Iterative Instance Segmentation

Ke Li
UC Berkeley
ke.li@eecs.berkeley.edu

Bharath Hariharan
Facebook AI Research
bharathh@fb.com

Jitendra Malik
UC Berkeley
malik@eecs.berkeley.edu

Abstract

Existing methods for pixel-wise labelling tasks generally disregard the underlying structure of labellings, often leading to predictions that are visually implausible. While incorporating structure into the model should improve prediction quality, doing so is challenging – manually specifying the form of structural constraints may be impractical and inference often becomes intractable even if structural constraints are given. We sidestep this problem by reducing structured prediction to a sequence of unconstrained prediction problems and demonstrate that this approach is capable of automatically discovering priors on shape, contiguity of region predictions and smoothness of region contours from data without any a priori specification. On the instance segmentation task, this method outperforms the state-of-the-art, achieving a mean AP^r *of* 63.6% *at* 50% *overlap and* 43.3% *at* 70% *overlap.*

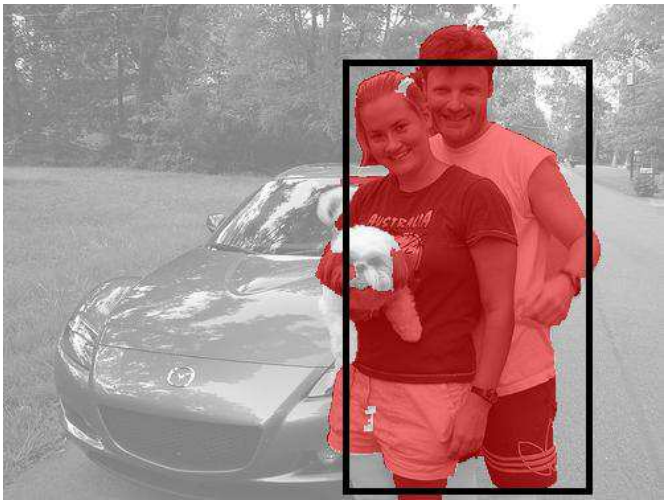

Figure 1: A challenging image in which object instances are segmented incorrectly. While pixels belonging to the category are identified correctly, they are not correctly separated into instances.

1. Introduction

In computer vision, the objective of many tasks is to predict a pixel-wise labelling of the input image. While the intrinsic structure of images constrains the space of sensible labellings, existing approaches typically eschew leveraging such cues and instead predict the label for each pixel independently. Consequently, the resulting predictions may not be visually plausible. To mitigate this, a common strategy is to perform post-processing on the predictions using superpixel projections [16] or conditional random fields (CRFs) [19], which ensures the final predictions are consistent with local appearance cues like colour and texture but fails to account for global object-level cues like shape.

Despite its obvious shortcomings, this strategy enjoys popularity, partly because incorporating global cues requires introducing higher-order potentials in the graphical model and often makes inference intractable. Because inference in general graphical models is NP-hard, extensive work on structured prediction has focused on devising efficient inference algorithms in special cases where the higher-order potentials take on a particular form. Unfortunately, this restricts the expressive power of the model. As a result, care must be taken to formulate the cues of interest as higher-order potentials of the desired form, which may not be possible. Moreover, low-energy configurations of the potentials often need to be specified manually a priori, which may not be practical when the cues of interest are complex and abstract concepts like shape.

In this paper, we devise a method that learns implicit shape priors and use them to improve the quality of the predicted pixel-wise labelling. Instead of attempting to capture shape using explicit constraints, we would like to model shape implicitly and allow the concept of shape to emerge from data automatically. To this end, we draw inspiration from iterative approaches like auto-context [32], inference machines [26] and iterative error feedback (IEF) [6]. Rather than learning a model to predict the target in one step, we decompose the prediction process into multiple steps and allow the model to make mistakes in intermediate steps as long as it is able to correct them in subsequent steps. By

learning to correct previous mistakes, the model must learn the underlying structure in the output implicitly in order to use it to make corrections.

To evaluate if the method is successful in learning shape constraints, a perfect testbed is the task of instance segmentation, the goal of which is to identify the pixels that belong to each individual object instance in an image. Because the unit of interest is an object instance rather than an entire object category, methods that leverage only local cues have difficulty in identifying the instance a pixel belongs to in scenes with multiple object instances of the same category that are adjacent to one another, as illustrated in Figure 1. We demonstrate that the proposed method is able to successfully learn a category-specific shape prior and correctly suppresses pixels belonging to other instances. It is also able to automatically discover a prior favouring contiguity of region predictions and smoothness of region contours despite these being not explicitly specified in the model. Quantitatively, it outperforms the state-of-the-art and achieves a mean AP^r of 63.6% at 50% overlap and 43.3% at 70% overlap.

2. Related Work

Yang et al. [33] first described the task of segmenting out individual instances of a category. The metrics we use in this paper were detailed by Tighe et al. [30], who proposed non-parametric transfer of instance masks from the training set to detected objects, and by Hariharan et al. [14] who used convolutional neural nets (CNNs) [20] to classify region proposals. We use the terminology and metrics proposed by the latter in this paper. Dai et al. [8] used ideas from [17] to speed up the CNN-based proposal classification significantly.

A simple way of tackling this task is to run an object detector and segment out each detected instance. The notion of segmenting out detected objects has a long history in computer vision. Usually this idea has been used to aid semantic segmentation, or the task of labeling pixels in an image with category labels. Borenstein and Ullman [3] first suggested using category-specific information to improve the accuracy of segmentation. Yang et al. [33] start from object detections from the deformable parts model [10] and paste figure-ground masks for each detected object. Similarly, Brox et al. [5] and Arbeláez et al. [1] paste figure-ground masks for poselet detections [4]. Recent advances in computer vision have all but replaced early detectors such as DPM and poselets with ones based on CNNs [20, 12, 11] and produced dramatic improvements in performance in the process. In the CNN era, Hariharan et al. [16] used features from CNNs to segment out R-CNN detections [12].

When producing figure-ground masks for detections, most of these approaches predict every pixel independently. However, this disregards the fact that pixels in the image are hardly independent of each other, and a figure-ground labeling has to satisfy certain constraints. Some of these constraints can be simply encoded as local smoothness: nearby pixels of similar color should be labeled similarly. This can be achieved simply by aligning the predicted segmentation to image contours [5] or projecting to superpixels [16]. More sophisticated approaches model the problem using CRFs with unary and pairwise potentials [27, 24, 19]. Later work considers extending these models by incorporating higher-order potentials of specific forms for which inference is tractable [18, 21]. A related line of work explores learning a generative model of masks [9] using a deep Boltzmann machine [28]. Zheng et al. [35] show that inference in CRFs can be viewed as recurrent neural nets and trained together with a CNN to label pixels, resulting in large gains. Another alternative is to use eigenvectors obtained from normalized cuts as an embedding for pixels [23, 22].

However, images contain more structure than just local appearance-dependent smoothness. For instance, one high informative form of global cue is shape; in the case of persons, it encodes important constraints like two heads cannot be part of the same person, the head must be above the torso and so on. There has been prior work on handling such constraints in the pose estimation task by using graphical models defined over keypoint locations [34, 31]. However, in many applications, keypoint locations are unknown and such constraints must be enforced on raw pixels. Explicitly specifying these constraints on pixels is impractical, since it would require formulating potentials that are capable of localizing different parts of an object, which itself is a challenging task. Even if this could be done, the potentials that are induced would be higher order (which arises from the relative position constraints among multiple parts of an object) and non-submodular (due to mutual exclusivity constraints between pixels belonging to two different heads). This makes exact inference and training in these graphical models intractable.

Auto-context [32] and inference machines [26] take advantage of the observation that performing accurate inference does not necessarily require modelling the posterior distribution explicitly. Instead, these approaches devise efficient iterative inference procedures that directly approximate message passing. By doing so, they are able to leverage information from distant spatial locations when making predictions while remaining computationally efficient. In a similar spirit, other methods model the iterative process as recurrent neural nets [25, 35]. IEF [6] uses a related approach on the task of human pose estimation by directly refining the prediction rather than approximating message passing in each iteration. While this approach shows promise when the predictions lie in a low-dimensional space of possible 2D locations of human joints,

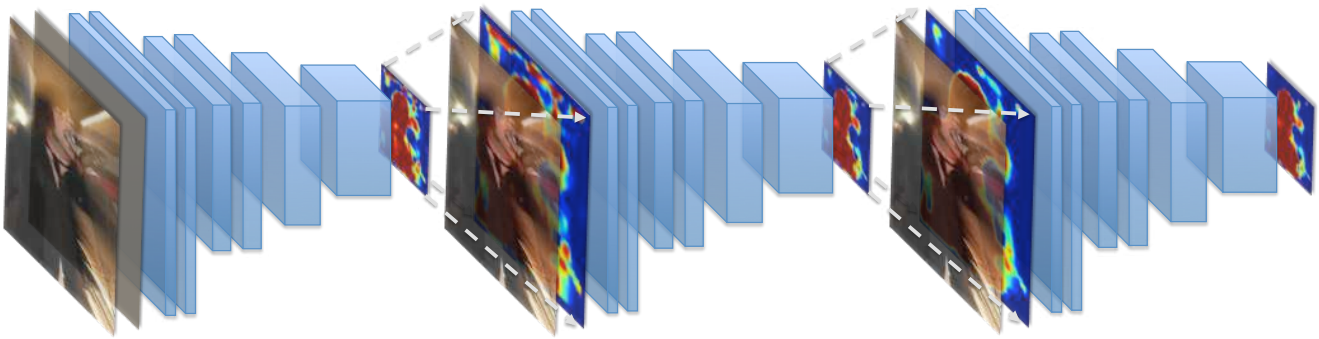

Figure 2: The proposed method decomposes the prediction process into multiple steps, each of which consists of performing unconstrained inference on the input image and the prediction from the preceding step. The diagram above illustrates a three-step prediction procedure when a convolutional neural net is used as the underlying model, as is the case with our method when applied to instance segmentation.

it is unclear if it will be effective when the output is high-dimensional and embeds complex structure like shape, as is the case with tasks that require a pixel-wise labelling of the input. In this paper, we devise an iterative method that supports prediction in high-dimensional spaces without a natural distance metric for measuring conformity to structure.

3. Method

3.1. Task and Setting

The objective of the instance segmentation task, also known as simultaneous detection and segmentation (SDS), is to predict the segmentation mask for each object instance in an image. Typically, an object detection system is run in the first stage of the pipeline, which generates a set of candidate bounding boxes along with the associated detection scores and category labels. Next, non-maximum suppression (NMS) is applied to these detections, which are then fed into the segmentation system, which predicts a heatmap for each bounding box representing the probability of each pixel inside the bounding box belonging to the foreground object of interest. The heatmaps then optionally undergo some form of post-processing, such as projection to superpixels. Finally, they are binarized by applying a threshold, yielding the final segmentation mask predictions. We use fast R-CNN [11] trained on MCG [2] bounding box proposals as our detection system and focus on designing the segmentation system in this paper.

3.2. Segmentation System

For our segmentation system, we use a CNN that takes a 224×224 patch as input and outputs a 50×50 heatmap prediction. The architecture is based on that of the hypercolmumn net proposed by Hariharan et al. [16], which is designed to be sensitive to image features at finer scales and relative locations of feature activations within the bounding box. Specifically, we use the architecture based on the VGG 16-layer net [29] (referred to as "O-Net" in [16]), in which heatmaps are computed from the concatenation of upsampled feature maps from multiple intermediate layers, known as the hypercolumn representation. The CNN is trained end-to-end on the PASCAL VOC 2012 training set with ground truth instance segmentation masks from the Semantic Boundaries Dataset (SBD) [13] starting from an initialization from the weights of a net finetuned for the detection task using R-CNN [12].

3.3. Algorithm

We would like to incorporate global cues like shape when making predictions. Shape encodes important structural constraints, such as the fact that a person cannot have two heads, which is why humans are capable of recognizing the category of an object from its silhouette almost effortlessly. So, leveraging shape enables us to disambiguate region hypotheses that all correctly cover pixels belonging to the category of interest but may group pixels into instances incorrectly.

Producing a heatmap prediction that is consistent with shape cues is a structured prediction problem, with the structure being shape constraints. The proposed algorithm works by reducing the structured prediction problem to a sequence of unconstrained prediction problems. Instead of forcing the model to produce a prediction that is consistent with both the input and the structure in a single step, we allow the model to disregard structure initially and train it to correct its mistakes arising from disregarding structure over multiple steps, while ensuring consistency of the prediction with the input in each step. The final prediction is therefore consistent with both the input and the structure. Later, we demonstrate that this procedure is capable to learning

a shape prior, a contiguity prior and a contour smoothness prior purely from data without any a priori specification to bias the learning towards finding these priors.

At test time, in each step, we feed the input image and the prediction from the previous step, which defaults to constant prediction of $1/2$ in the initial step, into the model and take the prediction from the last step as our final prediction. In our setting, the model takes the form of a CNN. Please see Figure 2 for a conceptual illustration of this procedure.

Algorithm 1 Training Procedure

Require: D is a training set consisting of (x, y) pairs, where x and y denote the instance and the ground truth labelling respectively, and f is the model

function TRAIN(D, f)

// $p_x^{(t)}$ is the predicted labelling of x in the t^{th} stage

$p_x^{(0)} \leftarrow \begin{pmatrix} 1/2 & \cdots & 1/2 \end{pmatrix}^T \ \forall\,(x, y) \in D$

for $t = 1$ **to** N **do**

// Training set for the current stage

$T \leftarrow \left\{ \left(\begin{pmatrix} x \\ p_x^{(i)} \end{pmatrix}, y \right) \middle| \, (x, y) \in D, i < t \right\}$

Train model f on T starting from the current parameters of f

$p_x^{(t)} \leftarrow f \begin{pmatrix} x \\ p_x^{(t-1)} \end{pmatrix} \ \forall\,(x, y) \in D$

end for

return f

end function

Algorithm 2 Testing Procedure

Require: f is the model and x is an instance

function TEST(f, x)

// $\hat{y}^{(t)}$ is the predicted labelling of x after t iterations

$\hat{y}^{(0)} \leftarrow \begin{pmatrix} 1/2 & \cdots & 1/2 \end{pmatrix}^T$

for $t = 1$ **to** M **do**

$\hat{y}^{(t)} \leftarrow f \begin{pmatrix} x \\ \hat{y}^{(t-1)} \end{pmatrix}$

end for

return $\hat{y}^{(M)}$

end function

Training the model is straightforward and is done in stages: in the first stage, the model is trained to predict the ground truth segmentation mask with the previous heatmap prediction set to $1/2$ for all pixels and the predictions of the model at the end of training are stored for later use. In each subsequent stage, the model is trained starting from the parameter values at the end of the previous stage to predict the ground truth segmentation mask from the input image and a prediction for the image generated during any of the preceding stages.

Pseudocode of the training and testing procedures are shown in Algorithms 1 and 2.

3.4. Discussion

Modelling shape constraints using traditional structured prediction approaches would be challenging for three reasons. First, because the notion of shape is highly abstract, it is difficult to explicitly formulate the set of structural constraints it imposes on the output. Furthermore, even if it could be done, manual specification would introduce biases that favour human preconceptions and lead to inaccuracies in the predictions. Therefore, manually engineering the form of structural constraints is neither feasible or desirable. Hence, the structural constraints are unknown and must be learned from data automatically. Second, because shape imposes constraints on the relationship between different parts of the object, such as the fact that a person cannot have two heads, it is dependent on the semantics of the image. As a result, the potentials must be capable of representing high-level semantic concepts like "head" and would need to have complex non-linear dependence on the input image, which would complicate learning. Finally, because shape simultaneously constrains the labels of many pixels and enforce mutual exclusivity between competing region hypotheses, the potentials would need to be of higher order and non-submodular, often making inference intractable.

Compared to the traditional single-step structured prediction paradigm, the proposed multi-step prediction procedure is more powerful because it is easier to model local corrections than the global structure. This can be viewed geometrically – a single-step prediction procedure effectively attempts to model the manifold defined by the structure directly, the geometry of which could be very complex. In contrast, our multi-step procedure learns to model the gradient of an implicit function whose level set defines the manifold, which tends to have much simpler geometry. Because it is possible to recover the manifold, which is a level set of an implicit function, from the gradient of the function, learning the gradient suffices for modelling structure.

3.5. Implementation Details

We modify the architecture introduced by Hariharan et al. [16] as follows. Because shape is only expected to be consistent for objects in the same category, we make the weights of the first layer category-dependent by adding twenty channels to the input layer, each corresponding to a different object category. The channel that corresponds to the category given by the detection system contains the heatmap prediction from the previous step, and channels corresponding to other categories are filled with zeros. To prepare the input to the CNN, patches inside the bounding boxes generated by the detection system are extracted and anisotropically scaled to 224×224 and the ground truth segmentation mask is transformed accordingly. Because the heatmap prediction from the preceding step is 50×50, we upsample it to 224×224 using bilinear interpolation before feeding it in as input. To ensure learning is well-conditioned, the heatmap prediction is rescaled and centred element-wise to lie in the range $[-127, 128]$ and the weights corresponding to the additional channels are ini-

tialized randomly with the same standard deviation as that of the weights corresponding to the colour channels.

The training set includes all detection boxes that overlap with the ground truth bounding boxes by more than 70%. At training time, boxes are uniformly sampled by category, and the weights for upsampled patches are set proportionally to their original areas for the purposes of computing the loss. The weights for all layers that are present in the VGG 16-layer architecture are initialized from the weights fine-tuned on the detection task and the weights for all other layers are initialized randomly. The loss function is the sum of the pixel-wise negative log likelihoods of the ground truth. The net is trained end-to-end using SGD on mini-batches of 32 patches with a learning rate of 5×10^{-5} and momentum of 0.9. We perform four stages of training and train for 30K, 42.5K, 50K and 20K iterations in stages one, two, three and four respectively. We find that the inference procedure typically converges after three steps and so we use three iterations at test time.

We can optionally perform post-processing by projecting to superpixels. To generate region predictions from heatmaps, we colour in a pixel or superpixel if the mean heat intensity inside a pixel or superpixel is greater than 40%. Finally, we can rescore the detections in the same manner as [16] by training support vector machines (SVMs) on features computed on the bounding box and the region predictions. To construct the training set, we take all bounding box detections that pass non-maximum suppression (NMS) using a bounding box overlap threshold of 70% and include those that overlap with the ground truth by more than 70% as positive instances and those by less than 50% as negative instances. To compute the features, we feed in the original image patch and the patch with the region background masked out to two CNNs trained as described in [15]. To obtain the final set of detections, we compute scores using the trained SVMs and apply NMS using a region overlap threshold of 30%.

3.6. Evaluation

We evaluate the proposed method in terms of region average precision (AP^r), which is introduced by [15]. Region average precision is defined in the same way as the standard average precision metric used for the detection task, with the difference being the computation of overlap between the prediction and the ground truth. For instance segmentation, overlap is defined as the pixel-wise intersection-over-union (IoU) of the region prediction and the ground truth segmentation mask, instead of the IoU of their respective bounding boxes. We evaluate against the SBD instance segmentation annotations on the PASCAL VOC 2012 validation set.

4. Experiments

First, we visualize the improvement in prediction accuracy as training progresses. In Figure 3, we show the pixel-wise heatmap predictions on image patches from the PASCAL VOC 2012 validation set after each stage of training. As shown, prediction quality steadily improves with each successive stage of training. Initially, the model is only able to identify some parts of the object; with each stage of training, it learns to recover additional parts of the object that were previously missed. After four stages of training, the model is able to correctly identify most parts belonging to the object. This indicates that the model is able to learn to make local corrections to its predictions in each stage. After four stages of training, the predictions are reasonably visually coherent and consistent with the underlying structure of the output space. Interestingly, the model gradually learns to suppress parts of other objects, as shown by the predictions on the bicycle and horse images, where the model learns to suppress parts of the pole and the other horse in later stages.

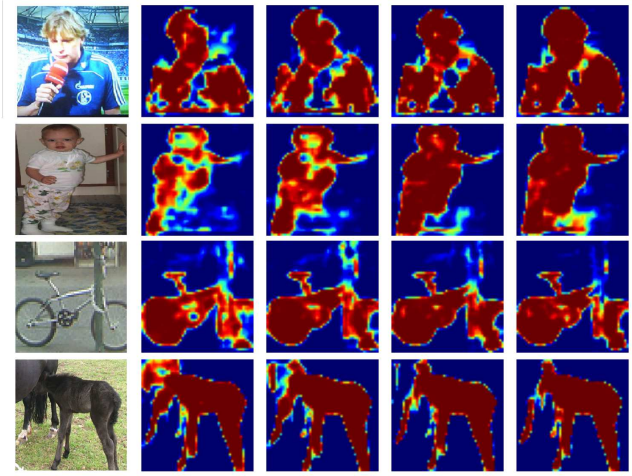

Figure 3: Heatmap predictions on images from the PASCAL VOC 2012 validation set after each stage of training. Best viewed in colour.

Next, we compare the performance of the proposed method with that of existing methods. As shown in Table 1, the proposed method outperforms all existing methods in terms of mean AP^r at both 50% and 70%. We analyze performance at a more granular level by comparing the proposed method to the state-of-the-art method, the hypercolumn net [16], under three settings: without superpixel projection, with superpixel projection and with superpixel projection and rescoring. As shown in Table 2, the proposed method achieves higher mean AP^r at 50% and 70% than the state-of-the-art in each setting. In particular, the proposed method achieves an 9.3-point gain over the state-of-

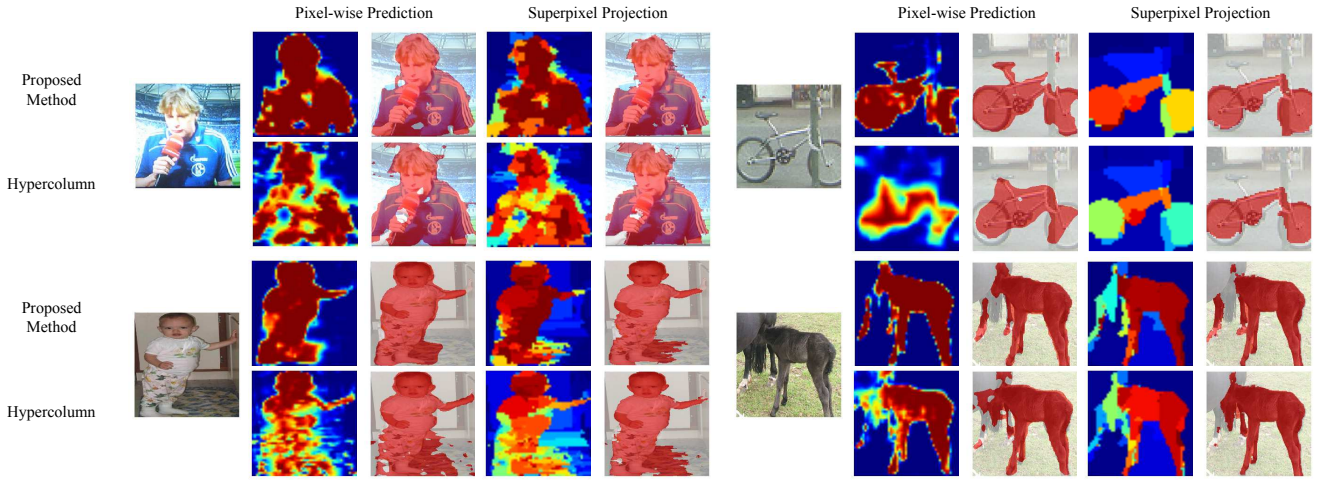

Figure 4: Comparison of heatmap and region predictions produced by the proposed method and the vanilla hypercolumn net on images from the PASCAL VOC 2012 validation set. Best viewed in colour.

Method	mAP^r at 50%	mAP^r at 70%
O_2P [7]	25.2	–
SDS [15]	49.7	25.3
CFM [8]	60.7	39.6
Hypercolumn [16]	62.4	39.4
Proposed Method	**63.6**	**43.3**

Table 1: Performance of the proposed method compared to existing methods.

Method and Setting	mAP^r at 50%	mAP^r at 70%
Raw pixel-wise prediction:		
Hypercolumn [16]	56.1	29.4
Proposed Method	**60.1**	**38.7**
With superpixel projection:		
Hypercolumn [16]	58.6	36.4
Proposed Method	**60.3**	**40.2**
With superpixel projection and rescoring:		
Hypercolumn [16]	62.4	39.4
Proposed Method	**63.6**	**43.3**

Table 2: Performance comparison of the proposed method and the state-of-the-art under different settings.

the-art in terms of its raw pixel-wise prediction performance at 70% overlap. This indicates the raw heatmaps produced by the proposed method are more accurate than those produced by the vanilla hypercolumn net. As a result, the proposed method requires less reliance on post-processing. We confirm this intuition by visualizing the heatmaps in Figure 4. When superpixel projection is applied, the proposed method improves performance by 1.7 points and 3.8 points at 50% and 70% overlaps respectively. With rescoring, the proposed method obtains a mean AP^r of 63.6% at 50% overlap and 43.3% at 70% overlap, which represent the best performance on the instance segmentation task to date. We break down performance by category under each setting in the supplementary material.

We examine heatmap and region predictions of the proposed method and the vanilla hypercolumn net, both with and without applying superpixel projection. As shown in Figure 4, the pixel-wise heatmap predictions produced by the proposed method are generally more visually coherent than those produced by the vanilla hypercolumn net. In particular, the proposed method predicts regions that are more consistent with shape. For example, the heatmap predictions produced by the proposed method for the sportscaster and the toddler images contain less noise and correctly identify most foreground pixels with high confidence. In contrast, the heatmap predictions produced by the hypercolumn net are both noisy and inconsistent with the typical shape of persons. On the bicycle image, the proposed method is able to produce a fairly accurate segmentation, whereas the hypercolumn net largely fails to find the contours of the bicycle. On the horse image, the proposed method correctly identifies the body and the legs of the horse. It also incorrectly hallucinates the head of the horse, which is actually occluded; this mistake is reasonable given the similar appearance of adjacent horses. This effect provides some evidence that the method is able to learn a shape prior successfully; because the shape prior discounts the probability of seeing a headless horse, it causes the model to hallucinate a head. On the other hand, the hypercolumn net chooses to hedge its bets on the possible locations of the head and so the resulting region prediction is noisy in the area near

the expected location of the head. Notably, the region predictions generated by the proposed method also tend to contain fewer holes and have smoother contours than those produced by the hypercolumn net, which is apparent in the case of the sportscaster and toddler images. This suggests that the model is able to learn a prior favouring the contiguity of regions and smoothness of region contours. More examples of heatmap and region predictions can be found in the supplementary material.

Applying superpixel projection significantly improves the region predictions of the vanilla hypercolumn net. It effectively smoothes out noise in the raw heatmap predictions by averaging the heat intensities over all pixels in a superpixel. As a result, the region predictions contain fewer holes after applying superpixel projection, as shown by the predictions on the sportscaster and toddler images. Superpixel projection also ensures that the region predictions conform to the edge contours in the image, which can result in a significant improvement if the raw pixel-wise region prediction is very poor, as is the case on the bicycle image. On the other hand, because the raw pixel-wise predictions of the proposed method are generally less noisy and have more accurate contours than those of the hypercolumn net, superpixel projection does not improve the quality of predictions as significantly. In some cases, it may lead to a performance drop, as pixel-wise prediction may capture details that are missed by the superpixel segmentation. As an example, on the bicycle image, the seat is originally segmented correctly in the pixel-wise prediction, but is completely missed after applying superpixel projection. Therefore, superpixel projection has the effect of masking prediction errors and limits performance when the quality of pixel-wise predictions becomes better than that of the superpixel segmentation.

We find that the proposed method is able to avoid some of the mistakes made by the vanilla hypercolumn net on images with challenging scene configurations, such as those depicting groups of people or animals. On such images, the hypercolumn net sometimes includes parts of adjacent persons in region predictions. Several examples are shown in Figure 5, in which region predictions contain parts from different people or animals. The proposed method is able to suppress parts of adjacent objects and correctly exclude them from region predictions, suggesting that the learned shape prior is able to help the model disambiguate region hypotheses that are otherwise consistent with local appearance cues.

We now analyze the improvement in overlap between region predictions and the ground truth segmentation masks at the level of individual detections. In Figure 6, we plot the maximum overlap of the pixel-wise region prediction produced by the proposed method with the ground truth against that of the region prediction generated by the vanilla hypercolumn net for each of the top 200 detections in each

Figure 5: Region predictions on images with challenging scene configurations.

category. So, in this plot, any data point above the diagonal represents a detection for which the proposed method produces a more accurate region prediction than the hypercolumn net. We find overlap with ground truth improves for 76% of the detections, degrades for 15.6% of the detections and remains the same for the rest. This is reflected in the plot, where the vast majority of data points lie above the diagonal, indicating that the proposed method improves the accuracy of region predictions for most detections.

Remarkably, for detections on which reasonably good overlap is achieved using the vanilla hypercolumn net, which tend to correspond to bounding boxes that are well-localized, the proposed method can improve overlap by 15% in many cases. Furthermore, the increase in overlap tends to be the greatest for detections on which the hypercolumn net achieves 75% overlap; when the proposed method is used, overlap for these detections at times reach more than 90%. This is particularly surprising given that improving upon good predictions is typically challenging.

Such a performance gain is conceptually difficult to achieve without leveraging structure in the output. This suggests that the proposed method is able to use the priors it learned to further refine region predictions that are already very accurate.

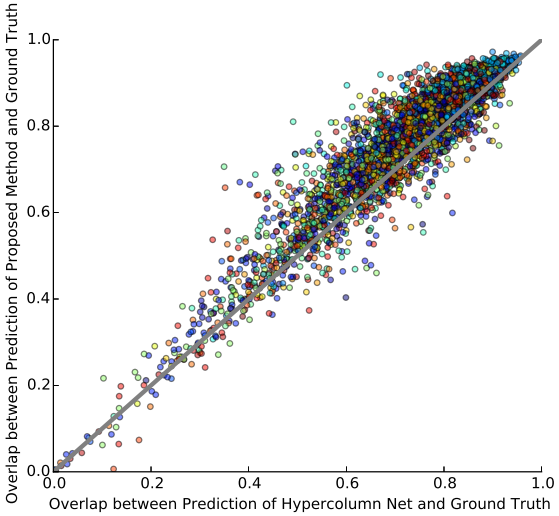

Figure 6: Comparison of maximum overlap of region predictions produced by the vanilla hypercolumn net and the proposed method with the ground truth. Each data point corresponds to a bounding box detection and the colour of each data point denotes the category of the detection. Points that lie above the diagonal represent detections for which the region predictions produced by the proposed method are more accurate than those produced by the hypercolumn net.

Finally, we conduct an experiment to test whether the proposed method is indeed able to learn a shape prior more directly. To this end, we select an image patch from the PASCAL VOC 2012 validation set that contains little visually distinctive features, so that it does not resemble an object from any of the categories. We then feed the patch into the model along with an arbitrary category label, which essentially forces the model to try to interpret the image as that of an object of the particular category. We are interested in examining if the model is able to hallucinate a region that is both consistent with the input image and resembles an object from the specified category.

Figure 7 shows the input image and the resulting heatmap predictions under different settings of category. As shown, when the category is set to bird, the heatmap prediction resembles the body and the wing of a bird. When the category is set to horse, the model hallucinates the body and the legs of a horse. Interestingly, the wing of the bird and the legs of the horse are hallucinated even though there are no corresponding contours that resemble these parts in the input image. When the category is set to bicycle, the model interprets the edges in the input image as the frame of a bicycle, which contrasts with the heatmap prediction when the category is set to television, which is not sensitive to thin edges in the input image and instead contains a large contiguous box that resembles the shape of a television set.

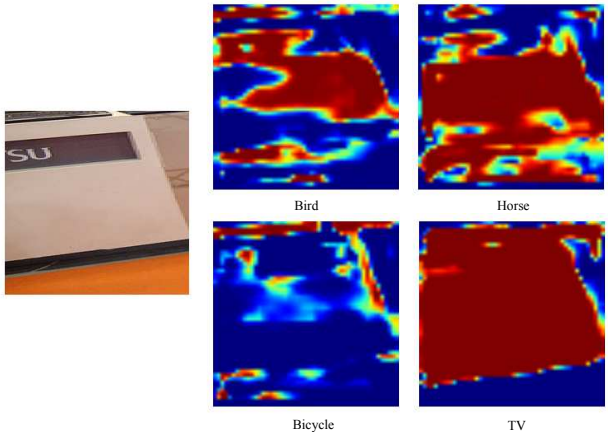

Figure 7: Heatmap predictions of the proposed method under different settings of category. As shown, the model is able to hallucinate plausible shapes that correspond to the specified categories.

5. Conclusion

We presented a method that is able to take advantage of the implicit structure that underlies the output space when making predictions. The method does not require manual specification of the form of the structure a priori and is able to discover salient structure from the data automatically. We applied the method to the instance segmentation task and showed that the method automatically learns a prior on shape, contiguity of regions and smoothness of region contours. We also demonstrated state-of-the-art performance using the method, which achieves a mean AP^r of 63.6% and 43.3% at 50% and 70% overlaps respectively. The method is generally applicable to all tasks that require the prediction of a pixel-wise labelling of the input image; we hope the success we demonstrated on instance segmentation will encourage application to other such tasks and further exploration of the method.

Acknowledgements. This work was supported by ONR MURI N00014-09-1-1051 and ONR MURI N00014-14-1-0671. Ke Li thanks the Natural Sciences and Engineering Research Council of Canada (NSERC) for fellowship support. The authors also thank NVIDIA Corporation for the donation of GPUs used for this research.

References

[1] P. Arbeláez, B. Hariharan, C. Gu, S. Gupta, and J. Malik. Semantic segmentation using regions and parts. In *CVPR*, 2012. 2

[2] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 328–335. IEEE, 2014. 3

[3] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *ECCV*. 2002. 2

[4] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*. 2010. 2

[5] T. Brox, L. Bourdev, S. Maji, and J. Malik. Object segmentation by alignment of poselet activations to image contours. In *CVPR*, 2011. 2

[6] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. *arXiv preprint arXiv:1507.06550*, 2015. 1, 2

[7] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*. 2012. 6

[8] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *CVPR*, 2015. 2, 6

[9] S. A. Eslami, N. Heess, C. K. Williams, and J. Winn. The shape boltzmann machine: a strong model of object shape. *International Journal of Computer Vision*, 107(2):155–176, 2014. 2

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9), 2010. 2

[11] R. Girshick. Fast R-CNN. In *ICCV*, 2015. 2, 3

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2, 3

[13] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 991–998. IEEE, 2011. 3

[14] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, 2014. 2

[15] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *Computer Vision–ECCV 2014*, pages 297–312. Springer, 2014. 5, 6

[16] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 2015. 1, 2, 3, 4, 5, 6

[17] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014. 2

[18] P. Kohli, A. Osokin, and S. Jegelka. A principled deep random field model for image segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1971–1978. IEEE, 2013. 2

[19] P. Krähenbühl and V. Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*, 2011. 1, 2

[20] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4), 1989. 2

[21] Y. Li, D. Tarlow, and R. Zemel. Exploring compositional high order pattern potentials for structured output learning. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 49–56. IEEE, 2013. 2

[22] M. Maire, S. X. Yu, and P. Perona. Object detection and segmentation from joint embedding of parts and pixels. In *ICCV*, 2011. 2

[23] S. Maji, N. K. Vishnoi, and J. Malik. Biased normalized cuts. In *CVPR*, 2011. 2

[24] O. M. Parkhi, A. Vedaldi, C. Jawahar, and A. Zisserman. The truth about cats and dogs. In *ICCV*, 2011. 2

[25] P. H. O. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In *ICML*, 2014. 2

[26] S. Ross, D. Munoz, M. Hebert, and J. A. Bagnell. Learning message-passing inference machines for structured prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 1, 2

[27] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG)*, 23(3):309–314, 2004. 2

[28] R. Salakhutdinov and G. E. Hinton. Deep boltzmann machines. In *International conference on artificial intelligence and statistics*, pages 448–455, 2009. 2

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[30] J. Tighe, M. Niethammer, and S. Lazebnik. Scene parsing with object instances and occlusion handling. In *ECCV*, 2010. 2

[31] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014. 2

[32] Z. Tu and X. Bai. Auto-context and its application to high-level vision tasks and 3D brain image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(10):1744–1757, 2010. 1, 2

[33] Y. Yang, S. Hallman, D. Ramanan, and C. C. Fowlkes. Layered object models for image segmentation. *TPAMI*, 34(9), 2012. 2

[34] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *TPAMI*, 35(12), 2013. 2

[35] 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. *arXiv preprint arXiv:1502.03240*, 2015. 2