

EFFICIENT LIKELIHOOD LEARNING OF A GENERIC CNN-CRF MODEL FOR SEMANTIC SEGMENTATION

Alexander Kirillov¹ Dmitrij Schlesinger¹ Walter Forkel¹ Anatoly Zelenin¹

Shuai Zheng² Philip Torr² Carsten Rother¹

¹Dresden University of Technology

²University of Oxford

ABSTRACT

Deep models, such as Convolutional Neural Networks (CNNs), are omnipresent in computer vision, as well as structured models, such as Conditional Random Fields (CRFs). Combining them brings many advantages, foremost the ability to incorporate prior knowledge into CNNs, e.g. by explicitly modelling the dependencies between output variables. In this work, we present a CRF model where unary factors are dependent on a CNN. Our main contribution is an efficient and scalable, maximum likelihood-based, learning procedure to infer all model parameters jointly. Previous work either concentrated on piecewise training, or maximum likelihood learning of restricted model families, such as Gaussian CRFs or CRFs with a few variables only. In contrast, we are the first to perform maximum likelihood learning for large-sized factor graphs with non-parametric potentials. We have applied our model to the task of semantic labeling of body parts in depth images. We show that it is superior to selected competing models and learning strategies. Furthermore, we empirically observe that our model can capture shape and context information of relating body parts.

1 INTRODUCTION

Deep models had tremendous success in the last few years in many areas of computational science. In computer vision, Convolutional Neural Networks were successfully applied for a wide range of applications – ranging from low-level vision, like segmentation and optical flow, to high-level vision, like scene understanding and semantic segmentation. For instance in the VOC2012 object segmentation challenge¹ the use of CNNs has pushed the Intersection-over-Union score by around 28% (from around 50% to currently around 78% (Lin et al., 2015)). The main contribution of CNNs is the ability to adaptively fine-tune millions of features to achieve best performance for the task at hand. However, CNNs have also their shortcomings. One limitation is that often a large corpus of labeled training images is necessary. Secondly, it is difficult to incorporate prior knowledge into a CNN architecture. To this end, graphical models like Conditional Random Fields (CRFs) (Lafferty et al., 2001) overcome these two limitations. CRFs have been used to model geometric properties, such as object shape, spatial relationship between objects, global properties like object connectivity, and many others. Furthermore, by designing CRFs, based on e.g. physical properties, they are able to achieve good results even with few training images. For these reasons, a recent trend has been to explore the combination of these two modeling paradigms. By doing so, CRFs are able to use the incredible power of CNNs, to fine-tune all model features. On the other hand, CNNs can more easily capture global properties such as object shape and contextual information. The study of this fruitful combination is the main focus of our work. This combination is also sometimes called “deep structured models”, introduced by (Chen et al., 2015). To this end, we present a new generic CRF model, where unary factors are based on a CNN. In this context, “generic” means that our factors are of a non-parametric form, in contrast to e.g. (Zheng et al., 2015) which considers Gaussian

¹<http://host.robots.ox.ac.uk:8080/leaderboard>

pairwise potentials. Our main contribution is an efficient and scalable maximum-likelihood learning procedure which jointly trains all model parameters of a large factor. For this we introduce an idea that reminds of Persistent Contrastive-Divergence (Tieleman, 2008), which in our case means to perform one sampling-pass over the image for computing the gradient. This helps to avoid the expensive inference step. The main practical benefit is the applicability to large factor graphs, e.g. we use factor graphs with 60K variables and 1.2M factors. In contrast, the two maximum-likelihood-based approaches suggested in (Chen et al., 2015) do not scale to large factor graphs, since it either performs slow inference or requires an practically infeasible size of working memory for storing auxiliary information (inference-related variables).

Related work. The idea of making CRF models more powerful by allowing factors to depend on many parameters has been explored extensively over the last decade. One example is the Decision Tree Field approach (Nowozin et al., 2011) where factors are dependent on Decision Trees. In this work, we are interested in making the factors dependent on CNNs. Note that one advantage of CNNs over Decision Trees is that CNNs learn the appropriate features for the task at hand, while Decision Trees, and many other classifiers, only combine and select from a pool of simple features, see e.g. (Sethi, 1990; Richmond et al., 2015) for a discussion on the relationship between CNNs and Decision Trees. We now describe the most relevant works that combine CNN and CRFs.

Since CNNs have been used for semantic segmentation, this field has made a big leap forward, as mentioned above, see e.g. (Long et al., 2014; Farabet et al., 2013). Recently, the advantages of additionally integrating a CRF model has given a further boost in performance, as demonstrated by many works. To the extent that the work (Lin et al., 2015) is currently leading the VOC2012 object segmentation challenge, as discussed below. In (Chen et al., 2014) a fully connected Gaussian CRF model (Krähenbühl & Koltun, 2011) was used, where the respective unaries are supplied by a CNN. The CRF inference was done with a Mean Field approximation. This separate training procedure was recently improved by (Zheng et al., 2015) with an end-to-end learning algorithm. To achieve this, they represent the Mean Field iterations by a Recurrent Neural Network. The same idea was published in (Schwing & Urtasun, 2015). In (Zheng et al., 2015), the Recurrent Network is made efficient by using a so-called permutohedral lattice approximation (Adams et al., 2010) for Gaussian filters. However, this approach allows for a special class of pairwise potentials only. In contrast, we do not assume any parametric form of our potentials. Hence, in our approach pairwise potentials can become “repulsive”². Repulsive potentials are especially useful in applications where the spatial relation between objects and object parts is important, like body part labelling or semantic segmentation. Indeed, we observe empirically that our model is superior to (Zheng et al., 2015). Besides the approach of (Zheng et al., 2015) and (Schwing & Urtasun, 2015), there are many other works that consider the idea of backpropagation with a so-called unrolled CRF-inference scheme, such as (Domke, 2013; Kiefel & Gehler, 2014; Barbu, 2009; Ross et al., 2011; Stoyanov et al., 2011; Tompson et al., 2014; Liu et al., 2015). These inference steps mostly correspond to message passing operations of e.g. Mean Field updates or Belief Propagation. In our work, we do not express the CRF inference as an unrolled inference, especially we do not perform any inference during training.

Another related approach for combining CNNs and CRFs was presented by (Lin et al., 2015). They deal with non-parametric pairwise terms. In particular, both unary and pairwise terms are dependent on CNNs. Note that our approach can also be extended to handle CNN-based pairwise terms. They apply their method to semantic segmentation of images, i.e. large factor graphs, and are currently leading the VOC2012 object segmentation competition. In order to deal with large factor graphs and large number of training images³ they have to avoid the expensive inference step during training, as we do. To apply maximum likelihood-based learning they consider two model approximations: a) pseudo-likelihood and b) piecewise training, which is simply the un-conditioned product of all factors. They found the latter to give empirically better results. It is important to note that since CNN-CRF models are highly none log-linear there are no guarantees on the approximation with respect to the true maximum likelihood estimate. In contrast, we do not approximate the model, but instead use stochastic maximization of the true likelihood. Hence, in the limit we will converge to the exact gradient ascent of the true likelihood.

²There is no common definition of repulsive potentials for more than two labels. Informally, such pairwise potentials suppress equal labels in contrast to e.g. the Potts model, which promotes spatial smoothness.

³They use VOC2012 and COCO dataset, i.e. a total of 133K images.

The closest work to ours is by (Chen et al., 2015). As in our approach it utilizes Maximum Likelihood learning, without model approximation. However, the key difference is that their approach to learning does not scale to large factor graphs. They demonstrate the application of image tagging and word prediction, where the maximum factor graph has 38 nodes and 700 pairwise factors. In contrast, we show results for 60K variables and 1.2M pairwise factors. Scaling our approach to even larger graphs is straightforward, especially when moving our highly parallelizable and low memory approach from CPU to GPU. In detail, (Chen et al., 2015) propose to substitute the true marginals by local beliefs, which are computed via Loopy Belief Propagation. This requires the computation of local beliefs at each iteration of the gradient update which can be very computational expensive, especially for problems like semantic segmentation where pixel-level graphical models are used. They also propose an elegant way to overcome this drawback by blending message passing iterations of Loopy BP with gradient updates. However, this approach requires storing dual variables of local beliefs for each sample in the training dataset. For semantic segmentation the storage space would be 10 – 100 times larger than one required for the dataset itself. In contrast, our approach only requires additional storage to store the current labelling for each training image.

Contribution. We present a generic CRF model where unary factors are based on a CNN. Our main contribution is an efficient and scalable, maximum likelihood-based learning procedure to infer all model parameters jointly. In contrast to related work, we are the first to perform maximum likelihood learning for large-sized factor graphs with non-parametric potentials. In particular, our learning procedure is highly parallelizable with a low memory footprint, and hence is ideal for a future GPU-based implementation. We have applied our method to the task of semantic labeling of body parts in depth images. We demonstrate highly competitive results, and observe the ability of our model to capture shape and contextual information of relating body parts.

2 THE MODEL

Our model is a Conditional Random Field (CRF) (Lafferty et al., 2001) with unary potentials that depend on the image content through a Convolutional Neural Network (CNN). Let $G = (R, E)$ be a graph with the node set R and the edge set E . In our application the nodes correspond to image pixels. Each pixel takes a label l from a pre-defined, finite, discrete label set L . For instance, in body part labelling the labels have values such as “head”, “left hand”, “torso” etc. The task is to assign a label to each image pixel, i.e. to obtain a mapping $y : R \rightarrow L$. We denote by $y_i \in L$ the label chosen for node $i \in R$, and more general, y_A denotes the restriction of the labelling y to a subset of nodes $A \subseteq R$. To model the posterior probability distribution of labellings y , given images x , we use a pairwise CRF model with associated energy

$$E(x, y, \theta) = \sum_{i \in R} \psi_i(y_i, x, \theta) + \sum_c \sum_{ij \in E_c} \psi_c(y_i, y_j, \theta). \quad (1)$$

The unary potentials $\psi_i : L \rightarrow \mathbb{R}$ assign values to each label $l \in L$ of pixel $i \in R$, depending on the image content x . For this we use a CNN, which is described in more detail later. Concerning the pairwise potentials, the CRF architecture is similar to various previous work, such as (Flach & Schlesinger, 2011; Nowozin et al., 2011), i.e. we use a densely connected neighbourhood structure. The set of all edges is divided into “classes”, indexed by c in (1). Thereby one class is characterized by the translational vector that connects the corresponding pixels. For example, one particular class may consist of all edges connecting pixels that are neighbours in the horizontal direction in the image grid. Another class might be the set of all edges which connects pixels that are 2 pixels horizontally and 3 pixels vertically apart, etc. A subset of edges for one class is denoted by E_c . The pairwise potentials $\psi_c : L \times L \rightarrow \mathbb{R}$ are class specific, i.e. all edges of the same class share parameters. The free model parameters are the network weights w and the pairwise potentials ψ_c for all c . They are summarized by θ in (1). The corresponding probability distribution is defined as⁴

$$p(y|x; \theta) = \frac{1}{Z(x, \theta)} \exp[-E(x, y, \theta)], \quad (2)$$

with the image specific partition function

$$Z(x, \theta) = \sum_y \exp[-E(x, y, \theta)]. \quad (3)$$

⁴The parameters are separated from random variables by semicolon.

3 LEARNING AND INFERENCE

Given a training sample $((x_1, y_1), (x_2, y_2) \dots (x_T, y_T))$ our goal is to maximize its conditional log-likelihood with respect to the unknown parameters θ :

$$(\arg) \max_{\theta} \sum_{t=1}^T \ln p(y_t | x_t; \theta) = (\arg) \max_{\theta} \sum_{t=1}^T [-\ln Z(x_t, \theta) - E(x_t, y_t, \theta)]. \quad (4)$$

The gradient of the likelihood for a particular example t (for many examples their gradients should be summed up) is

$$\nabla_{\theta} = -\frac{\partial \ln Z(x_t, \theta)}{\partial \theta} - \frac{\partial E(x_t, y_t, \theta)}{\partial \theta} = \mathbb{E}_{p(y|x_t; \theta)} \left[\frac{\partial E(x_t, y, \theta)}{\partial \theta} \right] - \frac{\partial E(x_t, y_t, \theta)}{\partial \theta} \quad (5)$$

where $\mathbb{E}_p[\xi]$ denotes the expectation of a random variable ξ with the probability distribution p . Since the computation of this expectation is not tractable, we use a stochastic approximation, i.e. we substitute the expectation of a random variable by its realization (one sample) as follows:

1. Sample a labelling \hat{y} according to the current probability distribution $p(y|x_t; \theta)$,
2. Compute stochastic gradient as

$$\nabla_{\theta} = \frac{\partial E(x_t, \hat{y}, \theta)}{\partial \theta} - \frac{\partial E(x_t, y_t, \theta)}{\partial \theta}. \quad (6)$$

Accelerating the sampling procedure. We use Gibbs Sampling (Geman & Geman, 1984) to sample \hat{y} . Starting from an arbitrary labelling, we sample a new label in each node i according to the probability distribution in this particular node, keeping all other labels fixed:

$$p_i(y_i=l|x, y_{R \setminus i}; \theta) \propto \exp \left[-\psi_i(l) - \sum_c (\psi_c(l, y_{j'}) + \psi_c(y_{j''}, l)) \right]. \quad (7)$$

Note, that according to our CRF architecture there are exactly two edges (apart from pixels close to the image-grid border) in each edge class c that are incident to a given node i , the corresponding nodes are denoted by j' and j'' in (7). We will call one scan over the whole image (i.e. performing (7) for all i) a sampling iteration.

In the theory, one should perform many sampling iterations starting from an arbitrary labelling in order to overcome the so-called burn-in phase of Gibbs Sampling, i.e. to obtain a sample that does not depend on the initialization. Moreover, this should be done for each gradient computation (6). In order to accelerate the complete learning procedure, we use the following trick. Assume that we have already a labelling y that is sampled according to the current probability distribution (the burn-in phase is over). After applying the gradient step, the model parameters (and hence, the probability distribution) have not changed very much, since the gradient is usually applied with a small step-size. Therefore, it is sufficient to perform just one sampling iteration in order to obtain a sample that matches the change in the probability distribution adequately, although the labelling obtained in such a manner obviously depends on the previous one. Note, that the same trick is used in the Persistent Contrastive-Divergence scheme.

However, in contrast to Persistent Contrastive-Divergence we have a different target probability distribution. As we show next, this simplifies our learning process even further, by circumventing any burn-phase of our sampler. In Persistent Contrastive-Divergence one has to draw samples from the *joint* distribution $p(y, x)$, while we have to draw samples from the *posterior* distribution $p(y|x)$. The posterior distribution is conditionally independent at the beginning of the learning, since we set all pairwise potentials to zero. Note that the label sampled according to (7) for a node i depends on the previous labelling only through the labels in the neighbouring nodes. If all pairwise potentials are zero then there is no dependency at all. Hence, for an independent model it is enough to perform just one sampling iteration in order to obtain a perfect sample from the target probability distribution. On the other hand, at the end of the learning process gradients are very small (ideally zero). Consequently, the model parameters do not change much in each gradient step. Hence, one sampling iteration is also sufficient. This gives an important practical benefit, namely the possibility to start from a conditionally independent, posterior probability distribution $p(y|x)$. In particular we can use a Neural Network that is pre-trained by standard methods, without a CRF.

Computing the gradient. It remains to explain how the gradient $\partial E(x, y, \theta) / \partial \theta$ in (6) is computed. Let us first consider the parameters of the unary potentials. As mentioned above, we use a CNN to model unary terms. The output of the CNN is a $d = |R| \cdot |L|$ -dimensional vector of “scores”. Let us denote this vector by $F(x, w) \in \mathbb{R}^d$, where w are the network weights. The labeling y can also be represented by the indicator vector $\phi(y)$ of the same dimension – each component of this vector corresponds to a pair $(i \in R, l \in L)$ and the values are 1 if $y_i = l$ and 0 otherwise. Hence, the energy part that corresponds to the unary potentials in (1) can be written as $\langle \phi(y), F(w, x) \rangle$, where $\langle \cdot \rangle$ denotes the inner product⁵. Given this, the derivative of the energy with respect to the network weights can be written according to the chain rule as

$$\frac{\partial E(x, y, w)}{\partial w} = \frac{\partial \langle \phi(y), F(w, x) \rangle}{\partial F(x, w)} \cdot \frac{\partial F(x, w)}{\partial w} = \phi(y) \cdot \frac{\partial F(x, w)}{\partial w}. \quad (8)$$

Consequently, the stochastic gradient (6) is

$$\nabla_w = \phi(\hat{y}) \cdot \frac{\partial F(x_t, w)}{\partial w} - \phi(y_t) \cdot \frac{\partial F(x_t, w)}{\partial w} = [\phi(\hat{y}) - \phi(y_t)] \cdot \frac{\partial F(x_t, w)}{\partial w}, \quad (9)$$

and hence can be computed via standard Error Back Propagation, where the difference $[\phi(\hat{y}) - \phi(y_t)]$ is the error that is propagated through the network.

The gradient (6) with respect to the pairwise potentials can be represented in a similar manner. Let us denote by $\phi'(y)$ the vector that consists of $C \cdot |L| \cdot |L|$ components (C is the number of edge classes), where each component corresponds to a particular pairwise potential, i.e. an edge class c (see (1)) and a label pair (l, l') . The value of each component of $\phi'(y)$ is the number of occurrences of the corresponding triple (c, l, l') in the labelling y . Denoting by ψ_c the vector of all unknown pairwise potentials, the second term in (1) is again the inner product $\langle \phi'(y), \psi_c \rangle$. Hence, the stochastic gradient (6) with respect to the unknown pairwise potentials ψ_c is simply $\nabla_{\psi_c} = \phi'(\hat{y}) - \phi'(y_t)$.

Algorithm 1 Computation of the stochastic gradient for one image

Input: Training example x_t , pre-computed training data statistics $\phi(y_t)$ and $\phi'(y_t)$,
current model parameters: pairwise potentials ψ_c for CRF and network weights
 w , the previously sampled labelling \hat{y} (random at the beginning)

Output: Gradients ∇_{ψ_c} and ∇_w of the conditional log-likelihood with respect to the
pairwise potentials and the network weights respectively

- 1) Perform a forward pass in the network to compute $F(x_t, w)$, i.e. the unary potentials $\psi_i(x_t, l)$
 - 2) Do one sampling iteration for each node (7) to obtain a new \hat{y} (starting from current labelling)
 - 3) Compute the model statistics $\phi(\hat{y})$ and $\phi'(\hat{y})$
 - 4) Compute the network error as $\phi(\hat{y}) - \phi(y_t)$
 - 5) Obtain the gradient ∇_w by propagating the error through the network (Error Back Propagation)
 - 6) Compute the gradient $\nabla_{\psi_c} = \phi'(\hat{y}) - \phi'(y_t)$
-

In summary, the computation of the stochastic gradient for one training sample (x_t, y_t) is presented in Alg. 1. Note that the training data statistics $\phi(y_t)$ and $\phi'(y_t)$ do not change during the learning procedure. Hence, they can be computed in advance. The gradient for all images (or for all images in a batch) is obtained by averaging.

Properties of the learning algorithm. We would like to discuss some relevant properties of the proposed learning algorithm. Firstly, doing a forward pass of the network and performing Error Back Propagation are standard operations. Hence, we are able to use sophisticated third-party methods for this. In particular, we use the Caffe-framework (Jia et al., 2014) in our experiments. Secondly,

⁵Such a representation is sometimes called “over-complete representation” and is often used e.g. in LP-relaxation-based inference methods, representing a distribution as a log-linear model, etc. In fact, the vectors $\phi(y)$ and $\phi'(y)$ (see later) are sufficient statistics, when representing the CRF as a member of the exponential family.

our sampling process is very fast in contrast to other inference techniques such as e.g. Belief Propagation. It is also easy to parallelize the sampling. Note that our learning scheme does not involve any inference. In some sense this is obvious since the aim is to learn the statistical model, i.e. to fit a probability distribution to the given training data, rather than to optimize for some results. Finally, although we use data from the previous iteration for each new sample, the amount of data to be stored is quite low – from iteration to iteration we only need to store a single labelling for each image. Both aspects, the ability to parallelize and the low memory footprint makes this procedure ideal for a GPU-based implementation, which we have not done yet.

Inference. There are different procedures for performing inference with our model. One option is the Maximum A-Posteriori decision (MAP) strategy. This leads to the problem of minimizing the energy (1) with respect to unknown labelling y . Another option is the Maximum-Marginal decision (MMARG). This leads to computing the node-wise posterior marginal probabilities, followed by choosing the labelling which has the best marginal in each node. In this work, we utilized the latter option, i.e. MMARG. The marginals are computed using Gibbs Sampling, as during learning.

We use the Maximum-Marginal decision strategy for various reasons. Firstly, since our potentials can take an arbitrary form it is often challenging to perform energy minimization. Graph-cut based solvers, such as α -expansion, are not applicable and one has to resort to general purpose solvers. For this, LP-relaxation-based solvers, such as TRW-S (Kolmogorov, 2006), are among the best and most popular solvers. We applied TRW-S to many instances of our problem, however, no reasonable results were achieved in reasonable time. The reason for this is that dual LP-based solvers typically have problems when the energy-values have “large numbers”, e.g. a pairwise term has a large weight that prohibits a certain configuration. In Sec. 4 we observe this behaviour with our model. Secondly, according to Bayesian Decision Theory, MAP follows the simple delta-loss, which does not reflect the similarity between labellings. On the other hand, very often the result of the inference is assessed by the pixel-accuracy, which corresponds to the Hamming loss. The latter leads to the Maximum-Marginal decision. Finally, it has been observed in previous works that it is an advantage when learning and inference follow the same computational scheme, to some extent. In our case the learning is based on marginals – i.e. the gradients and the network error (that should be propagated back through the network) are obtained from marginals. Hence, it seems to be reasonable to use a marginal-based decision strategy for inference as well.

4 EXPERIMENTS

In the experimental evaluation we compare our approach to alternative techniques and also variants of our approach. Additionally, we want to analyze the trained CRF model, in order to understand whether it can capture object shape and contextual information.

Dataset and evaluation. We applied our approach to the challenging task of predicting human body parts from depth images. To the best of our knowledge, there is no publically available dataset for this task which works with real depth images. For this reason (Denil et al., 2013) introduced a set of synthetically rendered depth images along with the corresponding ground truth labelling (see examples in Fig. 1 (left columns)). In total there are 19 different body part labels, and one additional label for the background. The dataset is split into 2.000 images for training and 500 images for testing. As quality measure they use the averaged per-pixel accuracy for body parts labelling, i.e. excluding the background. This makes sense since the background is easily identified from the depth map.

Baselines. We compare our approach to the method of (Denil et al., 2013), which introduced this dataset. Their approach is based on a random forest model. Unfortunately, we were not able to compare to the recent work (Ren et al., 2015), which extended (Denil et al., 2013), and is also based on random forests. The reason is that (Ren et al., 2015) use their own evaluation measure. This means that they only evaluate the accuracy of a very small subset of pixels. This subset is chosen in such a way that each of the 20 classes is represented by the same number of pixels. We do not think that this is a good measure since a very small subset is more likely to have a bias. Furthermore, we did not have this subset at our disposal. Since our main aim is to evaluate CNN-based CRF models, we compare to the approach (Zheng et al., 2015). As described above, they incorporate a densely connected Gaussian CRF model into the CNN as a Recurrent Neuronal Network of the

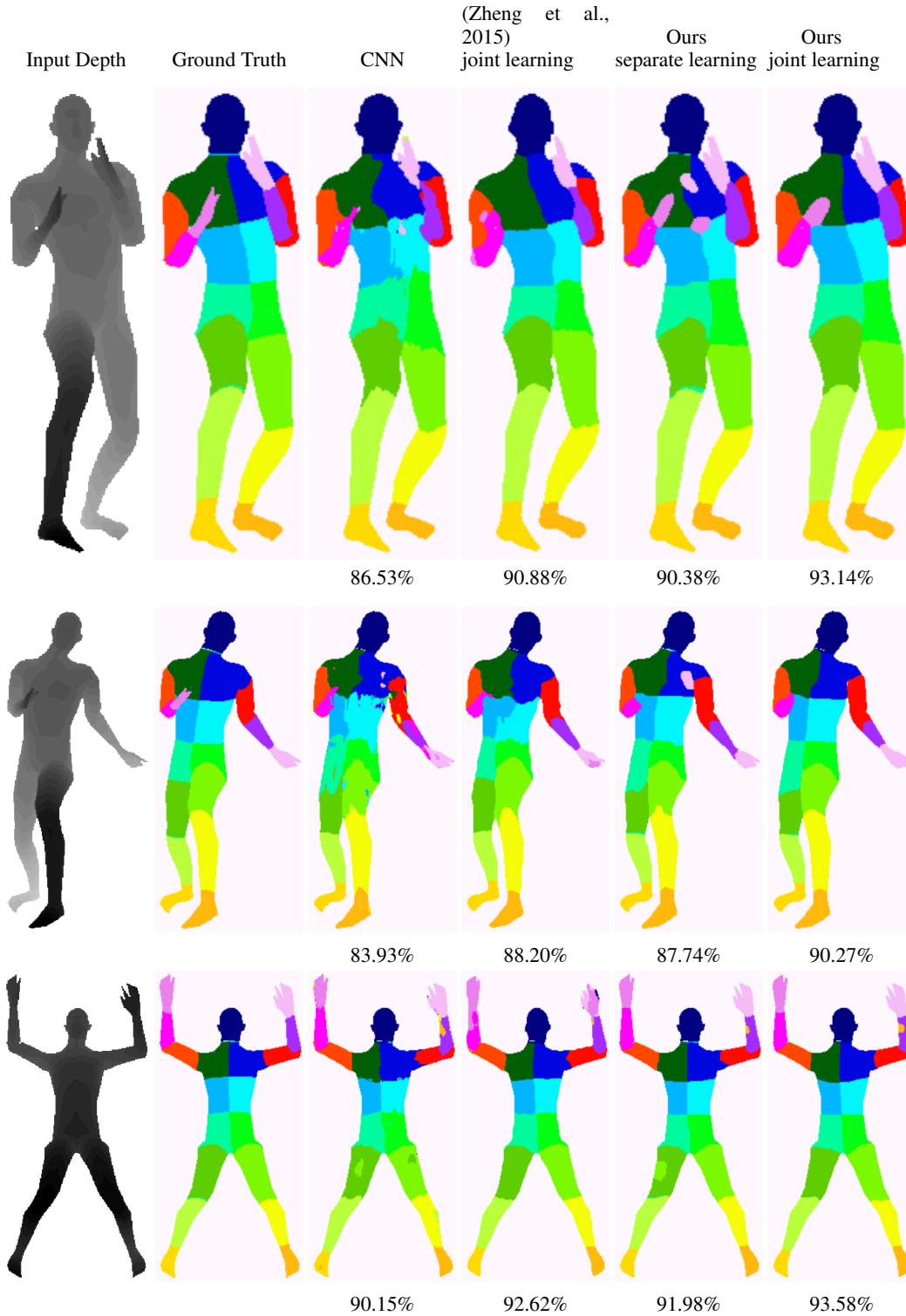


Figure 1: **Results.** (From left to right). The input depth image. The corresponding ground truth labeling for all body parts. The result of a trained CNN model. The result of (Zheng et al., 2015) using an end-to-end training procedure. Our results with separate learning and joint learning, respectively. Below each result we give the averaged pixel-wise accuracy for all body parts.

Method	Learning	Accuracy
Online Random Forest (Denil et al., 2013)	-	$\approx 79.0\%$
CNN	-	84.47
CNN + CRF (Zheng et al., 2015)	separate	86.55%
CNN + CRF (Zheng et al., 2015)	joint	88.17%
CNN + CRF (ours)	separate	87.62%
CNN + CRF (ours)	joint	89.01%

Table 1: Average per-pixel accuracy for all foreground parts. *Separate* learning means that weights of the respective CNN were trained prior to CRF parameters. In contrast, *joint* training means that all weights were learned jointly, starting with a pre-trained CNN. We observe that joint training is superior to separate training, and furthermore that the restricted model of (Zheng et al., 2015), which is based on a dense Gaussian CRF, is inferior to our generic CRF model.

corresponding Mean Field inference steps. This approach has recently been the state-of-the-art in the VOC2012 object segmentation challenge.

Implementation details of our model. As mentioned above, the unary terms of our CRF model depend on the image via a CNN. For the CNN we use our own, fully convolutional, architecture and train it from scratch, since most existing pre-trained CNNs use RGB images as input. Moreover, since some body parts, such as the hand, are relatively small and capture fine details we use a network architecture which does not reduce the image resolution in intermediate layers. All intermediate layers have 50 output channels and a stride of one. The final layer has 20 output channels that correspond to the output labels. In summary, our architecture looks like this: $\text{conv}(41 \times 41) \rightarrow \text{ReLU} \rightarrow \text{conv}(17 \times 17) \rightarrow \text{ReLU} \rightarrow \text{conv}(11 \times 11) \rightarrow \text{max-pool}(3 \times 3) \rightarrow \text{ReLU} \rightarrow \text{conv}(5 \times 5)$. We use softmax for calculating cross-entropy as a loss. The CNN was trained using stochastic gradient descent with momentum 0.99. We apply a weight-update after calculating the gradient for each sample.

For the CRF part (1) we use 32 classes of pairwise potentials. The neighborhood structure of our CRF is visualized in Fig. 2 (b). Each pixel is connected to 64 neighbours. As it was mentioned above, opposite edges are part of the same class. Note that our pairwise potentials are not image depended, which we leave as an extension for future work. As we see later, the potentials can nicely capture the shape of body parts and also their geometric relationship.

In our experiments, we consider two learning scenarios: *separate* learning and *joint* (end-to-end) learning. In both cases we start the learning procedure from the same pre-trained CNN, as described above. For separate learning only the CRF parameters are updated, whereas the CNN weights are kept fixed. In contrast, for joint (end-to-end) learning all parameters are updated. During test-time inference we empirically observed that starting Gibbs Sampling from a random labelling can lead to extremely long runtimes. To speed-up the burn-in-phase, we use the marginal distribution of the CNN without CRF. This means that the first sample is drawn from the marginal distribution of the pre-trained CNN.

Results. Qualitative results are shown in Fig. 1, and quantitative results are given in Table 1. We see that our model with joint learning is performing best. The CNN+CRF approach of (Zheng et al., 2015) is inferior. Note, that an accuracy difference of 1% can mean that e.g. a complete hand is incorrectly labeled. We attribute this to the fact that for this task the spatial layout of body parts is of particular importance. The underlying dense Gaussian CRF model of (Zheng et al., 2015) is rotational invariant and cannot capture contextual information such as “the head has to be above the torso”. Our approach is able to capture this, which we explain in detail in Fig. 2 and 3 in the appendix. We expect that even higher levels of accuracy can be achieved by exploring different network designs and learning strategies, which is part of future work.

5 DISCUSSION AND FUTURE WORK

We have presented a generic CRF model, where unary factors are modelled by a CNN. We have introduced an efficient and scalable maximum likelihood learning procedure to jointly train all model

parameters. By doing so, we were able to train and test on large-size factor graphs. We have demonstrated a performance gain over competing techniques for semantic labelling of body parts. We have observed that our generic CRF model is able to capture shape and context information of relating body parts.

There are many exciting avenues for future research. As mentioned above, we plan to move our training procedure to GPU in order to perform large-scale training. We also plan to apply our method to other application scenarios, such a semantic segmentation of RGB images. In this context it would be interesting to combine the dense CRF model of (Zheng et al., 2015) with our generic CRF model. Note that a dense CRF is implicitly modelling the property that objects have a compact color distribution, see (Cheng et al., 2015), which is a complementary modelling power to our generic CRF model. Finally, we may further improve our model by conditioning all factors on CNNs.

REFERENCES

- Adams, A., Baek, J., and Davis, M. A. Fast high-dimensional filtering using the permutohedral lattice. In *Computer Graphics Forum*, volume 29. Wiley Online Library, 2010.
- Barbu, A. Training an active random field for real-time image denoising. *Image Processing, IEEE Transactions on*, 18(11):2451–2462, 2009.
- Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *preprint arXiv:1412.7062*, 2014.
- Chen, L.-C., Schwing, A. G., Yuille, A. L., and Urtasun, R. Learning deep structured models. In *ICML*, pp. 1785–1794, 2015.
- Cheng, M.-M., Prisacariu, V. A., Zheng, S., Torr, P. H. S., and Rother, C. Densecut: Densely connected crfs for realtime grabcut. *Computer Graphics Forum*, 34(7), 2015.
- Denil, M., Matheson, D., and de Freitas, N. Consistency of online random forests. In *ICML*, 2013.
- Domke, J. Learning graphical model parameters with approximate marginal inference. *TPAMI*, 2013.
- Farabet, C., Couprie, C., Najman, L., and LeCun, Y. Learning hierarchical features for scene labeling. *TPAMI*, 35(8), 2013.
- Flach, B. and Schlesinger, D. Modelling composite shapes by gibbs random fields. In *CVPR*, 2011.
- Geman, S. and Geman, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *TPAMI*, 6(6), 1984.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. Caffe: Convolutional architecture for fast feature embedding. *preprint arXiv:1408.5093*, 2014.
- Kiefel, M. and Gehler, P. V. Human pose estimation with fields of parts. In *ECCV*, 2014.
- Kolmogorov, V. Convergent tree-reweighted message passing for energy minimization. *TPAMI*, 28(10), 2006.
- Krähenbühl, P. and Koltun, V. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.
- Lafferty, J., McCallum, A., and Pereira, F. CN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.
- Lin, G., Shen, C., Reid, I. D., and van den Hengel, A. Efficient piecewise training of deep structured models for semantic segmentation. *preprint arXiv:1504.01013*, 2015.
- Liu, Z., Li, X., Luo, P., Loy, C. C., and Tang, X. Semantic image segmentation via deep parsing network. In *ICCV*, 2015.
- Long, Jonathan, Shelhamer, Evan, and Darrell, Trevor. Fully convolutional networks for semantic segmentation. *preprint arXiv:1411.4038*, 2014.

- Nowozin, S., Rother, C., Bagon, S., Sharp, T., Yao, B., and Kohli, P. Decision tree fields. In *ICCV*, 2011.
- Ren, S., Cao, X., Wei, Y., and Sun, J. Global refinement of random forest. In *CVPR*, 2015.
- Richmond, D. L., Kainmueller, D., Yang, M. Y., Myers, E. W., and Rother, C. Relating cascaded random forests to deep convolutional neural networks for semantic segmentation. *preprint arXiv:1507.07583*, 2015.
- Ross, S., Munoz, D., Hebert, M., and Bagnell, J. A. Learning message-passing inference machines for structured prediction. In *CVPR*, 2011.
- Schwing, A. G. and Urtasun, R. Fully connected deep structured networks. *preprint arXiv:1503.02351*, 2015.
- Sethi, I. K. Entropy nets: from decision trees to neural networks. *Proceedings of the IEEE*, 78(10): 1605–1613, 1990.
- Stoyanov, V., Ropson, A., and Eisner, J. Empirical risk minimization of graphical model parameters given approximate inference, decoding, and model structure. In *AISTATS*, 2011.
- Tieleman, T. Training restricted boltzmann machines using approximations to the likelihood gradient. In *ICML*. ACM, 2008.
- Tompson, J. J., J., A., LeCun, Y., and Bregler, C. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014.
- Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., and Torr, P. H. S. Conditional random fields as recurrent neural networks. In *ICCV*, 2015.

6 APPENDIX

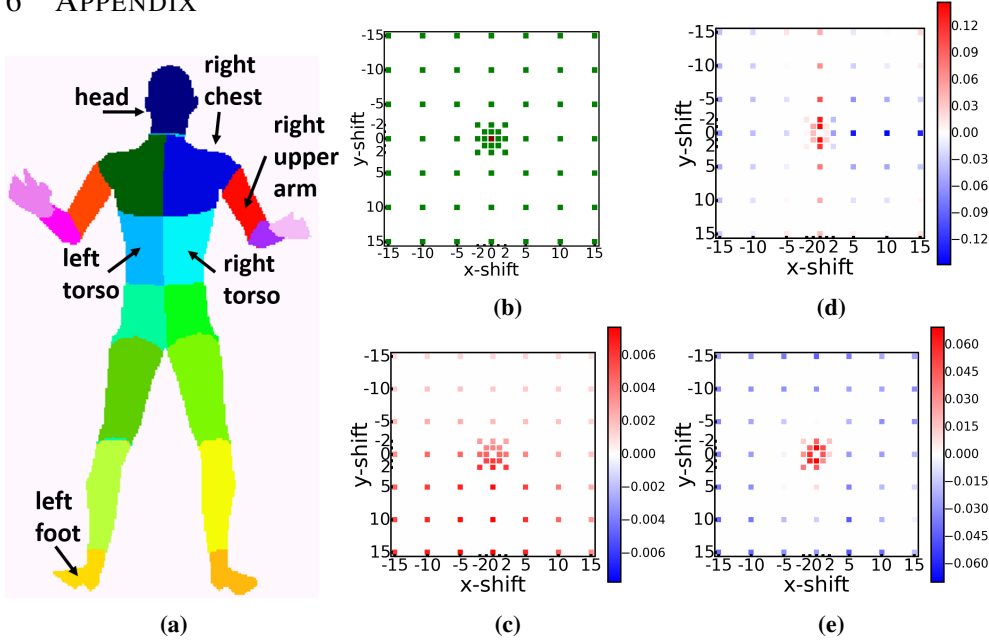


Figure 2: **Model Insights.** (a) Illustrating the 19 body parts of a human. (b) Neighbourhood structure for pairwise factors. The center pixel (red) is connected via pairwise factors to all green pixels. Note that “opposite” edges share same weights, e.g. the edge with x, y -shift (5, 10) has the same weights as the edge with x, y -shift (−5, −10). (c) Weights for pairwise potentials that connect the label “head” with the label “foot”. Red means a high energy value, i.e. a discouraged configuration, while blue means the opposite. Since there is no sample in the training dataset where a foot is close to a head, all edges are negative or close to 0. Note that white pixels (zero weights) can occur even for very unlikely configurations. The reason is that during CRF training these unlikely configurations did not occur. (d) Weights for pairwise potentials that connect the label “left torso” with the label “right torso”. The potentials enforce a straight, vertical border between the two labels, i.e. there is a large penalty for “left torso” on top (or below) of “right torso” (x -shift 0, y -shift arbitrary). Also, it is encouraged that “right torso” is to the right of the “left torso” (Positive x -shift and y -shift 0). (e) Weights for pairwise potentials that connect the label “right chest” with the label “right upper arm”. It is discouraged that the “right upper arm” appears close to “right chest”, but this configuration can occur at a certain distance. Since the training images have no preferred arm-chest configurations, all directions have similar weights.

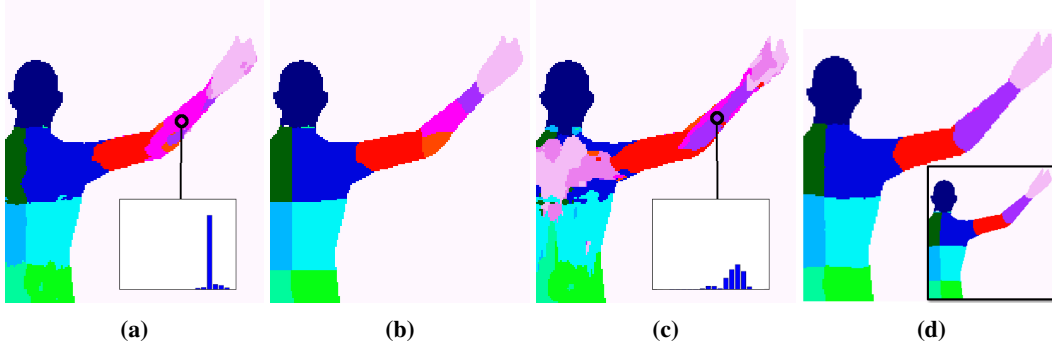


Figure 3: **Model Insights.** (a) Most likely labeling of a separately trained CNN. For circled pixel the local marginal distribution is shown. (b) Max marginal labeling of a separately trained CRF, which uses the CNN unaries from (a), i.e. our approach with separate learning procedure. We observe that unaries are spatially smoothed-out. (c) Most likely labeling of a CNN that was jointly trained with the CRF. The labeling looks less good than (a). However, the main observation is that pixel-wise marginal distributions are more ambiguous than in (a), see circled pixel. (d) The final, max-marginal labeling of the jointly trained CRF model, which is considerably better than the result in (b). The reason is that due to the ambiguity in the local unary marginals, the CRF is given more power to find the correct body part configuration. The inset shows the ground truth labelling.