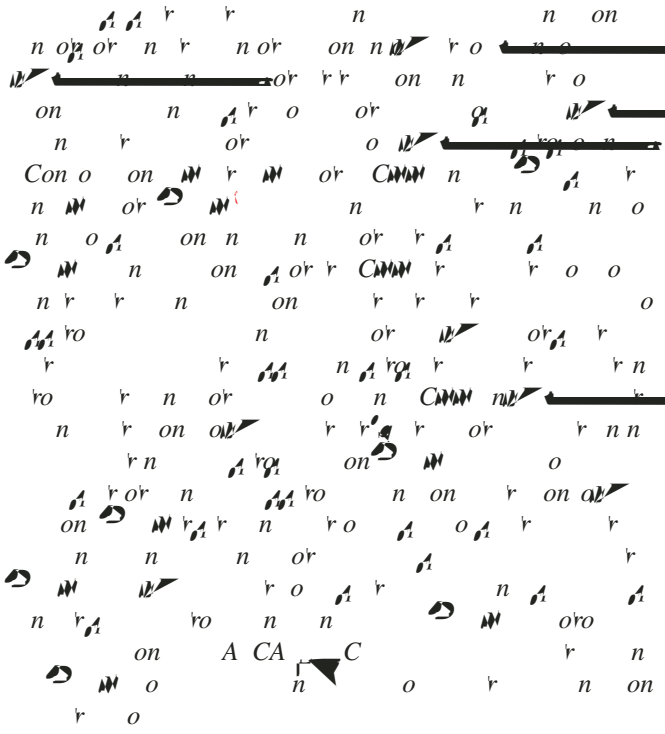
Semantic Image Segmentation via Deep Parsing Network

Ziwei Liu* Xiaoxiao Li* Ping Luo Chen Change Loy Xiaoou Tang
Depart ent of Infor ation Engineering The Chinese ni ersity of Hong Kong
{lz013,lx015,pluo,ccloy,xtang}@ie.cuhk.edu.hk

Abstract



1. Introduction

Mar o Rando Fie d MRF or Conditiona Rando Fie d CRF has achie ed great successes in se antic i age seg entation which is one of the ost cha enging prob e s in co puter ision Existing wor s such as can be genera y categorized into two groups based on their de+nitions of the unary and pairwise te s of MRF

In the +rst group researchers i pro ed abe ing ac curacy by exp oring rich infor ation to de+ne the pair wise functions inc uding ong range dependencies high order potentia s and se antic abe contexts For exa p e Krahenbuh attained accurate seg entation boundary by inferring on a fu y connected graph ineet extended by de ining both high order and ong range ter s between pixe s G oba or oca se antic contexts between abe s were a so in estigated by A though they acco p ished pro ising resu ts they ode ed the unary ter s as S M or Adaboost whose earning capacity beco es a bott enec The earning and inference of co p ex pairwise ter s are often expensi e

In the second group peop e earned a strong unary c as si+er by e eraging the recent ad ances of deep earning such as the Con o utiona Neura Networ CNN ith deep ode s these wor s de onstrated encouraging resu ts using si p e de+nition of the pairwise function or e en ignore it For instance Long transfor ed fu y connected ayers of CNN into con o utiona ayers a ing accurate per pixe c assi+ca tion possib e using the conte porary CNN architectures that were pre trained on I ageNet Chen i pro ed by feeding the outputs of CNN into a MRF with si p e pairwise potentia s but it treated CNN and MRF as separated co ponents A recent ad ance was obtained by which joint y trained CNN and MRF by passing the error of MRF inference bac ward into CNN but iterati e inference of MRF such as the ean +e d a gorith MF is required for each training i age during bac propagation BP Zheng further showed that the procedure of MF inference can be represented as a Recurrent Neura Networ RNN but their co putationa costs are si i ar e found that direct y co bing CNN and MRF as abo e is inef+cient because CNN typica y has i ions of para eters whi e MRF infers thousands of atent ariab es and e en worse incorporating co p ex pairwise ter s into MRF beco es i practica i iting the perfor ance of the entire syste

This wor proposes a no e Deep Parsing Networ DPN which is ab e to joint y train CNN and co p ex pairwise ter s DPN has se era appea ing **properties**

*indicates shared +rst authorship

Project page http://personal.ie.cuhk.edu.hk/~lz013/projects/DPN.html For ore technica detai s p ease contact the corresponding author Ping Luo ia p uo hi g ai co

DPN solves MRF with a single feed forward pass reducing computational cost and meanwhile maintaining high performance. Specifically, DPN models unary terms by extending the GG network. GG_{16} pre-trained on ImageNet, while additional layers are carefully designed to model complex pairwise terms. Learning of these terms is transformed into deterministic end to end computation by BP instead of embedding MF into BP as did. Although MF can be represented by RNN, it needs to recurrently compute the forward pass so as to achieve good performance and thus is time consuming, each forward pass contains hundred thousands of weights. DPN approximates MF by using only one iteration. This is made possible by joint learning strong unary terms and rich pairwise information. Pairwise terms determine the graphical structure. In previous works, if the former is changed, so is the latter as well as its inference procedure. But with DPN, modifying the complexity of pairwise terms, range of pixels and contexts, is as simple as modifying the receptive fields of convolutions, without varying BP. DPN is able to represent multiple types of pairwise terms, making any previous works such as, as its special cases. DPN approximates MF with convolutional and pooling operations, which can be speeded up by low rank approximation and easily parallelized in a Graphical Processing Unit, GPU.

Our **contributions** are summarized as below. (1) A novel DPN is proposed to jointly train GG_{16} and rich pairwise information. (2) Our model can be trained end-to-end. (3) Compared to existing deep models, DPN can approximate MF with only one iteration, reducing computational cost but still maintaining high performance. (4) We discover that DPN represents multiple types of MRFs, making any previous works such as RNN and DeepLab as its special cases. (5) Extensive experiments investigate which component of DPN is crucial to achieve high performance. A single DPN model achieves a new state of the art accuracy of on the PASCAL VOC test set. (6) We analyze the time complexity of DPN on GPU.

2. Our Approach

DPN learns MRF by extending GG_{16} to model unary terms and additional layers are carefully designed for pairwise terms.

Overview MRF is an undirected graph where each node represents a pixel in an image $\mathbf{I}$, and each edge represents relation between pixels. Each node is associated with a binary latent variable $y_u^i \in \{0, 1\}$, indicating whether a pixel i has label u. We have $\forall u \in L = \{1, 2, ..., l\}$, representing a set of l labels. The energy function of MRF is written as

$$E(\mathbf{y}) = \sum_{i \in \mathbf{V}} \psi_u(y_i^u) + \sum_{i,j \in \mathbf{E}} \psi_p(y_i^u, y_j^v), \quad (1)$$

where $\mathbf{y}$, $\mathcal{V}$ and $\mathcal{E}$ denote a set of latent variables, nodes and edges, respectively. $\psi_u(y_i^u)$ is the unary term measuring the cost of assigning label u to the i-th pixel. For instance, if pixel i belongs to the first category other than the second one, we should have $\psi_u(y_i^1) < \psi_u(y_i^2)$. Moreover, $\psi_p(y_i^u, y_j^v)$ is the pairwise term that measures the penalty of assigning labels u, v to pixels i, j respectively.

Intuitively, the unary terms represent per-pixel classifications, while the pairwise terms represent a set of smoothness constraints. The unary term in Eqn. (1) is typically defined as

$$\psi_u(y_i^u) = -\ln p(y_i^u = 1|\mathbf{I}),$$

where $p(y_i^u = 1|\mathbf{I})$ indicates the probability of the presence of label u at pixel i, modeling by GG_{16}. To simplify discussions, we abbreviate it as p_i^u. The smoothness term can be formulated as

$$\psi_p(y_i^u, y_j^v) = \mu(u, v)\mathrm{d}(i, j),$$

where the first term learns the penalty of global co-occurrence between any pair of labels, the output value of $\mu(u, v)$ is large if u and v should not coexist, while the second term calculates the distances between pixels, $\mathrm{d}(i, j) = \omega_1\|\mathbf{I}_i - \mathbf{I}_j\|^2 + \omega_2\|[x_i\ y_i] - [x_j\ y_j]\|^2$. Here $\mathbf{I}_i$ indicates a feature vector such as RGB values extracted from the i-th pixel, x, y denote coordinates of pixels positions, and ω_1, ω_2 are the constant weights. Eqn. (3) implies that if two pixels are close and look similar, they are encouraged to have labels that are compatible. It has been adopted by most of the recent deep models for semantic image segmentation.

However, Eqn. (3) has two main drawbacks. First, its first term captures the co-occurrence frequency of two labels in the training data, but neglects the spatial context between objects. For example, person may appear beside table, but not at its bottom. This spatial context is a mixture of patterns, as different object configurations may appear in different images. Second, it defines only the pairwise relations between pixels, missing their high order interactions.

To resolve these issues, we define the smoothness term by leveraging rich information between pixels, which is one of the **advantages** of DPN over existing deep models. We have

$$\psi_p(y_i^u, y_j^v) = \sum_{k=1}^{K} \lambda_k\ \kappa_k(i, u, j, v) \sum_{z \in \mathbf{N}_j} \mathrm{d}(j, z)p_z^v. \quad (4)$$

The first term in Eqn. (4) learns a **mixture of local label contexts** penalizing label assignment in a local region

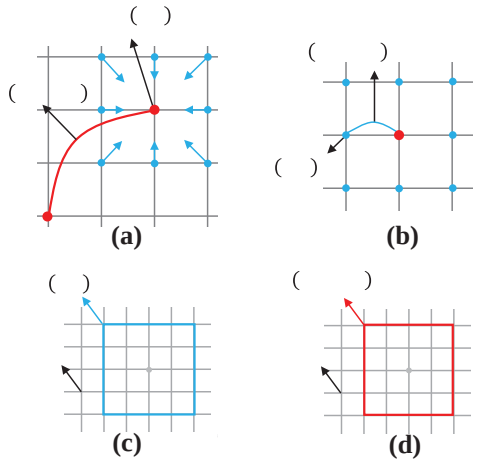


Figure . a I ustration of the pairwise ter s in DPN. b exp ains the abe contexts. c and d show that ean +e d update of DPN corresponds to con o utions.

where K is the nu ber of co ponents in ixture and λ_k is an indicator deter ining which co ponent is acti ated. e de+ne $\lambda_k \in \{0, 1\}$ and $\sum_{k=1}^{K} \lambda_k = 1$. An intuiti e i ustration is gi en in Fig. b, where the dots in red and b ue represent a center pixe i and its neighboring pixe s j. $j \in \mathcal{N}_i$ and (i, u) indicates assigning abe u to pixe i. Here (i, u, j, v) outputs abe ing cost between (i, u) and (j, v) with respect to their re ati e positions. For instance, if u, v represent person and tab e, the earned pena ties of positions j that are at the botto of center i shou d be arge. The second ter basica y ode s a **triple penalty** which in o es pixe s i, j, and j s neighbors, i p ying that if (i, u) and (j, v) are co patib e, then (i, u) shou d be a so co patib e with j s nearby pixe s (z, v), $\forall z \in \mathcal{N}_j$, as shown in Fig. a.

Learning para eters. weights of GG**16** and costs of abe contexts in Eqn. is to ini ize the distances between ground truth abe ap and y, which needs to be inferred subject to the s oothness constraints.

Inference Overview Inference of Eqn. can be obtained by the ean +e d MF a gorith , which esti ates the joint distribution of MRF $P(\mathbf{y}) = \frac{1}{Z}\exp\{-E(\mathbf{y})\}$ by using a fu y factorized proposa distribution $Q(\mathbf{y}) = \prod_{i \in \mathbf{V}} \prod_{u \in L} q_i^u$, where each q_i^u is a ariab e we need to esti ate, indicating the predicted probabi ity of assigning abe u to pixe i. To si p ify the discussion, we denote (y_i^u) and (y_i^u, y_j^v) as ^u_i and $^{uv}_{ij}$, respecti e y. $Q(\mathbf{y})$ is typica y opti ized by ini izing a free energy function of MRF

$$F(Q) = \sum_{i \in \mathbf{V}} \sum_{u \in L} q_i^u \, ^u_i + \sum_{i,j \in \mathbf{E}} \sum_{u \in L} \sum_{v \in L} q_i^u q_j^v \, ^{uv}_{ij} + \sum_{i \in \mathbf{V}} \sum_{u \in L} q_i^u \ln q_i^u. \quad (4)$$

Speci+ca y, the +rst ter in Eqn. characterizes the cost of each pixe s predictions, whi e the second ter char acterizes the consistencies of predictions between pixe s. The ast ter is the entropy easuring the con+dences of predictions. To esti ate q_i^u, we differentiate Eqn. with respect to it and equate the resu ting expression to zero. e then ha e a c osed for expression

$$q_i^u \propto \exp\Big\{-\big(^u_i + \sum_{j \in \mathbf{N}_i} \sum_{v \in L} q_j^v \, ^{uv}_{ij}\big)\Big\}, \quad (5)$$

such that the predictions for each pixe is independent y attained by repeating Eqn. , which i p ies whether pixe i ha e abe u is proportiona to the esti ated probabi ities of a its neighboring pixe s weighted by their corresponding s oothness pena ties. Substituting Eqn. into , we ha e

$$q_i^u \propto \exp\Big\{-\big(^u_i - \sum_{k=1}^{K} \lambda_k \sum_{v \in L} \sum_{j \in \mathbf{N}_i} {}_k(i, u, j, v) \sum_{z \in \mathbf{N}_j} \mathrm{d}(j, z) q_j^v q_z^v\big)\Big\},$$

where each q_i^u is initia ized by the corresponding p_i^u in Eqn. , which is the unary prediction of GG**16**. Eqn. satis+es the s oothness constraints.

In the fo owing, DPN approxi ates one iteration of Eqn. by deco posing it into two steps. Let Q^v be a predicted abe ap of the v th category. In the +rst step as shown in Fig. c, we ca cu ate the trip e pena ty ter in. by app ying a $m \times m$ + ter on each position j, where each e e ent of this + ter equa s $\mathrm{d}(j, z) q_j^v$, resu ting in $Q^{v\prime}$. Apparent y, this step s oothes the prediction of pixe j with respect to the distances between it and its neighborhood. In the second step as i ustrated in. d, the abe ing contexts can be obtained by con o ving $Q^{v\prime}$ with a $n \times n$ + ter, each e e ent of which equa s ${}_k(i, u, j, v)$, pena izing the trip e re ations as shown in. a.

3. Deep Parsing Network

This section describes the i p e entation of Eq. in a Deep Parsing Networ . DPN. DPN extends GG**16** as unary ter and additiona ayers are designed to approxi ate one iteration of MF inference as the pairwise ter . The hyper para eters of GG**16** and DPN are co pared in Tab e.

VGG16 As isted in Tab e. a, the +rst row represents the n of ayer and x y in the second row represents the of the recepti e +e d and the r of con o ution, respecti e y. For instance, in the con o utiona ayer i p ies that the recepti e +e d of each + ter is $\times$ and it is app ied on e ery sing e pixe of an input feature ap, whi e in the ax poo ing ayer indicates each feature

(a) VGG16:

layer	×conv	max	×conv	max	×conv	max	×conv	max	×conv	max	×fc	fc
filter stride												
channel												
activation	relu	idn	relu	idn	relu	idn	relu	idn	relu	idn	relu	soft
size												

(b) DPN:

layer	×conv	max	×conv	max	×conv	max	×conv	×conv	conv	conv	conv	lconv	conv	bmin	sum
filter stride															
channel															
activation	relu	idn	relu	idn	relu	idn	relu	relu	relu	relu	sigm	lin	lin	idn	soft
size															

Table 1. The comparisons between the network architectures of GG_{16} and DPN, as shown in (a) and (b) respectively. Each table contains five rows representing the **name of layer**, **receptive field of filter – stride**, **number of output feature maps**, **activation function** and **size of output feature maps**, respectively. Further more, **conv**, **lconv**, **max**, **bmin**, **fc**, and **sum** represent the convolution, local convolution, max pooling, bilinear pooling, fully connection and summation, respectively. Moreover, relu, idn, soft, sigm, and lin represent the activation functions, including rectified linear unit (relu), identity, softmax, sigmoid and linear, respectively.

map is pooled over every other pixel within a 3×3 local region. The last three rows show the number of the output feature maps, activation functions, and the size of output feature maps, respectively. As summarized in Table 1(a), GG_{16} contains thirteen convolutional layers, five max pooling layers, and three fully connected layers. These layers can be partitioned into twelve groups, each of which covers one or more homogenous layers. For example, the first group comprises two convolutional layers with 3×3 receptive field and 64 output feature maps, each of which is 224×224.

3.1. Modeling Unary Terms

To make full use of GG_{16}, which is pre-trained by ImageNet, we adopt all its parameters to initialize the filters of the first ten groups of DPN. To simplify the discussions, we take PASCAL VOC 2012 as an example. Note that DPN can be easily adapted to any other semantic image segmentation dataset by modifying its hyper parameters. VOC 2012 contains 21 categories and each image is rescaled to 512×512 in training. Therefore, DPN needs to predict total $512 \times 512 \times 21$ labels, one label for each pixel. To this end, we extends GG_{16} in two aspects.

In particular, let a^i and b^i denote the i-th group in Table 1(a) and (b), respectively. First, we **increase resolution** of GG_{16} by removing its max pooling layers at a^4 and a^5, because most of the information is lost after pooling. a^5 reduces the input size by 16 times (from 512×512 to 32×32). As a result, the smallest size of feature map in DPN is 64×64, keeping much more information compared with GG_{16}. Note that the filters of b^1 are initialized as the filters of a^1, but the 3×3 receptive field is padded into 5×5 as shown in Fig. 2(a), where the cells in white are the original values of the 3×3 filter and the cells in gray are zeros. This is done because a^1 is not presented in DPN, such that each filter in a^2 should be convolved on every other pixel of a^1. To maintain the convolution with one stride, we pad the filters with zeros. Furthermore, the feature maps in b^1 are upsampled to 512×512 by bilinear interpolation. Since DPN is trained with labels of the entire images, the missing information in the preceding layers of b^1 can be recovered by BP.

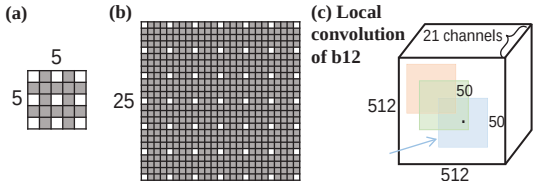


Figure 2. (a) and (b) show the padding of the filters. (c) illustrates local convolution of b^{12}.

Second, two fully connected layers at a^6 are transformed to two convolutional layers at b^6 and b^7, respectively. As shown in Table 1(a), the first fc layer learns $4096 \times 512 \times 7 \times 7$ parameters, which can be alternatively regarded as 4096 filters in b^6, each of which is $512 \times 7 \times 7$. Since a^6 and a^7 have been reduced the 3×3 receptive field is padded into 7×7 similar as above and shown in Fig. 2(b). The second fc layer learns a $4096 \times 4096 \times 1 \times 1$ weight matrix, corresponding to 4096 filters in b^7. Each filter is $4096 \times 1 \times 1$.

Operator b^8 generates the unary labeling results, producing twenty one 512×512 feature maps, each of which represents the probabilistic label map of each category.

3.2. Modeling Smoothness Terms

The last four layers of DPN, from b^8 to b^{11}, are carefully designed to smooth the unary labeling results.

- **b^{12}** As listed in Table 1(b), conv in b^{12} indicates a **locally convolutional layer**, which is widely used in face recognition [3, 4, 5] to capture different information from different facial positions. Similarly, distinct spatial positions of b^{12} have different filters, and each filter is shared across 21 input channels, as shown in Fig. 2(c). It

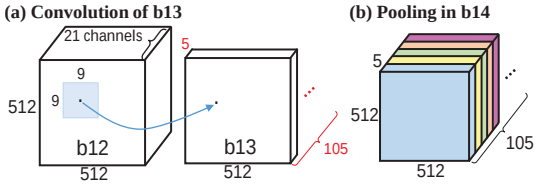


Figure . a and . b i ustrates the con o utions of b and the poo ings in b .

can be for u ated as

$$\mathbf{12}_{(j,v)} = \mathrm{lin}(\mathbf{k}_{(j,v)} * \mathbf{11}_{(j,v)}),$$

where $\mathrm{lin}(x) = ax + b$ representing the inear acti ation function, $*$ is the con o utiona operator, and $\mathbf{k}_{(j,v)}$ is a × × + ter at position j of channe v. Ideal e have $\mathbf{k}_{(j,1)} = \mathbf{k}_{(j,2)} = ... = \mathbf{k}_{(j,21)}$ shared across channe s. $\mathbf{11}_{(j,v)}$ indicates a oca patch in b , whi e $\mathbf{12}_{(j,v)}$ is the corresponding output of b . Since b has stride one, the resu t of $\mathbf{k}_j * \mathbf{11}_{(j,v)}$ is a sca ar. In su ary, b has × different + ters and produces output feature aps.

Eqn. i p e ents the **triple penalty** of Eqn. . Re ca that each output feature ap of b indicates a proba bi istic abe ap of a speci+c object appearing in the i age. As a resu t, Eqn. suggests that the probabi ity of object v presented at position j is updated by weighted a eraging o er the probabi ities at its nearby positions. Thus, as shown in Fig . c, $\mathbf{11}_{(j,v)}$ corresponds to a patch of Q^v centered at j, which has a ues $p_z^v, \forall z \in \mathcal{N}_j^{\mathbf{50\times50}}$. Si i ar y, $\mathbf{k}_{(j,v)}$ is initia ized by $\mathrm{d}(j,z)p_z^v$, i p ying each + ter captures dissi i arities between positions. These + ters re ain +xed during BP, other than earned as in con entiona CNN .

- **b13** As shown in Tab e . b and Fig . a, b is a con o utiona ayer that generates feature aps by using + ters of size × × . For exa p e, the a ue of $(i,u = 1)$ is attained by app ying a × × + ter at positions $\{(j, v = 1, ..., 21)\}$. In other words, b earns a + ter for each category to pena ize the probabi istic abe aps of b , corresponding to the **local label contexts** in Eqn. by assu ing $K = 5$ and $n = 9$, as shown in Fig . d .
- **b14** As i ustrated in Tab e and Fig . b, b is a b oc in poo ing ayer that poo s o er e ery × region with one stride across e ery input channe s, eading to output channe s. ÷ = . b activates the contextua pattern with the s a est pena ty.

- **b15** This ayer co bines both the unary and s ooth ness ter s by su ing the outputs of b and b in an e e ent wise anner si i ar to Eqn. .

$$\mathbf{15}_{(i,u)} = \frac{\exp\left[\ln(\mathbf{11}_{(i,u)}) - \mathbf{14}_{(i,u)}\right]}{\sum_{u=1}^{\mathbf{21}} \exp\left[\ln(\mathbf{11}_{(i,u)}) - \mathbf{14}_{(i,u)}\right]},$$

where probabi ity of assigning abe u to pixe i is nor a ized o er a the abe s.

Relation to Previous Deep Models Many existing deep ode s such as , , , e p oyed Eqn. as the pairwise ter s, which are the specia cases of Eqn. . To see this, et K= and j=, the right hand side of . reduces to

$$\exp\{-\,{}^{u}_{i} - \sum_{v\in L} \lambda_{\mathbf{1}}\, \mathbf{1}(i,u,i,v) \sum_{z\in \mathbf{N}_i} \mathrm{d}(i,z)p_i^v p_z^v\}$$
$$= \exp\{-\,{}^{u}_{i} - \sum_{v\in L} (u,v) \sum_{z\in \mathbf{N}_i, z\neq i} \mathrm{d}(i,z)p_z^v\}, \quad (.$$

where (u,v) and $\mathrm{d}(i,z)$ represent the g oba abe co occurrence and pairwise pixe si i arity of Eqn. , respec ti e y. This is because $\lambda_{\mathbf{1}}$ is a constant, $\mathrm{d}(i,i) = 0$, and $(i,u,i,v) = (u,v)$. Eqn. is the corresponding MF update equation of .

3.3. Learning Algorithms

Learning The +rst ten groups of DPN are initia ized by $\mathrm{GG}_{\mathbf{16}}$, whi e the ast four groups can be initia ized rando y. DPN is then +ne tuned in an incre enta anner with four stages. During +ne tuning a these stages so e the pixe wise soft ax oss , but updating different sets of para eters.

First, we add a oss function to b and +ne tune the weights fro b to b without the ast four groups, in order to earn the unary ter s. Second, to earn the trip e re ations, we stac b on top of b and update its para eters , $\omega_{\mathbf{1}}, \omega_{\mathbf{2}}$ in the distance easure, but the weights of the preceding groups , b ∼b are +xed. Third, b and b are stac ed onto b and si i ar y, their weights are updated with a the preceding para eters +xed, so as to earn the oca abe contexts. Fina y, a the para eters are joint y +ne tuned.

Implementation DPN transfor s Eqn. into con o utions and poo ings in the groups fro b to b , such that + tering at each pixe can be perfor ed in a para e anner. Assu e we ha e f input and f' output feature aps, $N \times N$ pixe s, + ters with $s \times s$ recepti e +e d, and a ini batch with M sa p es. b ta es a tota $f \cdot N^{\mathbf{2}} \cdot s^{\mathbf{2}} \cdot M$ operations, b ta es $f \cdot f' \cdot N^{\mathbf{2}} \cdot s^{\mathbf{2}} \cdot M$ operations, whi e both b and b require $f \cdot N^{\mathbf{2}} \cdot M$ operations.

Each + ter in b actua y represents a distance etric between pixe s in a speci+c region. In OC , the patterns of a the training i ages in a speci+c region are hetero geneous because of arious object shapes. Therefore, we initia ize each + ter with Euc idean distance. Ne ertheless, Eqn. is a ore genera for than the trip e pena ty in Eqn. , + ters in can be auto atica y earned fro data if the patterns in a speci+c region are ho ogenous, such as face or hu an i ages, which ha e ore regu ar shapes than i ages in OC .

e use the re eased GG_{16} ode , which is pub ic a ai ab e at http://www.robots.ox.ac.uk/~vgg/research/very_deep/

For exa p e, when $M=$ as in our experi ent, we ha e $\times 2\times 2\times =\times 11$ operations in b , which has the highest co p exity in DPN. e para e ize these operations using atrix u tip ication on GP as did b can be co puted within s. The tota unti e of the ast four ayers of DPN is s. Note that con o utions in DPN can be further speeded up by ow ran deco positions of the + ters and ode co pressions .

Howe er, direct ca cu ation of Eqn. is acce erated by fast Gaussian + tering . For a ini batch of ten × i ages, a recent y opti ized i p e entation ta es seconds on CP to co pute one iteration of . Therefore, DPN a es easier to be para e ized and speeded up.

4. Experiments

Dataset. e e a uate the proposed approach on the PAS CAL OC (OC dataset, which contains object categories and one bac ground category. Fo owing pre ious wor s such as , , , we e p oy 10,582 i ages for training, 1,449 i ages for a idation, and 1,456 i ages for testing.

Evaluation Metrics. A existing wor s e p oyed ean pixe wise intersection o er union, denoted as Io , to e a uate their perfor ance. To fu y exa ine the effec ti eness of DPN, we introduce another three etrics, in c uding tagging accuracy, TA, oca ization accuracy, LA, and boundary accuracy, BA, . TA co pares the predict ed i age e e tags with the ground truth tags, ca cu ating the accuracy of u ti c ass i age c assi+cation. LA e a uates the Io between the predicted object bounding boxes and the ground truth bounding boxes, denoted as bIo , easuring the precision of object oca ization. For those objects that ha e been correct y oca ized, we co pare the predicted object boundary with the ground truth boundary, easuring the precision of se antic bound ary si i ar to .

Comparisons. DPN is co pared with the best perfor ing ethods on OC , inc uding FCN , Zoo out , DeepLab , SSL , BoxSup , Piecewise , and RNN . A these ethods are based on CNNs and MRFs, and trained on OC data fo owing . They can be grouped according to different aspects. **joint-train.** Piecewise and RNN . **w/o joint-train.** DeepLab, SSL, FCN, and BoxSup . **pre-train on COCO.** RNN, SSL, and BoxSup. The +rst and the second groups are the ethods with and without joint training CNNs and MRFs, respecti e y. Methods in the ast group a so e p oyed MS COCO to pre train deep ode s. To conduct a co prehensi e co parison, the perfor ance of DPN are reported on both settings, with and without pre training on COCO.

They are the bounding boxes of the predicted seg entation regions.

Recepti e Fie d	base ine	×	×	×
Io			**64.7**	

a Co parisons between different recepti e +e ds of b .

Recepti e Fie d	×	×	×	ixtures
Io				**66.5**

b Co parisons between different recepti e +e ds of b .

Pairwise Ter s	DSN	DeepLab	DPN
i pro e ent.			**5.4**

c Co paring pairwise ter s of different ethods.

Tab e Ab ation study of hyper para eters.

In the fo owing Sec. i estigates the effecti eness of different co ponents of DPN on the OC *on* Sec. co pares DPN with the state of the art ethods on the OC .

4.1. Effectiveness of DPN

A the ode s e a uated in this section are trained and tested on OC .

Triple Penalty. The recepti e +e d of b indicates the range of trip e re ations for each pixe . e exa ine different settings of the recepti e +e ds, inc uding × , × , and × , as shown in Tab e a, where × achie es the best Io , which is sight y better than × . For a × i age, this resu t i p ies that × neighborhood is suf+cient to capture re ations between pixe s, whi e s a er or arger regions tend to under +t or o er +t the training data. Moreo er, a ode s of trip e re ations outperfor the base ine ethod that ode s dense pairwise re ations / GG denseCRF .

Label Contexts. Recepti e +e d of b indicates the range of oca abe context. To e a uate its effecti eness, we +x the recepti e +e d of b as × . As su arized in Tab e b, × ixtures i pro es preceding settings by , , , and percent respecti e y. e obser e arge gaps exist between × and × . Note that the × recepti e +e d of b corresponds to earning a g oba abe co occurrence without considering oca spatia contexts. Tab e c shows that the pairwise ter s of DPN are ore effecti e than DSN and DeepLab.

More i portant y, Io of a the categories can be i pro ed when increasing the size of recepti e +e d and earning a ixture. Speci+ca y, for each category, the i pro e ents of the ast three settings in Tab e b o er the +rst one are ± , ± , and ± respecti e y. e a so isua ize the earned abe co patibi ities and contexts in Fig. a and b respecti e y. a is obtained by su ing each + ter in b o er × region, indicating

The other deep ode s such as RNN and Piecewise did not report the exact i pro e ents after co bining unary and pairwise ter s.

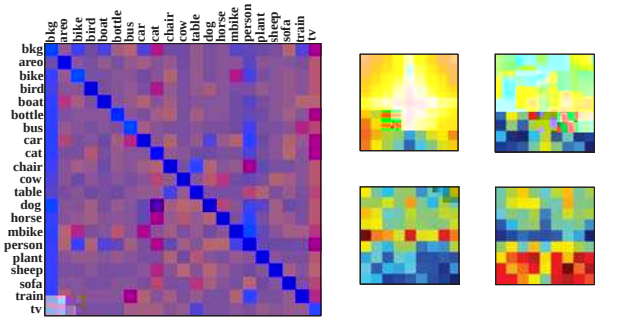


Figure isua ization of a earned abe co patibi ity b earned contextua infor ation **(Best viewed in color)**

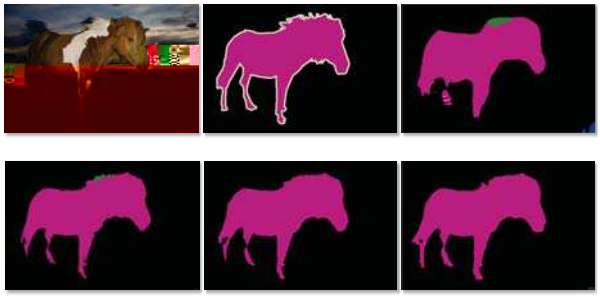


Figure Step by step isua ization of DPN **(Best viewed in color)**

how i e y a co u n object wou d present when a row object is presented B ue represents high possibi ity a is non sy etry For exa p e when horse is presented person is ore i e y to present than the other objects A so chair is co patib e with tab e and b g is co patib e with a the objects b isua izes so e contextua patterns where A B indicates that when A is presented where B is ore i e y to present For exa p e b g is around train otor bi e is be ow person and person is sitting on chair

Incremental Learning As discussed in Sec DPN is trained in an incre enta anner The right hand side of Ta b e a de onstrates that each stage eads to perfor ance gain co pared to its pre ious stage For instance trip e pena ty i pro es unary ter by percent whi e abe contexts i pro es trip e pena ty by percent More i portant y joint ne tuning a the co ponents unary ter s and pairwise ter s in DPN achie es another gain of percent A step by step isua ization is pro ided in Fig

e a so co pare incre enta earning with joint earning which ne tunes a the co ponents of DPN at the sa e ti e The training cur es of the are p otted in Fig a showing that the for er eads to higher and ore stab e accuracies with respect to different iterations whi e the atter ay get stuc at oca ini a This difference

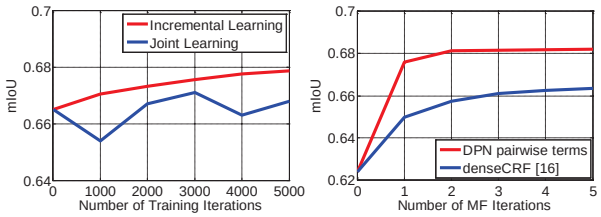


Figure Ab ation study of a training strategy b required MF iterations **(Best viewed in color)**

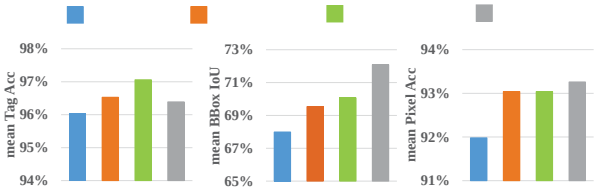


Figure Stage wise ana ysis of a ean tagging accuracy b ean oca ization accuracy c ean boundary accuracy

is easy to understand because incre enta earning on y introduces new para eters unti a existing para eters ha e been ne tuned

One-iteration MF DPN approxi ates one iteration of MF Fig b i ustrates that DPN reaches a good accuracy with one MF iteration A CRF with dense pairwise edges needs ore than iterations to con erge It a so has a arge gap co pared to DPN Note that the existing deep ode s such as required iterations to con erge as we

Different Components Modeling Different Information e further e a uate DPN using three etrics The resu ts are gi en in Fig For exa p e a i ustrates that the tagging accuracy can be i pro ed in the third stage as it captures abe co occurrence with a ixture of contextua patterns Howe er TA is decreased a itt e after the na stage Since joint tuning axi izes seg entation accu racies by opti izing a co ponents together extre e y s a objects which rare y occur in OC training set are discarded As shown in b accuracies of object oca ization are signi cant y i pro ed in the second and the na stages This is intuiti e because the unary prediction can be re ned by ong range and high order pixe re ations and joint training further i pro es resu ts c disc oses that the second stage a so captures object boundary since it easures dissi i arities between pixe s

Per-class Analysis Tab e a reports the per c ass accuracies of four e a uation etrics where the rst four rows represent the Io of four stages whi e the ast three rows represent TA LA and BA respecti e y e ha e se era a uab e obser ations which oti ate future researches Joint training bene ts ost of the categories

	areo	bi e	bird	boat	bott e	bus	car	cat	chair	cow	tab e	dog	horse	bi e	person	p ant	sheep	sofa	train	t	Avg.
nary Ter . Io																					
[Trip e Pena ty																					
[Labe Contexts			82.6					82.6		67.9			70.3					77.8			
[Joint Tuning	84.8	37.5		66.3	67.5	84.2	76.4		33.8		50.4	76.8		74.9	81.1	48.3		41.8	76.6	60.4	67.8
TA . tagging Acc																					
LA . bIo																					
BA . boundary Acc																					

. a Per c ass resu ts on OC

	areo	bi e	bird	boat	bott e	bus	car	cat	chair	cow	tab e	dog	horse	bi e	person	p ant	sheep	sofa	train	t	Io
FCN																					
Zoo out																					
Piecewise																					
DeepLab																					
RNN																					
SSL†																					
RNN†														86.4							
BoxSup†	89.8		89.2	68.9				87.7			67.1								81.2	70.7	
DPN																					
DPN†		61.6			74.7	91.2	84.3		36.5	86.3		84.4	87.8		85.4	63.6	87.3	61.3			77.5

. b Per c ass resu ts on OC . The approaches pre trained on COCO are ar ed with †.

Tab e . Per c ass resu ts on OC .

except ani a s such as bird, cat, and cow . So e instances of these categories are extre e y s a so that joint training discards the for s oother resu ts . Training DPN with pixe wise abe aps i p icit y ode s i age e e tags since it achie es a high a eraged TA of . Object oca ization a ways he ps. Howe er, for the object with co p ex boundary such as bi e, its Io is ow e en it can be oca ized . bi e has high LA but ow BA and Io . Fai ures of different categories ha e different factors. ith these three etrics, they can be easi y identi ed. For exa p e, the fai ures of chair, tab e, and p ant are caused by the dif cu ties to accura e y capture their bounding b oxes and boundaries. A though bott e and t are a so dif cu t to oca ize, they achie e oderate Io because of their regu ar shapes. In other words, Io of bott e and t can be signi cant y i pro ed if they can be accurate y oca ized.

4.2. Overall Performance

As shown in Tab e . b, we co pare DPN with the best perfor ing ethods on OC test set based on two settings, with and without pre training on COCO. The approaches pre trained on COCO are ar ed with †. e e a uate DPN on se era sca es of the i ages and then a erage the resu ts fo owing .

DPN outperfor s a the existing ethods that were trained on OC , but DPN needs on y one MF iteration to so e MRF, other than iterations of RNN, DeepLab, and Piecewise. By a eraging the resu ts of two DPNs, we achie e accuracy on OC without outside training data. As discussed in Sec. , MF iteration is the ost co p ex step e en when it is i p e ented as con o utions. Therefore, DPN at east reduces × runti e co pared to pre ious wor s.

Fo owing , we pre train DPN with COCO, where object categories that are a so presented in OC are se ected for training. A sing e DPN† has achie ed Io on OC test set. As shown in Tab e . b, we obser e that DPN† achie es best perfor ances on ore than ha f of the object c asses.

5. Conclusion

e proposed Deep Parsing Networ (DPN) to address se antic i age seg entation, which has se era appea ing properties. First, DPN uni es the inference and earning of unary ter and pairwise ter s in a sing e con o utiona networ . No iterati e inference are required during bac propagation. Second, high order re ations and ixtures of abe contexts are incorporated to its pairwise ter s ode ing, a ing existing wor s ser e as specia cases. Third, DPN is bui t upon con entiona operations of CNN, thus easy to be para e ized and speeded up.

DPN achie es state of the art perfor ance on OC , and u tip e pa uab e facts about se antic i age seg en tion are re ea ed through extensi e experi ents. Future directions inc ude in estigating the genera izabi ity of DPN to ore cha enging scenarios, arge nu ber of object c asses and substantia appearance sca e ariations.

References

[1] A. Ada s, J. Bae, and M. A. Da is. Fast high di ensiona filtering using the per utohedra attice. In *Co* *or* pages .

` The resu ts of these ethods were presented in either the pub ished papers or arXi pre prints.

P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation.

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *C*

S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer. cudnn: Efficient primitives for deep learning. In

J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. *rXiv*, May

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *C*

M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *C*

C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling.

P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *C*

T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *C*

B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *CV*

B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *CV*

G. E. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In

M. Jaderberg, A. Vedaldi, and A. Zisserman. Speeding up convolutional neural networks with low rank expansions. In *C*

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models.

P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials.

P. Krähenbühl and V. Koltun. Parameter learning and convergent inference for dense random fields. In

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In

G. Lin, C. Shen, I. Reid, and A. Hengel. Efficient piecewise training of deep structured models for semantic segmentation. *rXiv*, Apr

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *CV*

C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer.

J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *C*

P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. In *C*

P. Luo, X. Wang, and X. Tang. Pedestrian parsing via deep decompositional network. In *CV*

M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *C*

R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *C*

M. Opper, O. Winther, et al. From naive mean field theory to the tap equations.

G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly- and semi-supervised learning of a dcnn for semantic image segmentation. *rXiv*, May

X. Ren and J. Malik. Learning a classification model for segmentation. In *CV*

A. G. Schwing and R. Urtasun. Fully connected deep structured networks. *rXiv*, Mar

J. Shi and J. Malik. Normalized cuts and image segmentation.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *C*

Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In

M. Szummer, P. Kohli, and D. Hoiem. Learning crfs using graph cuts. In *CV*

Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *C*

Vineet, G. Sheasby, J. Warrell, and P. H. Torr. Posefield: An efficient mean-field based method for joint estimation of human pose, segmentation, and depth. In *n*

on Co

pages. Springer.

Vineet, J. Warrell, and P. H. Torr. Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. In *CV*

J. Yang, B. Price, S. Cohen, and M.-H. Yang. Context driven scene parsing with attention to rare classes. In *C*

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. *rXiv*, Apr