Improving Panoptic Segmentation at All Scales

Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder
Facebook
{porzi,rotabulo,pkontschieder}@fb.com

Figure 1: Panoptic segmentation on high-resolution natural images is challenged with recognizing objects at a wide range of scales. Standard approaches (left) can struggle when dealing with very small (zoomed detail) or very large objects (bus on the left). By introducing a novel instance scale-uniform sampling strategy and a crop-aware bounding box loss, we are able to improve panoptic segmentation results at all scales (right).

Abstract

*Crop-based training strategies decouple training resolution from GPU memory consumption, allowing the use of large-capacity panoptic segmentation networks on multi-megapixel images. Using crops, however, can introduce a bias towards truncating or missing large objects. To address this, we propose a novel crop-aware bounding box regression loss (*CABB *loss), which promotes predictions to be consistent with the visible parts of the cropped objects, while not over-penalizing them for extending outside of the crop. We further introduce a novel data sampling and augmentation strategy which improves generalization across scales by counteracting the imbalanced distribution of object sizes. Combining these two contributions with a carefully designed, top-down panoptic segmentation architecture, we obtain new state-of-the-art results on the challenging Mapillary Vistas (MVD), Indian Driving and Cityscapes datasets, surpassing the previously best approach on MVD by +4.5% PQ and +5.2% mAP.*

1. Introduction

Panoptic segmentation [16] is the task of generating per-pixel, semantic labels for an image, together with object-specific segmentation masks. It is thus a combination of semantic segmentation and instance segmentation, *i.e.* two long-standing tasks in computer vision that have been traditionally tackled separately. Due to its importance for tasks like autonomous driving or scene understanding it has recently attracted a lot of interest in the research community.

The majority of deep-learning based panoptic segmentation architectures [15, 23, 17, 29, 21] proposed a combination of specialized segmentation branches – one for conventional semantic segmentation and another one for instance segmentation – followed by a combination strategy to generate a final panoptic segmentation result. Instance segmentation branches in top-down panoptic architectures are dominantly designed on top of Mask R-CNN [12], *i.e.* a segmentation extension of Faster R-CNN [24] generating state-of-the-art mask predictions for given bounding boxes. In contrast and more recently, bottom-up panoptic architectures [6, 26] have emerged but still lag behind in terms of instance segmentation performance.

Panoptic segmentation networks are typically solving multiple tasks (object detection, instance segmentation and semantic segmentation), and are trained on batches of full-sized images. However, with increasing complexity of tasks and growing capacity of the network backbone, full-image training is quickly inhibited by available GPU memory, despite availability of memory-saving strategies during training like [25, 20, 11, 14]. Obvious mitigation strategies include a reduction of training batch size, downsizing of high-resolution training images, or building on backbones with lower capacity. These workarounds unfortunately introduce other limitations: i) Small batch sizes can lead to higher variance in the gradients which will reduce the effectiveness of Batch Normalization [13] and consequently the performance of the resulting model. ii) Reducing the image resolution leads to a loss of fine structures which are known to strongly correlate with objects belonging to the

long tail of the label distribution. Downsampling the images is consequently amplifying already existing performance issues on small and usually underrepresented classes. iii) A number of recent works [28, 5, 31] have shown that larger backbones with sophisticated strategies of maintaining high-resolution features are boosting panoptic segmentation results in comparison to those with reduced capacity.

A possible strategy to overcome the aforementioned issues is to move from full-image-based training to crop-based training. This was successfully used for conventional semantic segmentation [25, 3, 2], which is however an easier problem as the task is limited to a per-pixel classification problem. By fixing a certain crop size the details of fine structures can be preserved, and at a given memory budget, multiple crops can be stacked to form reasonably sized training batches. For more complex tasks like panoptic segmentation, the simple cropping strategy also affects the performance on object detection and consequently on instance segmentation. In particular, extracting fixed-size crops from images during training introduces a bias towards truncating large objects, with the likely consequence of underestimating their actual bounding box sizes during inference on full images (see, *e.g*. Fig. 1 left). Indeed, Fig. 2 (left) shows that the distribution of box sizes during crop-based training on the high-resolution Mapillary Vistas [22] dataset does not match with the one derived from full-image training data. In addition, Fig. 2 (right) shows that large objects (based on # pixels) are drastically underrepresented, which may lead to over-fitting and thus further harming generalization.

In this paper we overcome these issue by introducing two novel contributions: 1) A crop-based training strategy exploiting a *crop-aware loss* function (CABB) to address the problem of cropping large objects, and 2) Instance scale-uniform (ISUS) sampling as data augmentation strategy to combat the imbalance of object scales in the training data. Our solution enjoys all benefits from crop-based training as discussed above. In addition, our crop-aware loss incentivizes the model to predict bounding boxes to be consistent with the visible parts of cropped objects, while not over-penalizing predictions outside of the crop. The underlying intuition is simple: Even if an object bounding box size was modified through cropping, the actual object bounding boxes may be larger than what is visible to the network during training. By not penalizing hypothetical predictions beyond the visible area of a crop but still within their actual sizes, we can better model the bounding box size distribution given by the original training data. With ISUS we introduce an effective data augmentation strategy to improve feature-pyramid like representations as used for object detection at multiple scales. It aims at more evenly distributing supervision of object instances during training across pyramid scales, leading to improved recognition performance of instances at all scales during inference. In the experimental analyses we find that our crop-aware loss function is particularly effective on high-resolution images as available in the challenging Mapillary Vistas [22], Indian Driving [27], or Cityscapes [8] datasets.

Contributions. We summarize our contributions to the panoptic segmentation research community as follows.

- We introduce a novel, crop-aware training loss applicable to improving bounding box detection in panoptic segmentation networks when training them in a crop-based way. At negligible computational overhead ($\sim$10ms per batch) we show how our new loss addresses issues of crop-based training, considerably improving the performance on disproportionately often truncated bounding boxes.

- We describe a novel Instance Scale-Uniform Sampling approach to smooth the distribution of object sizes observed by a network at training time, improving its generalization across scales.

- We significantly push the state-of-the-art results on the high-resolution Mapillary Vistas dataset, improving on multiple evaluation metrics like Panoptic Quality [16] (+4.5%) and mean average precision (mAP) for mask segmentation (+5.2%). We also obtain remarkable performance gains on IDD and Cityscapes, improving PQ by +0.6% and mAP by +4.1% and +1.5%, respectively.

2. Technical Contributions

In this section we present our main methodological contributions. In particular, in Sec. 2.1 we describe a novel Instance-Scale Uniform Sampling (ISUS) approach aimed at reducing the object scale imbalance inherent in high-resolution panoptic datasets. Sections 2.2, 2.3 and 2.4 describe the Crop-Aware Bounding Box (CABB) loss, which we propose as a mitigation to the bias imposed by crop-based training on the detection of large objects.

2.1. Instance Scale-Uniform Sampling (ISUS)

Most top-down panoptic segmentation networks build on top of backbones that produce a "pyramid" of features at multiple scales. At training time, some heuristic rule [15] is applied to split the ground truth instances across the available scales, such that the network is trained to detect small objects using high-resolution features and large objects using low-resolution features. By sharing the parameters of the prediction modules (*e.g*. the RPN and ROI heads of [23]) across all scales, the network is incentivized to learn scale-invariant features. When dealing with high-resolution images, however, this approach encounters two major issues: i) the range of object scales can greatly exceed the range of scales available in the feature pyramid, and ii) the

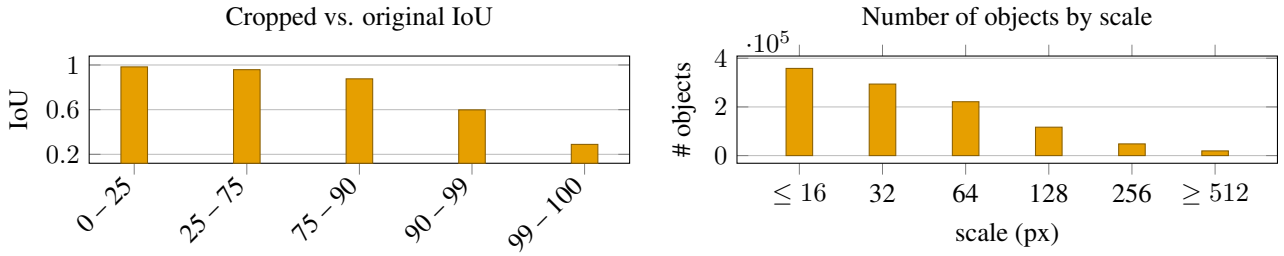

Figure 2: Left: average intersection over union of cropped bounding boxes w.r.t. their original extent, computed using the Mapillary Vistas training settings in Sec. 4.1. Right: distribution of object scales in the Mapillary Vistas training set.

distribution of object scales is markedly non-uniform (see Fig. 2). While (i) can be partially addressed by adding more feature scales, at the cost of increased memory and computation, (ii) will lead to a strong imbalance in the amount of supervision received by each level of the feature pyramid.

In order to mitigate this imbalance, we propose an extension to the Class-Uniform Sampling (CUS) approach introduced in [25] we coin Instance Scale-Uniform Sampling (ISUS). The standard CUS data preparation process follows four steps: 1) sample a semantic class with uniform probability; 2) load an image that contains that class and re-scale it such that its shortest side matches a predefined size s_0; 3) apply any data augmentation (*e.g.* flipping, random scaling); and 4) produce a random crop from an area of the image where the selected class is visible. In ISUS, we follow the same steps as in CUS, except that the scale augmentation procedure is made instance-aware. In particular, when a "thing" class is selected in step 1 and after completing step 2, we also sample a random instance of that class from the image and a random feature pyramid level. Then, in step 3 we compute a scaling factor σ such that the selected instance will be assigned to the selected level according to the heuristic adopted by the network being trained. In order to avoid excessively large or small scale factors, we clamp σ to a limited range r_{th}. Conversely, when a "stuff" class is selected in step 1, we follow the standard scale augmentation procedure, *i.e.* uniformly sample σ from a range r_{st}. In the long run, ISUS will have the effect of smoothing out the object scale distribution, providing more uniform supervision across all scales.

2.2. Bounding box regression

Most top-down panoptic segmentation approaches encode object bounding boxes in terms of offsets with respect to a set of reference boxes [17, 23, 21]. These reference boxes can be fixed, e.g. the "anchors" in the region proposal stage, or be the output of a different network section, e.g. the "proposals" in the detection stage. The goal of a network component that predicts bounding boxes is to regress these offset values given the input image (or derived features thereof).

A ground-truth bounding box $\mathtt{G}$ is encoded in terms of a center $\boldsymbol{c}_\mathtt{G} \in \mathbb{R}^2$ and dimensions $\boldsymbol{d}_\mathtt{G} \in \mathbb{R}^2$. Each ground-truth box is assigned a reference (or anchor) bounding box $\mathtt{A}$ with center $\boldsymbol{c}_\mathtt{A} \in \mathbb{R}^2$ and dimensions $\boldsymbol{d}_\mathtt{A} \in \mathbb{R}^2$. The ground truth for the training procedure is then encoded in relative terms and specifically given by $\Delta_\mathtt{G} = (\boldsymbol{\delta}_\mathtt{G}, \boldsymbol{\omega}_\mathtt{G})$ where

$$\boldsymbol{\delta}_\mathtt{G} = \frac{\boldsymbol{c}_\mathtt{G} - \boldsymbol{c}_\mathtt{A}}{\boldsymbol{d}_\mathtt{A}} \in \mathbb{R}^2 \qquad \text{and} \qquad \boldsymbol{\omega}_\mathtt{G} = \frac{\boldsymbol{d}_\mathtt{G}}{\boldsymbol{d}_\mathtt{A}} \in \mathbb{R}^2 .$$

Here and later, we implicitly assume for notational convenience that operations and functions applied to vectors work element-wise unless otherwise stated. We will also use the notation $\ominus$ to denote the operation above that returns $\Delta_\mathtt{G}$ given bounding boxes $\mathtt{G}$ and $\mathtt{A}$, *i.e.* $\Delta_\mathtt{G} = \mathtt{G} \ominus \mathtt{A}$.

Similarly, given an anchor bounding box $\mathtt{A}$ and $\Delta_\mathtt{P} = (\boldsymbol{\delta}_\mathtt{P}, \boldsymbol{\omega}_\mathtt{P})$, we can recover the predicted bounding box $\mathtt{P}$ with center $\boldsymbol{c}_\mathtt{P}$ and dimensions $\boldsymbol{d}_\mathtt{P}$ as

$$\boldsymbol{c}_\mathtt{P} = \boldsymbol{c}_\mathtt{A} + \boldsymbol{\delta}_\mathtt{P} \boldsymbol{d}_\mathtt{A} \qquad \text{and} \qquad \boldsymbol{d}_\mathtt{P} = \boldsymbol{\omega}_\mathtt{P} \boldsymbol{d}_\mathtt{A} .$$

Standard bounding box loss [24]. To train the network, the following per-box loss is minimized over the training dataset:

$$L_{\mathtt{BB}}(\Delta_\mathtt{P}; \Delta_\mathtt{G}) = \|\ell_\beta(\boldsymbol{\delta}_\mathtt{P} - \boldsymbol{\delta}_\mathtt{G}) + \ell_\beta(\log \boldsymbol{\omega}_\mathtt{P} - \log \boldsymbol{\omega}_\mathtt{G})\|_1 , \quad (1)$$

where $\|\cdot\|_1$ is the 1-norm and ℓ_β denotes the Huber (*a.k.a.* smooth-L1) norm with parameter $\beta > 0$, *i.e.*

$$\ell_\beta(z) = \begin{cases} \frac{1}{2\beta} z^2 & |z| \leq \beta \\ |z| - \frac{\beta}{2} & \text{otherwise,} \end{cases}$$

and $|z|$ gives the absolute value of z.

2.3. Crop-Aware Bounding Box (CABB)

In a standard crop-based training, a ground-truth bounding box $\mathtt{G}$ from the original image that overlaps with the cropping area $\mathtt{C}$ is typically cropped yielding a new bounding box denoted by $\mathtt{G}|_\mathtt{C}$. [1] Accordingly, the actual ground-truth $\Delta_\mathtt{G}$ that is used in the loss (1) is the result of $\Delta_\mathtt{G} =$

[1] When masks are available like in instance or panoptic segmentation, the cropping operation is performed at the mask level and the bounding box is recomputed a posteriori. We implicitly assume that this is the case if a ground-truth mask is available for $\mathtt{G}$.

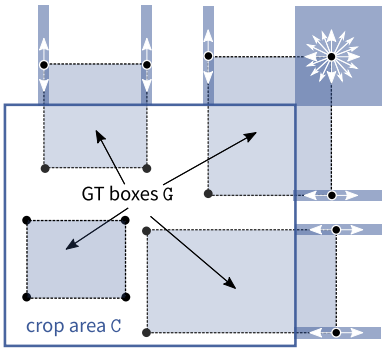

Figure 3: Example of Crop-Aware Bounding Boxes (CABB). We show 4 ground-truth boxes, three of which fall partially outside the crop area. The corresponding set $\rho(\mathtt{G},\mathtt{C})$, *a.k.a.* CABB, consists of all rectangular bounding boxes that can be formed by moving the white-bordered corners within the feasible areas (depicted in blue). Note that the areas extend to infinity but are truncated here.

$\mathtt{G}|_{\mathtt{C}} \ominus \mathtt{A}$. Training with this modified ground-truth, however, poses some issues, namely a bias towards cutting or missing big objects at inference time (see, *e.g.*, Fig. 1 and 6).

The solution we propose in this work consists in relaxing the notion of ground-truth bounding box $\mathtt{G}$ into a set of ground-truth boxes that coincide with $\mathtt{G}|_{\mathtt{C}}$ after the cropping operation. We denote by $\rho(\mathtt{G},\mathtt{C})$ the function that computes this set for given ground-truth box $\mathtt{G}$ and cropping area $\mathtt{C}$, *i.e.*

$$\rho(\mathtt{G},\mathtt{C}) = \{\mathtt{X} \in \mathcal{B} \,:\, \mathtt{X}|_{\mathtt{C}} = \mathtt{G}|_{\mathtt{C}}\}\,,$$

where $\mathtt{X}$ runs over all possible bounding boxes $\mathcal{B}$. We refer to $\rho(\mathtt{G},\mathtt{C})$ as a Crop-Aware Bounding Box (CABB) that in fact is a set of bounding boxes (see also Fig. 3). If the ground-truth bounding box $\mathtt{G}$ is strictly contained in the crop area then our CABB boils down to the original ground truth, for $\rho(\mathtt{G},\mathtt{C}) = \{\mathtt{G}\}$ in that case.[2] Since we will use a representation for bounding boxes relative to some anchor box $\mathtt{A}$ we introduce also the notation $\rho_{\mathtt{A}}(\mathtt{G},\mathtt{C})$, which returns the same set as above but with elements expressed relative to $\mathtt{A}$, *i.e.* $\rho_{\mathtt{A}}(\mathtt{G},\mathtt{C}) = \{\mathtt{X} \ominus \mathtt{A} \,:\, \mathtt{X} \in \rho(\mathtt{G},\mathtt{C})\}$.

Crop-aware bounding box loss. In order to exploit the proposed, relaxed notion of ground-truth bounding box, we introduce the following new loss function for a given ground-truth box $\mathtt{G}$, anchor box $\mathtt{A}$ and crop area $\mathtt{C}$:

$$\begin{aligned} L_{\mathtt{CABB}}(\Delta_{\mathtt{P}}) = &\min_{\Delta} L_{\mathtt{BB}}(\Delta_{\mathtt{P}};\Delta)\,, \\ &\text{s.t} \quad \Delta \in \rho_{\mathtt{A}}(\mathtt{G},\mathtt{C})\,. \end{aligned} \tag{2}$$

Any bounding box in $\rho(\mathtt{G},\mathtt{C})$ is compatible with the cropped ground-truth box we observe and thus could be potentially a valid prediction. To disambiguate, our new loss favours the solution closer to the actual prediction from the network in order to enforce a smoother training dynamic. Since the ground-truth box that is typically adopted for the standard loss in (1) belongs to the feasible set of the minimization in our new loss, we have that $L_{\mathtt{CABB}}$ lower bounds $L_{\mathtt{BB}}$.

[2]To simplify the description, we deliberately neglect the fact that a bounding box strictly contained in the original image and touching the boundary of the crop area should not be extended beyond the crop. However, our approach can be easily adapted to address these edge cases.

2.4. Computational Aspects

This section focuses on the computational aspects of our new loss. In particular, we will address the problem of evaluating it by solving the internal minimization as well as computing the gradient.

The minimization problem that is nested into our new loss has no straightforward solution, since it is neither convex nor quasi-convex and in general, local, non-global solutions might exist. Its feasible set is convex in $\Delta = (\boldsymbol{\delta}, \boldsymbol{\omega})$ since it can be written in terms of linear equalities and inequalities. Each dimension gives rise to an independent set of constraints and since also the objective function is separable with respect to dimension-specific variables, we have that the whole minimization problem can be separated into two independent minimization problems involving only dimension-specific variables.

Feasible set. Assume without loss of generality that the cropping area $\mathtt{C}$ is a box with top-left coordinate $(0,0)$ and bottom-right coordinate $\boldsymbol{d}_{\mathtt{C}} \in \mathbb{R}^2$. Then the feasible set of each dimension-specific minimization problem can be written as:

- $\delta - \frac{\omega}{2} \leq -\frac{c_{\mathtt{A}}}{d_{\mathtt{A}}}$ if $c_{\mathtt{G}} \leq \frac{d_{\mathtt{G}}}{2}$ else $\delta - \frac{\omega}{2} = \delta_{\mathtt{G}} - \frac{\omega_{\mathtt{G}}}{2}$ and
- $\delta + \frac{\omega}{2} \geq \frac{d_{\mathtt{C}} - c_{\mathtt{A}}}{d_{\mathtt{A}}}$ if $c_{\mathtt{G}} \geq d_{\mathtt{C}} - \frac{d_{\mathtt{G}}}{2}$ else $\delta + \frac{\omega}{2} = \delta_{\mathtt{G}} + \frac{\omega_{\mathtt{G}}}{2}$,

where we dropped the boldface style from the vector-valued variables to emphasize that the constraint is specified for a single dimension.

Optimization problem. We will now enumerate the different cases characterizing the feasible set and for each of them we will provide the dimension-specific optimization problem that should be solved. Akin to the feasible set above, all variables involved from here on refer implicitly to a single dimension.

- If $\frac{d_{\mathtt{G}}}{2} < c_{\mathtt{G}} < d_{\mathtt{C}} - \frac{d_{\mathtt{G}}}{2}$ then $\Delta^\star = (\delta_{\mathtt{G}}, \omega_{\mathtt{G}})$ is the solution to the minimization problem in (2) for the dimension under consideration, since the feasible set is singleton in this case.
- If $c_{\mathtt{G}} > \frac{d_{\mathtt{G}}}{2}$ and $c_{\mathtt{G}} \geq d_{\mathtt{C}} - \frac{d_{\mathtt{G}}}{2}$, we obtain an optimization problem in the variable ω of the form

$$\begin{aligned} \min_{\omega} \quad & \ell_\beta\left(\frac{\omega - \hat{\omega}}{2}\right) + \ell_\beta(\log(\omega) - \log(\omega_{\mathtt{P}})) \\ \text{s.t.} \quad & \omega \geq b_1 - a_1\,, \end{aligned} \tag{O_1}$$

where $a_1 = \delta_{\mathtt{G}} - \frac{\omega_{\mathtt{G}}}{2}$, $b_1 = \frac{d_{\mathtt{C}} - c_{\mathtt{A}}}{d_{\mathtt{A}}}$ and $\hat{\omega} = 2(\delta_{\mathtt{P}} - a_1)$. If $w^\star$ is a solution to (O_1) then $\Delta^\star = (a_1 + \frac{\omega^\star}{2}, \omega^\star)$ is a solution to the minimization problem in (2) for the dimension under consideration.

- If $c_{\mathtt{G}} \leq \frac{d_{\mathtt{G}}}{2}$ and $c_{\mathtt{G}} < d_{\mathtt{C}} - \frac{d_{\mathtt{G}}}{2}$, we obtain an optimization problem like (O_1) but with $a_1 = -\frac{c_{\mathtt{A}}}{d_{\mathtt{A}}}$, $b_1 = \delta_{\mathtt{G}} + \frac{\omega_{\mathtt{G}}}{2}$ and $\hat{\omega} = 2(b_1 - \delta_{\mathtt{P}})$. If $w^\star$ is a solution to (O_1) under this parametrization then $\Delta^\star = (b_1 - \frac{\omega^\star}{2}, \omega^\star)$ is a solution to the minimization problem in (2) for the dimension under consideration.

- If $d_{\mathtt{C}} - \frac{d_{\mathtt{G}}}{2} \leq c_{\mathtt{G}} \leq \frac{d_{\mathtt{G}}}{2}$ then we obtain an optimization problem of the form

$$\begin{aligned} \min_{\delta,\omega} \quad & \ell_\beta(\delta - \delta_{\mathtt{P}}) + \ell_\beta(\log(\omega) - \log(\omega_{\mathtt{P}})) \\ \text{s.t.} \quad & \delta - \frac{\omega}{2} \leq a_2\,, \quad \delta + \frac{\omega}{2} \geq b_2\,, \end{aligned} \tag{O_2}$$

where $a_2 = -\frac{c_{\mathtt{A}}}{d_{\mathtt{A}}}$ and $b_2 = \frac{d_{\mathtt{C}} - c_{\mathtt{A}}}{d_{\mathtt{A}}}$. Solutions to ($O_2$) map directly to solutions to (2) for the dimension under consideration.

We focus now on finding the solution to the optimization problems (O_1) and (O_2).

Solution to (O_1). As mentioned before, the optimization problem in (2) is in general non-convex and might have multiple local minima. The same holds true for the problem in (O_1) despite having a single variable. Nonetheless, we devised an ad-hoc solver for this problem that allows to quickly converge to a global solution under the desired precision. We skip the details due to lack of space, but we provide them in the supplementary material (see Alg. 1).

Solution to (O_2). To solve this problem we break it down into cases. We start by noting that the solution to the unconstrained optimization problem is trivially given by $\delta^\star = \delta_{\mathtt{P}}$ and $\omega^\star = \omega_{\mathtt{P}}$, because 0 is the minimizer of ℓ_β. The solution $\Delta^\star = (\delta^\star, \omega^\star)$ is valid for (O_2) if it satisfies the constraints, but this is easy to check by substitution. If this is the case, we found the solution, otherwise no solution exists in the interior of the feasible set (see Prop. 1 in supplementary material), but lies along the boundary of the feasible set. Accordingly, we start by forcing the first constraint to be active. This yields an instance of (O_1) with $a_1 = a_2$, $b_1 = b_2$ and $\hat{\omega} = 2(\delta_{\mathtt{P}} - a_2)$, which can be solved using the algorithm from the supplementary material, yielding $\omega_1^\star$. By substituting it into the activated constraint we obtain the other variable $\delta_1^\star = a_2 + \frac{\omega_1^\star}{2}$. Next, we move to activating the second constraint. This yields again an instance of the same optimization problem with the only difference being $\hat{\omega} = 2(b_2 - \delta_{\mathtt{P}})$. Again we solve it obtaining $\omega_2^\star$ and by substitution into the activated constraint we get $\delta_2^\star = b_2 - \frac{\omega_2^\star}{2}$. We finally retain the solution among $(\delta_1^\star, \omega_1^\star)$ and $(\delta_2^\star, \omega_2^\star)$ yielding the lowest objective. See Alg. 2 in supplementary material for further details.

Gradient. For the sake of training a neural network, we are interested in computing gradients of the new loss function, which exhibits a nested optimization problem. The following result shows that the derivative of the new loss function is equivalent to the derivative of the original one, with the ground-truth box replaced (as a constant) by the solution to the internal minimization problem. In general the solution to the internal minimization problem is a function of $\Delta_{\mathtt{P}}$ but the following result states that no gradient term is originated from this dependency. This is indeed a direct consequence of the envelope theorem [1].

Proposition 1. *Let ϕ be a function returning the minimizer in (2) given $\Delta_{\mathtt{P}}$,* i.e. *$L_{\mathtt{CABB}}(\Delta_{\mathtt{P}}) = L_{\mathtt{BB}}(\Delta_{\mathtt{P}}, \phi(\Delta_{\mathtt{P}}))$ holds for any $\Delta_{\mathtt{P}}$. Then*

$$\frac{d}{d\Delta_{\mathtt{P}}} L_{\mathtt{CABB}}(\Delta_{\mathtt{P}}) = \frac{\partial}{\partial \Delta_{\mathtt{P}}} L_{\mathtt{BB}}(\Delta_{\mathtt{P}}, \Delta)\Big|_{\Delta = \phi(\Delta_{\mathtt{P}})}\,.$$

3. Related Works

After scrutinizing the literature, we have found no other work directly addressing the specific challenges of training panoptic segmentation networks on high-resolution data, nor the bias introduced by crop-based training. Indeed, to our knowledge, we are tackling these issues for the first time. In the literature we find several methods for panoptic segmentation that are architecture-wise compatible with our CABB loss and ISUS, among which we have EfficientPS [21], AUNet [18], TASCNet [17], Panoptic-FPN [15], UPSNet [29] and Seamless Scene Segmentation [23], to mention a few. Indeed, those approaches rely on the computation of bounding boxes at some stage, and employ network backbones that produce multi-scale feature pyramids. Among them, only the first two report crop-based training results in the original work, while the remaining ones report full-image training results. This however does not mean that the latter approaches would not benefit from crop-based trainings. Indeed, in this work, we perform experiments using Seamless Scene Segmentation as baseline and show that there is significant improvements deriving from a crop-based training protocol. Other panoptic segmentation methods that benefit from crop-based training are AdaptIS [26], DeeperLab [30] SSAP [10] and Panoptic-Deeplab [6]. The latter approaches however are neither based on bounding boxes nor employ feature pyramids, thus our contributions do not directly apply to them. More broadly, recent works dealing with high-resolution image data include RefineNet [19] or CascadePSP [7], which however address the task of conventional semantic segmentation rather than Panoptic segmentation.

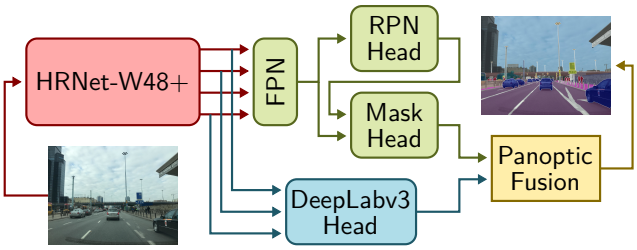

Figure 4: Overview of the main functional blocks of our network. Red: network body, *i.e*. HRNet-W48+. Green: instance segmentation section, composed of an FPN module followed by a Region Proposal Head (RPH) and a mask segmentation head. Blue: semantic segmentation section, *i.e*. DeepLabv3 head. Yellow: final panoptic fusion step.

4. Experimental Results

We evaluate our proposed CABB loss on the three largest publicly available, high-resolution panoptic segmentation datasets: Mapillary Vistas [22] (MVD), the Indian Driving Dataset [27] (IDD) and Cityscapes [9] (CS). MVD comprises 18k training, 2k validation, and 5k testing images, with resolutions ranging from 2 to 22 Mpixels and averaging 8.8 Mpixels, and annotations covering 65 semantic classes, 37 of which instance-specific. IDD comprises 7k training, 1k validation, and 2k testing images, most captured at a 2 Mpixels resolution and annotated with 26 semantic classes, 9 of them instance-specific. Cityscapes comprises 3k training, 500 validation, and 1.5k testing images, captured at 2 Mpixels resolutions and annotated with 19 classes, 8 of which instance-specific. Next, we present detailed ablation studies and a comparison with recent state-of-the-art panoptic segmentation approaches.

4.1. Network and Training Details

Our CABB loss and ISUS, described in Sec. 2.3, can be used in most top-down panoptic segmentation networks. To evaluate their effects, however, we focus our attention on a specific architecture, carefully crafted to achieve state-of-the-art performance on high-resolution datasets already *without* using either. In particular, we follow the general framework of Seamless-Scene-Segmentation [23], with several modifications described below (see Fig. 4). First, we replace the ResNet-50 "body" with HRNetV2-W48+ [28, 6], a specialized backbone which preserves high-resolution information from the image to the final stages of the network. Second, we replace the "Mini-DL" segmentation head from [23] with a DeepLabV3+ [4] module, connected to the HRNetV2-W48+ body as described in [6]. As in [23], we apply synchronized InPlace-ABN [25] throughout the network. Finally, CABB loss is used to replace the standard bounding box regression loss both in the region proposal and object detection modules.

We train our networks with stochastic gradient descent on 8 NVidia V100 GPUs with 32GB of memory. The HRNetV2-W48+ backbone is initialized from an ImageNet pre-training in the MVD and IDD experiments, while the Cityscapes networks are fine-tuned from their MVD-trained counterparts. We fix the crop size to 1024×1024 for MVD, and to 512×512 for IDD and Cityscapes due to their lower resolution, while inference is always performed on full images. Average inference time on MVD is ~ 1.2s per image. To reduce inter-run variability and obtain more comparable results, we fix all sources of randomness that can be easily controlled, resulting in the same sequence of images and initial network weights across all our trainings. For a detailed breakdown of the training hyper-parameters refer to Sec. D of the supplementary material.

4.2. Comparison with State of the Art

We provide a comparison of results in Table 1, with baselines including methods trained on full images (TASCNet [17], Seamless [23]) and crops (AdaptIS [26], EfficientPS [21], Panoptic Deeplab [6]), as well as multiple different backbones (EfficientNet in EffcientPS, ResNet-50 in Seamless and TASCNet, ResNeXt-101 in AdaptIS, Xception-71 and HRNet-W48+ in Panoptic Deeplab). We consider several different variants of our network: (i) one using the standard bounding box regression loss and CUS, trained either on full images (FULL) or crops (CROP); (ii) one using our CABB loss and CUS, trained on crops (CROP + CABB); (iii) one using the standard bounding box regression loss and ISUS, trained on crops (CROP + ISUS); and finally (iv) one using both our CABB loss and ISUS, trained on crops (CROP + CABB + ISUS).

The MVD results on top in Table 1 show that CROP outperforms FULL on all metrics, attesting to the advantages of crop-based training. Both our CABB loss and ISUS separately lead to consistent improvements w.r.t. CROP on all aggregate and pure recognition metrics. The effects of CABB and ISUS will be explored in more detail in Sec. 4.3. We also see that even the weakest among our network variants surpasses all PQ baselines, the only exception being the HRNet-W48-based version of Panoptic Deeplab. After introducing all of our contributions in CROP + CABB + ISUS, we establish a new state of the art on Mapillary Vistas, surpassing existing approaches by very wide margins (*e.g*. +4.5% PQ, +5.2% mAP).

The IDD experiments in the middle of Table 1 show similar results: CROP outperforms FULL in most metrics, while CABB + ISUS bring further improvements, most pronounced in PC. Compared to prior works, we observe much improved mAP scores and state of the art PQ, while segmentation metrics lag a bit behind. One possible explanation could be the advanced panoptic fusion strategy adopted in EfficientPS, which particularly aims at improving instance

Network	C	Pre-training	PQ	PQ^{th}	PQ^{st}	mAP	mIoU	PC	PC^{th}	PC^{st}	$PQ^{\dagger}$
TASCNet [17]	✗	I	32.6	31.1	34.4	18.6	–	–	–	–	–
AdaptIS [26]	✓	I	35.9	31.5	–	–	–	–	–	–	–
Seamless [23]	✗	I	37.7	33.8	42.9	16.4	50.4	–	–	–	–
Deeplab, X71 [6]	✓	I	37.7	30.4	**47.4**	14.9	55.3	–	–	–	–
EfficientPS [21]	✓	I	38.3	33.9	44.2	18.7	52.6	–	–	–	–
Deeplab, HR48 [6]	✓	I	40.6	–	–	17.8	**57.6**	–	–	–	–
Seamless [23] + CROP	✓	I	39.2	36.5	42.8	19.0	50.8	48.8	41.2	59.0	41.5
Seamless [23] + CABB + ISUS	✓	I	40.5	38.0	43.7	19.4	51.0	50.7	43.1	60.8	42.9
FULL	✗	I	39.4	34.0	46.5	16.2	54.4	55.2	49.7	62.4	39.5
CROP	✓	I	43.6	41.9	45.9	22.3	54.9	56.2	52.4	61.2	45.7
CROP + CABB	✓	I	44.5	42.5	47.0	23.0	55.4	57.4	54.2	61.6	46.3
CROP + ISUS	✓	I	44.7	43.1	46.9	23.0	56.3	59.4	56.1	63.7	46.9
CROP + CABB + ISUS	✓	I	**45.1**	**43.4**	**47.4**	**23.9**	56.3	**60.4**	**57.2**	**64.6**	**47.2**
Seamless [23]	✗	I	47.7	48.9	47.1	30.1	69.6	–	–	–	–
EfficientPS [21]	✓	I	50.1	50.7	**49.8**	31.6	**71.3**	–	–	–	–
FULL	✗	I	49.1	51.0	48.1	32.3	69.0	71.0	76.2	68.3	50.5
CROP	✓	I	50.3	52.5	49.1	35.3	69.7	70.8	73.8	69.2	51.4
CROP + CABB + ISUS	✓	I	**50.7**	**52.9**	49.5	**35.7**	70.4	**72.8**	**78.1**	**70.0**	**51.9**
Seamless [23]	✗	I, V	65.0	60.7	68.0	–	80.7	–	–	–	–
Deeplab, X71 [6]	✓	I, V	65.3	–	–	38.8	82.5	–	–	–	–
EfficientPS [21]	✓	I, V	66.1	**62.7**	68.5	41.9	81.0	–	–	–	–
FULL	✗	I, V	66.0	61.7	69.1	39.5	64.2	80.8	79.9	81.4	64.2
CROP	✓	I, V	66.6	61.1	69.5	42.2	81.7	81.3	80.0	82.3	64.4
CROP + CABB + ISUS	✓	I, V	**66.7**	62.4	**69.9**	**43.4**	**82.6**	**82.6**	**82.4**	**82.7**	**65.1**

Table 1: State of the art results on Mapillary Vistas (top), the Indian Driving Dataset (middle), and Cityscapes (bottom) compared with variants of our network. A ✓symbol in column "C" indicates crop-based training. "Deeplab" abbreviates Panoptic Deeplab [6]. "I" and "V" are used to indicate pre-training on ImageNet and Mapillary Vistas, respectively.

segmentation. We observe the same trends in the Cityscapes results reported in the bottom of Table 1, although with reduced margins. While Cityscapes is smaller than IDD and MVD, and some metrics are already quite saturated, we still get notable +1.5% gain for mAP in our CROP + CABB + ISUS setting over previous state-of-the-art.

4.3. Detailed Analysis

After showing our new high-scores for MVD, IDD and Cityscapes in the previous section we provde in-depth analyses for CABB and ISUS next. First, to validate the generality of our proposals, we evaluate crop-based training, our CABB loss, and ISUS when applied to the approach of Porzi *et al*. [23]. We report the results in Table 1 under two settings, both trained on 1024×1024 crops: the unmodified network from [23], reproduced from their original code (Seamless + CROP), and the same network combined with our CABB loss and ISUS (Seamless + CABB + ISUS). Consistent with our other results, the introduction of crop-based training brings consistent improvements over the baseline, particularly in detection metrics, while the CABB loss and ISUS further boost the scores achieving a +2.8% improvement in PQ w.r.t. Seamless. Further ablations on ISUS are reported in Sec. E of the supplementary material.

As discussed in Sec. 1 and 2, we expect crop-based training to have a negative impact on large objects, which we aim to mitigate with our CABB loss, while our ISUS should bring improvements across all scales by smoothing out the object size imbalance. To verify this, in Fig. 5 we plot box (left) and mask (right) mAP scores as a function of object size (*i.e*. area), splitting the validation instances into five categories according to size percentiles. As expected, CROP outperforms FULL by a wide margin on smaller objects, as it is able to work on almost double the input resolution. On the other hand, the gap between CROP and FULL shrinks as object size increases, with FULL finally surpassing CROP on the largest objects. By adding CABB the crop-based network is able to fill the gap with FULL when dealing with objects in the 99th size percentile, while maintaining strong performance in all other size categories. ISUS brings generalized improvements over CROP at most scales, with the exception of the smallest one. More surprisingly, ISUS seems to be similarly beneficial as CABB on the largest objects. A possible explanation is that, by increasing generalization

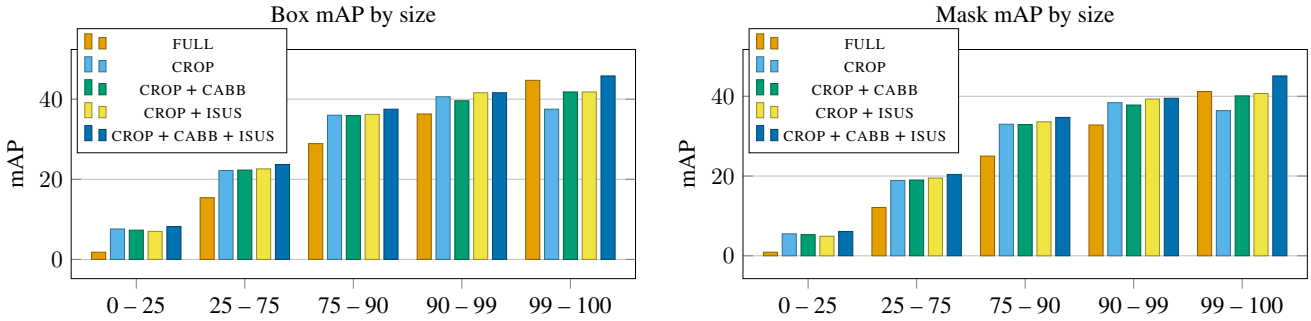

Figure 5: Mean Average Precision results on Mapillary Vistas, averaged over different size-based subdivisions of the validation instances. The reported ranges are percentiles of the distribution of instance areas in the validation set.

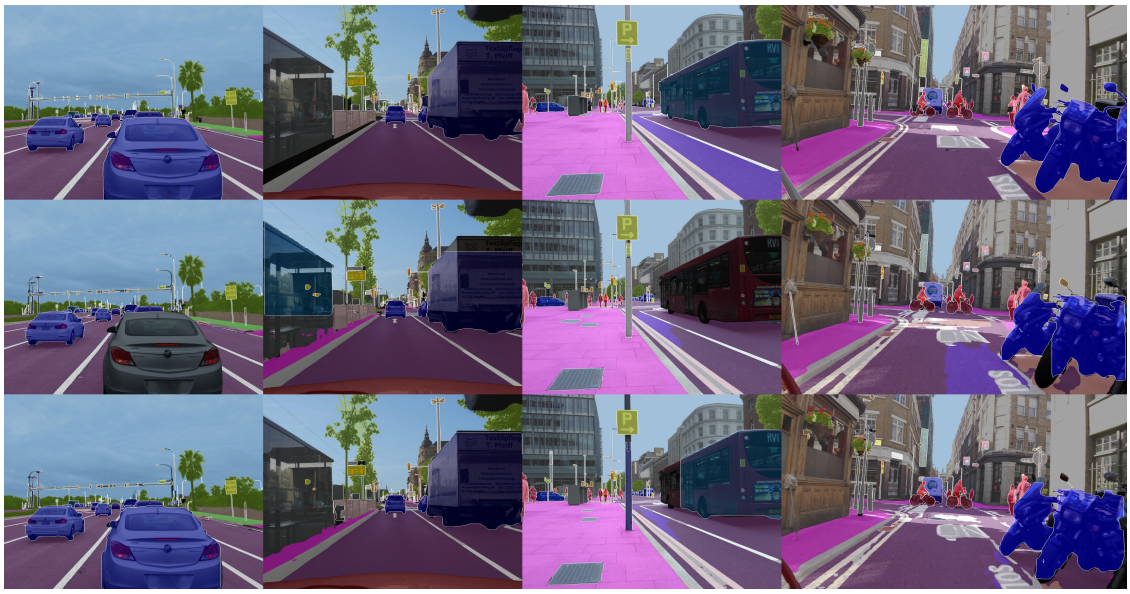

Figure 6: Ground truth (first row) and panoptic segmentation results on Mapillary Vistas' validation set obtained with CROP (second row) and CROP + CABB + ISUS (third row). Notice how CROP + CABB + ISUS is able to detect very big instances which are completely missed by CROP. This figure is best viewed on screen and at magnification.

across scales, ISUS allows the network to properly infer the sizes and positions of objects that are bigger than the training crop. Finally, when CABB and ISUS are combined, we observe consistent improvements on all sizes.

In Table 1 we report additional comparisons between our network variants, based on PC and $PQ^{\dagger}$ (see Sec. C in the supplementary material). In all datasets, we observe a clear improvement in these metrics when the CABB loss and ISUS are introduced in the network. In particular, the gap between CROP and CROP + CABB + ISUS in PC^{th} is markedly larger than in PQ^{th}. This is unsurprising, as the PC metrics weight image segments proportionally to size, clearly highlighting how the CABB loss is able to boost the network's accuracy on large instances. This is also visible from the qualitative results in Fig. 6, showing a comparison between the outputs of CROP and CROP + CABB + ISUS on 12Mpixels Mapillary Vistas validation images featuring large objects.

5. Conclusions

In this paper we have tackled the problem of training panoptic segmentation networks on high resolution images, using crop-based training strategies to enable the use of modern, high-capacity architectures. Training on crops has a negative impact on the detection of large objects, which we addressed by introducing a novel crop-aware bounding box regression loss. To counteract the imbalanced distribution of objects sizes, we further proposed a novel data sampling and augmentation strategy which we have shown to improve generalization across scales. By combining these with a state-of-the-art panoptic segmentation architecture we achieved new top scores on the Mapillary Vistas dataset, surpassing the previous best performing approaches by +4.5% PQ and +5.2% mAP. We also showed state of the art results on the Indian Driving and Cityscapes datasets on multiple detection and segmentation metrics.

References

[1] S. N. Afriat. Theory of maxima and the method of lagrange. *SIAM J. Appl. Math.*, 20(3):343–357, 1971. 5

[2] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *(PAMI)*, 40(4):834–848, 2018. 2

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, September 2018. 2

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018. 6

[5] Bowen Cheng, Maxwell D. Collins, Yukun Zhu, Ting Liu, Thomas S. Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab. *arXiv:1910.04751*, 2019. 2

[6] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12485, 2020. 1, 5, 6, 7

[7] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *(CVPR)*, 2020. 5

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6

[10] N. Gao, Y. Shan, Y. Wang, X/ Zhao, Y. Yu, M. Yang, and K. Huang. SSAP: Single-shot instance segmentation with affinity pyramid. In *(ICCV)*, 2019. 5

[11] Aidan N. Gomez, Mengye Ren, Raquel Urtasun, and Roger B. Grosse. The reversible residual network: Backpropagation without storing activations. In *(NIPS)*, December 2017. 1

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. 1

[14] Paras Jain, Ajay Jain, Aniruddha Nrusimha, Amir Gholami, Pieter Abbeel, Joseph Gonzalez, Kurt Keutzer, and Ion Stoica. Breaking the memory wall with optimal tensor rematerialization. In *Proceedings of Machine Learning and Systems 2020*. 2020. 1

[15] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *(CVPR)*, pages 6399–6408, 2019. 1, 2, 5

[16] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *(CVPR)*, pages 9404–9413, 2019. 1, 2

[17] Jie Li, Allan Raventos, Arjun Bhargava, Takaaki Tagawa, and Adrien Gaidon. Learning to fuse things and stuff. *CoRR*, abs/1812.01192, 2018. 1, 3, 5, 6, 7

[18] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *(CVPR)*, 2019. 5

[19] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *(CVPR)*, 2017. 5

[20] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *(ICLR)*, 2018. 1

[21] Rohit Mohan and Abhinav Valada. Efficientps: Efficient panoptic segmentation. *arXiv preprint arXiv:2004.02307*, 2020. 1, 3, 5, 6, 7

[22] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The Mapillary Vistas dataset for semantic understanding of street scenes. In *(ICCV)*, 2017. 2, 6

[23] Lorenzo Porzi, Samuel Rota Bulò, Aleksander Colovic, and Peter Kontschieder. Seamless scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 3, 5, 6, 7

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *(NIPS)*, 2015. 1, 3

[25] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. In-place activated batchnorm for memory-optimized training of DNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 3, 6

[26] Konstantin Sofiiuk, Olga Barinova, and Anton Konushin. Adaptis: Adaptive instance selection network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7355–7363, 2019. 1, 5, 6, 7

[27] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and C V Jawahar. Indian driving dataset (IDD): A dataset for exploring problems of autonomous navigation in unconstrained environments. In *(WACV)*, 2019. 2, 6

[28] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019. 2, 6

[29] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8818–8826, 2019. 1, 5

[30] Tien-Ju Yang, Maxwell D. Collins, Yukun Zhu, Jyh-Jing Hwang, Ting Liu, Xiao Zhang, Vivienne Sze, George Papandreou, and Liang-Chieh Chen. Deeperlab: Single-shot image parser. *CoRR*, abs/1902.05093, 2019. 5

[31] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. *arXiv:1909.11065*, 2020. 2