as the hidden high-level coding, as shown in Fig. 1(e). Such a model builds up the missing link from segmentation to high-level feature utilizing the unsupervised learning methodology.

We theoretically derive the training objective of our CVAE model, then design the structure of neural networks based on the state-of-the-art FCN parsing network for implementation. It can be notably shown that even without the local features provided by FCN, our high-level feature can reconstruct the general shape of object (Fig. 1(c)). Combined with the local features, the proposed CVAE produces globally consistent prediction as shown in Fig. 1(d). In addition, we also show that our model can be combined with CRF-based post-precessing for better segmentation.

We note that two concurrent works [10, 23] on ArXiv try to integrate high level prior into deep neural networks for semantic segmentation as well. However, the former one is semi-supervised and requires elaboration of additional annotation of dataset, while the later one only uses the aver-

aged feature extracted from the last FCN feature map as the global feature. Comparably, we theoretically explain global feature as a high-level coding from a generative perspective, and show how to generate more complex global feature with additional trainable layers.

We believe that the proposed method can inspire future work aiming for better network designing for semantic segmentation that utilizes global priors. The code of this work is submitted as supplemental material and will be made public.

2. Related Work

With the emergence of deep learning [18, 31, 24] techniques, the trend and prospect of using deep neural networks for solving the long time vision problem on semantic segmentation becomes more and more clear. In this section, we mainly review deep neural networks for dense per-pixel labeling and recent works on semantic segmentation. As the proposed Conditional Variational Auto-encoder (CVAE) model is inspired from the Variational Auto-encoder (VAE) [16, 27] model, we briefly summarize the key technology of VAE as well.

Fully Convolutional Network (FCN) [24] is a special type of convolutional neural networks which replaces the fully-connected layers of CNN by convolutional layers with 1×1 kernels. With such modification, FCN efficiently outputs classification map at every spatial location.

To overcome the potentially inconsistent prediction of FCN, graphical models such as Conditional Random Field (CRF) are merged into the framework of neural networks. Specifically, Zheng *et al.* [39] and Schwing *et al.* [30] apply FCN to generate the unary term of CRF, then simulate the mean-field message passing inference of dense CRF by a specially designed recurrent neural network. Other CRF-based semantic segmentation methods such as Lin *et al.* [22] and Ross *et al.* [28] simulate a general message passing inference process with neural network. Lin *et al.* [21] train a neural network for extracting unary and piecewise potentials by applying a piecewise training strategy.

To further introduce high-level information, a recent work by Hong *et al.* [10] uses semi-supervised learning for semantic segmentation. With bounding-box annotations on PASCAL VOC dataset, FCN is combined with object detection for better performance. However, such semi-supervised approach requires elaboration of additional annotation of dataset, and is hard to apply for dataset without bounding-box annotation. Another concurrent attempt by Liu *et al.* [23] applies average pooling to the last feature map of FCN, then uses the averaged feature as global feature. Though empirical experiments show simple feature averaging does improve FCN’s prediction, different from this paper, we theoretically explain global feature as a high-

level coding from generative perspective, and show how to generate more complex global feature with additional trainable layers.

Semantic Segmentation extensively studied in the last 10 years, merges segmentation with recognition to produce per pixel semantic labeling. From discriminative perspective, one challenge is how to design image features and learning method that best discriminate different labels. At early stage, discriminative feature is usually hand-engineered, such works include [25, 19, 38] while classifiers vary from linear model, support vector machine to random forest. Recently, as CNN shows its power in discriminative vision tasks, people starts to use CNN for semantic segmentation [4, 2, 6, 26].

The other challenge is how to design model that incorporates shape prior from segmentation itself. Superpixel [19, 4], CRF [37, 19], region proposal [5] and advanced graphical model is combined with the mentioned discriminative method. High-level prior under graphical models has been largely discussed thanks to the development of unsupervised graphical models. The work such as Eslami *et al.* [3], Yang *et al.* [36], Kae *et al.* [14] and Li *et al.* [20] utilize unsupervised models such as Boltzmann Machine [9], Deep Belief Network [8] and Deep Boltzmann Machine [29] to introduce global constraint for segmentation. However, in these works, graphical models are combined with hand-engineered image features, which are usually not as discriminative as features learned from neural networks. In the experiment section of this paper, we even observe that FCN predictions alone may achieve comparable or better results compared with graphical model based methods.

Variational Auto-encoder (VAE) [16, 27] is recently brought up as a neural network based unsupervised generative model for tasks such as representation learning and data generation. It uses a two-layer hierarchical generative model, and assumes data points \mathbf{x} being generated from a random process that involves in an unobserved coding \mathbf{z} .

The generation process consists two steps: First, value $\hat{\mathbf{z}}$ is generated from some prior coding distribution $p(\mathbf{z})$. Second, value $\hat{\mathbf{x}}$ is generated from some conditional likelihood distribution $p(\mathbf{x}|\mathbf{z})$. To maximize the marginal likelihood of data point $\log p(\mathbf{x})$ with the intractable latent variable $\mathbf{z} \sim p(\mathbf{z}|\mathbf{x})$, a probabilistic *encoder* $q(\mathbf{z}|\mathbf{x})$ is introduced to approximate the true posterior $p(\mathbf{z}|\mathbf{x})$, and is used to further derive the lower-bound of the marginal likelihood:

$$\log p(\mathbf{x}) \geq -D_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) + \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}) \log p(\mathbf{x}|\mathbf{z}), \quad (1)$$

where the first term is the negative KL-divergence from prior approximation $q(z|x)$ to true prior $p(z)$, and the second term is expected reconstruction error from the coding $z \sim q(z|x)$. Then the maximization of the marginal like-

lihood is relaxed to the maximization of the above lower-bound.

Unlike most generative graphical models, the encoding distribution $q(\mathbf{z}|\mathbf{x})$ and the decoding distribution $p(\mathbf{x}|\mathbf{z})$ are parameterized by neural networks (a usual choice for distribution is multivariate Gaussian where mean and covariance are decided by neural networks). The networks implementation allows the parameters in model $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{x}|\mathbf{z})$ to be trained by stochastic gradient descent method with the unbiased SGVB gradient estimator [16].

In contrast, our proposed Conditional Variational Autoencoder (CVAE) adopts a three-layer hierarchical structure containing the high-level coding \mathbf{z} , mid-level semantic segmentation \mathbf{s} and low-level image \mathbf{x} . In the CVAE model, we derive a supervised training objective function $p(\mathbf{s}|\mathbf{x})$ and maximize this conditional marginal likelihood, see Section 3. Compared with unsupervised VAE, the proposed CVAE model enables structured supervised learning such as semantic segmentation. It can also be implemented in any networks including FCN, see Section 4.

3. Conditional Variational Auto-encoder

Our proposed generative model CVAE consists of three layers as shown in Fig. 1(e). The natural image, denoted as \mathbf{x} is considered as the given input for semantic segmentation. Given an image \mathbf{x} , the corresponding high-level coding \mathbf{z} is generated from conditional distribution $p(\mathbf{z}|\mathbf{x})$. Given both image \mathbf{x} and the corresponding high-level coding \mathbf{z} , the semantic segmentation is generated from conditional distribution $p(\mathbf{s}|\mathbf{z}, \mathbf{x})$. For convenience, we name the conditional distribution $p(\mathbf{z}|\mathbf{x})$ as the *image encoder* and the conditional distribution $p(\mathbf{s}|\mathbf{z}, \mathbf{x})$ as the *hybrid decoder*.

Apparently, such generative model indicates that the task of semantic segmentation is to maximize the conditional log probability, i.e., $\log p(\mathbf{s}|\mathbf{x})$, which involves in the marginalization of intractable hidden variable z . Similarity to VAE where such intractability is resolved by relaxing the original target function to a lower-bound function, here we try to derive the variational lower-bound of our target function. By additionally introducing a *segmentation encoder* $q(\mathbf{z}|\mathbf{s})$ to extract coding \mathbf{z} from \mathbf{s} , target function $\log p(\mathbf{s}|\mathbf{x})$ can be represented as

$$\begin{aligned}
& \log p(\mathbf{s}|\mathbf{x}) \\
&= \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{s}) \log p(\mathbf{s}|\mathbf{x}) \\
&= \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{s}) \log \frac{p(\mathbf{s}, \mathbf{z}|\mathbf{x})}{p(\mathbf{z}|\mathbf{s}, \mathbf{x})} \\
&= \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{s}) \log \frac{p(\mathbf{s}, \mathbf{z}|\mathbf{x})}{p(\mathbf{z}|\mathbf{s})} \quad (2) \\
&= \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{s}) \left(\log \frac{q(\mathbf{z}|\mathbf{s})}{p(\mathbf{z}|\mathbf{s})} + \log \frac{p(\mathbf{s}, \mathbf{z}|\mathbf{x})}{p(\mathbf{z}|\mathbf{s})} \right) \\
&= \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{s}) (\log p(\mathbf{s}, \mathbf{z}|\mathbf{x}) - \log q(\mathbf{z}|\mathbf{s}) + \log \frac{q(\mathbf{z}|\mathbf{s})}{p(\mathbf{z}|\mathbf{s})}).
\end{aligned}$$

Here, the $p(\mathbf{z}|\mathbf{s})$ in the last term is an intractable component, but the whole last term is exactly the KL-divergence from $q(\mathbf{z}|\mathbf{s})$ to $p(\mathbf{z}|\mathbf{s})$ and is always no less than zero, i.e.,

$$\sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{s}) \log \frac{q(\mathbf{z}|\mathbf{s})}{p(\mathbf{z}|\mathbf{s})} = D_{KL}(q(\mathbf{z}|\mathbf{s})||p(\mathbf{z}|\mathbf{s})) \geq 0, \quad (3)$$

so we have

$$\begin{aligned}
& \log p(\mathbf{s}|\mathbf{x}) \\
&\geq \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{s}) (\log p(\mathbf{s}, \mathbf{z}|\mathbf{x})) - \log q(\mathbf{z}|\mathbf{s}) \\
&= \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{s}) (\log p(\mathbf{z}|\mathbf{x}) + \log p(\mathbf{s}|\mathbf{z}, \mathbf{x}) - \log q(\mathbf{z}|\mathbf{s})) \quad (4) \\
&= \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{s}) \left(-\log \frac{q(\mathbf{z}|\mathbf{s})}{p(\mathbf{z}|\mathbf{x})} + \log p(\mathbf{s}|\mathbf{z}, \mathbf{x}) \right),
\end{aligned}$$

where the equality in second row holds by Bayes' rule, i.e., $\log p(\mathbf{s}, \mathbf{z}|\mathbf{x}) = \log p(\mathbf{z}|\mathbf{x}) + \log p(\mathbf{s}|\mathbf{z}, \mathbf{x})$. Here note that the first term in Eqn. 4 is a KL-divergence from $q(\mathbf{z}|\mathbf{s})$ to $p(\mathbf{z}|\mathbf{x})$, which comes

$$\begin{aligned}
& \log p(\mathbf{s}|\mathbf{x}) \\
&\geq -D_{KL}(q(\mathbf{z}|\mathbf{s})||p(\mathbf{z}|\mathbf{x})) + \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{s}) \log p(\mathbf{s}|\mathbf{z}, \mathbf{x}). \quad (5)
\end{aligned}$$

Eqn. (5) indicates the variational lower bound of log probability $\log p(\mathbf{s}|\mathbf{x})$ can be represented by our segmentation encoder $q(\mathbf{z}|\mathbf{s})$, image encoder $p(\mathbf{z}|\mathbf{x})$, and hybrid decoder $p(\mathbf{s}|\mathbf{z}, \mathbf{x})$. Our objective function then becomes

$$\max_{\theta} -D_{KL}(q(\mathbf{z}|\mathbf{s})||p(\mathbf{z}|\mathbf{x})) + \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{s}) \log p(\mathbf{s}|\mathbf{z}, \mathbf{x}), \quad (6)$$

where θ are parameters for the encoders and decoder.

With the neural network implementation of the encoders and the decoder (Section .4), the optimization of Eqn. (6) can be deployed by first solving its gradient, followed with

network optimization by a gradient-based stochastic optimizer ADAM ([15]). Specifically, the gradient of the left KL-divergence can be calculated analytically while the gradient of the right expectation term can be calculated by using SGVB (Stochastic Gradient Variational Bayes[16]) estimator. During SGVB estimation, the gradient can be calculated by sampling vector \mathbf{z} from the segmentation encoder $q(\mathbf{z}|\mathbf{s})$ for L times, then taking the averaged gradient to estimate the expectation of gradient. Note that when the batch size is large enough (approximately 50), the sampling time L can be 1.

In summary, during the training stage, image \mathbf{x} and segmentation \mathbf{s} are fed into image-encoder and segmentation-encoder respectively, generating Gaussian distribution $p(\mathbf{z}|\mathbf{x})$ and $q(\mathbf{z}|\mathbf{x})$ for computing the KL term. Then a sample \mathbf{z} following $\mathbf{z} \sim q(\mathbf{z}|\mathbf{s})$ is passed through the hybrid decoder, generating the distribution of semantic segmentation. Finally, parameters of $p(\mathbf{z}|\mathbf{x})$, $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{s}|\mathbf{z}, \mathbf{x})$ are upgraded by ADAM.

In the testing stage, we have no \mathbf{s} , but only \mathbf{x} to obtain the high-level information \mathbf{z} . That means we get \mathbf{z} from image decoder $p(\mathbf{z}|\mathbf{x})$, then passes it along with image \mathbf{x} to the hybrid decoder $p(\mathbf{s}|\mathbf{x}, \mathbf{z})$ to get segmentation \mathbf{s}

Note that unlike the concurrent work [23] that uses a simple average feature pooling to generate global feature, the formulation Eqn. (6) allows us to train a model $p(\mathbf{z}|\mathbf{x})$ that generates global features from the given image.

4. Network Implementation

Given the design of the CVAE generative model, the mentioned probabilistic encoders and decoder should then be implemented using neural network. The structure of neural network adopted is flexible. But generally, the one with better learning capability will achieve better performance for CVAE, and we therefore choose FCN in this work.

Image Encoder: While the encoder distribution can be in any form, for simplicity, we assume the image encoder distribution $p(\mathbf{z}|\mathbf{x})$ following a multivariate Gaussian with diagonal covariance. Since CNN is proven to be effective for extracting feature from image, it is used to parameterize the image encoder distribution. In this way, the mean and logarithm of diagonal covariance of the Gaussian distribution are provided by the final parallel fully-connected layers of a CNN that takes \mathbf{x} as input (as illustrated in Fig. 2(a)).

Segmentation Encoder: Similarly, $q(\mathbf{z}|\mathbf{s})$ is chosen to be multivariate Gaussian distribution as well, where the mean and logarithm of diagonal covariance are produced by another CNN. Different from the above image encoder, the segmentation encoder takes a slightly different input: because our the goal is to extract high-level coding that describes the general shape of segmentation, we resize the segmentation to be of small scale. Then the small segmen-

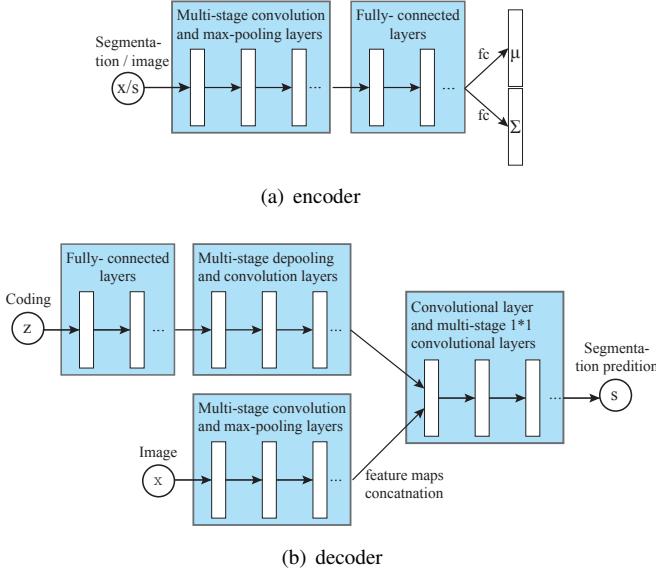


Figure 2. Network implementation of our model. (a) Image encoder $p(\mathbf{z}|\mathbf{x})$ and segment encoder $q(\mathbf{z}|\mathbf{s})$ are chosen to be multivariate where the mean and logarithm of diagonal covariance is produced by a CNN; (b) the hybrid decoder $p(\mathbf{s}|\mathbf{z}, \mathbf{x})$ decodes global feature map from \mathbf{z} by a deconvolutional network, and decodes local feature map from \mathbf{z} by a FCN, then further produces segmentation result.

tation is converted into an one-hot representation. The network implementation of segmentation encoder is illustrated in Fig. 2(a).

Hybrid Decoder: Taking \mathbf{z} and \mathbf{x} as inputs, $p(\mathbf{s}|\mathbf{z}, \mathbf{x})$ is expected to utilize the global constraint provided by \mathbf{z} and local information provided by \mathbf{x} . In our implementation, image \mathbf{x} is converted to a local feature map by standard FCN as shown at the bottom of Fig. 2(b). At the same time, high-level coding \mathbf{z} is converted to a global feature map by fully connected layers and the consequential unpooling layers (using nearest neighbor upsampling) and the convolution layers. Afterwards, the concatenation of global/local feature map is passed through one normal convolutional layer and several 1×1 convolution layers for producing the final semantic segmentation. The overall structure of the hybrid decoder is depicted in Fig. 2(b).

As both hybrid decoder and image encoder contain convolution/pooling layers to extract image features, to avoid over-fitting, a weight sharing strategy between the top layers of image encoder and hybrid decoder is adopted. Specifically, the first two convolution/pooling layers are shared between the hybrid decoder and the image encoder, while more convolution/pooling layers are applied to the image encoder for further extracting more abstract features for high-level coding. The illustration of weight sharing with real designing of the entire model is shown in Fig. 3, where

the convolution and pooling layers (i) that connected with \mathbf{x} are shared among image encoded and hybrid decoder. To improve generalization, the state-of-the-art VGG network [31] structure is used in our shared module.

After multiple pooling is performed, the resolution of feature maps gradually becomes lower, leading to the low resolution of our final prediction. Simply upsampling the prediction bi-linearly causes rough results. Inspired by the network design from [24, 7], after the entire model being trained, we optionally add an upsampling network at the end of our model (as illustrated in the supplementary material), where the low-resolution feature map at the end of network is sequentially upsampled and concatenated with the high-resolution feature map at lower layers, to produce high-resolution segmentation result.

Pretraining: In our training step, one step of segmentation encoder training is fast but takes many iteration to converge. On the other hand, one iteration VGG FCN training is slow, although takes much fewer iterations to converge. Thus when directly jointly train the model, we need to synchronize these two encoder training and requires many iterations. Thus, much of training time is wasted on the VGG FCN network waiting for segmentation encoder to converge. As noted in most recent deep neural networks such as VGG [31], GoogLeNet [32] and FCN [24] that module-wise pretraining plays a crucial role for network training, we propose to pretrain several different modules before a joint training, as shown in different colors sequentially in Fig. 3. Specifically,

- The convolution and pooling layers (i) module in Fig. 3 is shared among image encoding model and decoding model. To pretrain this module, a FCN initialized by imagenet VGG parameters is trained (as shown in Fig. 4(a)) for producing semantic segmentation.
- Meanwhile, the segment encoder is pretrained by training a VAE for generating semantic segmentation (as shown in Fig. 4(b)). The VAE that has a standard Gaussian prior ($p(\mathbf{z}) = N(z; 0, I)$) aims to learn a proper coding of segmentation. Our segmentation encoding model is then initialized by the encoding part of the trained VAE .
- To pretrain image coding model, the fixed Gaussian prior of above VAE is replaced by the image encoding model (as illustrated in Fig. 4(c)). We freeze the weight of the trained VAE model and train an image encoding model, aiming to learn an image-encoder which extracts similar coding to the trained segmentation-encoder. The training target of this model is similar to VAE model, except for the KL term is changed to the KL divergence from segmentation-encoding distribution to image-encoding distribution.

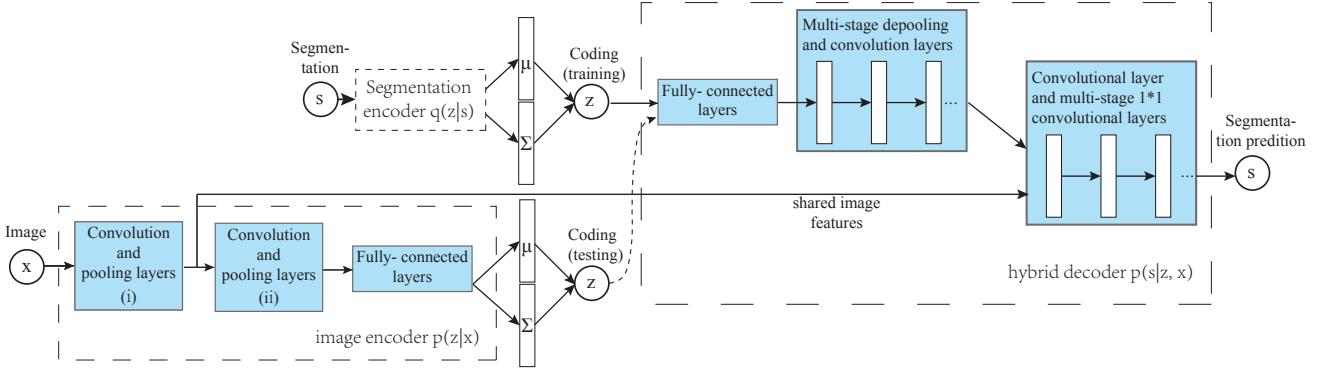


Figure 3. The overall structure our model contains three parts: an image encoder, a segmentation encoder and a hybrid decoder. The top convolution/pooling layers are shared among image encoder and hybrid decoder. During training stage, segmentation s is passed through segmentation encoder to produce $p(z|s)$ while image x is passed through image encoder to produce $p(z|x)$. Then sample $z \sim p(z|x)$ is passed through hybrid decoder to produce final segmentation result. Finally, parameters of the entire model is updated by ADAM algorithm [15]. During testing stage, image x is passed through image encoder to produce $p(z|x)$, then $z \sim p(z|x)$ is passed through hybrid decoder to produce semantic segmentation.

After the pre-training of all above modules, the entire model is jointly trained. The segmentation results produced by pretrained FCN model (Fig. 4(a)) and pretrained image coding model (Fig. 4(c)) are presented and discussed in our experiments (Section 5).

5. Experiments and Discussions

In this section, we evaluate our method on several datasets including: Labeled Faces in the Wild (LFW) that contains more than 13,000 faces [12]; Caltech-UCSD Birds 200 dataset that contains 6033 images of 200 bird species [35]; and Penn-Fudan Pedestrians dataset¹ that consists 170 images with one or more pedestrians on the background [34]. The corresponding qualitative and quantitative assessments are reported and compared to the state-of-art including CRF, CHPOPPs [20], GLOC [14] and MMBM [36]. All the quantitative evaluation results of these methods in Table 1 are given by the original papers. We also compare our method with post-processing (Our HR network + denseCRF) with MMBM + GraphCut [36] method. To fully evaluate the influence of each network on our overall scheme, we specifically implement intermediate networks and denoted as: our pretrained FCN network (Fig. 4(a)), our pretrained image-encoder network (Fig. 4(c)), our low-resolution (LR) network (Fig. 3) and our high-resolution (HR) network, respectively.

Labeled Faces in the Wild (LFW) dataset: Our model is

¹In the experiment, we use LFW Part Labels Database which contains the labeling of 2927 face images into Hair/Skin/Background labels. For Caltech-UCSD Birds 200 and Penn-Fudan Pedestrians Dataset, we use the cropped version of dataset with foreground/background annotation and train/test split provided by [36].

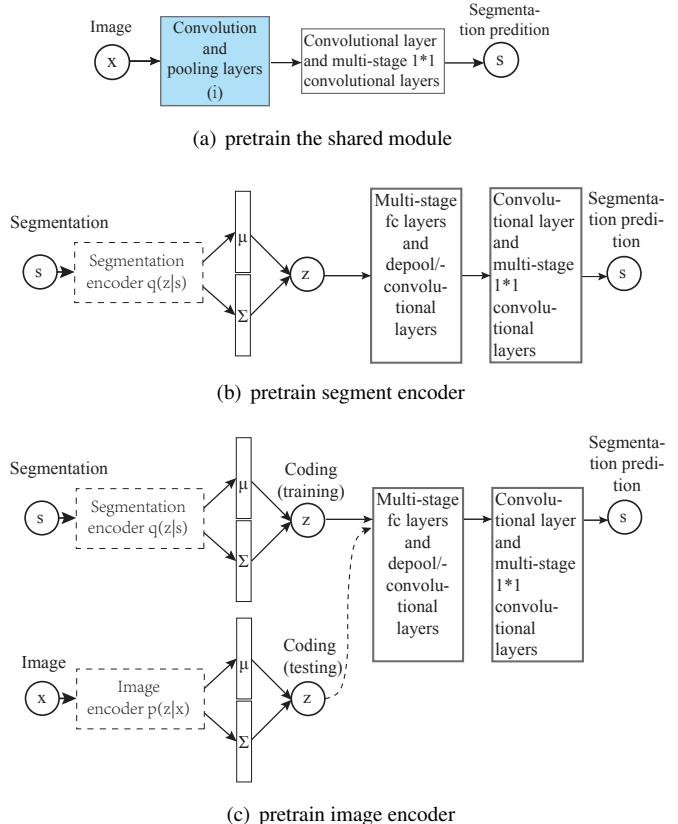


Figure 4. Illustration of how to pretrain several modules.

trained on the 1500 training images, and validated on the 500 images². For fair comparison, we employ the Super-

²As images and segmentations from the dataset are of size 250×250 ,

Method	LFW - SAP	Bird-AP	Bird-IoU	Penn-AP	Penn-IoU
CRF	93.23	83.50	38.45	84.87	68.35
CHOPPs [20]	-	74.52	48.84	86.55	71.33
MMBM1 (case1) [36]	-	80.96	60.37	82.66	64.80
MMBM1 (case2) [36]	-	87.73	72.45	85.27	69.20
MMBM1 (case3) [36]	-	75.73	63.22	83.35	65.78
MMBM1 (case4) [36]	-	88.07	72.96	89.91	76.92
MMBM2 [36]	-	86.38	69.87	89.74	77.30
Spatial CRF [14]	93.95	-	-	-	-
CRBM [14]	94.10	-	-	-	-
GLOC [14]	94.95	-	-	-	-
Our pretrained FCN [24]	94.79	89.79	77.61	90.27	76.36
Our pretrained image-encoder	90.46	84.17	67.78	86.82	69.59
Our LR network	95.88	90.86	80.08	91.46	79.34
Our HR network	96.59	91.41	81.18	91.61	79.54
Method + Post-processing	LFW-SAP	Bird-AP	Bird-IoU	Penn-AP	Penn-IoU
MMBM1 (case4) + GC [36]	-	90.42	75.92	90.42	77.97
MMBM2 (case4) + GC [36]	-	90.77	72.40	90.77	79.42
Our HR network + denseCRF	-	92.37	81.24	92.37	81.24

Table 1. Evaluation of concerned methods using Superpixel Average Precision (SAP), Average Precision (AP) and IoU (with or without post-processing) on three datasets: Labeled Faces in the Wild (denoted as LFW) [12], Caltech-UCSD Birds 200 (denoted as Bird) [35] and Penn-Fudan Pedestrians (denoted as Penn) [34].

pixel Average Precision (SAP) as noted in [14]. As our trained models provide pixel-wise prediction with different resolutions (models without upsampling layers lead to 32×32 low resolution prediction and models with upsampling layers produce 256×256 high resolution prediction), we adapt a simple scheme to predict the label of every superpixel: the segmentation result is firstly resized to be of size 250×250 by bilinear interpolation. Then for every superpixel, the number of pixels inside the superpixel is counted followed by performing a max-voting. The quantitative comparisons of concerned methods on LFW dataset are presented in Table 1, and we have the following observations:

- Among the state-of-art, GLOC [14] achieves outstanding performance mainly due to the RBM-CRF modeling and specially designed face features from [11].
- The pretrained FCN alone is quite robust, and achieves 94.79% superpixel accuracy. With the aid of global information, our network (low-resolution) boosts the FCN result from 94.79% to 95.88%. After upsampling layers that combine local image feature, the accuracy is further boosted from 95.88% to 96.59%.
- Additionally, we show that with the global vector merely, our pretrained image encoding network can predict the general shape of segmentation, and achieves 90.46% super-pixel accuracy.

Qualitative comparisons of concerned methods are illustrated in Fig. 5, where first and second columns represent

for ease of our network implementation, they are resized to 256×256 for performing our segmentation method.

the input testing image (Fig. 5(a)) and the ground truth segmentation (Fig. 5(b)), respectively. For the testing images in first and second rows, the pretrained FCN (Fig. 5(d)) alone outperforms GLOC (Fig. 5(c)), since FCN works better in extracting discriminative features than hand-crafted features that used in GLOC. However, for the other testing images, FCN fails to distinguish the object faces due to the confusing backgrounds. On the contrary, with the help of high-level coding, our network effectively utilizes the prior of faces and produces more pleasing results, as illustrated in Fig. 5(f), 5(g) and 5(h). In particular, from column (e), with high-level coding merely, the pretrained image encoder roughly reconstructs the general shape of faces, implying that the image encoder is capable to distinguish key properties of faces such as looking left/middle/right.

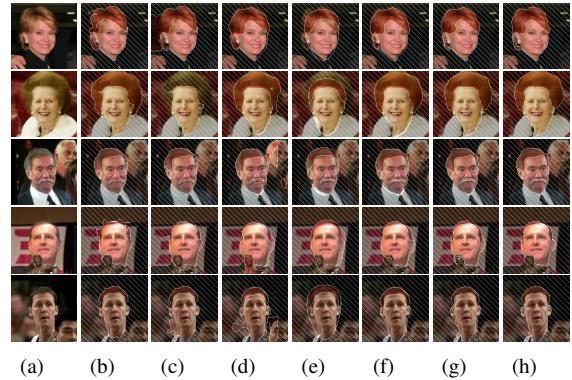


Figure 5. Qualitative results on LFW dataset. a) Input image, b) Ground truth, c) GLOC, d) pretrained FCN, e) pretrained image-encoder, f) Our LR network, g) Our HR network, h) Our HR network + denseCRF.

Caltech-UCSD Birds 200 Dataset: Table 1 reveals that pretrained FCN along achieves 89.79% average pixel accuracy, outperforming previous non-neural network global segmentation methods, since image features extracted by VGG network are more robust compared to hand-crafted features. Our LR model boosts the average precision to 90.86%, which is further boosted up to 91.41% after up-sampling. By applying a denseCRF method [17] to our prediction (denoted as ‘Our HR model + denseCRF’), we show that additional CRF-based post-processing further improves the result³. It outperforms ‘MMBM + GraphCut’ approaches [36] by boosting the accuracy to be 92.37%.

Qualitative comparisons are illustrated in Fig. 6(d). FCN has a relative small receptive field, thus the center of foreground is sometimes misclassified as background. After combining FCN with global information decoded from high-level vector, our model (Fig. 6(f)) produces notably better segmentation results. As we expected, with only high-level coding, our pretrained image-encoder is able to predict the rough semantic results (global shape of segmentations), as verified in Fig. 6(e).

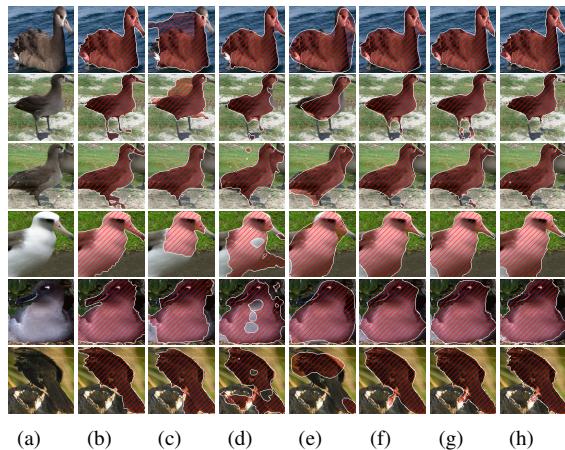


Figure 6. Qualitative comparisons on Caltech-UCSD Birds 200 dataset. a) Input image, b) Ground truth, c) GLOC, d) pretrained FCN, e) pretrained image-encoder, f) Our LR network, g) Our HR network, h) Our HR network + denseCRF.

Penn-Fudan Pedestrians Dataset: It can be shown in Table 1 that the pretrained FCN achieves 90.27% average pixel accuracy, outperforming previous global segmentation methods, as FCN features are more discriminative than hand-crafted features. With the high-level coding, our LR model and HR model achieve much higher accuracy with 91.46 and 91.61% respectively. By applying a denseCRF method [17] to our prediction (denoted as ‘Our LR net-

³Note that our model is in principle jointly trainable with other post-processing CRF networks like CRF-RNN [39] for better results. Yet due to discrepancy of implementation platforms, we did not test the joint training with CRF-RNN.

work + denseCRF’), we show that additional CRF-based post-processing further improves the result. It outperforms ‘MMBM + GraphCut’ approach [36] by boosting the accuracy to be 92.37%.

Accordingly, from qualitative comparisons in Fig. 7(d), while FCN prediction may fail due to the lack of global structure, our encoding-decoding model is able to recognize the global structure. By combining the global clues and local clues, our model produces better semantic segmentation.

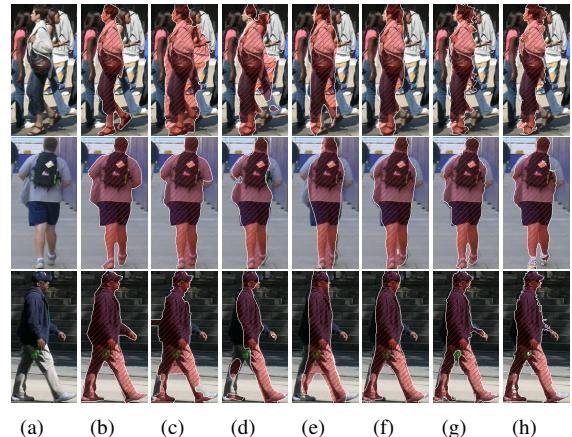


Figure 7. Qualitative comparisons on Penn-Fudan Pedestrians dataset. a) Input image; b) Ground truth; c) MMBM+graph cut; d) pretrained FCN; e) pretrained image-encoder; f) Our LR network; g) Our HR network; h) Our HR network + denseCRF.

Extensive experiments shows that, the global constraint provided by the image encoder (rather than the segmentation encoder) dominates the overall shape prior of the final result, implying that it is critical to refrain from overfitting during the training of the image encoder. For this, we adapt two strategies to reduce overfitting, i.e., to reduce the model capacity of the image / segment encoder, and to apply data augmentation techniques such as flipping and small image shifting. Undoubtedly, the former strategy will to some extent bring down the performance of the pretrained models both in train/validation set. Fortunately, our decoding network will learn to adapt the coding from training set, thus it is actually better to choose a coding with less overfitting than a coding that performs well but severely overfitted.

6. Limitations

One may notice that we did not verify our proposed model on datasets such as VOC2011/2012. While VOC datasets are much more varied, variations like shifting, scaling, rotation and overlapping prevent a trivial convolution/deconvolution structured VAE to extract good coding. However, with the rapid research advancement in advanced network structure such as [33, 13, 1] that introduces visual attention or scale invariance, a more compact cod-

ing/decoding network implementation is capable to handle the variation, which will be addressed as our future work.

7. Conclusion

Integrating global prior with local texture information is crucial for semantic segmentation. In this paper, we have proposed a conditional variational auto-encoder (CVAE) model for semantic segmentation and designed a general neural network structure to extract and utilize global information for semantic segmentation. Extensive experiments demonstrate that our proposed method outperforms available methods on several representative datasets, which shows the fact that combining global prior for semantic segmentation is feasible and promising under the deep neural networks framework.

References

- [1] S. Chintala, M. Ranzato, A. Szlam, Y. Tian, M. Tygert, and W. Zaremba. Scale-invariant learning and convolutional networks. *arXiv preprint arXiv:1506.08230*, 2015. 8
- [2] C. Couprie, C. Farabet, L. Najman, and Y. LeCun. Indoor semantic segmentation using depth information. *arXiv preprint :1301.3572*, 2013. 3
- [3] S. A. Eslami, N. Heess, C. K. Williams, and J. Winn. The shape boltzmann machine: a strong model of object shape. *IJCV*, 107(2):155–176, 2014. 3
- [4] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *PAMI*, 35(8):1915–1929, 2013. 3
- [5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 3
- [6] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, pages 297–312. 2014. 3
- [7] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 5
- [8] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006. 3
- [9] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. 3
- [10] S. Hong, H. Noh, and B. Han. Decoupled deep neural network for semi-supervised semantic segmentation. *arXiv preprint :1506.04924*, 2015. 2
- [11] G. B. Huang, M. Narayana, and E. Learned-Miller. Towards unconstrained face recognition. In *CVPR*, 2008. 7
- [12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report. 6, 7
- [13] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. *arXiv preprint :1506.02025*, 2015. 8
- [14] A. Kae, K. Sohn, H. Lee, and E. Learned-Miller. Augmenting crfs with boltzmann machine shape priors for image labeling. In *CVPR*, pages 2019–2026, 2013. 3, 6, 7
- [15] D. Kingma and B. Jimmy. Adam: A method for stochastic optimization. 4, 6
- [16] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint :1312.6114*, 2013. 2, 3, 4
- [17] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, pages 109–117, 2011. 1, 8
- [18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 1, 2
- [19] V. Lempitsky, A. Vedaldi, and A. Zisserman. Pylon model for semantic segmentation. In *NIPS*, pages 1485–1493, 2011. 3
- [20] Y. Li, D. Tarlow, and R. Zemel. Exploring compositional high order pattern potentials for structured output learning. In *CVPR*, pages 49–56, 2013. 3, 6, 7
- [21] G. Lin, C. Shen, I. Reid, et al. Efficient piecewise training of deep structured models for semantic segmentation. *arXiv preprint :1504.01013*, 2015. 1, 2
- [22] G. Lin, C. Shen, I. Reid, and A. van den Hengel. Deeply learning the messages in message passing inference. In *NIPS*, 2015. 1, 2
- [23] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *arXiv preprint :1506.04579*, 2015. 2, 4
- [24] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. 2015. 1, 2, 5, 7
- [25] A. Montillo, J. Shotton, J. Winn, J. E. Iglesias, D. Metaxas, and A. Criminisi. Entangled decision forests and their application for semantic segmentation of ct images. In *Information Processing in Medical Imaging*, pages 184–196, 2011. 3
- [26] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features. *arXiv preprint :1412.0774*, 2014. 3
- [27] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, pages 1278–1286. JMLR Workshop and Conference Proceedings, 2014. 2, 3
- [28] S. Ross, D. Munoz, M. Hebert, and J. A. Bagnell. Learning message-passing inference machines for structured prediction. In *CVPR*, pages 2737–2744, 2011. 1, 2
- [29] R. Salakhutdinov and G. E. Hinton. Deep boltzmann machines. In *ICAIS*, pages 448–455, 2009. 3
- [30] A. G. Schwing and R. Urtasun. Fully connected deep structured networks. *arXiv preprint :1503.02351*, 2015. 1, 2
- [31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2, 5
- [32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint :1409.4842*, 2014. 1, 5

- [33] Y. Tang, N. Srivastava, and R. R. Salakhutdinov. Learning generative models with visual attention. In *NIPS*, pages 1808–1816, 2014. 8
- [34] L. Wang, J. Shi, G. Song, and I.-F. Shen. Object detection combining recognition and segmentation. In *ACCV*, pages 189–199, 2007. 6, 7
- [35] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-ucsd birds 200. 2010. 6, 7
- [36] J. Yang, S. Safar, and M.-H. Yang. Max-margin boltzmann machines for object segmentation. In *CVPR*, pages 320–327, 2014. 3, 6, 7, 8
- [37] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 702–709. IEEE, 2012. 3
- [38] C. Zhang, L. Wang, and R. Yang. Semantic segmentation of urban scenes using dense depth maps. In *ECCV*, pages 708–721, 2010. 3
- [39] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. In *CVPR*, 2015. 1, 2, 8