Zoom Better to See Clearer: Human and Object Parsing with Hierarchical Auto-Zoom Net

Fangting Xia(✉), Peng Wang, Liang-Chieh Chen, and Alan L. Yuille

University of California, Los Angeles, USA
{sukixia,jerrykingpku,lcchen,yuille}@ucla.edu, alan.yuille@jhu.edu

Abstract. Parsing articulated objects, *e.g.* humans and animals, into semantic parts (*e.g.* head, body and arms, *etc.*) from natural images is a challenging and fundamental problem in computer vision. A big difficulty is the large variability of scale and location for objects and their corresponding parts. Even limited mistakes in estimating scale and location will degrade the parsing output and cause errors in boundary details. To tackle this difficulty, we propose a "Hierarchical Auto-Zoom Net" (HAZN) for object part parsing which adapts to the local scales of objects and parts. HAZN is a sequence of two "Auto-Zoom Nets" (AZNs), each employing fully convolutional networks for two tasks: (1) predict the locations and scales of object instances (the first AZN) or their parts (the second AZN); (2) estimate the part scores for predicted object instance or part regions. Our model can adaptively "zoom" (resize) predicted image regions into their proper scales to refine the parsing. We conduct extensive experiments over the PASCAL part datasets on humans, horses, and cows. In all the three categories, our approach significantly outperforms alternative state-of-the-arts by more than 5 % mIOU and is especially better at segmenting small instances and small parts. In summary, our strategy of first zooming into objects and then zooming into parts is very effective. It also enables us to process different regions of the image at different scales adaptively so that we do not need to waste computational resources scaling the entire image.

Keywords: Human parsing · Part segmentation · Multi-scale modeling

1 Introduction

When people look at natural images, they often first locate regions that contain objects, and then perform the more detailed task of object parsing, *i.e.* decomposing each object instance into its semantic parts. In computer vision, object parsing plays a key role in the real understanding of objects in images and helps for many visual tasks, *e.g.* segmentation [9,30], pose estimation [8],

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46454-1_39) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part V, LNCS 9909, pp. 648–663, 2016.
DOI: 10.1007/978-3-319-46454-1_39

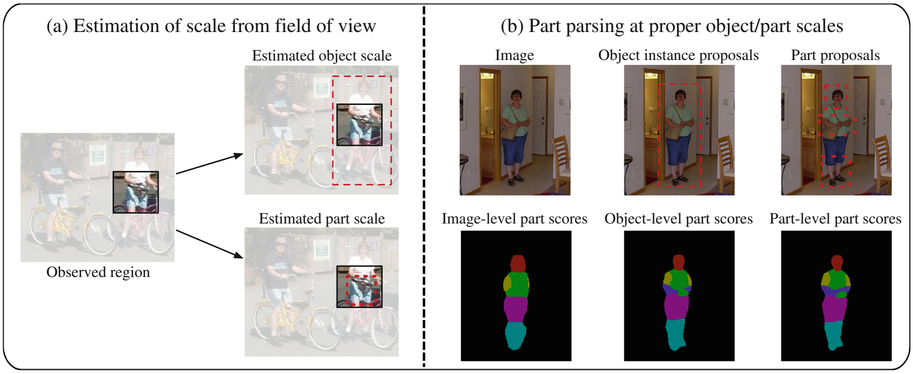


Fig. 1. Intuition of Hierarchical Auto-Zoom Net (HAZN). (a) The scale and location of an object and its parts (the red dashed boxes) can be estimated from the observed field of view (the black solid box) of a neural network. (b) Part parsing can be more accurate by using proper object and part scales. At the top row, we show our estimated object and part scales. In the bottom row, our part parsing results gradually become better by increasingly utilizing the estimated object and part scales. (Color figure online)

and fine-grained recognition [35]. It also has many industrial applications such as robotics and image descriptions for the blind.

There has been a growing literature on the related task of object semantic segmentation due to the availability of benchmarks such as PASCAL VOC [10] and MS-COCO [20]. There has been work on human parsing, *i.e.* segmenting humans into their semantic parts, but this has mainly been studied under constrained conditions which pre-suppose known scale, fairly accurate localization, clear appearances, and/or relatively simple poses [3,8,9,21,34,36]. There are few works done on parsing animals, like cows and horses, yet these also face similar restrictions, *e.g.* roughly known size and location [29,30].

In this paper, we address the task of parsing objects, such as humans and animals, in "the wild" where there are large variations in scale, location, occlusion, and pose. This motivates us to work with PASCAL images [10] because these were chosen for studying multiple visual tasks, do not suffer from dataset design bias [18], and include large variations of objects, particularly of scale. Parsing humans in PASCAL is considerably more difficult than in other datasets like Fashionista [34], which were constructed solely to evaluate human parsing.

Recently, deep learning methods have led to big improvements on object parsing [13,30], with the emergence of fully convolutional nets (FCNs) [23] and the availability of object part annotations on large-scale datasets, *e.g.* PASCAL [6]. However, these methods can still make mistakes on small or large scale objects and, in particular, they have no mechanism to adapt to the size of the object.

In this paper, we present a hierarchical method for object parsing that performs scale estimation and object parsing jointly and is able to adapt its scale to objects and parts. It is partly motivated by the proposal-free end-to-end

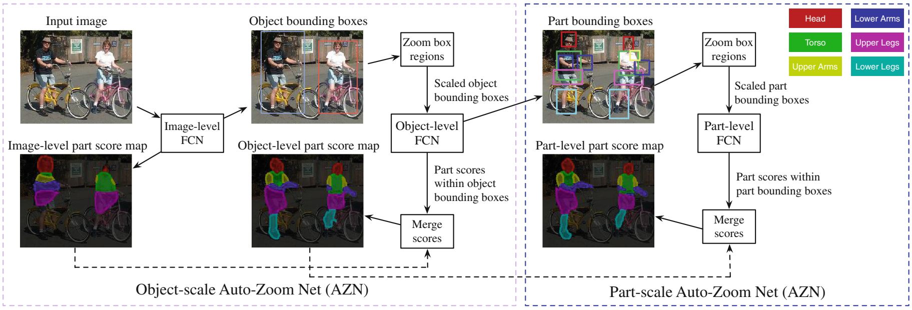

Input image
Object bounding boxes
Zoom box regions
Scaled object bounding boxes
Image-level FCN
Object-level FCN
Image-level part score map
Object-level part score map
Part scores within object bounding boxes
Merge scores
Object-scale Auto-Zoom Net (AZN)
Part bounding boxes
Zoom box regions
Scaled part bounding boxes
Part-level FCN
Part-level part score map
Part scores within part bounding boxes
Merge scores
Part-scale Auto-Zoom Net (AZN)
Head
Lower Arms
Torso
Upper Legs
Upper Arms
Lower Legs

Fig. 2. Testing framework of HAZN. We address object part parsing by adapting to the sizes of objects (object-scale AZN) and parts (part-scale AZN). The part scores are predicted and refined by three FCNs, over three levels of granularity, *i.e.* image-level, object-level, and part-level. At each level, the FCN outputs the part score map for the current level, and estimates the locations and scales for the next level. The details of parts are gradually discovered and improved along the proposed auto-zoom process (*i.e.* location/scale estimation, region zooming, and part score re-estimation).

detection strategies [15,19,26,27], which prove that the scale and location of a target object, and of its corresponding parts, can be estimated accurately from the field-of-view (FOV) window by applying a deep net (Fig. 1(a)). We call our approach "Hierarchical Auto-Zoom Net" (HAZN) which parses the objects at three levels of granularity, namely image-level, object-level, and part-level, gradually giving clearer and better parsing results (see Fig. 1(b)). The HAZN sequentially combines two "Auto-Zoom Nets" (AZNs), each of which predicts the locations and scales for objects (the first AZN) or parts (the second AZN), properly zooms (resizes) the predicted image regions, and refines the object parsing results for those image regions (see Fig. 2). The HAZN uses three FCNs [23] that share the same structure. The first FCN acts directly on the image to estimate a finite set of possible locations and sizes of objects (*e.g.* bounding boxes) with confidence scores, together with a part score map of the image. The part score map is similar to that proposed by previous deep-learned methods. The object bounding boxes are scaled to a fixed size by zooming in or zooming out (as applicable) and the image and part score maps within the boxes are also scaled by bilinear interpolation for zooming in or downsampling for zooming out. Then the second FCN is applied to the scaled object bounding boxes to make proposals (bounding boxes) for the parts, with confidence values, and to re-estimate the part scores within the object bounding boxes. This yields improved part scores. We then apply the third FCN to the scaled part bounding boxes to produce new estimates of the part scores and to combine all of them (for different object and part bounding boxes) to output final part scores, which are our parse of the object. This strategy is modified slightly so that we scale humans differently depending on whether we have detected a complete human or only the upper part of a human, which can be determined from the part score map.

For dealing with scale, the adaptiveness of our approach and the way it combines scale estimation with parsing give novel computational advantages over traditional multi-scale methods. Previous methods mainly select a fixed set of scales and then perform fusion on the outputs of a deep net at different layers. Computational requirements mean that the number of scales must be small and it is impractical to use very fine scales due to memory limitations. Our approach is considerably more flexible because we adaptively estimate scales at different regions in the image which allows us to search over a large range of scales. In particular, we can use very fine scales because we will probably only need to do this within small image regions. For example, our largest zooming ratio is 2.5 (at part level) on PASCAL while that number is 1.5 if we have to zoom the whole image. This is a big advantage when trying to detect small parts, such as the tail of a cow, as is shown by the experiments.

We report extensive experimental results for parsing humans on the challenging PASCAL-Person-Part dataset [6] and for parsing animals on a horse-cow dataset [29]. Our approach outperforms previous state-of-the-arts by a large margin. We are particulary good at detecting small object parts.

2 Background

The study of human part parsing has been largely restricted to constrained environments, where a human instance in an image is well localized and has a relatively simple pose like standing or walking [3,8,9,21,33,34,36]. These shape-based or appearance-based models (with hand-crafted features or bottom-up segments) are limited when applied to parsing human instances in the wild because humans in real-world images are often in various poses, scales, and may be occluded or highly deformed.

Over the past few years, with the powerful deep convolutional neural networks (DCNNs) [17] and big data, researchers have made significant performance improvement for semantic object segmentation in the wild [4,7,22,24,25,28,31], showing that DCNNs can also be applied to segment object parts in the wild. These deep segmentation models work on the whole image, regarding each semantic part as a class label. But this strategy suffers from the large scale variation of objects and parts, and many details can be easily missed. [13] proposed to sequentially perform object detection, object segmentation and part segmentation, in which the object is first localized by a RCNN [12], then the object (in the form of a bounding box) is segmented by a FCN [23] to produce an object mask, and finally part segmentation is performed by partitioning the mask. The process has two potential drawbacks: (1) it is complex to train all components of the model; (2) the error from object masks, *e.g.* local confusion and inaccurate edges, propagates to the part segments. Our model follows this general coarse-to-fine strategy, but is more unified (with all three FCNs employing the same structure) and more importantly, we do not make premature decisions. In order to better discover object details and use object-level context, [30] employed a two-stream FCN to jointly infer object and part segmentations for animals, where the part stream was performed to discover part-level details and the object stream was performed to find object-level context. Although this work discovers object-level context to help part parsing, it only uses a single-scale network for both object and part score prediction, where small-scale objects might be missed at the beginning and the scale variation of parts still remains unsolved.

Many studies in computer vision have addressed the scale issue to improve recognition or segmentation. These include exploiting multiple cues [14], hierarchical region grouping [2,11], and applying general or salient object proposals combined with iterative localization [1,32,37]. However, most of these works either adopted low-level features or only considered constrained scene layouts, making it hard to handle wild scene variations and difficult to unify with DCNNs. Some recent works try to handle the scale issue within a DCNN structure. They commonly use multi-scale features from intermediate layers, and perform late fusion on them [4,13,23] in order to achieve scale invariance. Most recently, [5] proposed a scale attention model, which learns pixel-wise weights for merging the outputs from three fixed scales. These approaches, though developed on powerful DCNNs, are all limited by the number of scales they can select and the possibility that the scales they select may not cover a proper one. Our model avoids the scale selection error by directly regressing the bounding boxes for objects/parts and zooming the regions into proper scales. In addition, this mechanism allows us to explore a broader range of scales, contributing to the discovery of missing objects and the accuracy of part boundaries.

3 The Model

As shown in Fig. 2, our Hierarchical Auto-Zoom model (HAZN) has three levels of granularity for tackling scale variation in object parsing, *i.e.* image-level, object-level, and part-level. At each level, a fully convolutional neural network (FCN) is used to perform scale/location estimation and part parsing simultaneously. The three levels of FCNs are all built on the same network structure, a modified FCN called DeepLab-LargeFOV [4]. This network structure is one of the most effective FCNs in segmentation, so we also treat it as our baseline for final performance comparison.

To handle scale variation in objects and parts, the HAZN concatenates two Auto-Zoom Nets (AZNs), namely object-scale AZN and part-scale AZN, into a unified network. The object-scale AZN refines the image-level part score map with object bounding box proposals while the part-scale AZN further refines the object-level part score map with part bounding box proposals. Each AZN employs an auto-zoom process: first estimates the region of interest (ROI), then properly resizes the predicted regions, and finally refines the part scores within the resized regions.

3.1 Object-Scale Auto-Zoom Net (AZN)

For the task of object part parsing, we are provided with n training examples $\{\mathbf{I}_i, \mathbf{L}_i\}_{i=1}^n$, where $\mathbf{I}$ is the given image and $\mathbf{L}$ is the pixel-wise semantic part

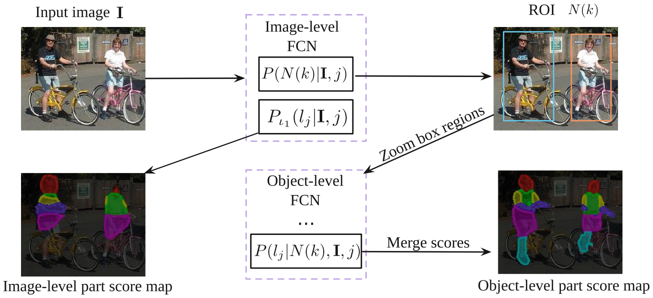


Fig. 3. Object-scale Auto-Zoom Net from a probabilistic view, which predicts ROI region $N(k)$ at object-scale, and then refines part scores based on the properly zoomed region $N(k)$. Details are in Sect. 3.1.

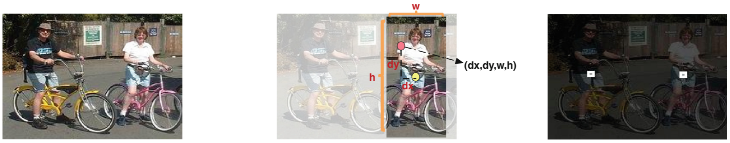
labels. Our target is to learn the posterior distribution $P(l_j|\mathbf{I}, j)$ for each pixel j of an image $\mathbf{I}$, which is approximated by our object-scale AZN (see Fig. 3).

We first use the image-level FCN (see Fig. 2) to produce the image-level part score map $P_{\iota_1}(l_j|\mathbf{I}, j)$, which gives comparable performance to our baseline method (DeepLab-LargeFOV). This is a normal ***part parsing network*** that uses the original image as input and outputs the pixel-wise part score map. Our object-scale AZN aims to refine this part score map with consideration of object instance scales. To do so, we add a second component to the image-level FCN, performing regression to estimate the size and location of an object bounding box (or ROI) for each pixel, together with a confidence map indicating the likelihood that the box is an object. This component is called a ***scale estimation network*** (**SEN**), which shares the first few layers with the part parsing network in the image-level FCN. In math, the SEN corresponds to a probabilistic model $P(b_j|\mathbf{I}, j)$, where b_j is the estimated bounding box for pixel j, and $P(b_j|...)$ is the confidence score of b_j.

After getting $\{b_j|\forall j \in \mathbf{I}\}$, we threshold the confidence map and perform non-maximum suppresion to yield a finite set of object ROIs (typically 5–10 per image, with some overlap): $\{b_k|k \in \mathbf{I}\}$. Each b_k, the bounding box estimated from pixel k, is associated with a confidence score $P(b_k)$. As shown in Fig. 2, a **region zooming** operation is then performed on each b_k, resizing b_k to a standard-sized ROI $N(k)$. Specifically, this zooming operation computes a zooming ratio for bounding box b_k, and then enlarges or shrinks the image within b_k by the zooming ratio. We will discuss how to compute the zooming ratio in Sect. 4.

Now we have a set of zoomed ROI proposals $\{N(k)|k \in \mathbf{I}\}$, each $N(k)$ associated with score $P(b_k)$. We learn another probabilistic model $P(l_j|N(k), \mathbf{I}, j)$, which re-estimates the part label for each pixel j within the zoomed ROI $N(k)$. This probabilistic model corresponds to the part parsing network in the object-level FCN (see Fig. 2), which takes as input the zoomed object bounding boxes and outputs the part scores within those object bounding boxes.

The new part scores for the zoomed ROIs need to be merged to produce the object-level part score map for the whole image. Since there may be multiple



Input image

Object bounding box regression map

Object confidence seeds

Fig. 4. Ground truth regression target for training the scale estimation network (SEN) in the image-level FCN. Details in Sect. 3.3.

ROIs that cover a pixel j, we define the neighbouring region set for pixel j as $\mathcal{Q}(j) = \{N(k) | j \in N(k), k \in \mathbf{I}\}$. Under this definition of $\mathcal{Q}(j)$, the **score merging** process can be expressed as Eq. 1, which essentially computes the weighted sum of part scores for pixel j, from the zoomed ROIs that cover j. For a pixel that is not covered by any zoomed ROI, we simply use its image-level part score as the current part score. Formally, the object-level part score $P_{\iota_2}(l_j|\mathbf{I}, j)$, is computed as,

$$P_{\iota_2}(l_j|\mathbf{I}, j) = \sum\nolimits_{N(k) \in \mathcal{Q}(j)} P(l_j|N(k), \mathbf{I}, j) P(N(k)|\mathbf{I}, j);$$
$$P(N(k)|\mathbf{I}, j) = P(b_k) / \sum\nolimits_{k: N(k) \in \mathcal{Q}(j)} P(b_k) \tag{1}$$

3.2 Hierarchical Auto-Zoom Net (HAZN)

The scale of object parts can also vary considerably even if the scale of the object is fixed. This leads to a hierarchical strategy with multiple stages, called the Hierarchical Auto-Zoom Net (HAZN), which applies AZNs to images to find objects and then on objects to find parts, followed by a part score refinement stage. As shown in Fig. 2, we add the part-scale AZN to the end of the object-scale AZN. Specifically, we add a second component (*i.e.* SEN) to the object-level FCN, to estimate the size and location of part bounding boxes, together with confidence maps for every pixel within a zoomed object ROI. Again the confidence map is thresholded, and non-maximal suppresion is applied, to yield a finite set of part ROIs (typically 5–30 per image, with some overlap). Each part ROI is zoomed to a fixed size. Then, we re-estimate the part scores within each zoomed part ROI using the part parsing network in the part-level FCN. The part parsing network is the only component of the part-level FCN, which takes the zoomed part ROI and the zoomed object-level part scores (within the part ROI) as inputs. After getting the part scores within each zoomed Part ROI, the score merging process is the same as in the object-scale AZN.

It's worth mentioning that we can easily extend our HAZN to include more AZNs at finer scale levels if we focus on smaller object parts such as human eyes.

3.3 Training and Testing Phases for Object-Scale AZN

We use **DeepLab-LargeFOV** [4] as the basic network structure for both the scale estimation network (SEN) and the part parsing network. The two networks, serving as components of a multi-tasking FCN, share the first three layers.

Training the SEN. The SEN aims to regress the region of interest (ROI) for each pixel j in the form of a bounding box, b_j. Here we borrow the idea of DenseBox [15] for scale estimation, since it is simple and performs well enough for our task. In detail, at object level, the ROI of pixel j corresponds to the object instance box that pixel j belongs to. For training the SEN, two output label maps are needed as visualized in Fig. 4. The first one is the bounding box regression map $\mathbf{L}_b$, which is a four-channel output for each pixel j to represent its ROI b_j: $\mathbf{l}_{bj} = \{dx_j, dy_j, w_j, h_j\}$. Here (dx_j, dy_j) is the relative position from pixel j to the center of b_j; h_j and w_j are the height and width of b_j. We then re-scale the outputs by dividing them with 400. The other target output map is a binary confidence seed map $\mathbf{L}_c$, in which $\mathbf{l}_{cj} \in \{0, 1\}$ is the ROI selection indicator at pixel j. It indicates the preferred pixels for us to use for ROI prediction, which helps the algorithm prevent many false positives. In practice, we choose the central pixels of each object instance as the confidence seeds, which tend to predict the object bounding boxes more accurately than those pixels at the boundary of an object instance region.

Given the ground-truth label maps of object part parsing, we can easily derive the training examples for the SEN: $\mathcal{H} = \{\mathbf{I}_i, \mathbf{L}_{bi}, \mathbf{L}_{ci}\}_{i=1}^{n}$, where n is the number of training instances. We minimize the negative log likelihood to learn the weights $\mathbf{W}$ for the SEN, and the loss l_{SEN} is defined in Eq. 2.

$$\begin{aligned} l_{SEN}(\mathcal{H}|\mathbf{W}) &= \frac{1}{n}\sum\nolimits_i (l_b(\mathbf{I}_i, \mathbf{L}_{bi}|\mathbf{W}) + \lambda l_c(\mathbf{I}_i, \mathbf{L}_{ci}|\mathbf{W})); \\ l_c(\mathbf{I}, \mathbf{L}_c|\mathbf{W}) &= -\beta \sum_{j:l_{cj}=1} \log P(l^*_{cj} = 1|\mathbf{I}, \mathbf{W}) - (1-\beta) \sum_{j:l_{cj}=0} \log P(l^*_{cj} = 0|\mathbf{I}, \mathbf{W}); \\ l_b(\mathbf{I}, \mathbf{L}_b|\mathbf{W}) &= \frac{1}{|\mathbf{L}^+_{cj}|} \sum\nolimits_{j:l_{cj}=1} \|\mathbf{l}_{bj} - \mathbf{l}^*_{bj}\|^2 \end{aligned} \quad (2)$$

For the confidence seeds, we employ the balanced cross entropy loss, where l^*_{cj} and l_{cj} are the predicted value and ground truth value respectively. The probability is from a sigmoid function performing on the activation of the last layer of the CNN at pixel j. β is defined as the proportion of pixels with $l_{cj} = 0$ in the image, which is used to balance the positive and negative instances. The loss for bounding box regression is the Euclidean distance over the confidence seed points, and $|\mathbf{L}^+_{cj}|$ is the number of pixels with $l_{cj} = 1$.

Testing the SEN. The SEN outputs both the confidence score map $P(l^*_{cj} = 1|\mathbf{I}, \mathbf{W})$ and a four-dimensional bounding box $\mathbf{l}^*_{bj}$ for each pixel j. We regard a pixel j with confidence score higher than 0.5 to be reliable and output its bounding box $b_j = \mathbf{l}^*_{bj}$, associated with confidence score $P(b_j) = P(l^*_{cj} = 1|\mathbf{I}, \mathbf{W})$.

We perform non-maximum suppression (IOU threshold = 0.4) based on the confidence scores, yielding several candidate bounding boxes $\{\mathbf{b}_j | j \in \mathbf{I}\}$ with confidence scores $P(\mathbf{b}_j)$. Each b_j is then properly zoomed, becoming $N(j)$.

Training the part parsing. The training of the part parsing network is standard. For the object-level FCN, the part parsing network is trained based on all the zoomed image regions (ROIs), with the ground-truth part label maps $\mathcal{H}_p = \{\mathbf{L}_{pi}\}_{i=1}^n$ within the zoomed ROIs. For the image-level FCN, the part parsing network is trained based on the original training images. We merge the part parsing network with the SEN, yielding the image-level FCN with loss defined in Eq. 3. Here, $l_p(\mathbf{I}, \mathbf{L}_p)$ is the commonly used multinomial logistic regression loss for classification.

$$l_{AZN}(\mathcal{H}, \mathcal{H}_p | \mathbf{W}) = \frac{1}{n} \sum\nolimits_i l_p(\mathbf{I}_i, \mathbf{L}_{pi}) + l_{SEN}(\mathcal{H} | \mathbf{W}); \tag{3}$$

Testing the part parsing. For testing the object-scale AZN, we first run the image-level FCN, yielding part score maps at the image level and bounding boxes for the object level. Then we zoom onto the bounding boxes and parse these regions based on the object-level FCN, yielding part score maps at the object level. By merging the part score maps from the two levels, we get better parsing results for the whole image.

4 Experiments

4.1 Implementation Details

Selection of confidence seeds. To train the scale estimation network (SEN), we need to select confidence seeds for object instances or parts. For human instances, we use the human instance masks from the PASCAL-Person-Part Dataset [6] and select the central 7×7 pixels within each instance mask as the confidence seeds. To get the confidence seeds for human parts, we first compute connected part segments from the groundtruth part label map, and then also select the central 7×7 pixels within each part segment. We present the details of our approach for humans because the extension to horses and cows is straightforward.

Zooming ratio of ROIs. The SEN networks in the FCNs provide a set of human/part bounding boxes (ROIs), $\{b_j | j \in \mathbf{I}\}$, which are then zoomed to a proper human/part scale. The zooming ratio of b_j, $f(b_j, L_p^{b_j})$, is decided based on the size of b_j and the previously computed part label map $L_p^{b_j}$ within b_j. We use slightly different strategies to compute the zooming ratio at the human and part levels. For the part level, we simply resize the bounding box to a fixed size, *i.e.* $f_p(b_j) = s_t / max(w_j, h_j)$, where $s_t = 255$ is the target size. Here w_j and h_j are the width and height of b_j. For the human level, we need to consider the frequently occurred truncation case when only the upper half of a human instance is visible. In practice, we use the image-level part label map $L_p^{b_j}$ within the box, and check the existence of legs to decide whether the full body is visible. If the full body is visible, we use the same strategy as parts. Otherwise, we change the target size s_t to 140, yielding relative smaller region than the full body visible case. We select the target size based on a validation set. Finally, we limit all zooming ratio $f_p(b_j)$ within the range $[0.4, 2.5]$ for both human and part bounding boxes to avoid artifacts from up or down sampling of images.

4.2 Experimental Protocol

Dataset. We conduct experiments on humans part parsing using the PASCAL-Person-Part dataset annotated by [6]. The dataset contains detailed part annotations for every person, *e.g.* head, torso, *etc.* We merge the annotations into six clases: Head, Torso, Upper/Lower Arms and Upper/Lower Legs (plus one background class). We only use those images containing humans for training (1716 images in the training set) and testing (1817 images in the validation set), the same as [5]. Note that parsing humans in PASCAL is challenging because it has larger variations in scale and pose than other human parsing datasets. In addtion, we also perform parsing experiments on the horse-cow dataset [29], which contains animal instances in a rough bounding box. In this dataset, we adopt the same experimental setting as in [30].

Training. We train the FCNs using stochastic gradient descent with mini-batches. Each mini-batch contains 30 images. The initial learning rate is 0.001 (0.01 for the final classifier layer) and is decreased by a factor of 0.1 after every 2000 iterations. We set the momentum to be 0.9 and the weight decay to be 0.0005. The initialization model is a modified VGG-16 network pre-trained on ImageNet. Fine-tuning our network on all the reported experiments takes about 30 h on a NVIDIA Tesla K40 GPU. After training, the average inference time for one PASCAL image is 1.3 s/image.

Evaluation metric. The object parsing results is evaluated in terms of mean IOU (mIOU). It is computed as the pixel intersection-over-union (IOU) averaged across classes [10], which is also adopted recently to evaluate parts [5,30]. In the supplementary material, we also evaluate the part parsing performance w.r.t. each object instance in terms of AP^r_{part} as defined in [13].

Network architecture. We use DeepLab-LargeFOV [4] as building blocks for the FCNs in our Hierarchical Auto-Zoom Net (HAZN).

4.3 Experimental Results on Parsing Humans in the Wild

Comparison with state-of-the-arts. As shown in Table 1, we compare our full model (HAZN) with four baselines. The first one is DeepLab-LargeFOV [4]. The second one is DeepLab-LargeFOV-CRF, which adds a post-processing step to DeepLab-LargeFOV by means of a fully-connected Conditional Random Field (CRF) [16]. CRFs are commonly used as postprocessing for object semantic segmentation to refine boundaries [4]. The third one is Multi-Scale Averaging, which feeds the DeepLab-LargeFOV model with images resized to three fixed scales (0.5, 1.0 and 1.5) and then takes the average of the three part score maps to produce the final parsing result. The fourth one is Multi-Scale Attention [5], a most recent work which uses a scale attention model to handle the scale variations in object parsing.

Our HAZN obtains the performance of 57.5 %, which is 5.8 % better than DeepLab-LargeFOV, and 4.5 % better than DeepLab-LargeFOV-CRF. Our model significantly improves the segmentation accuracy in all parts. Note we do not use any CRF for post processing. The CRF, though proven effective in refining boundaries in object segmentation, is not strong enough at recovering details of human parts as well as correcting the errors made by the DeepLab-LargeFOV.

The third baseline (Multi-Scale Averaging) enumerates multi-scale features which is commonly used to handle the scale variations, yet its performance is poorer than ours, indicating the effectiveness of our Auto-Zoom framework.

Our overall mIOU is 1.15 % better than the fourth baseline (Multi-Scale Attention), but we are much better in terms of detailed parts like upper legs (around 3 % improvement). In addition, we further analyze the scale-invariant ability in Table 2, which both methods aim to improve. We can see that our model surpasses Multi-Scale Attention in all instance sizes especially at size XS (9.5 %) and size S (5.5 %).

Importance of object and part scale. Table 1 also shows the effectiveness of the two scales in our HAZN. In practice, we remove either the object-scale AZN or the part-scale AZN from the full HAZN model, yielding two sub-models: (1) **HAZN (no object scale)**, which only handles the scale variation at part level; (2) **HAZN (no part scale)**, which only handles the scale variation at object instance level. Compared with our full model, removing the object-scale AZN causes 2.8 % mIOU degradation while removing the part-scale AZN results in 1 % mIOU degradation. We can see that the object-scale AZN, which handles the scale variation at object instance level, contributes a lot to our final parsing performance. The part-scale AZN further improves the parsing by refining the detailed part predictions, *e.g.* bringing around 3 % improvement on lower arms.

Table 1. Part parsing accuracy (%) on PASCAL-Person-Part in terms of mean IOU.

Method	head	torso	u-arms	l-arms	u-legs	l-legs	bg	Avg.
DeepLab-LargeFOV [4]	78.09	54.02	37.29	36.85	33.73	29.61	92.85	51.78
DeepLab-LargeFOV-CRF	80.13	55.56	36.43	38.72	35.50	30.82	93.52	52.95
Multi-Scale Averaging	79.89	57.40	40.57	41.14	37.66	34.31	93.43	54.91
Multi-Scale Attention [5]	**81.47**	59.06	44.15	42.50	38.28	35.62	93.65	56.39
HAZN (no object scale)	80.25	57.20	42.24	42.02	36.40	31.96	93.42	54.78
HAZN (no part scale)	79.83	59.72	43.84	40.84	40.49	37.23	93.55	56.50
HAZN (full model)	80.76	**60.50**	**45.65**	**43.11**	**41.21**	**37.74**	**93.78**	**57.54**

Part parsing accuracy w.r.t. size of human instance. Since we handle human with various sizes, it is important to check how our model performs w.r.t. the change of human size in images. We categorize all the ground truth human instances into four different sizes according to the bounding box area of each instance s_b (the square root of the bounding box area). Then we compute the mIOU within the bounding box for each of these four scales. The four sizes are defined as follows: (1) Size XS: $s_b \in [0, 80]$, where the human instance is extremely small in the image; (2) Size S: $s_b \in [80, 140]$; (3) Size M: $s_b \in [140, 220]$; (4) Size L: $s_b \in [220, 520]$, which usually corresponds to truncated human instances where the human's head or torso covers the majority of the image.

The results are given in Table 2. The baseline DeepLab-LargeFOV performs badly at size XS or S (usually only the head or the torso can be detected by the baseline), while our HAZN (full model) surpasses it significantly by 14.6 % for size XS and by 10.8 % for size S. This shows that HAZN is particularly good for small objects. For instances in size M and L, our model also significantly improve the baselines by around 5 %. In general, by using HAZN, we achieve much better scale invariant property to object size than a generally used FCN type of model. We also list the results for the other three baselines for reference. In addition, it is also important to jointly perform the two scale AZNs in a sequence. To show this, we additionally list the results from our model without object/part scale AZN in the 5_{th} and the 6_{th} row respectively. By jumping over object scale (HAZN no object scale), the performance becomes significantly worse at size XS, since the model can barely detect the object parts at the image-level when the object is too small. If we remove part scale instead (HAZN no part scale), the performance also dropped in all sizes. This is because using part-scale AZN can recover the part details much better than only using object scale.

Qualitative results. We qualitatively evaluate our model in Fig. 5. The baseline DeepLab-LargeFOV-CRF produces several errors due to lack of object and part scale information, *e.g.* background confusion (1_{st} row), human part confusion (3_{rd} row), important part missing (4_{th} row), *etc.* Our HAZN (no part scale), which only contains object-scale AZN, already successfully relieves the confu-

Table 2. Part parsing accuracy w.r.t. size of human instance (%) on PASCAL-Person-Part in terms of mean IOU.

Method	Size XS	Size S	Size M	Size L
DeepLab-LargeFOV [4]	32.5	44.5	50.7	50.9
DeepLab-LargeFOV-CRF	31.5	44.6	51.5	52.5
Multi-Scale Averaging	33.7	45.9	52.5	54.7
Multi-Scale Attention [5]	37.6	49.8	55.1	55.5
HAZN (no object scale)	38.2	51.0	55.1	53.4
HAZN (no part scale)	45.1	53.1	55.0	55.0
HAZN (full model)	**47.1**	**55.3**	**56.8**	**56.0**

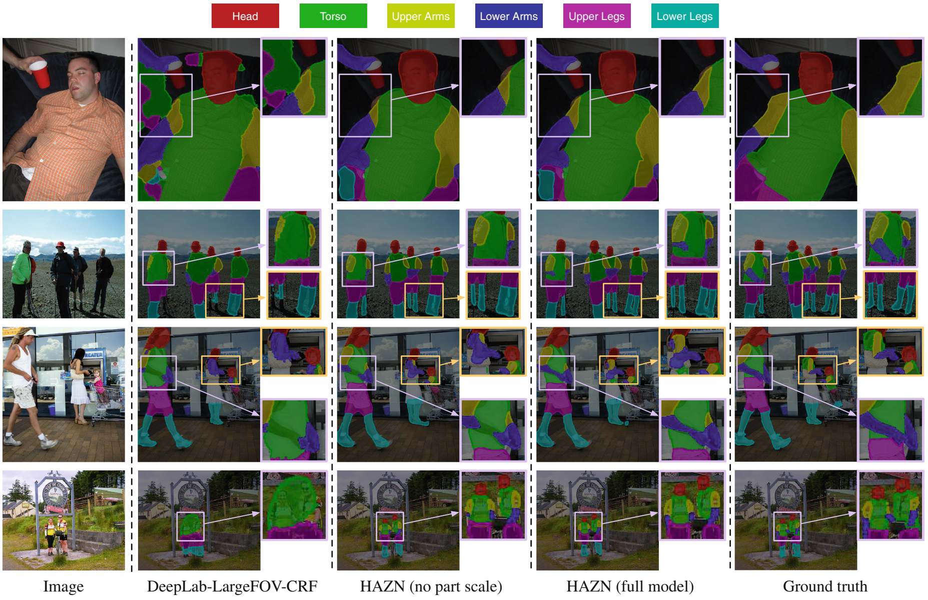


Fig. 5. Qualitative comparison on the PASCAL-Person-Part dataset. We compare with DeepLab-LargeFOV-CRF [4] and HAZN (no part scale). Our proposed HAZN models (the 3_{rd} and 4_{th} columns) attain better visual parsing results, especially for small scale human instances and small parts such as legs and arms.

sions for large scale human instances while recovers the parts for small scale human instances. By further introducing part scale, the part details and boundaries are recovered even more satisfactorily.

Failure cases. Figure 6 shows our typical failure modes. Compared with the baseline DeepLab-LargeFOV-CRF, our models give more reasonable parsing results with less local confusion, but they still suffer from heavy occlusion and unusual poses.

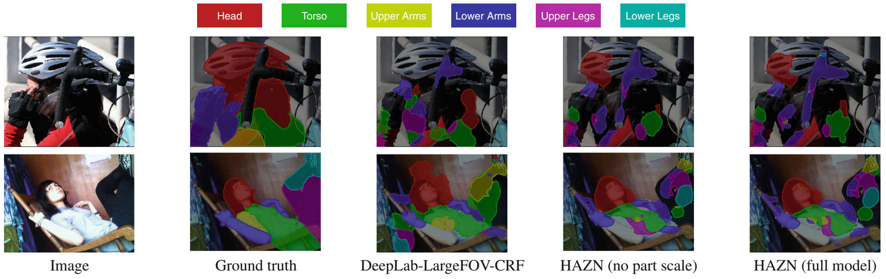


Fig. 6. Failure cases for both the baseline and our models.

Table 3. Mean IOU (mIOU) over the Horse-Cow dataset. We compare with the semantic part segmentation (SPS) [29], the Hypercolumn (HC*) [13] and the joint part and object (JPO) results [30]. We also list the performance of DeepLab-LargeFOV (LargeFOV) [4].

Horse							Cow					
Method	Bkg	head	body	leg	tail	Avg.	Bkg	head	body	leg	tail	Avg.
SPS [29]	79.14	47.64	69.74	38.85	-	-	78.00	40.55	61.65	36.32	-	-
HC* [13]	85.71	57.30	77.88	51.93	37.10	61.98	81.86	55.18	72.75	42.03	11.04	52.57
JPO [30]	87.34	60.02	77.52	58.35	**51.88**	67.02	85.68	58.04	76.04	51.12	15.00	57.18
LargeFOV	87.44	64.45	80.70	54.61	44.03	66.25	86.56	62.76	78.42	48.83	19.97	59.31
HAZN	**90.94**	**70.75**	**84.49**	**63.91**	51.73	**72.36**	**90.71**	**75.18**	**83.33**	**57.42**	**29.37**	**67.20**

4.4 Experiments on the Horse-Cow Dataset

Besides humans, we also applied our method to horses and cows presented in [29]. All the testing procedures are the same as those described above for humans. We copy the baseline numbers from [30], and give the evaluation results in Table 3. It shows that our baseline model, the DeepLab-LargeFOV [4], already achieves competative results with the state-of-the-arts, while our HAZN further provides a big improvement on both horses and cows. The improvement over the state-of-the-art method [30] is roughly 5 % mIOU. It is most noticeable for small parts, *e.g.* the improvement for detecting horse/cow head and cow tails is more than 10 %. This shows that our auto-zoom strategy can be effectively generalized to other objects for part parsing.

5 Conclusions

In this paper, we propose the "Hierarachical Auto-Zoom Net" (HAZN) to parse objects in the wild, yielding per-pixel segmentation of the object parts. It adaptably estimates the scales of objects, and their parts, by a two-stage process of Auto-Zoom Nets. We show that on the challenging PASCAL dataset, HAZN performs significantly better (by 5 % mIOU) than other state-of-the-art methods, when applied to humans, horses, and cows.

In the future, we would love to extend our HAZN to parse more detailed parts, such as human hand and human eyes. Also, the idea of our AZN can be applied to other tasks like pose estimation in the wild, to make further progress.

Acknowledgements. We would like to gratefully acknowledge support from NSF award CCF-1317376, and NSF STC award CCF-1231216. We also thank NVIDIA for providing us with free GPUs that are used to train deep models. Additionally, many thanks to Lingxi Xie, Zhou Ren, and Xianjie Chen for proofreading this paper and giving suggestions.

References

1. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. PAMI **34**(11), 2189–2202 (2012)
2. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. PAMI **33**(5), 898–916 (2011)
3. Bo, Y., Fowlkes, C.C.: Shape-based pedestrian parsing. In: CVPR (2011)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: ICLR (2015)
5. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: Scale-aware semantic image segmentation. arXiv:1511.03339 (2015)
6. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.L.: Detect what you can: Detecting and representing objects using holistic models and body parts. In: CVPR (2014)
7. Dai, J., He, K., Sun, J.: Boxsup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: ICCV (2015)
8. Dong, J., Chen, Q., Shen, X., Yang, J., Yan, S.: Towards unified human parsing and pose estimation. In: CVPR (2014)
9. Eslami, S.M.A., Williams, C.K.I.: A generative model for parts-based object segmentation. In: NIPS (2012)
10. Everingham, M., Eslami, S.A., Gool, L.V., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. IJCV **111**(1), 98–136 (2014)
11. Florack, L., Romeny, B.T.H., Viergever, M., Koenderink, J.: The gaussian scale-space paradigm and the multiscale local jet. IJCV **18**(1), 61–75 (1996)
12. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)
13. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: CVPR (2015)
14. Hoiem, D., Efros, A.A., Hebert, M.: Putting objects in perspective. IJCV **80**(1), 3–15 (2008)
15. Huang, L., Yang, Y., Deng, Y., Yu, Y.: Densebox: unifying landmark localization with end to end object detection. arXiv:1509.04874 (2015)
16. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFs with gaussian edge potentials. In: NIPS (2011)
17. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
18. Li, Y., Hou, X., Koch, C., Rehg, J., Yuille, A.: The secrets of salient object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 280–287 (2014)
19. Liang, X., Wei, Y., Shen, X., Yang, J., Lin, L., Yan, S.: Proposal-free network for instance-level object segmentation. CoRR abs/1509.02636 (2015)
20. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 740–755. Springer, Heidelberg (2014)
21. Liu, S., Liang, X., Liu, L., Shen, X., Yang, J., Xu, C., Lin, L., Cao, X., Yan, S.: Matching-CNN meets KNN: quasi-parametric human parsing. In: CVPR (2015)

22. Liu, Z., Li, X., Luo, P., Loy, C.C., Tang, X.: Semantic image segmentation via deep parsing network. In: ICCV (2015)
23. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
24. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. arXiv:1505.04366 (2015)
25. Papandreou, G., Chen, L.C., Murphy, K., Yuille, A.L.: Weakly- and semi-supervised learning of a dcnn for semantic image segmentation. In: ICCV (2015)
26. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. CoRR abs/1506.02640 (2015)
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. arXiv:1506.01497 (2015)
28. Tsogkas, S., Kokkinos, I., Papandreou, G., Vedaldi, A.: Semantic part segmentation with deep learning. arXiv:1505.02438 (2015)
29. Wang, J., Yuille, A.: Semantic part segmentation using compositional model combining shape and appearance. In: CVPR (2015)
30. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.: Joint object and part segmentation using deep learned potentials. In: ICCV (2015)
31. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.L.: Towards unified depth and semantic prediction from a single image. In: CVPR (2015)
32. Wang, P., Wang, J., Zeng, G., Feng, J., Zha, H., Li, S.: Salient object detection for searched web images via global saliency. In: CVPR, pp. 3194–3201 (2012)
33. Xia, F., Zhu, J., Wang, P., Yuille, A.L.: Pose-guided human parsing with deep learned features. AAAI abs/1508.03881 (2016)
34. Yamaguchi, K., Kiapour, M.H., Ortiz, L.E., Berg, T.L.: Parsing clothing in fashion photographs. In: CVPR (2012)
35. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based R-CNNs for fine-grained category detection. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part I. LNCS, vol. 8689, pp. 834–849. Springer, Heidelberg (2014)
36. Zhu, L.L., Chen, Y., Lin, C., Yuille, A.: Max margin learning of hierarchical configural deformable templates (hcdts) for efficient object parsing and pose estimation. IJCV **93**(1), 1–21 (2011)
37. Zhu, Y., Urtasun, R., Salakhutdinov, R., Fidler, S.: segDeepM: exploiting segmentation and context in deep neural networks for object detection. In: CVPR (2015)