Part-aware Panoptic Segmentation

Daan de Geus[1*] Panagiotis Meletis[1*] Chenyang Lu[1] Xiaoxiao Wen[2] Gijs Dubbelman[1]
[1]Eindhoven University of Technology [2]University of Amsterdam
{d.c.d.geus, p.c.meletis} @tue.nl

Abstract

In this work, we introduce the new scene understanding task of Part-aware Panoptic Segmentation (PPS), which aims to understand a scene at multiple levels of abstraction, and unifies the tasks of scene parsing and part parsing. For this novel task, we provide consistent annotations on two commonly used datasets: Cityscapes and Pascal VOC. Moreover, we present a single metric to evaluate PPS, called Part-aware Panoptic Quality (PartPQ). For this new task, using the metric and annotations, we set multiple baselines by merging results of existing state-of-the-art methods for panoptic segmentation and part segmentation. Finally, we conduct several experiments that evaluate the importance of the different levels of abstraction in this single task.

1. Introduction

Humans perceive and understand a scene at multiple levels of abstraction. Concretely, when observing a scene, we do not only see a single semantic label for each visual entity, such as *person* or *car*. We also distinguish the parts of entities, such as *person-leg* and *car-wheel*, and we are able to group together the parts that belong to a single individual entity. Currently, there is no computer vision task that aims at simultaneously understanding a scene holistically on both of these levels of abstraction: *scene parsing* and *part parsing*.

Instead, most methods focus on solving a task at a single level of abstraction. On the one hand, scene parsing aims to recognize and semantically segment all foreground objects (*things*) and background classes (*stuff*) in an image. Recently, this task has been formalized as *panoptic segmentation* [21], for which the goal is to predict 1) a class label and 2) an instance *id* for each pixel in an image. This formalization has resulted in a boost in research interest that advanced the state-of-the-art [5, 20, 27, 42, 44, 49]. On the other hand, part parsing takes over where scene parsing stops, as it aims to segment an image based on part-level semantics, *i.e.*, the parts constituting the scene-level classes. For this level of abstraction, there is a wide range of different task definitions, and resulting methods. Most methods focus on a single object class and are instance-agnostic, while only a few are instance-aware [15, 26, 56], or focus on multiple object classes [41, 57]. A more comprehensive overview of related work is provided in Section 2.

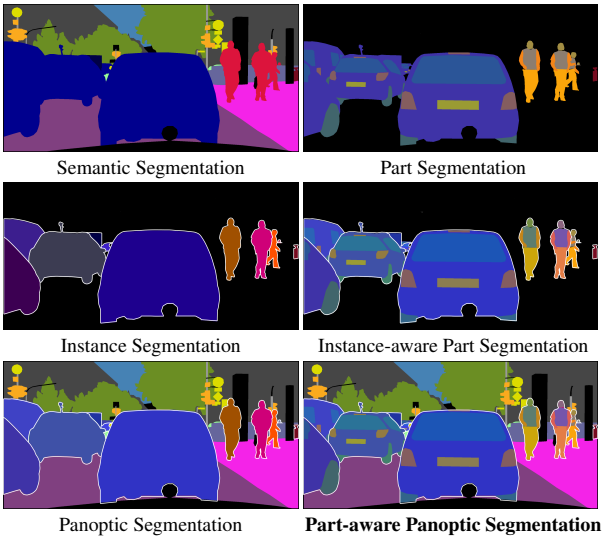

Figure 1. **Evolution of scene understanding tasks**: from semantic to panoptic (top to bottom) and from part-agnostic to part-aware (left to right). Colors indicate scene-level and part-level semantics. Instance-level boundaries are emphasized with a white contour.

To come closer to unified perception at multiple levels of abstraction, this work defines a task that combines scene parsing and part parsing in a single task. This task encompasses the ability to 1) apply per-pixel scene-level classification, 2) segment things classes into individual instances, and 3) segment stuff classes or things instances into their respective parts. We call this task *part-aware panoptic segmentation* (PPS); the conceptual differences with existing tasks are visualized in Figure 1. Together with this task, we also define a metric to evaluate it. This metric, *part-aware panoptic quality* (PartPQ), extends the *panoptic quality* metric [21] to cover part segmentation performance per detected things instance or stuff class. More details on the task and metric can be found in Section 3.

To allow for research on the new task of PPS, we intro-

*Both authors contributed equally.

duce consistent part-aware panoptic annotations for two commonly used datasets. For Cityscapes [6], we have labeled part classes for all 3.5k images of the train and validation set, which are consistent with the existing panoptic annotations. For Pascal VOC [13], we have combined the existing datasets for semantic segmentation [43] and instance-aware part segmentation [4] to generate a consistent annotation set for PPS. In Section 4, we provide further explanations and statistics on these datasets.

As there is no existing work on part-aware panoptic segmentation, we establish several benchmarks. We create baselines by generating state-of-the-art results on panoptic segmentation and part segmentation, and merging these to the PPS format using heuristics. As explained in Section 5, there are several design choices that need to be taken into account, when combining predictions at multiple levels of abstraction. Specifically: should we opt for a top-down method, where we prioritize the scene-level predictions from panoptic segmentation, and complement these with part predictions, or is it better to use a bottom-up approach, where we combine parts to generate scene-level predictions? To evaluate this, we conduct experiments to research the benefits of both types of strategies. Both these experiments and the baselines provide a direction for future research on multi-task training of PPS architectures where the different subtasks can benefit from each other.

To summarize, this work contains the following contributions:

- The introduction of the part-aware panoptic segmentation (PPS) task, unifying perception at multiple levels of abstraction.
- The PartPQ metric to evaluate this task.
- Coherent PPS annotations for two commonly used datasets, which are made available to the public.
- Baselines for the PPS task on two datasets.
- An analysis of the design choices for the new PPS task.

All annotations and the code are available at https://github.com/tue-mps/panoptic_parts.

2. Related work

Research on visual scene understanding aims to extract all-encompassing information from images with the long-term goal to mimic human visual-cognitive capabilities. So far, research has primarily focused on approaching scene understanding at a single level of abstraction. In this work, we propose a single coherent task for multiple levels of abstraction, which unifies the tasks of scene parsing and part parsing.

2.1. Scene parsing

We refer to scene parsing as the overall task to semantically understand an image at the class level, and to distinguish between individual things instances. Recently, this task has been formalized as panoptic segmentation [21], which is a unification of the typically distinct tasks of semantic segmentation and instance segmentation. In earlier forms, this task has been investigated in [47, 53].

Initially, most panoptic segmentation methods applied a multi-task network that trains and outputs instance segmentation and semantic segmentation in parallel, followed by a merging operation to generate panoptic segmentation results [8, 20, 28, 42, 44]. Recently, more methods are introduced that focus on optimizing the process of merging to panoptic segmentation [23, 36, 49, 52], or try to solve the task more holistically or efficiently [5, 9, 17, 27, 51].

Although panoptic segmentation allows for more holistic scene understanding than the earlier tasks of semantic segmentation and instance segmentation, it does not cover knowledge of part-level semantics of the identified segments. Such knowledge would provide a more comprehensive understanding of the scene, and would allow for more detailed downstream reasoning.

2.2. Part parsing

We refer to part parsing as the umbrella task of segmenting images based on part-level semantics. At a high level, we can distinguish two types of tasks: part segmentation and pose estimation. Part segmentation requires a pixel-level prediction for all identified parts, whereas pose estimation aims at detecting connecting keypoints between the parts for each object. Pose estimation is inherently instance-aware and is exhaustively researched, as is clear from the surveys in [7, 38].

However, dense part-level segmentation remained for a long time instance-agnostic, as it is usually treated as a semantic segmentation problem [14, 18, 19, 26, 32, 37, 39, 40, 41, 57]. In the trend of coming to more holistic tasks, a dense pose task was introduced in [1] and a unification of pose estimation and part segmentation is provided in [11]. Only recently, to the best of our knowledge, an instance-aware human part segmentation task was introduced and studied in [15, 26, 56]. Most research has focused on part segmentation for humans [12, 15, 22, 24, 25, 31, 30, 34, 46, 50, 56], but other parts have also received attention, *e.g.*, facial parts [33], and animal parts [4, 48]. A limited amount of papers have addressed multi-class part segmentation [41, 57], but so far these methods are not instance-aware. As a result, instance-aware part segmentation on a more general dataset, consisting of a wider range of classes and parts, remains unaddressed.

Moreover, although work has shown that learning on multiple levels of abstraction can improve the performance of a part segmentation network [41, 57], part parsing has not yet been merged with scene parsing into one holistic task, which can describe the image at multiple levels of abstraction. In our work, we aim to boost the interest in

this area by providing a unified task for *part-aware panoptic segmentation*, and accompanying metrics and annotations.

2.3. Datasets

In order to train and evaluate on the new PPS task, we need datasets that 1) have scene-level labels for panoptic segmentation and 2) have part-level labels for a set of those scene-level classes. Although a plethora of datasets exist for object detection and semantic segmentation, only few have labels compatible with the panoptic segmentation task (*e.g.*, [6, 35]). For part-level segmentation, the datasets are even more scarce. LIP [30], MHP [56] and CIHP [15] provide instance-aware, part-level annotations, but only for human parts. To the best of our knowledge, Pascal-Parts is the only dataset that has part-level annotations for a more general set of classes [4]. However, these annotations do not contain any information on classes without parts.

From this, we observe that there is no dataset that covers all the requirements for the PPS task. Therefore, we present consistent part-aware panoptic annotations on two datasets. For Cityscapes [6], a commonly used dataset for panoptic segmentation, we annotate parts for five different things classes. Moreover, we collect and arrange the different annotation sets for Pascal VOC [13] to generate a complete and consistent annotation set for 10k Pascal VOC images.

3. Part-aware Panoptic Segmentation

3.1. Task definition

The task of *Part-aware Panoptic Segmentation* (PPS) is an image understanding task that is designed to capture image understanding at multiple levels of abstraction. Specifically, it captures 1) scene-level semantics, 2) instance-level information, and 3) part-level semantics. To achieve this, we define PPS as a task that enriches panoptic segmentation [21] with part-level semantics.

A part-aware panoptic segmentation algorithm describes every pixel in an image with a set of semantic and instance-level information. This can be expressed for pixel i in the form $(l, p, z)_i$, where l represents the scene-level semantic class, p the part-level semantic class, and $z \in \mathbb{N}$ the instance *id*. The scene- and part-level semantic classes are predefined and usually correspond to the available semantic granularity of a dataset's labels, while the instance *id* is an unbounded integer separating, per image, distinct instances of the same scene-level semantic class.

The scene-level semantic class l is chosen from a predetermined set of $\mathcal{L} := \{l_1, \dots, l_L\}$ classes. For any of these classes a set of part-level semantic classes $\mathcal{P}_l = \{p_{l,1}, \dots, p_{l,P_l}\}$ containing P_l semantic parts may be defined. We denote the superset of all parts as $\mathfrak{P} = \cup_l \mathcal{P}_l,\ l \in \mathcal{L}$. The set $\mathcal{L}$ can be separated into disjoint subsets in two different ways. *Firstly*, $\mathcal{L} = \mathcal{L}^{\text{St}} \cup \mathcal{L}^{\text{Th}}$. The subset $\mathcal{L}^{\text{St}}$ consists of the stuff classes, *i.e.*, uncountable entities with amorphous shape (*e.g.*, sky, sea), and subset $\mathcal{L}^{\text{Th}}$ contains the things classes, which are classes for countable objects with well-defined shape (*e.g.*, car, person). *Secondly*, $\mathcal{L}$ can also be separated in a subset of scene-level classes that have parts (*e.g.*, limbs, car parts), $\mathcal{L}^{\text{parts}}$, and scene-level classes that do not have parts, $\mathcal{L}^{\text{no-parts}}$. Here, $\mathcal{L} = \mathcal{L}^{\text{parts}} \cup \mathcal{L}^{\text{no-parts}}$. We require that both $\mathcal{L}^{\text{St}} \cap \mathcal{L}^{\text{Th}} = \emptyset$ and $\mathcal{L}^{\text{parts}} \cap \mathcal{L}^{\text{no-parts}} = \emptyset$. The selection of classes belonging to the four subsets $\mathcal{L}^{\text{St}}$, $\mathcal{L}^{\text{Th}}$, $\mathcal{L}^{\text{parts}}$, $\mathcal{L}^{\text{no-parts}}$ is a design choice that is typically determined based on the requirements of the application, or the purpose of a dataset, as for [21].

A PPS algorithm makes a prediction that adheres to the following requirements: 1) a scene-level semantic class $\mathcal{L}$ must be assigned to all pixels, 2) a part-level semantic class must be assigned to – and only to – all pixels that are assigned a scene-level class from $\mathcal{L}^{\text{parts}}$, and 3) an instance-level *id* is provided for – and only for – pixels that are assigned a scene-level class from $\mathcal{L}^{\text{Th}}$. In summary, a pixel can be labeled with one of following combinations, where "$-$" denotes that the specific abstraction level is irrelevant, as a:

- Stuff class: $(l, -, -), l \in \mathcal{L}^{St}$
- Stuff class with parts: $(l, p, -), l \in \mathcal{L}^{St} \cap \mathcal{L}^{\text{parts}}, p \in \mathcal{P}_l$
- Things class: $(l, -, z), l \in \mathcal{L}^{Th}, z \in \mathbb{N}$
- Things class with parts: $(l, p, z), l \in \mathcal{L}^{Th} \cap \mathcal{L}^{\text{parts}}, p \in \mathcal{P}_l$

Finally, the PPS format accepts a special *void* label for scene-level and part-level semantics, which represents ambiguous pixels or concepts not included in any subset $\mathcal{L}$.

Relationship to other tasks. *Part-aware panoptic segmentation* (PPS) is related to and generalizes various per-pixel segmentation tasks. *Part segmentation* is specialized semantic segmentation focusing on segmenting object parts, but it does not require separating parts according to the object instance they belong to. In the PPS format it can be described as $(l, p, -)_i, l \in \mathcal{L}^{\text{parts}}, p \in \mathfrak{P}$. *Instance-aware part segmentation*, can be described as $(l, p, z)_i, l \in \mathcal{L}^{\text{Th}} \cap \mathcal{L}^{\text{parts}}$, $p \in \mathfrak{P}$, and pivots part parsing on an instance level, but treats any non-things pixel as background, losing environmental context. Finally, *panoptic segmentation*, $(l, -, z)_i, l \in \mathcal{L}$, includes no notion of part semantics.

3.2. Part-aware Panoptic Quality

With the proposed PPS task, that unifies perception at multiple levels of abstraction, we aim to quantify the performance of the methods for this task using a *single unified metric*. Inspired by the previous Panoptic Quality (PQ) metric [21], we propose *Part-aware Panoptic Quality* (PartPQ). The proposed PartPQ is designed to capture 1) the ability to identify and classify panoptic segments, *i.e.*, stuff regions and things instances, and 2) the part segmentation quality within the identified panoptic segments.

Dataset	Instance aware	Panoptic aware	Parts aware	Stuff classes	Things classes	Parts classes	#Images train / val	Average image size	Average #inst./img
PASCAL-Context [43]	-	-	-	459 (59)	-	-	4998 / 5105	387 × 470	-
LIP [30]	-	-	✓	-	1	20	30.5k / 10k	325 × 240	-
CIHP [15]	✓	-	✓	-	1	20	28.3k / 5k	484 × 578	3.4
MHP v2.0 [56]	✓	-	✓	-	1	59	15.4k / 5k	644 × 718	3
PASCAL-Person-Parts [4]	✓	-	✓	-	1	6	1716 / 1817	387 × 470	2.2
PASCAL-Parts [4]	✓	-	✓	-	20	194	4998 / 5105	387 × 470	2.5
Cityscapes [6]	✓	✓	-	23	8	-	2975 / 500	1024 × 2048	17.9
This work									
PASCAL Panoptic Parts	✓	✓	✓	80	20	194	4998 / 5105	387 × 470	2.5
Cityscapes Panoptic Parts	✓	✓	✓	23	8	23	2975 / 500	1024 × 2048	17.9

Table 1. **Dataset statistics** for related (part) segmentation datasets and our proposed datasets. *PASCAL-Context* has 459 semantic classes but only 59 of them are included in the official split.

The PartPQ per scene-level class l is formalized as

$$\text{PartPQ} = \frac{\sum_{(p,g)\in TP} \text{IOU}_\text{p}(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}. \quad (1)$$

As in the original PQ, we assess the ability to identify panoptic segments by counting the amount of true positive, TP, false positive, FP, and false negative, FN, segments, based on the Intersection Over Union (IOU) between a predicted segment p and a ground-truth segment g for a class l. A prediction is a TP if it has an overlap with a ground-truth segment with an IOU > 0.5. An FP is a predicted segment that is not matched with the ground-truth, and an FN is a ground-truth segment not matched with a prediction.

The part segmentation performance within matched segments is captured by the $\text{IOU}_\text{p}(p,g)$ term in Equation 1. To be compatible both with scene-level classes with parts ($\mathcal{L}^{\text{parts}}$), and without parts ($\mathcal{L}^{\text{no-parts}}$), we define two cases:

$$\text{IOU}_\text{p}(p,g) = \begin{cases} \text{mean IOU}_\text{part}(p,g), & l \in \mathcal{L}^{\text{parts}} \\ \text{IOU}_\text{inst}(p,g), & l \in \mathcal{L}^{\text{no-parts}} \end{cases} \quad (2)$$

For the classes $\mathcal{L}^{\text{parts}}$, we calculate the mean Intersection Over Union for all part classes in the two matched panoptic segments. This is the multi-class mean IOU where the region outside the two segments is labeled background. When computing this score, we allow the prediction to contain pixels with a *void* part label. In the mean IOU, those pixels will not be counted as false positives, but will be counted as false negatives (similar to scene-level *void* labels in PQ [21]). For the subset of classes without parts, $\mathcal{L}^{\text{no-parts}}$, the instance-level IOU is computed as in the original PQ.

In essence, the multi-class mean IOU_part term captures the quality of both the mask of the panoptic segment, and the part segmentation within this segment. Both the quality of the panoptic mask and the part segmentation within the mask need to be high in order to get a high score.

The overall PartPQ is calculated by averaging over all per-class PartPQ scores for scene-level classes $l \in \mathcal{L}$. In Section 5, we evaluate the performance using PartPQ on two datasets. We show that this metric exhibits a reliable performance measure of different approaches, and is consistent with other metrics commonly used for the subtasks combined in part-aware panoptic segmentation.

4. Datasets

We accompany the PPS task with two new datasets, Cityscapes Panoptic Parts (CPP) and PASCAL Panoptic Parts (PPP), which are based on the established scene understanding datasets Cityscapes [6] and PASCAL VOC [13], respectively. The introduced datasets include per-pixel annotations on multiple levels of visual abstraction: scene-level and part-level semantics, and instance-level information. As can be seen from Table 1, the existing datasets landscape is inadequate for PPS since no dataset features all of these levels of abstraction. If any combination of the existing datasets is used to achieve multi-level abstraction, conflicts would arise at the pixel level due to overlapping labels. Our datasets comprise a consistent set of annotations, which are free of such conflicts.

4.1. Cityscapes Panoptic Parts

Cityscapes Panoptic Parts (CPP) extends with part-level semantics the popular Cityscapes dataset [6] of urban scenes recorded in Germany and neighboring countries. We manually annotated with 23 part-level semantic classes the original publicly available 2975 training and 500 validation images. We employed a pipeline that takes advantage of original annotations to guide and hint annotators.

CPP is fully compatible with the original Cityscapes panoptic annotations and is, to the best of our knowledge, the first urban scenes dataset with annotations on scene-level, part-level and instance-level, on the same set of images.

Taking into consideration the complexity of scenes and the variety in number and pose of traffic participants we selected 5 scene-level semantic classes from the *human* and *vehicle* high-level categories to be annotated with parts, *i.e.*, $\mathcal{L}^{\text{parts}} = \{person, rider, car, truck, bus\}$. The *human* categories are annotated with $\mathcal{P}^{human} = \{torso, head, arm, leg\}$ and the *vehicle* categories with $\mathcal{P}^{vehicle} = \{chassis, window, wheel, light, license\ plate\}$. Statistics for CPP are presented in Table 1 and in Figure 2.

4.2. PASCAL Panoptic Parts

PASCAL Panoptic Parts (PPP) extends the PASCAL VOC 2010 benchmark [13] with part-level and scene-level semantics. The original PASCAL VOC dataset is labeled on scene-level semantics, and only partly on instance-level. A large number of subsequent extensions have been proposed with annotations over different levels of abstraction, leading to various inconsistencies between them at the pixel level. We created PPP by carefully merging PASCAL-Context [43] and PASCAL-Parts [4] to maintain high quality of annotations and solve any conflicts. As the PPP dataset solves conflicts between PASCAL-Context [43] and PASCAL-Parts [4], evaluations on PPP are not consistent with those on the aforementioned datasets.

PPP preserves the original splitting into 4998 training and 5105 validation images. On the scene-level abstraction PPP contains $\left|\mathcal{L}^{\text{Th}}\right| = 20$ classes with instance-level annotations and $\left|\mathcal{L}^{\text{St}}\right| = 80$ classes without instances. On the part-level abstraction it comprises $|\mathfrak{P}| = 194$ parts spanning $|\mathcal{L}^{\text{parts}}| = 16$ classes, and $\left|\mathcal{L}^{\text{Th}} \cap \mathcal{L}^{\text{parts}}\right| = 16$. For easier comparison with related methods we provide mappings from PPP to commonly used subsets: 7 parts for human part parsing on PASCAL-Person-Parts [4] and 58 parts for the reduced set used in [41, 57]. More statistics can be found in Table 1.

For both CPP and PPP, part-level classes are only defined for scene-level things classes. We anticipate that, in future work, designers of datasets also opt for assigning part classes to stuff classes. If so, this is fully compatible with our task definition and metric, as they already support this.

5. Experimental analysis

With the introduced task definition, annotations and metric, we now establish benchmarks for the part-aware panoptic segmentation task, and compare the PartPQ metric with existing metrics. The results are presented and explained in Section 5.1, and can serve as references for future research.

Secondly, to get insight into the difference in quality and relative importance of results on the different levels of abstraction in our unified task, and the design choices that play a role in this regard, we conduct several ablation experiments on these baselines in Section 5.2.

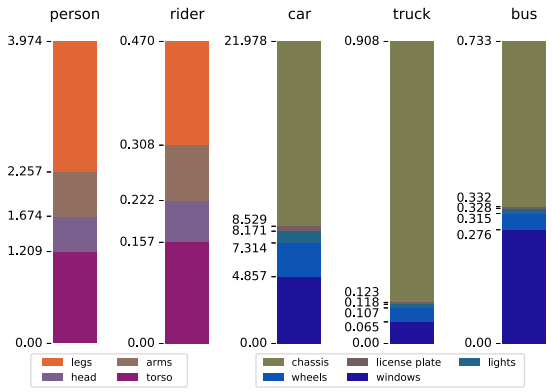

Figure 2. **Statistics CPP.** Absolute number of Cityscapes pixels ($\times 10^7$) that we annotated per scene-level semantic class.

5.1. Benchmarking

Since the part-aware panoptic segmentation task and the PartPQ metric are new, there are no methods for this task yet, and hence no results. To fill this gap, we establish baselines for PPS by merging results of methods for the subtasks of panoptic and part segmentation. For this process, we select both state-of-the-art and commonly used methods. The results for these subtasks are mostly generated using publicly available code, or provided to us by the authors of the respective methods. Only in select cases, when trained models are not publicly available, we train an existing network on the concerned data. If so, we indicate this.

All results are evaluated on the PartPQ metric introduced in Section 3.2. We also report on the PartPQ separately for scene-level classes that have parts ($\mathcal{L}^{\text{parts}}$) with PartPQ_{P} and those that do not have parts ($\mathcal{L}^{\text{no-parts}}$) with $\text{PartPQ}_{\text{NP}}$. To show the performance of the subtask methods before merging, we also report the performance on the Panoptic Quality (PQ) for panoptic segmentation [21] (which we also split in PQ_{P} and PQ_{NP}), Average Precision (AP) for instance segmentation, and mean Intersection Over Union (mIOU) for semantic segmentation and part segmentation.

5.1.1 Merging procedure

To get predictions that adhere to the PPS task defined in Section 3, we need to merge the results on the subtasks of panoptic segmentation and part segmentation. To achieve this, we maintain a straightforward top-down, rule-based merging approach. First, for scene-level semantic classes that do not have part classes ($l \in \mathcal{L}^{\text{no-parts}}$), no additional prediction is required, so we copy the predictions from panoptic segmentation. Secondly, for each segment in the panoptic segmentation prediction that does require an additional part label ($l \in \mathcal{L}^{\text{parts}}$), we identify and extract the part predictions for the pixels corresponding to this segment. If a part prediction contains a part class that does not correspond to the scene-level class (*e.g.*, a *head* pixel in a *bus* segment), we

Panoptic seg. method	Part seg. method	*Before merging* mIOU SemS	AP mask	PQ All	PQ P	PQ NP	mIOU PartS	*After merging* PartPQ All	PartPQ P	PartPQ NP
Cityscapes Panoptic Parts *val*										
UPSNet [49]	DeepLabv3+ [3]	75.2	33.3	59.1	57.3	59.7	75.6	55.1	42.3	59.7
DeepLabv3+ & Mask R-CNN* [3, 16]	DeepLabv3+ [3]	78.8	36.5	61.0	58.7	61.9	75.6	56.9	43.0	61.9
EfficientPS [42]	BSANet [57]	80.3	39.7	65.0	64.2	65.2	76.0	60.2	46.1	65.2
HRNet-OCR & PolyTransform* [54, 29]	BSANet [57]	81.6	44.6	66.2	64.2	67.0	76.0	61.4	45.8	67.0
Pascal Panoptic Parts *validation*										
DeepLabv3+ & Mask R-CNN [3, 16]	DeepLabv3+ [3]	47.1	38.5	35.0	61.5	26.0	53.9	31.4	47.2	26.0
DLv3-ResNeSt269 & DetectoRS [2, 55, 45]	BSANet [57]	55.1	44.8	42.0	66.0	33.8	58.6	38.3	51.6	33.8

Table 2. **Baselines.** Part-aware panoptic segmentation results for the baselines on the *Cityscapes Panoptic Parts* (CPP) and *Pascal Panoptic Parts* (PPP) datasets, generated using results from commonly used (top), and state-of-the-art methods (bottom) for semantic segmentation, instance segmentation, panoptic segmentation and part segmentation. For the results on CPP, $\text{mIOU}_{\text{PartS}}$ indicates the mean IOU for part segmentation on grouped parts (see Section 5.2.2). Metrics split into *P* and *NP* are evaluated on scene-level classes with and without parts, respectively (see Section 5.1). * Indicates pretraining on the COCO dataset [35].

set the part prediction for this pixel to the *void* label.

In Section 5.2.1, we show that this top-down merging strategy works better than a strategy that requires the predictions for both part segmentation and panoptic segmentation to agree. It is likely that there is a better, more complex way to construct or possibly learn this merging strategy, but we leave this for future work to address.

5.1.2 Cityscapes Panoptic Parts

Methods. For the baselines on Cityscapes [6], we generate part-aware panoptic segmentation results using both single network methods for panoptic segmentation [49, 42], and panoptic segmentation results generated from methods on semantic segmentation [3, 54] and instance segmentation [16, 29]. In the latter case, the panoptic segmentation results are created using the heuristic merging process described in [21]. For part segmentation, we trained two networks [3, 57] ourselves, since we are the first to introduce part labels on Cityscapes.

For all baselines, in order to have fair and consistent results, we use methods that are only trained on the Cityscapes `train` set without using the *coarse* labels, with pre-training only on ImageNet [10]. The only exceptions are the instance segmentation methods [16, 29], which are pre-trained on COCO [35], as has become common practice.

Results. With the aforementioned state-of-the-art methods and merging strategy, we set state-of-the-art baselines for part-level panoptic segmentation. The results for Cityscapes Panoptic Parts are reported in Table 2, and qualitative examples are shown in Figure 3. The results show that the scores on PartPQ are lower than the regular PQ, which is expectable, as we add complexity to the problem with part-level segmentation of segments, and the per-instance IOU of PQ is replaced with the part-level IOU in PartPQ. As expected, the scores for $\text{PartPQ}_{\text{NP}}$ are identical to PQ_{NP}, as the results and metric for $\mathcal{L}^{\text{no-parts}}$ are unchanged. When comparing the PartPQ to other metrics, we see that a difference in scores between methods is comparable to the existing metrics on the subtasks. This indicates that the metric captures the aspects covered by those metrics, while being a single metric for the unified task of part-aware panoptic segmentation.

5.1.3 Pascal Panoptic Parts

Methods. Due to a lack of existing work on panoptic segmentation for the Pascal VOC dataset [13], we generate panoptic segmentation results by fusing semantic segmentation [2, 55] and instance segmentation [16, 45] results, following [21]. Specifically, the semantic segmentation methods are generated using existing models trained on 59 classes of the Pascal-Context dataset [43], and we train the instance segmentation models on the 20 things classes of our Pascal Panoptic Parts dataset. For part segmentation, we generate state-of-the-art results using an existing model [57] trained on a dataset that includes 58 part classes from the Pascal-Parts dataset [4], and we train another commonly used model [3] on that same dataset. Despite the different annotations used for training, all models are trained on the same 4998 images in the Pascal VOC 2010 `training` split.

Results. The results for the baselines on the PPP dataset are reported in Table 2. From the table, it is clear that, again, scores for PartPQ increase proportionally to the existing metrics for the subtasks, and that $\text{PartPQ}_{\text{NP}}$ remains identical to PQ_{NP}. Qualitative examples are displayed in Figure 4.

5.2. Ablation experiments

5.2.1 Merging panoptic and part segmentation

Experiment. For the aforementioned baselines, we use a top-down merging strategy that effectively prioritizes panoptic segmentation over part segmentation, by taking the scene-level semantic label from the panoptic output. It is also possible to take a more conservative approach that also con-

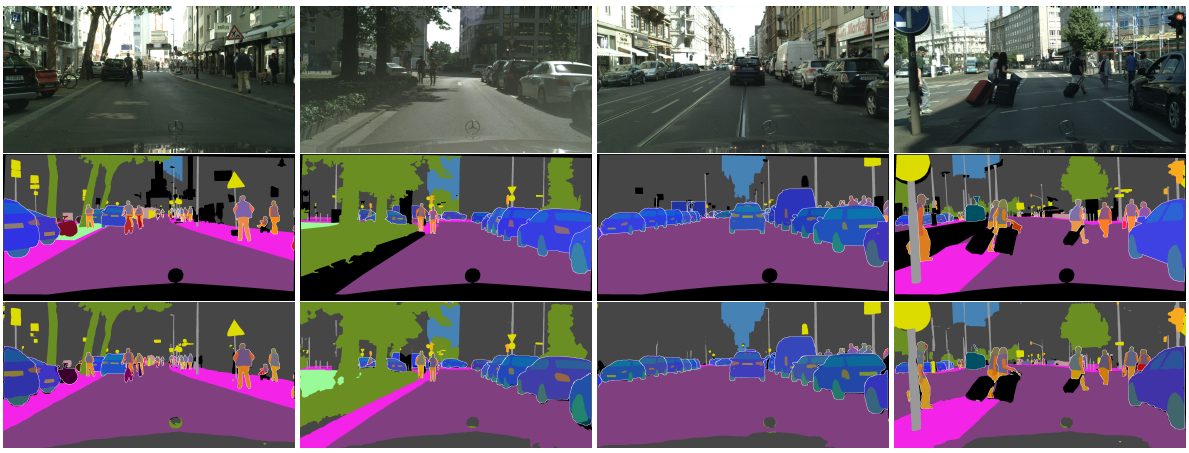

Figure 3. **Examples Cityscapes Panoptic Parts.** Top: input; middle: ground truth; bottom: predictions highest-scoring PPS baseline.

Figure 4. **Examples Pascal Panoptic Parts.** Top: input; middle: ground truth; bottom: predictions highest-scoring PPS baseline.

	Before merging		*After merging*	
Merging str.	**PQ**	**mIOU$_{\text{PartS}}$**	**PartPQ**	**PartPQ$_{\text{P}}$**
State-of-the-art results on Cityscapes Panoptic Parts				
original	66.2	67.2	**60.9**	**44.0**
alternative	66.2	67.2	60.2	41.3
State-of-the-art results on Pascal Panoptic Parts				
original	42.0	58.6	**38.3**	**51.6**
alternative	42.0	58.6	37.5	50.6

Table 3. The results of **different merging procedures**, on the `val` split of CPP and the `validation` split of PPP. The *original* merging strategy prioritizes panoptic segmentation; the *alternative* strategy requires both predictions to agree on the scene-level label.

siders bottom-up information, by requiring panoptic and part segmentation predictions to agree on the scene-level semantic label. For this alternative approach, a panoptic segment is compared with the part predictions at the corresponding pixels, and for each pixel, the panoptic prediction is only kept if the part prediction is possible for the scene-level label of that panoptic segment (*e.g.*, *truck-wheel* for a *truck* instance). Otherwise, the pixel is removed from the segment, and both the scene-level and the part-level predictions are set to *void*. This merging approach would lead to better results

Grouping	**PQ**	**mIOU**	**mIOU$_{\text{grouped}}$**	**PartPQ**	**PartPQ$_{\text{P}}$**
Commonly used methods for panoptic seg. and part seg.					
-	61.0	54.3	74.5	55.8	38.8
✓	61.0	*n/a*	**75.6**	**56.9**	**43.0**
State-of-the-art methods for panoptic seg. and part seg.					
-	66.2	67.2	75.3	60.9	44.0
✓	66.2	*n/a*	**76.0**	**61.4**	**45.8**

Table 4. **Grouping parts**. Trained on the Cityscapes Panoptic Parts set using grouped parts: 1) *car*, *bus* and *truck* parts, and 2) *person* and *rider* parts. Reported mIOU scores are for part-level semantics.

than the original, if the panoptic segmentation method frequently makes mistakes that the part segmentation method does not make, and if part segmentation predictions are not incorrect where panoptic segmentation is correct.

Results. The results, reported in Table 3, clearly show that the original merging method performs better. For the alternative approach, the PartPQ for classes with parts is consistently lower. This occurs as pixels are incorrectly removed from segments. These results clearly indicate that it is better to prioritize the scene-level label from panoptic segmentation over that from part segmentation.

Panoptic seg.		*Part seg.*		*Semantic information gain*	
mPA	**mIOU**	**mPA**	**mIOU**	**mSIG$_{\text{pan}\rightarrow\text{part}}$**	**mSIG$_{\text{part}\rightarrow\text{pan}}$**
91.6	85.9	88.6	82.5	54.1	39.4

Table 5. **Comparing performance on scene-level semantics** between *state-of-the-art* methods for panoptic segmentation and part segmentation, on Cityscapes Panoptic Parts `val`.

5.2.2 Grouping semantically similar parts

Experiment. The results from Section 5.2.1 suggest that methods trained on panoptic segmentation are better able to predict scene-level semantics than part segmentation methods, favoring a top-down approach to PPS. To further explore the potential benefits of a top-down approach, we conduct experiments where we train a part segmentation method on parts that are grouped by semantic similarity (*e.g.*, *bus-wheel* and *car-wheel* are grouped as *wheel*). This is likely to work because 1) there is more data per part class and 2) there is less ambiguity between the part classes. This favors a top-down approach because it means that, to get a prediction in the PPS format, the scene-level label needs to be extracted from panoptic segmentation, and that part segmentation is used to learn the specific parts only.

Results. We train part segmentation networks for which the parts for 1) *car*, *bus*, and *truck*, and 2) *person* and *rider*, are grouped, effectively reducing the amount of parts from 23 to 9. The results for this experiment are shown in Table 4, and they show that the PartPQ for classes with parts, PartPQ$_{\text{P}}$, increases with up to 4.2 points when parts are grouped. This supports our hypothesis.

5.2.3 Comparing levels of abstraction

In the previous experiments, we have seen results that indicate that it is sensible to approach PPS in a top-down manner, *i.e.*, to first predict the scene-level semantic label, and then look for parts within those regions. To further substantiate this hypothesis, and to assess what the main information source should be for scene-level semantics, we conduct an additional experiment that compares the scene-level performance of methods trained on panoptic and part segmentation.

Metrics. To assess the extent to which correct scene-level information is available in one method, but not in another, we introduce the *Semantic Information Gain* (SIG) metric, which quantifies the extent to which errors made by a given method *B* can be compensated for by the correct predictions of a method *A*. We define the the SIG of method A with respect to method B, $SIG_{A\rightarrow B}$, as

$$SIG_{A\rightarrow B} = \frac{1}{|X_{FP_B}|} \sum_{x \in X_{FP_B}} TP_{A,x} \times 100\%, \quad (3)$$

where X_{FP_B} is the set of pixels incorrectly predicted by method B, and $TP_{A,x} = 1$ if method A is correct at pixel x and $TP_{A,x} = 0$ otherwise. We evaluate the SIG per class in the ground truth, and report the mean SIG (mSIG) over all scene-level classes with parts, $\mathcal{L}^{\text{parts}}$. We also report on mean Pixel Accuracy (mPA) and mean Intersection Over Union (mIOU).

Results. When looking at the results in Table 5, it is clear that the panoptic segmentation method is considerably more accurate on the concerning five scene-level classes than part segmentation. Moreover, panoptic segmentation predictions can resolve, on average, 54.1% of the errors made by part segmentation. Specifically, these errors seem to occur for classes that have parts that could be confused with each other (*e.g.*, *bus* and *truck*). This supports the aforementioned hypothesis about a top-down approach being a good way to approach part-aware panoptic segmentation.

This does not mean, however, that the bottom-up alternative has no potential at all. Table 5 shows that, to a lesser degree, part segmentation can also solve errors made by the panoptic segmentation method. Therefore, it is likely that a future top-down method for PPS could be improved when enriched with specific bottom-up features.

6. Conclusion

In this work, we presented the novel task of part-aware panoptic segmentation (PPS), which takes the next step in holistic scene understanding by unifying scene parsing and part parsing. With the accompanying metric and datasets, we have generated state-of-the-art results for this task, constructed from state-of-the-art results on the underlying subtasks. We hope that this work will spark new innovations in the area of scene understanding, for which our results can serve as baselines.

Specifically, we hope to see innovations in single-network PPS methods that learn the levels of abstraction – *i.e.*, part-level and scene-level – jointly, to leverage the interaction between these levels during training. An important design choice is the way in which information from these levels of abstraction is combined. From the experiments conducted in this work, we observe that results suggest that it is best to maintain a top-down approach, where panoptic predictions are extended with part-level predictions.

To provide a foundation for future research, the code and data used to realize this work are shared with the research community.

Acknowledgements We thank the authors of EfficientPS [42] and PolyTransform [29] for providing us with the predictions by their networks. This work is supported by Eindhoven Engine, NXP Semiconductors and Brainport Eindhoven, and partly funded by the Netherlands Organization for Scientific Research (NWO) in the context of the i-CAVE programme.

References

[1] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 6

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *ECCV*, 2018. 6

[4] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014. 2, 3, 4, 5, 6

[5] 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 *CVPR*, 2020. 1, 2

[6] 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 *CVPR*, 2016. 2, 3, 4, 6

[7] Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology*, 24(6):663–676, 2019. 2

[8] Daan de Geus, Panagiotis Meletis, and Gijs Dubbelman. Single network panoptic segmentation for street scene understanding. In *IEEE Intelligent Vehicles Symposium (IV)*, 2019. 2

[9] Daan de Geus, Panagiotis Meletis, and Gijs Dubbelman. Fast Panoptic Segmentation Network. *IEEE Robotics and Automation Letters*, 5(2):1742–1749, 2020. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[11] Jian Dong, Qiang Chen, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. Towards unified human parsing and pose estimation. In *CVPR*, 2014. 2

[12] Jian Dong, Qiang Chen, Wei Xia, Zhongyang Huang, and Shuicheng Yan. A deformable mixture parsing model with parselets. In *ICCV*, 2013. 2

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 2, 3, 4, 5, 6

[14] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *CVPR*, 2019. 2

[15] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *ECCV*, 2018. 1, 2, 3, 4

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 6

[17] Rui Hou, Jie Li, Arjun Bhargava, Allan Raventos, Vitor Guizilini, Chao Fang, Jerome Lynch, and Adrien Gaidon. Real-time panoptic segmentation from dense detections. In *CVPR*, 2020. 2

[18] Yalong Jiang and Zheru Chi. A CNN Model for Semantic Person Part Segmentation With Capacity Optimization. *IEEE Transactions on Image Processing*, 28(5):2465–2478, 2018. 2

[19] Yalong Jiang and Zheru Chi. A CNN Model for Human Parsing Based on Capacity Optimization. *Applied Sciences*, 9(7):1330, 2019. 2

[20] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic Feature Pyramid Networks. In *CVPR*, 2019. 1, 2

[21] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic Segmentation. In *CVPR*, 2019. 1, 2, 3, 4, 5, 6

[22] Lubor Ladicky, Philip HS Torr, and Andrew Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *CVPR*, 2013. 2

[23] Justin Lazarow, Kwonjoon Lee, Kunyu Shi, and Zhuowen Tu. Learning Instance Occlusion for Panoptic Segmentation. In *CVPR*, 2020. 2

[24] Jianshu Li, Jian Zhao, Yunpeng Chen, Sujoy Roy, Shuicheng Yan, Jiashi Feng, and Terence Sim. Multi-human parsing machines. In *ACM Multimedia Conference on Multimedia Conference*, 2018. 2

[25] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[26] Qizhu Li, Anurag Arnab, and Philip HS Torr. Holistic, instance-level human parsing. In *BMVC*, 2017. 1, 2

[27] Qizhu Li, Xiaojuan Qi, and Philip H. S. Torr. Unifying training and inference for panoptic segmentation. In *CVPR*, 2020. 1, 2

[28] 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. 2

[29] Justin Liang, Namdar Homayounfar, Wei-Chiu Ma, Yuwen Xiong, Rui Hu, and Raquel Urtasun. PolyTransform: Deep Polygon Transformer for Instance Segmentation. In *CVPR*, 2020. 6, 8

[30] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):871–885, 2018. 2, 3, 4

[31] Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic Object Parsing with Graph LSTM. In *ECCV*, 2016. 2

[32] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. In *ICCV*, 2015. 2

[33] Jinpeng Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, and Lu Yuan. Face Parsing with RoI Tanh-Warping. In *CVPR*, 2019. 2

[34] Kevin Lin, Lijuan Wang, Kun Luo, Yinpeng Chen, Zicheng Liu, and Ming-Ting Sun. Cross-Domain Complementary Learning Using Pose for Multi-Person Part Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. 2

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 3, 6

[36] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. An end-to-end network for panoptic segmentation. In *CVPR*, 2020. 2

[37] Si Liu, Yao Sun, Defa Zhu, Guanghui Ren, Yu Chen, Jiashi Feng, and Jizhong Han. Cross-domain Human Parsing via Adversarial Feature and Label Adaptation. In *AAAI*, 2018. 2

[38] Zhao Liu, Jianke Zhu, Jiajun Bu, and Chun Chen. A survey of human pose estimation: The body parts parsing based methods. *Journal of Visual Communication and Image Representation*, 32:10–19, 2015. 2

[39] Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian parsing via deep decompositional network. In *ICCV*, 2013. 2

[40] Xianghui Luo, Zhuo Su, Jiaming Guo, Gengwei Zhang, and Xiangjian He. Trusted Guidance Pyramid Network for Human Parsing. In *ACM Multimedia Conference on Multimedia Conference*, 2018. 2

[41] Umberto Michieli, Edoardo Borsato, Luca Rossi, and Pietro Zanuttigh. GMNet: Graph Matching Network for Large Scale Part Semantic Segmentation in the Wild. In *ECCV*, 2020. 1, 2, 5

[42] Rohit Mohan and Abhinav Valada. EfficientPS: Efficient panoptic segmentation. *International Journal of Computer Vision*, 2021. 1, 2, 6, 8

[43] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *CVPR*, 2014. 2, 4, 5, 6

[44] Lorenzo Porzi, Samuel Rota Bulo, Aleksander Colovic, and Peter Kontschieder. Seamless scene segmentation. In *CVPR*, 2019. 1, 2

[45] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution. *arXiv preprint arXiv:2006.02334*, 2020. 6

[46] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. Devil in the details: Towards accurate single and multiple human parsing. In *AAAI*, 2019. 2

[47] Zhuowen Tu, Xiangrong Chen, Alan L Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of Computer Vision*, 63(2):113–140, 2005. 2

[48] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. Joint object and part segmentation using deep learned potentials. In *ICCV*, 2015. 2

[49] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. UPSNet: A unified panoptic segmentation network. In *CVPR*, 2019. 1, 2, 6

[50] Lu Yang, Qing Song, Zhihui Wang, and Ming Jiang. Parsing R-CNN for Instance-Level Human Analysis. In *CVPR*, 2019. 2

[51] 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. *arXiv preprint arXiv:1902.05093*, 2019. 2

[52] Yibo Yang, Hongyang Li, Xia Li, Qijie Zhao, Jianlong Wu, and Zhouchen Lin. SOGNet: Scene Overlap Graph Network for Panoptic Segmentation. In *AAAI*, 2020. 2

[53] Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: joint object detection. In *CVPR*, 2012. 2

[54] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-Contextual Representations for Semantic Segmentation. In *ECCV*, 2020. 6

[55] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Muller, R. Manmatha, Mu Li, and Alexander Smola. ResNeSt: Split-Attention Networks. *arXiv preprint arXiv:2004.08955*, 2020. 6

[56] Jian Zhao, Jianshu Li, Yu Cheng, Terence Sim, Shuicheng Yan, and Jiashi Feng. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In *ACM Multimedia Conference on Multimedia Conference*, 2018. 1, 2, 3, 4

[57] Yifan Zhao, Jia Li, Yu Zhang, and Yonghong Tian. Multi-Class Part Parsing With Joint Boundary-Semantic Awareness. In *ICCV*, 2019. 1, 2, 5, 6