**Паноптическая сегментация**

Alexander Kirillov1*,*2 Kaiming He1 Ross Girshick1 Carsten Rother2 Piotr Dolla´r1 1Facebook AI Research (FAIR) 2HCI/IWR, Heidelberg University, Germany

# Абстракт

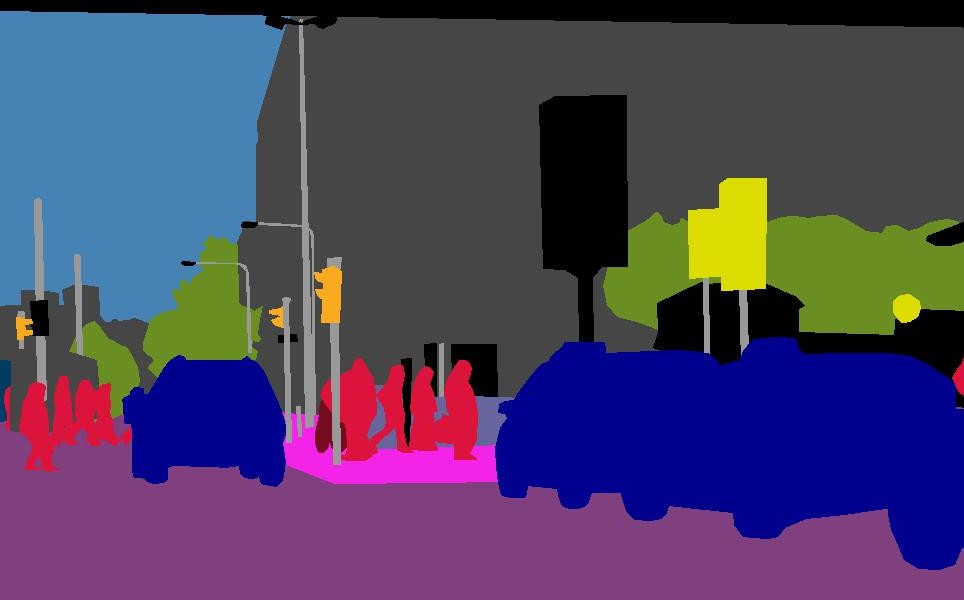
arXiv:1801.00868v3 [cs.CV] 10 Apr 2019

*Мы предлагаем и изучаем задачу, которую называем паноптической сегментацией (ПС). Паноптическая сегментация объединяет обычно различающиеся задачи семантической сегментации (присвоение метки класса каждому пикселю) и сегментации экземпляров (обнаружение и сегментация каждого экземпляра объекта). Предлагаемая задача требует генерации согласованной, полной и содержательной сегментации сцены, что является важным шагом на пути к созданию реальных систем машинного зрения. Хотя ранние работы в области компьютерного зрения были посвящены смежным задачам анализа изображений/сцен, в настоящее время они не пользуются популярностью, возможно, из-за отсутствия подходящих метрик или связанных с этим проблем распознавания. Для решения этой проблемы мы предлагаем новую метрику паноптического качества (ПК), которая позволяет интерпретировать и унифицировать показатели производительности для всех классов (вещей и объектов). Используя предложенную метрику, мы проводим тщательное исследование производительности человека и машины при решении задач ПС на трех существующих наборах данных, выявляя интересные закономерности в этой задаче. Цель нашей работы — возродить интерес сообщества к более унифицированному взгляду на сегментацию изображений.*

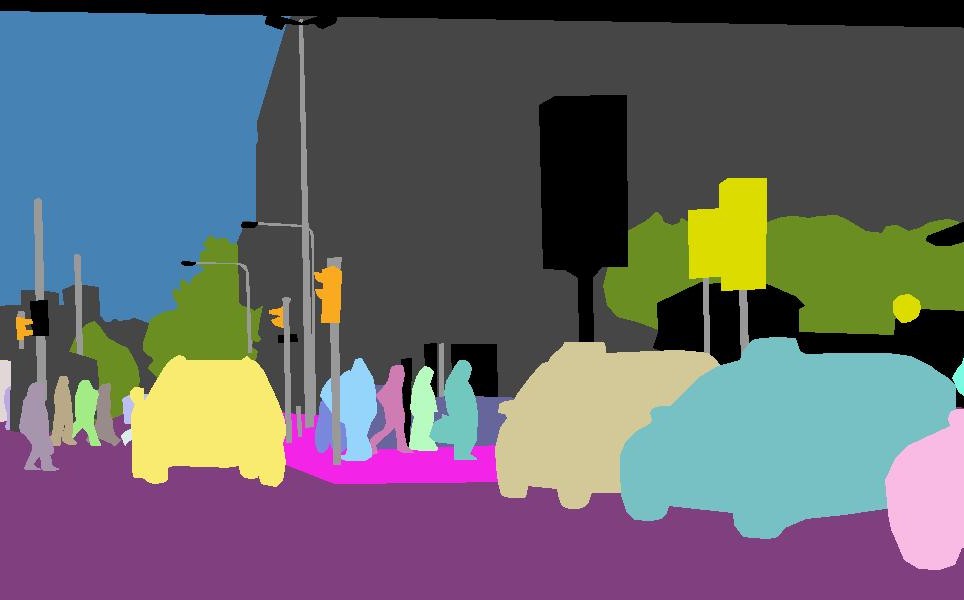
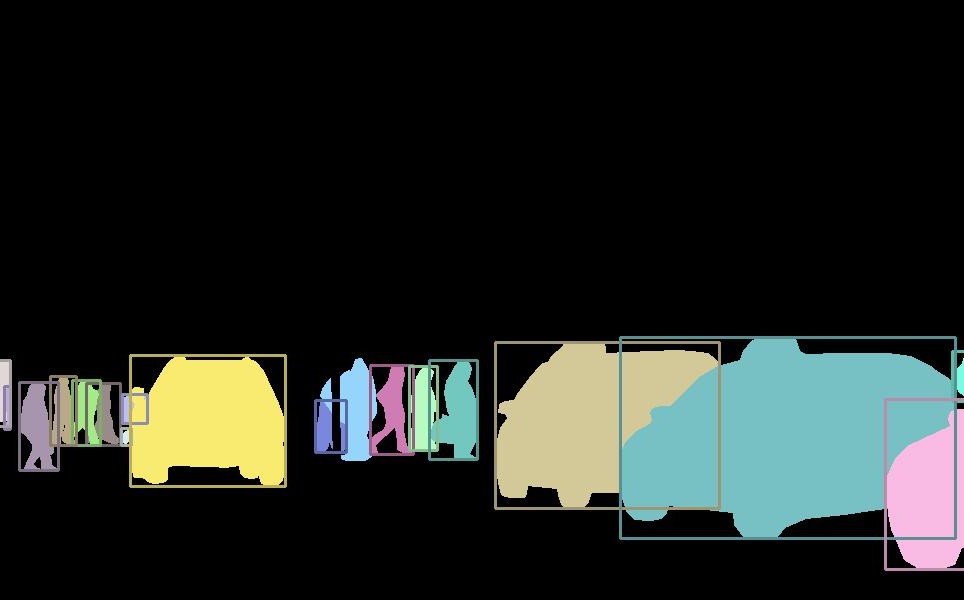
# Введение

На заре компьютерного зрения основное внимание уделялось *вещам* — поддающимся подсчету объектам, таким как люди, животные, инструменты. Ставя под сомнение целесообразность этой тенденции, Адельсон [1] повысил важность изучения систем, которые распознают *предметы* – аморфные области с похожей текстурой или материалом, такие как трава, небо, дорога. Это разделение между вещами и предметами сохраняется до сих пор, что отражается как в разделении задач визуального распознавания, так и в специализированных алгоритмах, разработанных для задач распознавания вещей и предметов.

Изучение материалов чаще всего формулируется как задача, известная как *семантическая сегментация* (см. рис. 1b). Поскольку объекты являются аморфными и неисчислимыми, эта задача определяется как простое присвоение метки класса каждому пикселю изображения (следует отметить, что семантическая сегментация рассматривает классы объектов как объекты). В противоположность этому, *изучение объектов* обычно формулируется как задача обнаружения объектов или *сегментации экземпляров*, где цель состоит в том, чтобы обнаружить каждый объект и обозначить его ограничивающей рамкой или маской сегментации соответственно (см. рисунок 1c). Хотя на первый взгляд они кажутся связанными, наборы данных, детали и метрики



(a) изображение (b) семантическая сегмент-ация



(c) сегментация экземпляров (d) паноптическая сегмент- ация

Рисунок 1: Для заданного (a) изображения мы показываем *эталонные данные* для: (b) семантической сегментации (метки классов для каждого пикселя), (c) сегментация экземпляров (маска для каждого объекта и метка класса) и (d) предлагаемая задача *паноптической сегментации* (метки класса и экземпляра для каждого пикселя). Задача ПС: (1) охватывает как классы предметов, так и классы вещей, (2) использует простой, но общий формат и (3) вводит единую метрику оценки для всех классов. Паноптическая сегментация обобщает как семантическую, так и сегментацию отдельных экземпляров, и мы ожидаем, что объединенная задача создаст новые проблемы и позволит разработать инновационные методы.

для этих двух задач визуального распознавания существенно различаются.

Разрыв между семантической сегментацией и сегментацией отдельных экземпляров привел к параллельному расколу в методах решения этих задач. Классификаторы вещей обычно строятся на основе полностью сверточных сетей [30] с дилатациями [52, 5], в то время как детекторы объектов часто используют предложения объектов [15] и основаны на регионах [37, 14]. В целом, за последнее десятилетие алгоритмический прогресс в решении этих задач был невероятным, однако, сосредоточившись на этих задачах изолированно, можно упустить из виду один важный момент.

Естественно, возникает вопрос: *возможно ли примирение между вещами и предметами*? И какова наиболее эффективная конструкция единой системы машинного зрения, генерирующей богатые и согласованные сегментации сцены? Эти вопросы особенно важны, учитывая их актуальность в реальных приложениях, таких как автономное вождение или дополненная реальность.

Интересно, что, хотя в настоящее время в работе доминируют семантическая сегментация и сегментация экземпляров, в эпоху до глубокого обучения существовал интерес к совместной задаче, описываемой с использованием различных названий, таких как *анализ сцены* [42], *анализ изображения* [43] или *целостное понимание сцены* [51]. Несмотря на свою практическую значимость, это общее направление в настоящее время не пользуется популярностью, возможно, из-за отсутствия подходящих метрик или проблем с распознаванием.

В нашей работе мы стремимся возродить это направление. Мы предлагаем задачу, которая: (1) *охватывает как классы вещей, так и классы предметов*, (2) *использует простой, но общий формат вывода* и (3) *вводит единую метрику оценки*. Чтобы четко разграничить результаты с предыдущими работами, мы будем называть полученную задачу *паноптической сегментацией* (ПС). Определение слова «паноптический» означает «включающий все видимое в одном ракурсе», в нашем контексте паноптический означает единый, глобальный подход к сегментации.

**Формат задачи**, который мы используем для паноптической сегментации, прост: каждому пикселю изображения необходимо присвоить семантическую метку и идентификатор экземпляра. Пиксели с одинаковой меткой и идентификатором принадлежат одному и тому же объекту; для меток объектов идентификатор экземпляра игнорируется. Визуализация представлена ​​на рисунке 1d. Этот формат уже использовался ранее, особенно методами, которые создают неперекрывающиеся сегментации экземпляров [18, 28, 2]. Мы используем его для нашей совместной задачи, которая включает в себя вещи и предметы.

Фундаментальным аспектом паноптической сегментации является **метрика задачи**, используемая для оценки. Хотя существует множество метрик, популярных как для семантической сегментации, так и для сегментации отдельных элементов, эти метрики лучше всего подходят либо для сегментации объектов, либо для сегментации отдельных элементов, соответственно, но не для обеих одновременно. Мы считаем, что использование непересекающихся метрик является одной из основных причин, по которой сообщество обычно изучает сегментацию объектов и вещей изолированно. Для решения этой проблемы в разделе 4 мы вводим метрику *паноптического качества* (ПК). ПК *прост* и *информативен*, и, что наиболее важно, может использоваться для измерения производительности как отдельных устройств, так и устройств в целом *единообразным* способом. Мы надеемся, что предложенный совместный показатель будет способствовать более широкому внедрению совместной задачи.

Задача панорамной сегментации включает в себя как семантическую, так и сегментацию отдельных объектов, но вводит новые алгоритмические проблемы. В отличие от семантической сегментации, она требует различения отдельных экземпляров объектов; это представляет собой проблему для полностью сверточных нейронных сетей. В отличие от сегментации экземпляров, сегменты объектов должны быть *неперекрывающимися*; это создает проблему для методов, основанных на регионах, которые работают с каждым объектом независимо. Создание согласованных сегментаций изображений, устраняющих несоответствия между объектами, является важным шагом на пути к практическому применению.

Поскольку как эталонные данные, так и формат алгоритма для ПС должны иметь одинаковую форму, мы можем провести детальное исследование согласованности *человеческого восприятия* при панорамной сегментации. Это позволяет нам более детально понять метрику PQ, включая подробный анализ производительности распознавания и сегментации, а также производительности объектов и вещей. Кроме того, измерение качества работы человека помогает лучше понять производительность машины. Это важно, поскольку позволит нам отслеживать насыщение производительности на различных наборах данных для анализа производительности.

Наконец, мы проводим предварительное исследование производительности машины для ПС. Для этого мы определяем простую и, вероятно, субоптимальную эвристику, которая объединяет выходные данные двух *независимых* систем для семантической и экземпляровой сегментации посредством ряда этапов постобработки, которые объединяют их выходные данные (по сути, сложная форма подавления немаксимумов). Наша эвристика устанавливает базовый уровень для ПС и позволяет нам понять основные алгоритмические проблемы, которые она представляет.

Мы изучаем производительность как человека, так и машины на трех популярных наборах данных для сегментации, содержащих аннотации как «вещи», так и «объекты». К ним относятся наборы данных Cityscapes [6], ADE20k [55] и Mapillary Vistas [35]. Для каждого из этих наборов данных мы получили результаты современных методов непосредственно от организаторов конкурса. В будущем мы расширим наш анализ на COCO [25], на котором аннотируются данные [4]. В совокупности наши результаты, полученные на этих наборах данных, формируют прочную основу для изучения как человеческой, так и машинной производительности при панорамной сегментации.

И COCO [25], и Mapillary Vistas [35] включили задачу паноптической сегментации в число направлений своих задач распознавания на ECCV 2018. Мы надеемся, что включение ПС наряду с задачами сегментации экземпляров и семантической сегментации в эти популярные наборы данных для распознавания поможет привести к более широкому внедрению предлагаемой совместной задачи.

# Связанные работы

Новые наборы данных и задачи играли ключевую роль на протяжении всей истории компьютерного зрения. Они помогают катализировать прогресс и позволяют совершать прорывы в нашей области, и, что не менее важно, они помогают нам измерять и признавать прогресс, которого добивается наше сообщество. Например, ImageNet [38] способствовал недавней популяризации методов глубокого обучения для визуального распознавания [20] и демонстрирует потенциальную преобразующую силу, которой могут обладать наборы данных и задачи. Наши цели при внедрении задачи паноптической сегментации схожи: бросить вызов нашему сообществу, стимулировать исследования в новых направлениях и способствовать как ожидаемым, так и неожиданным инновациям. Далее мы рассмотрим смежные задачи.

**Задачи обнаружения объектов.** Ранние работы по обнаружению лиц с использованием специальных наборов данных (например, [44, 46]) способствовали популяризации обнаружения объектов с помощью ограничивающих рамок. Позже наборы данных для обнаружения пешеходов [8] способствовали прогрессу в этой области. Набор данных PAS-CAL VOC [9] расширил задачу, включив в нее более разнообразный набор общих классов объектов на более сложных изображениях. Совсем недавно набор данных COCO [25] сместил задачу обнаружения в сторону сегментации экземпляров. Сформулировав эту задачу и предоставив высококачественный набор данных, COCO помог определить новое и захватывающее направление исследований и привел ко многим недавним прорывам в сегментации экземпляров [36, 24, 14]. Наши общие цели для паноптической сегментации схожи.

**Задачи семантической сегментации.** Наборы данных для семантической сегментации имеют богатую историю [39, 26, 9] и способствовали ключевым инновациям (например, были разработаны полностью сверточные сети [30] с использованием [26, 9]). Эти наборы данных содержат как классы предметов, так и классы вещей, но не различают отдельные экземпляры объектов. В последнее время в этой области появилось множество новых наборов данных для сегментации, включая Cityscapes [6], ADE20k [55] и Mapillary Vistas [35]. Эти наборы данных фактически поддерживают как семантическую, так и сегментацию экземпляров, и для каждого из них был выбран отдельный путь для решения этих двух задач. Важно отметить, что они содержат всю необходимую информацию для PS. Другими словами, *задачу паноптической сегментации можно начать с этих наборов данных без сбора новых данных.*

**Многозадачное обучение.** Благодаря успеху глубокого обучения во многих задачах визуального распознавания возник значительный интерес к подходам к *многозадачному обучению*, обладающим широкой компетенцией и способным решать множество разнообразных задач зрения в рамках одной системы [19, 32, 34]. Например, UberNet [19] решает множество задач визуального распознавания от низкого до высокого уровня, включая обнаружение объектов и семантическую сегментацию, используя одну сеть. Несмотря на значительный интерес к этой области, мы подчеркиваем, что паноптическая сегментация — это *не* многозадачная задача, а скорее единый, *унифицированный* подход к сегментации изображений. В частности, многозадачный режим позволяет получать независимые и потенциально противоречивые результаты для объектов и предметов, в то время как паноптическая сегментация требует единой согласованной сегментации сцены.

**Задачи совместной сегментации.** В эпоху до появления глубокого обучения существовал значительный интерес к генерации согласованных интерпретаций сцен. В основополагающей работе по анализу изображений [43] была предложена общая байесовская структура для совместного моделирования сегментации, обнаружения и распознавания. Позже подходы, основанные на графических моделях, изучали согласованную сегментацию предметов и вещей [51, 41, 42, 40]. Хотя эти методы имели общую мотивацию, не было согласованного определения задачи, и использовались различные форматы вывода и различные метрики оценки, включая отдельные метрики для оценки результатов по классам предметов и вещей. В последние годы это направление стало менее популярным, возможно, по этим причинам.

В нашей работе мы стремимся возродить это общее направление, но в отличие от более ранних работ, мы фокусируемся на самой задаче. В частности, как обсуждалось ранее, ПС: (1) рассматривает как классы вещей, так и классы предметов, (2) использует простой формат и (3) вводит единую метрику как для вещей, так и для предметов. В предыдущих работах по совместной сегментации использовались различные форматы и непересекающиеся метрики для оценки объектов и вещей. Методы, генерирующие непересекающиеся сегментации экземпляров [18, 3, 28, 2], используют тот же формат, что и ПС, но эти методы обычно рассматривают только классы объектов. Рассматривая как объекты, так и вещи, используя простой формат и вводя единую метрику, мы надеемся способствовать более широкому внедрению совместной задачи.

**Задача амодальной сегментации.** В [56] объекты аннотируются *амодально*: отмечается полный размер каждой области, а не только видимая. Наша работа сосредоточена на сегментации всех *видимых* областей, но расширение паноптической сегментации на амодальную среду является интересным направлением для будущих исследований.

# Формат паноптической сегментации

**Формат задачи.** Формат для паноптической сегментации легко определить. Дано заданное множество из *L* семантических классов, закодированных как L := {0*, . . . , L* − 1}, задача требует использования *алгоритма панорамной сегментации* для сопоставления каждого пикселя *i* изображения с парой пикселей (*li, zi*) ∈ L × N, где *li* обозначает семантический класс пикселя *i,* а *zi* обозначает идентификатор его экземпляра. *zi*’s группирует пиксели одного класса в отдельные сегменты. Аннотации эталонных данных кодируются одинаково. Неоднозначным или не относящимся к классу пикселям может быть присвоена специальная пустая метка; то есть не все пиксели должны иметь семантическую метку.

**Метки «вещи» и «предметы».** Семантический набор меток состоит из подмножествLSt и LTh, таких что L = LSt ∪ LTh и LSt ∩ LTh = ∅. Эти подмножества соответствуют меткам «вещи» и «предметы» соответственно. Когда пиксель помечен меткой *li* ∈ LSt, соответствующий ему идентификатор экземпляра *zi* не имеет значения. То есть, для классов «вещи» все пиксели принадлежат одному и тому же экземпляру (например, одному и тому же небу). В противном случае, все пиксели с одинаковым присвоением (*li, zi*), где *li* ∈ LTh, принадлежат одному и тому же экземпляру (например, одной и той же машине), и наоборот, все пиксели, принадлежащие одному экземпляру, должны иметь одинаковые (*li, zi*). Выбор классов «вещи» и «предметы» — это проектное решение, оставленное на усмотрение создателя набора данных, как и в предыдущих наборах данных.

**Связь с семантической сегментацией.** Формат задачи ПС является строгим обобщением формата для семантической сегментации. Действительно, обе задачи требуют присвоения семантической метки каждому пикселю изображения. Если в эталонных данных не указаны экземпляры, или все классы являются объектами, то форматы задач идентичны (хотя метрики задач различаются). Кроме того, включение классов объектов, которые могут иметь несколько экземпляров на изображении, отличает эти задачи.

**Связь с сегментацией экземпляров.** Задача сегментации экземпляров требует метода сегментации каждого экземпляра объекта на изображении. Однако она допускает перекрывающиеся сегменты, тогда как задача паноптической сегментации позволяет присваивать каждому пикселю только одну семантическую метку и один идентификатор экземпляра. Следовательно, для PS перекрытия по определению невозможны. В следующем разделе мы покажем, что это различие играет важную роль в оценке производительности.

**Оценки достоверности.** Подобно семантической сегментации, но в отличие от сегментации отдельных экземпляров, для паноптической сегментации нам *не* требуются оценки достоверности, связанные с каждым сегментом. Это делает задачу паноптической сегментации *симметричной* по отношению к людям и машинам: и те, и другие должны генерировать один и тот же тип аннотации изображений. Это также упрощает оценку согласованности действий человека при паноптической сегментации. Это отличается от сегментации отдельных экземпляров, которая не так легко поддается подобному исследованию, поскольку аннотаторы-люди не предоставляют явных оценок достоверности (хотя может быть измерена одна точка точности/полноты). Мы отмечаем, что оценки достоверности предоставляют последующим системам больше информации, что может быть полезно, поэтому в определенных условиях может быть желательно, чтобы алгоритм паноптической сегментации генерировал оценки достоверности.

# Panoptic Segmentation Metric

In this section we introduce a new metric for panoptic segmentation. We begin by noting that existing metrics are specialized for either semantic or instance segmentation and cannot be used to evaluate the joint task involving both stuff and thing classes. Previous work on joint segmenta- tion sidestepped this issue by evaluating stuff and thing per- formance using independent metrics (*e.g*. [[51](#_bookmark82), [41](#_bookmark72), [42](#_bookmark73), [40](#_bookmark71)]). However, this introduces challenges in algorithm develop- ment, makes comparisons more difficult, and hinders com- munication. We hope that introducing a unified metric for stuff and things will encourage the study of the unified task.

Before going into further details, we start by identifying the following desiderata for a suitable metric for PS:

**Completeness.** The metric should treat stuff and thing classes in a uniform way, capturing all aspects of the task.

**Interpretability.** We seek a metric with identifiable meaning that facilitates communication and understanding. **Simplicity.** In addition, the metric should be simple to define and implement. This improves transparency and al- lows for easy reimplementation. Related to this, the metric

should be efficient to compute to enable rapid evaluation.

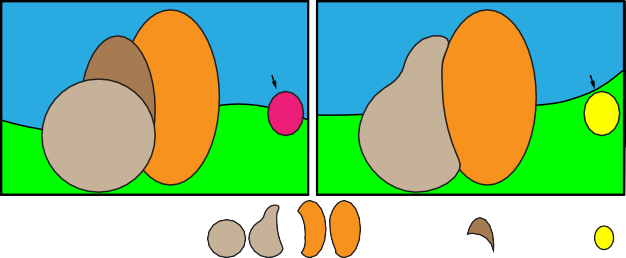
Guided by these principles, we propose a new *panoptic quality* (PQ) metric. PQ measures the quality of a predicted panoptic segmentation relative to the ground truth. It in- volves two steps: (1) segment matching and (2) PQ compu- tation given the matches. We describe each step next then return to a comparison to existing metrics.

## Segment Matching

We specify that a predicted segment and a ground truth segment can match only if their intersection over union (IoU) is strictly greater than 0.5. This requirement, together with the non-overlapping property of a panoptic segmenta- tion, gives a *unique matching*: there can be at most one pre- dicted segment matched with each ground truth segment.

**Theorem 1.** *Given a predicted and ground truth panoptic segmentation of an image, each ground truth segment can have at most one corresponding predicted segment with IoU strictly greater than 0.5 and vice verse.*

*Proof.* Let *g* be a ground truth segment and *p*1 and *p*2 be two predicted segments. By definition, *p*1 ∩ *p*2 = ∅ (they



sky

person person

person

sky

person

dog

person

person

grass

Person — TP: {

Ground Truth

ggrraassss

,

}; FN: { }; FP: { }

Prediction

Figure 2: Toy illustration of ground truth and predicted panoptic segmentations of an image. Pairs of segments of the same color have IoU larger than 0.5 and are therefore matched. We show how the segments for the *person* class are partitioned into true positives *TP*, false negatives *FN* , and false positives *FP*.

Therefore, if IoU(*p*1*, g*) *>* 0*.*5, then IoU(*p*2*, g*) has to be smaller than 0.5. Reversing the role of *p* and *g* can be used to prove that only one ground truth segment can have IoU with a predicted segment strictly greater than 0.5. 

The requirement that matches must have IoU greater than 0.5, which in turn yields the unique matching theorem, achieves two of our desired properties. First, it is *simple* and efficient as correspondences are unique and trivial to obtain. Second, it is *interpretable* and easy to understand (and does not require solving a complex matching problem as is commonly the case for these types of metrics [[13](#_bookmark44), [50](#_bookmark81)]). Note that due to the uniqueness property, for IoU *>* 0.5, any reasonable matching strategy (including greedy and op- timal) will yield an identical matching. For smaller IoU other matching techniques would be required; however, in the experiments we will show that lower thresholds are un-

necessary as matches with IoU ≤ 0.5 are rare in practice.

## PQ Computation

We calculate PQ for each class independently and aver- age over classes. This makes PQ insensitive to class im- balance. For each class, the unique matching splits the predicted and ground truth segments into three sets: true positives (*TP*), false positives (*FP* ), and false negatives (*FN* ), representing matched pairs of segments, unmatched predicted segments, and unmatched ground truth segments, respectively. An example is illustrated in Figure [2](#_bookmark8). Given these three sets, PQ is defined as:

Σ

do not overlap). Since |*pi* ∪ *g*| ≥ |*g*|, we get the following:

PQ =

(*p,g*)∈*TP*

IoU(*p, g*)

*.* (1)

|*pi* ∩ *g*|

|*pi* ∩ *g*|

|*TP*| + 1 |*FP* | + 1 |*FN* |

IoU(*pi, g*) = ≤ for *i* ∈ {1*,* 2} *.* 2 2

|*pi* ∪ *g*|

|*TP*|

|*g*|

PQ is intuitive after inspection: 1 Σ

(*p,g*)∈*TP*

IoU(*p, g*)

Summing over *i*, and since |*p*1 ∩ *g*| + |*p*2 ∩ *g*| ≤ |*g*| due to

the fact that *p*1 ∩ *p*2 = ∅, we get:

is simply the average IoU of matched segments, while

1 |*FP* | + 1 |*FN* | is added to the denominator to penalize

2 2

IoU(*p , g*) + IoU(*p , g*) ≤ |*p*1 ∩ *g*| + |*p*2 ∩ *g*| ≤ 1 *.*

1 2

|*g*|

segments without matches. Note that all segments receive equal importance regardless of their area. Furthermore, if we multiply and divide PQ by the size of the *TP* set, then

PQ can be seen as the multiplication of a *segmentation qual- ity* (SQ) term and a *recognition quality* (RQ) term:

Σ

to estimate a precision/recall curve. Note that while confi- dence scores are quite natural for object detection, they are

PQ =

(*p,g*)∈*TP*

IoU(*p, g*)

*×*

*|TP|*

*.* (2)

not used for semantic segmentation. Hence, AP cannot be

used for measuring the output of semantic segmentation, or

*|TP|*

`segmentatio˛n¸quality (SQ)x

*|TP|* + 1 *|FP|* + 1 *|FN|*

` recognition˛q¸uality (RQ) x

2

2

likewise of PS (see also the discussion of confidences in §[3](#_bookmark4)).

**Panoptic quality.** PQ treats all classes (stuff and things)

Written this way, RQ is the familiar *F*1 score [[45](#_bookmark76)] widely used for quality estimation in detection settings [[33](#_bookmark64)]. SQ is simply the average IoU of matched segments. We find the decomposition of PQ = SQ × RQ to provide insight for analysis. We note, however, that the two values are not inde-

pendent since SQ is measured only over matched segments. Our definition of PQ achieves our desiderata. It measures performance of all classes in a uniform way using a simple and interpretable formula. We conclude by discussing how

we handle void regions and groups of instances [[25](#_bookmark56)].

**Void labels.** There are two sources of void labels in the ground truth: (a) out of class pixels and (b) ambiguous or unknown pixels. As often we cannot differentiate these two cases, we don’t evaluate predictions for void pixels. Specifi- cally: (1) during matching, all pixels in a predicted segment that are labeled as void in the ground truth are removed from the prediction and do not affect IoU computation, and (2) after matching, unmatched predicted segments that contain a fraction of void pixels over the matching threshold are re- moved and do not count as false positives. Finally, outputs may also contain void pixels; these do not affect evaluation.

**Group labels.** A common annotation practice [[6](#_bookmark37), [25](#_bookmark56)] is to use a group label instead of instance ids for adjacent in- stances of the same semantic class if accurate delineation of each instance is difficult. For computing PQ: (1) during matching, group regions are not used, and (2) after match- ing, unmatched predicted segments that contain a fraction of pixels from a group of the same class over the matching threshold are removed and do not count as false positives.

## Comparison to Existing Metrics

We conclude by comparing PQ to existing metrics for semantic and instance segmentation.

**Semantic segmentation metrics.** Common metrics for semantic segmentation include pixel accuracy, mean accu- racy, and IoU [[30](#_bookmark60)]. These metrics are computed based only on pixel outputs/labels and completely ignore object-level labels. For example, IoU is the ratio between correctly pre- dicted pixels and total number of pixels in either the predic- tion or ground truth for each class. As these metrics ignore instance labels, they are not well suited for evaluating thing classes. Finally, please note that IoU for semantic segmen- tation is distinct from our segmentation quality (SQ), which is computed as the average IoU over *matched segments*.

**Instance segmentation metrics.** The standard metric for instance segmentation is Average Precision (AP) [[25](#_bookmark56), [13](#_bookmark44)]. AP requires each object segment to have a confidence score

in a uniform way. We note that while decomposing PQ into SQ and RQ is helpful with interpreting results, PQ is *not* a combination of semantic and instance segmentation metrics. Rather, SQ and RQ are computed for every class (stuff and things), and measure segmentation and recogni- tion quality, respectively. PQ thus unifies evaluation over all

classes. We support this claim with rigorous experimental evaluation of PQ in §[7](#_bookmark23), including comparisons to IoU and AP for semantic and instance segmentation, respectively.

# Panoptic Segmentation Datasets

To our knowledge only three public datasets have both dense semantic and instance segmentation annotations: Cityscapes [[6](#_bookmark37)], ADE20k [[55](#_bookmark86)], and Mapillary Vistas [[35](#_bookmark66)]. We use all three datasets for panoptic segmentation. In ad- dition, in the future we will extend our analysis to COCO

[[25](#_bookmark56)] on which stuff has been recently annotated [[4](#_bookmark35)][1](#_bookmark11).

**Cityscapes** [[6](#_bookmark37)] has 5000 images (2975 train, 500 val, and 1525 test) of ego-centric driving scenarios in urban settings. It has dense pixel annotations (97% coverage) of 19 classes among which 8 have instance-level segmentations.

**ADE20k** [[55](#_bookmark86)] has over 25k images (20k train, 2k val, 3k test) that are densely annotated with an open-dictionary label set. For the 2017 Places Challenge[2](#_bookmark12), 100 thing and 50 stuff classes that cover 89% of all pixels are selected. We use this closed vocabulary in our study.

**Mapillary Vistas** [[35](#_bookmark66)] has 25k street-view images (18k train, 2k val, 5k test) in a wide range of resolutions. The ‘research edition’ of the dataset is densely annotated (98% pixel coverage) with 28 stuff and 37 thing classes.

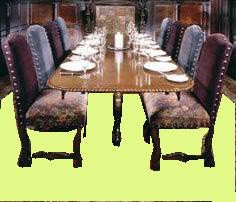
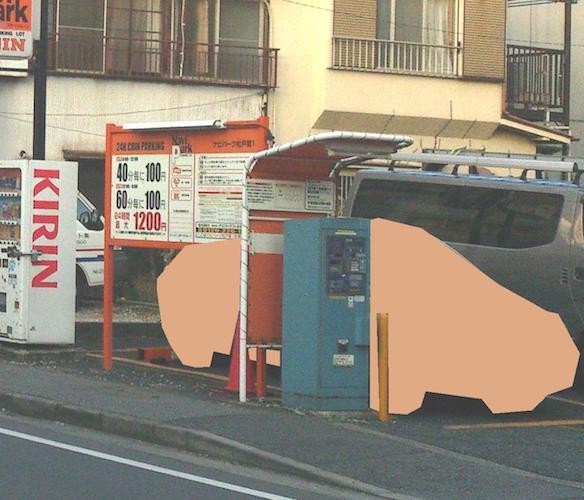
# Human Consistency Study

One advantage of panoptic segmentation is that it en- ables measuring human annotation consistency. Aside from this being interesting as an end in itself, human consistency studies allow us to understand the task in detail, including details of our proposed metric and breakdowns of human consistency along various axes. This gives us insight into intrinsic challenges posed by the task without biasing our

analysis by algorithmic choices. Furthermore, human stud- ies help ground machine performance (discussed in §[7](#_bookmark23)) and allow us to calibrate our understanding of the task.

1COCO instance segmentations contain overlaps. We collected depth ordering for all pairs of overlapping instances in COCO to resolve these overlaps: <http://cocodataset.org/#panoptic-2018>.

2[http://placeschallenge.csail.mit.edu](http://placeschallenge.csail.mit.edu/)



floor

rug ✔

building

tram ✔

Figure 3: **Segmentation flaws.** Images are zoomed and cropped. Top row (Vistas image): both annotators identify the object as a car, however, one splits the car into two cars. Bottom row (Cityscapes image): the segmentation is genuinely ambiguous.

Figure 4: **Classification flaws.** Images are zoomed and cropped. Top row (ADE20k image): simple misclassification. Bottom row (Cityscapes image): the scene is extremely difficult, tram is the correct class for the segment. Many errors are difficult to resolve.

PQ

Cityscapes 69.7

ADE20k 67.1

Vistas 57.5

PQSt 71.3

70.3

62.6

PQTh 67.4

65.9

53.4

SQ 84.2

85.8

79.5

SQSt SQTh

84.4 83.9

85.5 85.9

81.6 77.9

RQ 82.1

78.0

71.4

RQSt RQTh

83.4 80.2

82.4 76.4

76.0 67.7

PQS PQM PQL SQS SQM SQL RQS RQM RQL Cityscapes 35.1 62.3 84.8 67.8 81.0 89.9 51.5 76.5 94.1

ADE20k 49.9 69.4 79.0 78.0 84.0 87.8 64.2 82.5 89.8

Vistas 35.6 47.7 69.4 70.1 76.6 83.1 51.5 62.3 82.6

Table 1: **Human consistency for stuff *vs*. things**. Panoptic, seg- mentation, and recognition quality (PQ, SQ, RQ) averaged over classes (PQ=SQ*×*RQ per class) are reported as percentages. Per- haps surprisingly, we find that human consistency on each dataset is relatively similar for both stuff and things.

**Human annotations.** To enable human consistency anal- ysis, dataset creators graciously supplied us with 30 doubly annotated images for Cityscapes, 64 for ADE20k, and 46 for Vistas. For Cityscapes and Vistas, the images are an- notated independently by different annotators. ADE20k is annotated by a single well-trained annotator who labeled the same set of images with a gap of six months. To mea- sure panoptic quality (PQ) for human annotators, we treat one annotation for each image as ground truth and the other as the prediction. Note that the PQ is symmetric w.r.t. the ground truth and prediction, so order is unimportant.

**Human consistency.** First, Table [1](#_bookmark16) shows human con- sistency on each dataset, along with the decomposition of PQ into segmentation quality (SQ) and recognition qual- ity (RQ). As expected, humans are not perfect at this task, which is consistent with studies of annotation quality from [[6](#_bookmark37), [55](#_bookmark86), [35](#_bookmark66)]. Visualizations of human segmentation and clas- sification errors are shown in Figures [3](#_bookmark13) and [4](#_bookmark14), respectively. We note that Table [1](#_bookmark16) establishes a measure of annotator

agreement on each dataset, *not* an upper bound on human consistency. We further emphasize that numbers are not comparable across datasets and should not be used to assess dataset quality. The number of classes, percent of annotated pixels, and scene complexity vary across datasets, each of which significantly impacts annotation difficulty.

Table 2: **Human consistency *vs*. scale**, for small (S), medium (M) and large (L) objects. Scale plays a large role in determining hu- man consistency for panoptic segmentation. On large objects both SQ and RQ are above 80 on all datasets, while for small objects RQ drops precipitously. SQ for small objects is quite reasonable.

**Stuff *vs*. things.** PS requires segmentation of both stuff and things. In Table [1](#_bookmark16) we also show PQSt and PQTh which is the PQ averaged over stuff classes and thing classes, re- spectively. For Cityscapes and ADE20k human consistency for stuff and things are close, on Vistas the gap is a bit larger. Overall, this implies stuff and things have similar difficulty, although thing classes are somewhat harder. In Figure [5](#_bookmark18) we show PQ for every class in each dataset, sorted by PQ. Observe that stuff and things classes distribute fairly evenly. This implies that the proposed metric strikes a good balance and, indeed, is successful at unifying the stuff and things segmentation tasks without either dominating the error.

**Small *vs*. large objects.** To analyze how PQ varies with object size we partition the datasets into small (S), medium (M), and large (L) objects by considering the smallest 25%, middle 50%, and largest 25% of objects in each dataset, respectively. In Table [2](#_bookmark17), we see that for large objects hu- man consistency for all datasets is quite good. For small objects, RQ drops significantly implying human annotators often have a hard time finding small objects. However, if a small object is found, it is segmented relatively well.

**IoU threshold.** By enforcing an overlap greater than 0.5 IoU, we are given a unique matching by Theorem [1](#_bookmark7). How- ever, is the 0.5 threshold reasonable? An alternate strategy is to use no threshold and perform the matching by solving a maximum weighted bipartite matching problem [[47](#_bookmark78)]. The

**Cityscapes**

wall

terrain

bicycle

train

pole

motorcycle

fence

traffic light

rider

person

truck

car

traffic sign

sidewalk

bus

vegetation

sky

building

road

**ADE20k** 1

0.5

**CDF**

0

0.25 0.50 0.75

**IoU**

Figure 6: **Cumulative density functions of overlaps** for matched segments in three datasets when matches are computed by solv- ing a maximum weighted bipartite matching problem [[47](#_bookmark78)]. After matching, less than 16% of matched objects have IoU below 0.5.

Cityscapes ADE20k

Vistas

0 0.5 1

**PQ**

70

threshold=0.25  threshold=0.5  threshold=0.75

60

7

68.1

67.1

61.6

59.2

60.2

57.5

49.2

69.

71.3

**PQ**

50

Things Stuff

0 0.5 1 0 0.5 1

runway

blind toilet

refrigerator kitchen island bridge television

bed canopy sky chandelier river towel painting desk building mountain plaything wall sconce cushion ashcan grass clock stove chair

swimming pool

car hood pot blanket light

coffee table tree airplane glass bicycle poster bottle traffic light countertop flower

rock

streetlight basket step

shelf tray earth

**Vistas**

sky

rail-track road vegetation

ground-animal service-lane wheeled-slow building

car-mount on-rails bicyclist sidewalk car

fire-hydrant ego-vehicle truck

bus

crosswalk-zebra

terrain

junction-box

cctv-camera

bike-rack crosswalk-plain traffic-sign-frame curb-cut

utility-pole

guard-rail

bird

mountain

manhole

wall

pole

billboard

back

parking

ist

curb

motorcycl

fence

n

catch-basi

snow

t

street-ligh

bicycle

general

banner

e

trash-can

motorcycl

person

bike-lane

t

ier

traffic-ligh

other-barr

bridge

front

**PQ PQ**

Figure 5: **Per-Class Human consistency, sorted by PQ**. Thing classes are shown in red, stuff classes in orange (for ADE20k every other class is shown, classes without matches in the dual- annotated tests sets are omitted). Things and stuff are distributed fairly evenly, implying PQ balances their performance.

optimization will return a matching that maximizes the sum

**Cityscapes ADE20k Vistas**

Figure 7: **Human consistency for different IoU thresholds.** The difference in PQ using a matching threshold of 0.25 *vs*. 0.5 is rel- atively small. For IoU of 0.25 matching is obtained by solving a maximum weighted bipartite matching problem. For a threshold greater than 0.5 the matching is unique and much easier to obtain.

 Segmentation Quality (SQ)  Recognition Quality (RQ)



α=1

α= α=

1 1

2 4

68.7

0

78.

84.2

71.4

59.4

4

81.

79.5

7

85.

85.8

89.2

82.

1

72.5

2 4

1 1

α= α=

α=1

85

75

65

of IoUs of the matched segments. We perform the match- 1 1

ing using this optimization and plot the cumulative density functions of the match overlaps in Figure [6](#_bookmark20). Less than 16%

**Cityscapes**

**ADE20k**

α=1 α= α=

**Vistas**

2 4

of the matches have IoU overlap less than 0.5, indicating that relaxing the threshold should have minor effect.

To verify this intuition, in Figure [7](#_bookmark21) we show PQ com- puted for different IoU thresholds. Notably, the difference in PQ for IoU of 0.25 and 0.5 is relatively small, especially compared to the gap between IoU of 0.5 and 0.75, where the change in PQ is larger. Furthermore, many matches at lower IoU are false matches. Therefore, given that the matching for IoU of 0.5 is not only unique, but also simple and intu- itive, we believe that the default choice of 0.5 is reasonable.

**SQ *vs*. RQ balance.** Our RQ definition is equivalent to the *F*1 score. However, other choices are possible. Inspired by the generalized *Fβ* score [[45](#_bookmark76)], we can introduce a param- eter *α* that enables tuning the penalty for recognition errors:

Figure 8: **SQ *vs*. RQ** for different *α*, see ([3](#_bookmark24)). Lowering *α* reduces the penalty of unmatched segments and thus increases the reported RQ (SQ is not affected). We use *α* of 0.5 throughout but by tuning *α* one can balance the influence of SQ and RQ in the final metric.

of PS *vs*. RQ on the final PQ metric. In Figure [8](#_bookmark22) we show SQ and RQ for various *α*. The default *α* strikes a good bal- ance between SQ and RQ. In principle, altering *α* can be used to balance the influence of segmentation and recogni- tion errors on the final metric. In a similar spirit, one could also add a parameter *β* to balance influence of FPs *vs*. FNs.

# Machine Performance Baselines

We now present simple machine baselines for panoptic

RQ*α* = |*TP*|

|*TP*| + *α*|*FP* | + *α*|*FN* |

*.* (3)

segmentation. We are interested in three questions: (1) How do heuristic combinations of top-performing instance and semantic segmentation systems perform on panoptic seg-

By default *α* is 0.5. Lowering *α* reduces the penalty of unmatched segments and thus increases RQ (SQ is not af- fected). Since PQ=SQ×RQ, this changes the relative effect

mentation? (2) How does PQ compare to existing metrics like AP and IoU? (3) How do the machine results compare to the human results that we presented previously?

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cityscapes**  Mask R-CNN+COCO [[14](#_bookmark45)] | AP APNO  **36.4 33.1** | | PQTh SQTh RQTh **Cityscapes**  **54.0 79.4 67.8** PSPNet multi-scale [[54](#_bookmark85)] | | | | | IoU  **80.6** | PQSt SQSt RQSt  **66.6 82.2 79.3** | | |
| Mask R-CNN [[14](#_bookmark45)] | 31.5 | 28.0 | 49.6 | 78.7 | 63.0 | PSPNet single-scale [[54](#_bookmark85)] | | 79.6 | 65.2 | 81.6 | 78.0 |
| **ADE20k** | AP | APNO | PQTh | SQTh | RQTh | **ADE20k** | | IoU | PQSt | SQSt | RQSt |
| Megvii [[31](#_bookmark62)] | **30.1** | **24.8** | **41.1** | **81.6** | **49.6** |  | CASIA IVA JD [[12](#_bookmark43)] | **32.3** | **27.4** | **61.9** | **33.7** |
| G-RMI [[10](#_bookmark41)] | 24.6 | 20.6 | 35.3 | 79.3 | 43.2 |  | G-RMI [[11](#_bookmark42)] | 30.6 | 19.3 | 58.7 | 24.3 |

Table 3: **Machine results on instance segmentation** (stuff classes ignored). Non-overlapping predictions are obtained using the pro- posed heuristic. APNO is AP of the non-overlapping predictions. As expected, removing overlaps harms AP as detectors benefit from predicting multiple overlapping hypotheses. Methods with better AP also have better APNO and likewise improved PQ.

**Algorithms and data.** We want to understand panoptic segmentation in terms of existing well-established methods. Therefore, we create a basic PS system by applying reason- able heuristics (described shortly) to the output of existing top instance and semantic segmentation systems.

We obtained algorithm output for three datasets. For *Cityscapes*, we use the val set output generated by the cur- rent leading algorithms (PSPNet [[54](#_bookmark85)] and Mask R-CNN [[14](#_bookmark45)] for semantic and instance segmentation, respectively). For *ADE20k*, we received output for the winners of both the semantic [[12](#_bookmark43), [11](#_bookmark42)] and instance [[31](#_bookmark62), [10](#_bookmark41)] segmentation tracks on a 1k subset of test images from the 2017 Places Challenge. For *Vistas*, which is used for the LSUN’17 Seg- mentation Challenge, the organizers provide us with 1k test images and results from the winning entries for the instance and semantic segmentation tracks [[29](#_bookmark61), [53](#_bookmark84)].

Using this data, we start by analyzing PQ for the instance and semantic segmentation tasks separately, and then exam- ine the full panoptic segmentation task. Note that our ‘base- lines’ are very powerful and that simpler baselines may be more reasonable for fair comparison in papers on PS.

**Instance segmentation.** Instance segmentation algo- rithms produce overlapping segments. To measure PQ, we must first resolve these overlaps. To do so we develop a sim- ple non-maximum suppression (NMS)-like procedure. We first sort the predicted segments by their confidence scores and remove instances with low scores. Then, we iterate over sorted instances, starting from the most confident. For each instance we first remove pixels which have been assigned to previous segments, then, if a sufficient fraction of the seg- ment remains, we accept the non-overlapping portion, oth- erwise we discard the entire segment. All thresholds are se- lected by grid search to optimize PQ. Results on Cityscapes and ADE20k are shown in Table [3](#_bookmark25) (Vistas is omitted as it only had one entry to the 2017 instance challenge). Most importantly, AP and PQ track closely, and we expect im- provements in a detector’s AP will also improve its PQ.

**Semantic segmentation.** Semantic segmentations have no overlapping segments by design, and therefore we can

Table 4: **Machine results on semantic segmentation** (thing classes ignored). Methods with better mean IoU also show better PQ results. Note that G-RMI has quite low PQ. We found this is because it hallucinates many small patches of classes not present in an image. While this only slightly affects IoU which counts *pixel* errors it severely degrades PQ which counts *instance* errors.

directly compute PQ. In Table [4](#_bookmark27) we compare mean IoU, a standard metric for this task, to PQ. For Cityscapes, the PQ gap between methods corresponds to the IoU gap. For ADE20k, the gap is much larger. This is because whereas IoU counts correctly predicted pixel, PQ operates at the level of instances. See the Table [4](#_bookmark27) caption for details.

**Panoptic segmentation.** To produce algorithm outputs for PS, we start from the non-overlapping instance seg- ments from the NMS-like procedure described previously. Then, we combine those segments with semantic segmenta- tion results by resolving any overlap between thing and stuff classes in favor of the thing class (*i.e*., a pixel with a thing and stuff label is assigned the thing label and its instance id). This heuristic is imperfect but sufficient as a baseline.

Table [5](#_bookmark30) compares PQSt and PQTh computed on the com- bined (‘panoptic’) results to the performance achieved from the separate predictions discussed above. For these results we use the winning entries from each respective competi- tion for both the instance and semantic tasks. Since overlaps are resolved in favor of things, PQTh is constant while PQSt is slightly lower for the panoptic predictions. Visualizations of panoptic outputs are shown in Figure [9](#_bookmark28).

**Human *vs*. machine panoptic segmentation.** To com- pare human *vs*. machine PQ, we use the machine panoptic predictions described above. For human results, we use the dual-annotated images described in §[6](#_bookmark10) and use bootstrap- ping to obtain confidence intervals since these image sets

are small. These comparisons are imperfect as they use dif- ferent test images and are averaged over different classes (some classes without matches in the dual-annotated tests sets are omitted), but they can still give some useful signal. We present the comparison in Table [6](#_bookmark31). For SQ, ma- chines trail humans only slightly. On the other hand, ma- chine RQ is dramatically lower than human RQ, especially on ADE20k and Vistas. This implies that recognition, *i.e*., classification, is the main challenge for current methods. Overall, there is a significant gap between human and ma- chine performance. We hope that this gap will inspire future

research for the proposed panoptic segmentation task.

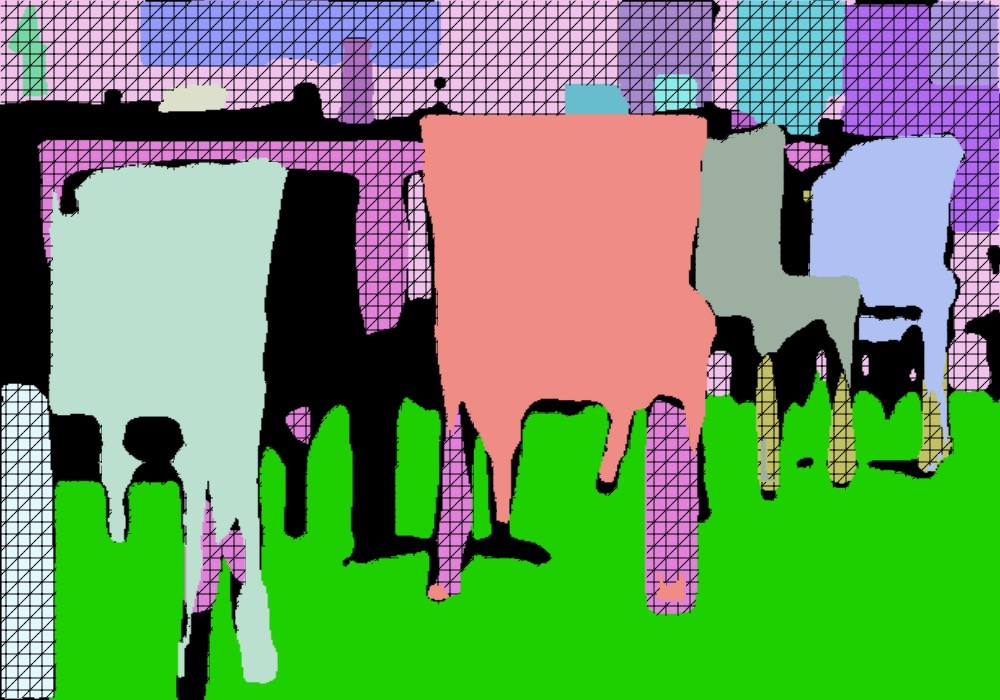
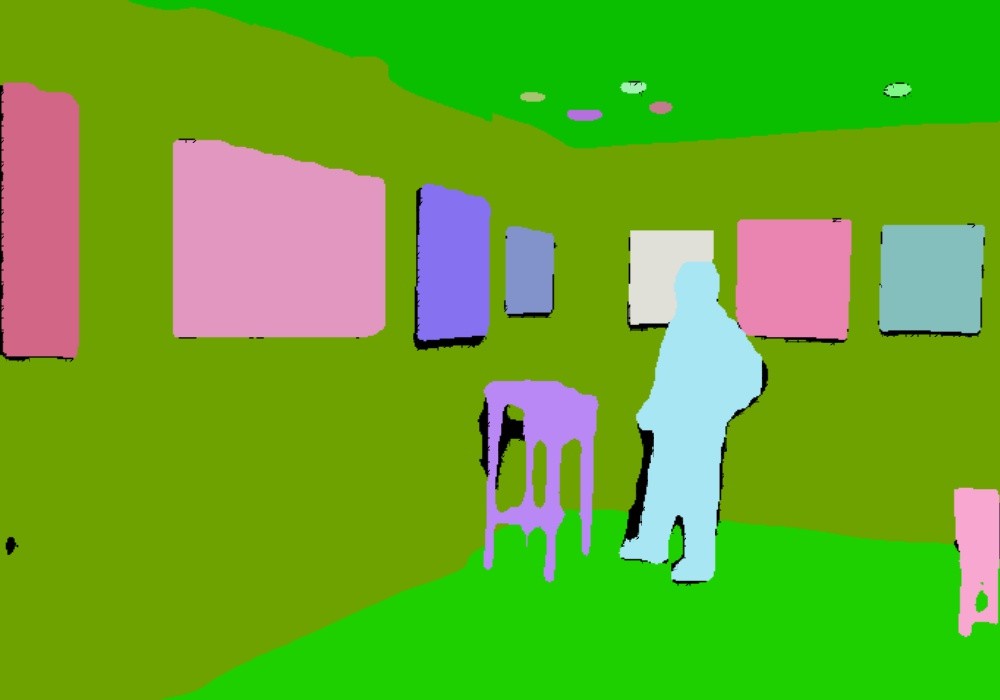
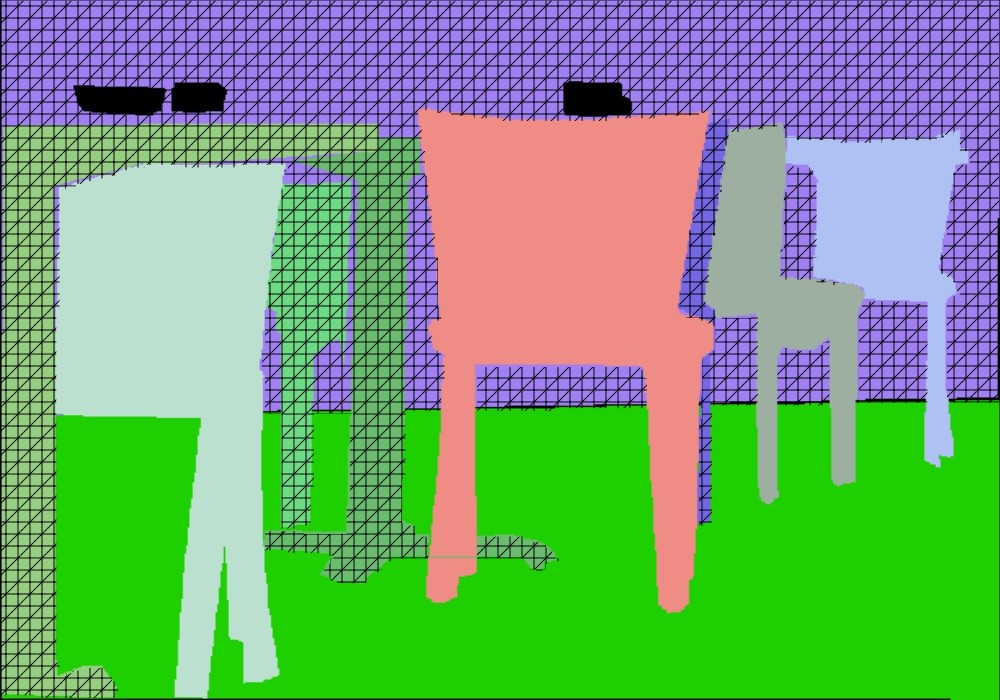


Figure 9: **Panoptic segmentation results** on Cityscapes (left two) and ADE20k (right three). Predictions are based on the merged outputs of state-of-the-art instance and semantic segmentation algorithms (see Tables [3](#_bookmark25) and [4](#_bookmark27)). Colors for matched segments (IoU*>*0.5) match (crosshatch pattern indicates unmatched regions and black indicates unlabeled regions). Best viewed in color and with zoom.

prediction

ground truth

**Cityscapes** PQ PQSt PQTh **Cityscapes** PQ SQ RQ PQSt PQTh

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| machine-separate | n/a | 66.6 | 54.0 |  | human | 69.6 | +2*.*5  −2*.*7 | 84.1 | +0*.*8  −0*.*8 | 82.0 | +2*.*7  −2*.*9 | 71.2 | +2*.*3  −2*.*5 | 67.4 | +4*.*6  −4*.*9 |
| machine-panoptic | 61.2 | 66.4 | 54.0 |  | machine | 61.2 |  | 80.9 |  | 74.4 |  | 66.4 |  | 54.0 |  |

**ADE20k** PQ PQSt PQTh **ADE20k** PQ SQ RQ PQSt PQTh

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| machine-separate | n/a | 27.4 | 41.1 |  | human | 67.6 | +2*.*0  −2*.*0 | 85.7 | +0*.*6  −0*.*6 | 78.6 | +2*.*1  −2*.*1 | 71.0 | +3*.*7  −3*.*2 | 66.4 | +2*.*3  −2*.*4 |
| machine-panoptic | 35.6 | 24.5 | 41.1 |  | machine | 35.6 |  | 74.4 |  | 43.2 |  | 24.5 |  | 41.1 |  |

**Vistas** PQ PQSt PQTh **Vistas** PQ SQ RQ PQSt PQTh

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| machine-separate | n/a | 43.7 | 35.7 |  | human | 57.7 | +1*.*9  −2*.*0 | 79.7 | +0*.*8  −0*.*7 | 71.6 | +2*.*2  −2*.*3 | 62.7 | +2*.*8  −2*.*8 | 53.6 | +2*.*7  −2*.*8 |
| machine-panoptic | 38.3 | 41.8 | 35.7 |  | machine | 38.3 |  | 73.6 |  | 47.7 |  | 41.8 |  | 35.7 |  |

Table 5: **Panoptic *vs*. independent predictions.** The ‘machine- separate’ rows show PQ of semantic and instance segmentation methods computed independently (see also Tables [3](#_bookmark25) and [4](#_bookmark27)). For ‘machine-panoptic’, we merge the non-overlapping thing and stuff predictions obtained from state-of-the-art methods into a true panoptic segmentation of the image. Due to the merging heuristic used, PQTh stays the same while PQSt is slightly degraded.

image

# Future of Panoptic Segmentation

Our goal is to drive research in novel directions by invit- ing the community to explore the new panoptic segmenta- tion task. We believe that the proposed task can lead to expected and unexpected innovations. We conclude by dis- cussing some of these possibilities and our future plans.

Motivated by simplicity, the PS ‘algorithm’ in this paper is based on the *heuristic* combination of outputs from top- performing instance and semantic segmentation systems. This approach is a basic first step, but we expect more inter- esting algorithms to be introduced. Specifically, we hope to see PS drive innovation in at least two areas: (1) Deeply in- tegrated end-to-end models that simultaneously address the

Table 6: **Human *vs*. machine performance.** On each of the con- sidered datasets human consistency is much higher than machine performance (approximate comparison, see text for details). This is especially true for RQ, while SQ is closer. The gap is largest on ADE20k and smallest on Cityscapes. Note that as only a small set of human annotations is available, we use bootstrapping and show the the 5th and 95th percentiles error ranges for human results.

dual stuff-and-thing nature of PS. A number of instance seg- mentation approaches including [[28](#_bookmark59), [2](#_bookmark33), [3](#_bookmark34), [18](#_bookmark49)] are designed to produce non-overlapping instance predictions and could serve as the foundation of such a system. (2) Since a PS cannot have overlapping segments, some form of higher- level ‘reasoning’ may be beneficial, for example, based on extending learnable NMS [[7](#_bookmark38), [16](#_bookmark47)] to PS. We hope that the panoptic segmentation task will invigorate research in these areas leading to exciting new breakthroughs in vision.

Finally we note that the panoptic segmentation task was featured as a challenge track by both the COCO [[25](#_bookmark56)] and Mapillary Vistas [[35](#_bookmark66)] recognition challenges and that the proposed task has already begun to gain traction in the com- munity (*e.g*. [[23](#_bookmark54), [48](#_bookmark79), [49](#_bookmark80), [27](#_bookmark58), [22](#_bookmark53), [21](#_bookmark52), [17](#_bookmark48)] address PS).

# References

1. E. H. Adelson. On seeing stuff: the perception of materials by hu- mans and machines. In *Human Vision and Electronic Imaging*, 2001. [1](#_bookmark0)
2. A. Arnab and P. H. Torr. Pixelwise instance segmentation with a dynamically instantiated network. In *CVPR*, 2017. [2](#_bookmark2), [3](#_bookmark3), [9](#_bookmark29)
3. M. Bai and R. Urtasun. Deep watershed transform for instance seg- mentation. In *CVPR*, 2017. [3](#_bookmark3), [9](#_bookmark29)
4. H. Caesar, J. Uijlings, and V. Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, 2018. [2](#_bookmark2), [5](#_bookmark9)
5. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab: Semantic image segmentation with deep convo- lutional nets, atrous convolution, and fully connected CRFs. *PAMI*, 2018. [1](#_bookmark0)
6. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Be- nenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. [2](#_bookmark2), [3](#_bookmark3), [5](#_bookmark9), [6](#_bookmark15)
7. C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *IJCV*, 2011. [9](#_bookmark29)
8. P. Dolla´r, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection:

An evaluation of the state of the art. *PAMI*, 2012. [2](#_bookmark2)

1. M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes challenge: A retrospective. *IJCV*, 2015. [2](#_bookmark2)
2. A. Fathi, N. Kanazawa, and K. Murphy. Places challenge 2017: in- stance segmentation, G-RMI team. 2017. [8](#_bookmark26)
3. A. Fathi, K. Yang, and K. Murphy. Places challenge 2017: scene parsing, G-RMI team. 2017. [8](#_bookmark26)
4. J. Fu, J. Liu, L. Guo, H. Tian, F. Liu, H. Lu, Y. Li, Y. Bao, and W. Yan. Places challenge 2017: scene parsing, CASIA IVA JD team. 2017. [8](#_bookmark26)
5. B. Hariharan, P. Arbela´ez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, 2014. [4](#_bookmark6), [5](#_bookmark9)
6. K. He, G. Gkioxari, P. Dolla´r, and R. Girshick. Mask R-CNN. In

*ICCV*, 2017. [1](#_bookmark0), [2](#_bookmark2), [8](#_bookmark26)

1. J. Hosang, R. Benenson, P. Dolla´r, and B. Schiele. What makes for effective detection proposals? *PAMI*, 2015. [1](#_bookmark0)
2. J. Hosang, R. Benenson, and B. Schiele. Learning non-maximum suppression. *PAMI*, 2017. [9](#_bookmark29)
3. A. Kirillov, R. Girshick, K. He, and P. Dolla´r. Panoptic feature pyra- mid networks. In *CVPR*, 2019. [9](#_bookmark29)
4. A. Kirillov, E. Levinkov, B. Andres, B. Savchynskyy, and C. Rother. InstanceCut: from edges to instances with multicut. In *CVPR*, 2017. [2](#_bookmark2), [3](#_bookmark3), [9](#_bookmark29)
5. I. Kokkinos. UberNet: Training a universal convolutional neural net- work for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*, 2017. [3](#_bookmark3)
6. A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. [2](#_bookmark2)
7. J. Li, A. Raventos, A. Bhargava, T. Tagawa, and A. Gaidon. Learning to fuse things and stuff. *arXiv:1812.01192*, 2018. [9](#_bookmark29)
8. Q. Li, A. Arnab, and P. H. Torr. Weakly-and semi-supervised panop- tic segmentation. In *ECCV*, 2018. [9](#_bookmark29)
9. Y. Li, X. Chen, Z. Zhu, L. Xie, G. Huang, D. Du, and

X. Wang. Attention-guided unified network for panoptic segmen- tation. *arXiv:1812.03904*, 2018. [9](#_bookmark29)

1. Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance- aware semantic segmentation. In *CVPR*, 2017. [2](#_bookmark2)
2. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan,

P. Dolla´r, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. [2](#_bookmark2), [5](#_bookmark9), [9](#_bookmark29)

1. C. Liu, J. Yuen, and A. Torralba. SIFT flow: Dense correspondence across scenes and its applications. *PAMI*, 2011. [2](#_bookmark2)
2. H. Liu, C. Peng, C. Yu, J. Wang, X. Liu, G. Yu, and W. Jiang. An end-to-end network for panoptic segmentation. *arXiv:1903.05027*, 2019. [9](#_bookmark29)
3. S. Liu, J. Jia, S. Fidler, and R. Urtasun. SGN: Sequential grouping networks for instance segmentation. In *CVPR*, 2017. [2](#_bookmark2), [3](#_bookmark3), [9](#_bookmark29)
4. S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. LSUN’17: insatnce segmen- tation task, UCenter winner team. 2017. [8](#_bookmark26)
5. J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. [1](#_bookmark0), [2](#_bookmark2), [5](#_bookmark9)
6. R. Luo, B. Jiang, T. Xiao, C. Peng, Y. Jiang, Z. Li, X. Zhang, G. Yu,

Y. Mu, and J. Sun. Places challenge 2017: instance segmentation,

Megvii (Face++) team. 2017. [8](#_bookmark26)

1. J. Malik, P. Arbela´ez, J. Carreira, K. Fragkiadaki, R. Girshick,

G. Gkioxari, S. Gupta, B. Hariharan, A. Kar, and S. Tulsiani. The three R’s of computer vision: Recognition, reconstruction and reor- ganization. *PRL*, 2016. [3](#_bookmark3)

1. D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natu- ral image boundaries using local brightness, color, and texture cues.

*PAMI*, 2004. [5](#_bookmark9)

1. I. Misra, A. Shrivastava, A. Gupta, and M. Hebert. Cross-stitch net- works for multi-task learning. In *CVPR*, 2016. [3](#_bookmark3)
2. G. Neuhold, T. Ollmann, S. Rota Bulo`, and P. Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes.

In *CVPR*, 2017. [2](#_bookmark2), [3](#_bookmark3), [5](#_bookmark9), [6](#_bookmark15), [9](#_bookmark29)

1. P. O. Pinheiro, R. Collobert, and P. Dolla´r. Learning to segment object candidates. In *NIPS*, 2015. [2](#_bookmark2)
2. S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real- time object detection with region proposal networks. In *NIPS*, 2015. [1](#_bookmark0)
3. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma,

Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and

L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge.

*IJCV*, 2015. [2](#_bookmark2)

1. J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recog. and segm. In *ECCV*, 2006. [2](#_bookmark2)
2. M. Sun, B. Kim, P. Kohli, and S. Savarese. Relating things and stuff via object property interactions. *PAMI*, 2014. [3](#_bookmark3), [4](#_bookmark6)
3. J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013. [3](#_bookmark3), [4](#_bookmark6)
4. J. Tighe, M. Niethammer, and S. Lazebnik. Scene parsing with object instances and occlusion ordering. In *CVPR*, 2014. [2](#_bookmark2), [3](#_bookmark3), [4](#_bookmark6)
5. Z. Tu, X. Chen, A. L. Yuille, and S.-C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. *IJCV*, 2005. [2](#_bookmark2), [3](#_bookmark3)
6. R. Vaillant, C. Monrocq, and Y. LeCun. Original approach for the localisation of objects in images. *IEE Proc. on Vision, Image, and* *Signal Processing*, 1994. [2](#_bookmark2)
7. C. Van Rijsbergen. *Information retrieval*. London: Butterworths, 1979. [5](#_bookmark9), [7](#_bookmark19)
8. P. Viola and M. Jones. Rapid object detection using a boosted cas- cade of simple features. In *CVPR*, 2001. [2](#_bookmark2)
9. D. B. West. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001. [7](#_bookmark19)
10. Y. Xiong, R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, and

R. Urtasun. UPSNet: A unified panoptic segmentation network.

*arXiv:1901.03784*, 2019. [9](#_bookmark29)

1. T.-J. Yang, M. D. Collins, Y. Zhu, J.-J. Hwang, T. Liu, X. Zhang,

V. Sze, G. Papandreou, and L.-C. Chen. DeeperLab: Single-shot image parser. *arXiv:1902.05093*, 2019. [9](#_bookmark29)

1. Y. Yang, S. Hallman, D. Ramanan, and C. C. Fowlkes. Layered object models for image segmentation. *PAMI*, 2012. [4](#_bookmark6)
2. J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmenta- tion. In *CVPR*, 2012. [2](#_bookmark2), [3](#_bookmark3), [4](#_bookmark6)
3. F. Yu and V. Koltun. Multi-scale context aggregation by dilated con- volutions. In *ICLR*, 2016. [1](#_bookmark0)
4. Y. Zhang, H. Zhao, and J. Shi. LSUN’17: semantic segmentation task, PSPNet winner team. 2017. [8](#_bookmark26)
5. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017. [8](#_bookmark26)
6. B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba.

Scene parsing through ADE20K dataset. In *CVPR*, 2017. [2](#_bookmark2), [3](#_bookmark3), [5](#_bookmark9), [6](#_bookmark15)

1. Y. Zhu, Y. Tian, D. Mexatas, and P. Dolla´r. Semantic amodal seg- mentation. In *CVPR*, 2017. [3](#_bookmark3)