

Convolutional Simplex Projection Network for Weakly Supervised Semantic Segmentation

Rania Briq¹
briq@iai.uni-bonn.de

Michael Moeller²
Michael.Moeller@uni-siegen.de

Juergen Gall¹
gall@iai.uni-bonn.de

¹ University of Bonn

² University of Siegen

Abstract

Weakly supervised semantic segmentation has been a subject of increased interest due to the scarcity of fully annotated images. We introduce a new approach for solving weakly supervised semantic segmentation with deep Convolutional Neural Networks (CNNs). The method introduces a novel layer which applies simplex projection on the output of a neural network using area constraints of class objects. The proposed method is general and can be seamlessly integrated into any CNN architecture. Moreover, the projection layer allows strongly supervised models to be adapted to weakly supervised models effortlessly by substituting ground truth labels. Our experiments have shown that applying such an operation on the output of a CNN improves the accuracy of semantic segmentation in a weakly supervised setting with image-level labels.

1 Introduction

The task of semantic image segmentation, which requires solving the problem of assigning a semantic class label to each pixel in a given image, is considered crucial in solving more complex tasks such as scene understanding. The most successful approaches for semantic image segmentation rely on convolutional neural networks (CNN). CNNs, however, require large training sets where each image is pixel-wise annotated. Since obtaining dense pixel annotations for training data is an expensive and tedious task, there has been a growing effort to learn CNNs using less supervision. In particular, a weakly supervised setting in which only image-level labels are available has been addressed in several works [1, 2, 3, 4, 5].

In order to guide the training process, a prior on the size of objects in an image is often used. For instance, Chen et al. [6] employ an expectation-maximization approach, which alternates between predicting the output of the network and estimating its parameters, which is refined by size priors. Instead of using a prior, Pathak et al. [7] introduce constraints on the size of the objects. While constraints have the advantage of reducing the set of feasible solutions and are a more principled way to integrate prior knowledge about object sizes during training of a CNN, current state-of-the-art methods for weakly supervised semantic image

segmentation do not rely on constraints. This is due to the way the approach in [10] enforces constraints on the output of a neural network. It proposes predicting a latent distribution for which violations of predefined linear constraints are penalized, and then optimizes the parameters of the CNN to follow the latent distribution as closely as possible by minimizing the Kullback-Leibler divergence. This means that an additional optimization step needs to be performed in each training iteration.

In this work, we provide a more practical approach that integrates constraints directly as a network layer and can therefore be added to any convolutional neural network. We term the layer *simplex projection layer* since it projects the output of the previous layer onto the simplex and enforces the network to satisfy the given constraints. Such an operation can be efficiently performed in linear time [9].

To demonstrate the benefit of the simplex projection layer, we integrate the layer into a state-of-the-art approach for weakly supervised semantic image segmentation [7]. In our experiments on the PASCAL VOC 2012 segmentation benchmark [4], we show that the layer substantially improves the segmentation accuracy of [7]. The source code is available at <https://github.com/briq/CSPN>.

2 Related Work

Over the past few years, CNNs have made significant advances in improving the results in semantic segmentation, particularly in a strongly supervised setting [2]. In such a setting, however, obtaining fully annotated data requires much effort and time and therefore limits their availability. To circumvent this limitation, several approaches that rely on weaker forms of supervision have been proposed. For instance, [3] propose using bounding boxes in their training data in a semi-supervised setting. Lin et al. [5] proposed using scribbles which mark the center of an object in a given image. Chaudhry et al. [1] combine saliency and attention maps that form pseudo ground truth during training.

So far, the least expensive methods rely on image-level labels only. Pinheiro and Collobert [11], for example, reformulate the problem as multiple instance learning (MIL), such that an image is considered a positive bag if it contains at least one pixel of a certain class and negative otherwise. Chen et al. [4] employ an Expectation Maximization (EM) algorithm which alternates between predicting the pixel labels and optimizing the network parameters. Shimoda and Yanai [16] employ a method based on back-propagation in which they compute the CNN derivatives with respect to the intermediate layers rather than the input images and then use subtraction to create class-specific saliency maps. Pathak et al. [13] propose a method that formulates MIL in a fully convolutional network, in which the loss is computed coarsely at maximum predictions in the heat map. Saleh et al. [15] create a foreground/background mask by exploiting the unit activations of some hidden layers in a fully convolutional CNN that had been pre-trained for object recognition. Qi et al. [14] combine semantic segmentation and object localization modules in the CNN that provide augmented feedback to each other for error correction. Kim et al. [8] train two Fully Convolutional Networks (FCN) for image-level classification in two phases, where in the second phase the second FCN is trained with the highlighted regions in the heat maps from the first FCN being suppressed. Other methods rely on using external modules to generate cues such as localization and size information [12, 6]. Such a module may either have been learned in a weakly supervised setting or trained on a different dataset in a more strongly supervised manner.

3 Convolutional Simplex Projection Network (CSPN)

The task of semantic image segmentation solves the problem of assigning semantic class labels to each pixel in an image, *i.e.*, for a given image X of an arbitrary size $m \times n$ and a set of possible semantic labels denoted by L , the goal is to determine a label $k \in L$ for each pixel $x_{ij} \in X$, $i \in [m]$, $j \in [n]$. Such a task can be solved by a CNN by estimating $Y = (y_{ij})$ with $y_{ij} = \operatorname{argmax}_{k \in L} (Q_k(X; \Theta))_{ij}$, which takes the image X as input and predicts for each pixel (i, j) the class label y_{ij} . $Q(X; \Theta)$ denotes the network output function and the variable Θ represents the network parameters being optimized during training.

In the context of weakly supervised learning, pixel-wise labels are not provided during training. Instead, the network parameters Θ are learned based on a set of images, in which each training sample is a pair (X, S) , where $S \subseteq L \setminus \{k_0\}$ denotes the semantic classes present in the image X and k_0 denotes the background label. In order to guide the training process, approaches for weakly supervised learning like [10] use size priors for semantic classes as additional weak supervision. Size priors were rephrased as constraints in the work of [11]. Their method uses a penalty function for violated linear constraints in order to estimate a latent distribution and optimizes the parameters of the CNN to follow the latent distribution as closely as possible. The approach alternates between updating the latent distribution using a dual formulation and updating the network parameters. In particular, the constraints are enforced during training only and cannot be guaranteed during inference. Furthermore, it is not straight forward to combine the approach with more advanced approaches for weakly supervised image segmentation, which outperform [11] by a large margin. In this work, we provide a more elegant approach that integrates constraints directly as a network layer and can therefore be added to any convolutional neural network.

We first introduce the simplex projection layer in Section 3.1. We then describe in Section 3.2 how such a layer can be applied for weakly supervised semantic segmentation using any CNN for image segmentation and finally describe an implementation, which adds the simplex projection layer to SEC [10].

3.1 Simplex Projection Layer

In general, projection onto the simplex is a minimization problem defined as:

$$\min_w \frac{1}{2} \|w - v\|_2^2 \quad \text{s.t.} \quad \sum_{i=1}^n w_i = v, \quad w_i \geq 0, \quad (1)$$

i.e. given a vector $v \in \mathbb{R}^n$ that does not satisfy a given equality constraint, it finds a new vector $w \in \mathbb{R}^n$ that satisfies the constraint and that is as close to v as possible. In the context of our problem, the goal is to train a network $Q(X; \Theta)$ that receives an image X as input, and outputs a pixel-dense semantic segmentation Y with $y_{ij} = \operatorname{argmax}_{k \in L} (Q_k(X; \Theta))_{ij}$. For integrating our approach into such a network, let us assume that the precise information about the size of the object $k \in L$ present in the image is v_k pixels. In Section 3.2, we describe how v can be obtained but the approach is agnostic to how such an estimate may be obtained.

The constraint that object k should have size v_k can be formalized by requiring the output $Q(X; \Theta)$ of the network for an object k to be in the constraint set C_Y defined as

$$C_Y = \{Y \mid \sum_{i,j} y_{k_{i,j}} = v_k\}, \quad (2)$$

i.e. only solutions where the number of pixels that are classified as class k is equal to v_k are valid. Since argmax is non-differentiable and the loss of the network is computed on $Q(X; \Theta)$, we instead approximate the constraint set by

$$C_Q = \{Q \mid \sum_{i,j} q_{k_{i,j}} = v_k\}, \quad (3)$$

i.e. we require that the sum of the elements of the heat map of class k be equal to the size v_k . In order to enforce such constraints, we introduce the projection onto the set C_Q ,

$$\pi_C(Q_k) = \operatorname{argmin}_{V_k \in C_Q} \|Q_k - V_k\|_F^2, \quad (4)$$

as a layer into the network, where Q_k is the output of the previous layer for class $k \in L$ and $\|Q_k - V_k\|_F^2 = \sum_{i,j} (q_{k_{i,j}} - v_{k_{i,j}})^2$ denotes the squared Frobenius norm.

The projection onto the simplex $\pi_C(Q)$ can be efficiently solved by Algorithm 1 [3]. In our context, the simplex projection layer takes as input a confidence map for class k and a constraint value v_k and outputs a refined map, which satisfies the constraint. In essence, the algorithm finds the elements of the heat map whose sum is used to calculate the difference from the true size. Once these elements are found, the difference is distributed evenly across the heat map. This solution is dictated by solving the Lagrangian formulation of the problem in (1). The operations of the simplex projection layer are very inexpensive and the expected time complexity is linear, as such, integrating such a projection layer in a CNN does not increase its time complexity. Furthermore, the projection layer is not limited to a certain network architecture and can in principle be applied in any weakly supervised setting. The algorithm, however, requires prior knowledge of the size of an object in a given image, and given a weakly supervised setting such a number can only be estimated as we will discuss in the following section.

3.2 CSPN for Weakly Supervised Semantic Segmentation

Figure 1 gives an overview of how the simplex projection layer can be applied in a weakly supervised setting, in which only image-level labels are available. In order to enforce some given constraints at the last layer, we introduce a softmax layer after the last fully convolutional layer in the network, which performs:

$$\sigma(Q_k(x_{ij})) = \frac{e^{Q_k(x_{ij})}}{\sum_{l=1}^L e^{Q_l(x_{ij})}} \quad (5)$$

for every class $k \in L$ and pixel in the output image. The softmax layer is followed by the simplex projection layer which takes as input the probability values from the softmax layer, the size of each object present in the image, and outputs the projection onto the constraint set C_Q . Figure 2 visualises the effect of the projection layer.

An argmax operation is applied to the output of the simplex projection layer to create the target label for the loss layer

$$\hat{Y} = \operatorname{argmax}_{k \in L} \pi_C(Q_k). \quad (6)$$

As a loss function, we use the standard cross-entropy loss given by:

$$L_{project} = - \sum_{ij} \sum_{k=1}^L \hat{y}_{k_{ij}} \log \left((Q_k(X; \Theta))_{ij} \right) \quad (7)$$

Input: A matrix $M \in \mathbb{R}^{m \times n}$ and a scalar $v > 0$
Initialize $U = [m] \times [n]$ $s = 0$ $\rho = 0$

- 1 While $U \neq \emptyset$
- 2 Pick $(g, h) \in U$ at random
- 3 Partition U :
- 4 $G = \{(i, j) \in U \mid M_{ij} \geq M_{g,h}\}$
- 5 $L = \{(i, j) \in U \mid M_{ij} < M_{g,h}\}$;
- 6 Calculate $\Delta\rho = |G|$; $\Delta s = \sum_{(i,j) \in G} M_{i,j}$
- 7 IF $(s + \Delta s) - (\rho + \Delta\rho)M_{g,h} < v$
- 8 $s = s + \Delta s$; $\rho = \rho + \Delta\rho$; $U \leftarrow L$
- 9 ELSE
- 10 $U \leftarrow G \setminus \{(g, h)\}$
- 11 ENDIF
- 12 SET $\theta = \frac{s-v}{\rho}$
- 13 OUTPUT $M_{ij} = \max\{M_{i,j} - \theta, 0\}$

Algorithm 1: Projection onto the simplex. The matrix M represents a heat map containing the softmax probabilities for a class k and v is a scalar denoting the object size of class k . The algorithm has resemblance to the median-finding algorithm in which the idea is to reduce the time complexity by avoiding sorting the elements.

where Θ is the set of the CNN parameters, $\hat{y}_{k_{ij}}$ is one if the pixel x_{ij} is assigned to class k after the projection (6) and zero otherwise, and $(Q_k(X; \Theta))_{ij}$ denotes the predicted class probability at pixel x_{ij} before the simplex projection layer.

Since we are tackling semantic segmentation in a weakly supervised setting, the object size v_k is not available at hand except if the class k is not present in the image in which case $v_k = 0$. To obtain an estimate of v_k , we use the approach in [16] that estimates class-specific saliency maps.

While negative values in the maps mark non-salient regions, positive values indicate potential salient regions. To obtain v_k we therefore count the number of pixels that have a score larger or equal to $\tau = \frac{1}{8}$. To avoid counting a pixel twice for different classes, we only count the pixels for class k with the highest saliency score. In our experiments, we show that the saliency maps estimated by [16] are not very accurate, but they provide a reasonable estimate for v_k . It is noteworthy that the approach [16] is trained in the same weakly supervised setting and does not require any additional supervision or additional training data. Even though we have used saliency maps to obtain these estimates, the approach is independent of what estimator may be used.

3.3 Simplex Projection Layer in SEC

While the previous section described how the simplex projection layer can be added to any CNN for weakly supervised image segmentation, we now describe its integration into the state-of-the-art SEC [7] approach. The loss of SEC consists of three terms, namely seed, expand and constrain (SEC), which encourage the network to meet localization cues for an object, predict its right extent and to meet its precise boundaries respectively. The overall

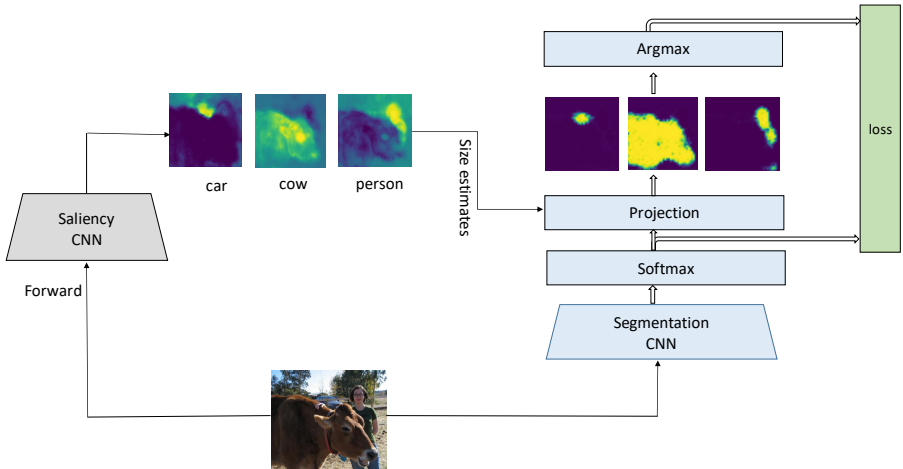


Figure 1: Illustration of the general approach with a projection layer in a weakly supervised setting. In our implementation, a saliency CNN provides saliency maps for each class in order to supply size estimates to the projection layer. These maps can be computed offline in advance. The projection layer outputs refined maps based on these estimates. The argmax operation is then applied to these heat maps thereby generating a mask that serves as ground truth according to which the loss function is computed. Note that for illustration purposes, the heat maps plotted after the projection layer have undergone a softmax operation, but this is not necessary in the network due to the argmax layer.

loss in SEC is given by:

$$L = \sum_{(X,S)} (L_{seed}(Q(X; \Theta), S) + L_{expand}(Q(X; \Theta), S) + L_{constrain}(Q(X; \Theta), S)). \quad (8)$$

For a detailed description of the tree terms, we refer to [1].

In our work, we replace the expand loss with our loss based on the simplex projection (7):

$$L = \sum_{(X,S)} (L_{seed}(Q(X; \Theta), S) + L_{project}(Q(X; \Theta), S) + L_{constrain}(Q(X; \Theta), S)). \quad (9)$$

To compute $L_{project}$, we append the simplex projection layer to the last fully convolutional layer fc8 in the CNN.

In principle, L_{expand} and $L_{project}$ can be combined but the terms are redundant and there is no benefit of adding both terms. While [1] uses the expand loss in order to expand a detected object to a reasonable size, the projection loss projects the solution onto the constraint set which ensures that each object has a reasonable size. $L_{project}$ can therefore be considered as a more strict loss function than L_{expand} . The difference between the two loss functions (8) and (9) is illustrated in Figure 3. The optimal performance was achieved after about 4 – 5 epochs. This fast convergence can be attributed to the fact that the space of feasible solutions is reduced to only those that are within the constraint set.

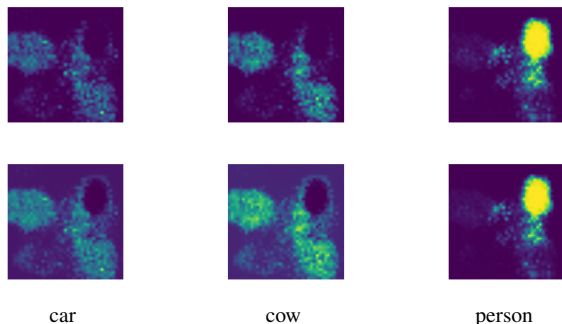


Figure 2: Visualisation of the effect of the projection layer at an early stage of training. The simplex projection is applied to the heat maps of the last convolution layer in the CNN, namely fc8. The first row is the heat map of fc8, and the second row is the projected heat map. As can be seen, the projection layer enhances the heat maps of the CNN such that the detected objects become more prominent in the projected heat map.



Figure 3: This figure shows how the loss function has been adapted from SEC to our proposed approach. The loss of SEC consists of three loss functions (left). In our approach (right), we omit the expansion loss and add the projection loss.

4 Evaluation

The proposed method has been trained and evaluated on the PASCAL VOC 2012 segmentation benchmark [1], which contains 10,582 training, 1449 validation and 1456 test images. The implementation is based on the Deeplab architecture [2], refined by Kolesnikov and Lampert [3] to integrate the SEC loss, and the Caffe deep learning framework [4]. Conforming with common practice, the results are refined by a CRF in a post-processing step.

4.1 Comparison with State-of-the-Art

Tables 1 and 2 compare the accuracy of our approach (CSPN) with the state-of-the-art for weakly supervised image segmentation for the validation and test sets respectively. The numbers of the other approaches are taken from [18]. On the validation set, CSPN outperforms DCSM [16] and SEC [7] by +10.4% and +3.8%, respectively, and on the test set by +10.4% and +3.8%. This shows that the simplex projection layer substantially improves the accuracy of SEC.

Our approach also outperforms by a large margin the expectation-maximization approach proposed by Papandreou et al. [9] and the constrained convolutional neural network (CCNN) [10]. Although Pathak et al. [10] showed in their work that CCNN improves [9] slightly by +1.5% on the validation set, the accuracy of CCNN is much lower compared to [9] in Table 1. This is due to the fact that CCNN builds on a preliminary version of [9]. In order

Method	mIoU	Additional supervision
DCSP-VGG16 [10]	58.6	supervised saliency
DCSP-ResNet-101 [10]	60.8	supervised saliency
AE w/o PSL [10]	50.9	supervised saliency
AE-PSL [10]	55.0	supervised saliency
MIL-FCN [10]	25.7	
EM-Adapt [9]	38.2	
CCNN [10]	35.3	
BFBP [15]	46.6	
DCSM [16]	44.1	
SEC [7]	50.7	
AF-SS [13]	52.6	
Two-phase [6]	53.1	
Roy and Todorovic [14]	52.8	
DCNA-VGG16 Zhang et al. [18]	55.4	
DCNA-ResNet-101 Zhang et al. [18]	58.2	
CSPN(ours)	54.5	

Table 1: Results on the validation set.

Method	mIoU
MIL-FCN [10]	24.9
EM-Adapt [9]	39.6
CCNN [10]	35.3
BFBP [15]	48.0
DCSM [16]	45.1
SEC [7]	51.7
AF-SS [13]	52.7
Two-phase [6]	53.8
Roy and Todorovic [14]	55.5
DCNA-VGG16 Zhang et al. [18]	56.4
DCNA-ResNet-101 Zhang et al. [18]	60.1
CSPN(ours)	55.5

Table 2: Results on the test set.

to evaluate how much improvement our approach would attain when integrated in [9], we added our simplex projection layer to the public available implementation of [9]. In this case, the accuracy is improved by +6%, which is clearly higher than the improvement which was reported in [10]. Since, however, both [9] and [10] are inferior to the state-of-the-art in weakly supervised image segmentation, the reported numbers of CSPN build on the integration of the simplex projection layer in SEC as discussed in Section 3.3.

When we compare our approach with the state-of-the-art approaches [6, 14] that use the same amount of supervision, we see that our approach is slightly better or on par with them. Only the very recent work of DCNA [18] outperforms our approach. Note that our model builds on SEC, which uses VGG16. We expect further improvements if ResNet-101 is used instead of VGG16. There are also approaches like DCSP [10] which achieve a higher accuracy, but these approaches use additional supervision for training a saliency model.

Loss	SEC+P	SC+P (CSPN)	S+P	SEC
mIoU	51.0	54.5	44.6	50.7

Table 3: Combining the projection loss (P) with additional loss functions from SEC.

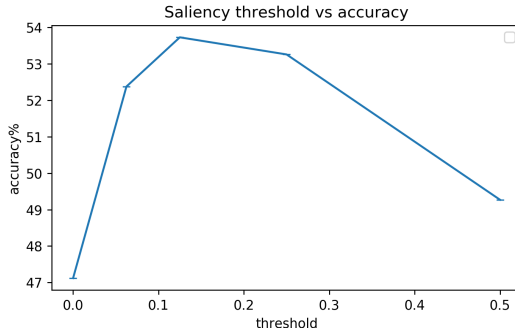


Figure 4: This plot examines the effect of the threshold value on the class-specific saliency. Values between $1/16$ and $1/4$ yield the most reliable values v_k to be used in the projection.

4.2 Ablation Studies

In Section 3.3, we replace the expansion loss of SEC by the projection loss. In Table 3, we report the accuracy if the expansion loss is added in addition or the constrain loss is omitted. Note that the seed loss is crucial for SEC and cannot be omitted. Adding the expansion loss to (9) decreases the accuracy. Since both the expansion loss and projection loss ensure that the segmented objects have a reasonable size, adding both of them has the undesired effect of increasing the impact of the size compared to the seed and constrain loss. If the constrain loss is omitted, the accuracy also drops. This is expected since the simplex projection does not take object boundaries into account, which the constrain loss does using a CRF model based on the CNN output and image pixels. As such, the projection loss cannot compensate for the constrain loss.

In order to show that the merit of the method is due to the projection layer rather than the mere fusion of [16] and [14], we used the saliency of [14] as the plain ground truth and obtained 51.7, which only slightly improves over the 50.7 of SEC, compared to 54.5 when applying the projection.

As described in Section 3.2, we use the threshold $\tau = 1/8$ to obtain v_k from the class-specific saliency maps [16]. The impact of the threshold is plotted in Figure 4. Note that the threshold has to be larger than zero since negative values or values around zero mark non-salient regions [16]. The plot shows that a threshold between $1/16$ and $1/4$ works well. If the threshold is too high, *i.e.* $1/2$ or larger, the size of the objects is underestimated and the accuracy decreases.

As illustrated in Figure 1, we perform an argmax operation after the simplex projection layer. As an alternative, one could also add a softmax layer after the projection layer to obtain target class probabilities and use them for the loss function instead of $\hat{y}_{k;j}$ (7). If we use softmax instead of argmax, the accuracy falls from 54.5% to 50.2%.

A few qualitative results including failures cases are shown in Figure 5.

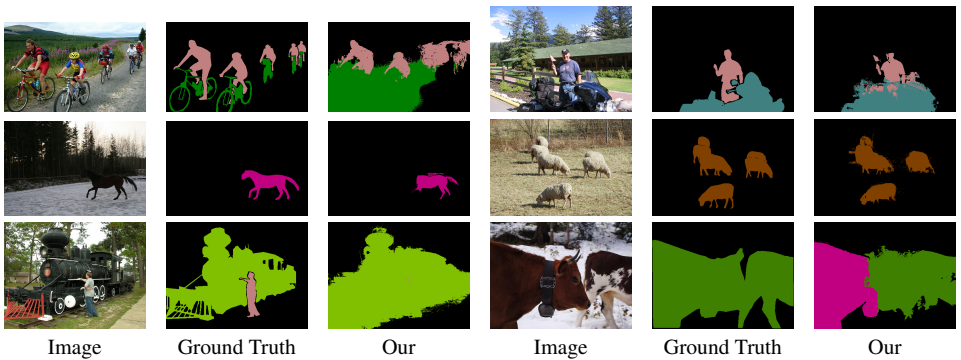


Figure 5: Above are some results obtained from evaluating our approach on the PASCAL VOC 2012 validation set, including two failure cases in the last row.

5 Conclusions

In this work, we have proposed the simplex projection layer which projects the output of a previous layer to the simplex. We have demonstrated the advantage of such layer for weakly supervised semantic segmentation where the projection layer is used to enforce constraints on the size of the objects. We integrated the layer in a state-of-the-art approach for weakly supervised segmentation and improved the segmentation accuracy substantially. As part of future work, we will investigate how the approach could be used for other tasks where the set of feasible solutions of a network can be reduced by constraints.

Acknowledgements

The work has been financially supported by the ERC Starting Grant ARCA (677650).

References

- [1] Arslan Chaudhry, Puneet K. Dokania, and Philip H. S. Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. *British Machine Vision Conference*, 2017.
- [2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.
- [3] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l_1 -ball for learning in high dimensions. In *International Conference on Machine Learning*, pages 272–279. ACM, 2008.
- [4] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

-
- [5] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.
- [6] Dahun Kim, Donghyeon Cho, and Donggeun Yoo. Two-phase learning for weakly supervised object localization. In *IEEE International Conference on Computer Vision*.
- [7] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision*, pages 695–711. Springer, 2016.
- [8] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3159–3167, 2016.
- [9] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *IEEE International Conference on Computer Vision*, pages 1742–1750, 2015.
- [10] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *IEEE International Conference on Computer Vision*, pages 1796–1804, 2015.
- [11] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. In *ICLR Workshop*, 2015.
- [12] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1713–1721, 2015.
- [13] Xiaojuan Qi, Zhengzhe Liu, Jianping Shi, Hengshuang Zhao, and Jiaya Jia. Augmented feedback in semantic segmentation under image level supervision. In *European Conference on Computer Vision*, pages 90–105. Springer, 2016.
- [14] Anirban Roy and Sinisa Todorovic. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3529–3538, 2017.
- [15] Fatemehsadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, Stephen Gould, and Jose M Alvarez. Built-in foreground/background prior for weakly-supervised semantic segmentation. In *European Conference on Computer Vision*, pages 413–432. Springer, 2016.
- [16] Wataru Shimoda and Keiji Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *European Conference on Computer Vision*, 2016.
- [17] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

- [18] Tianyi Zhang, Guosheng Lin, Jianfei Cai, Tong Shen, Chunhua Shen, and Alex C Kot. Decoupled spatial neural attention for weakly supervised semantic segmentation. *arXiv preprint arXiv:1803.02563*, 2018.