Contextual Image Parsing via Panoptic Segment Sorting

Jyh-Jing Hwang*, Tsung-Wei Ke*, Stella X. Yu*
{jyh,twke,stellayu}@berkeley.edu
UC Berkeley / ICSI
Berkeley, CA, USA

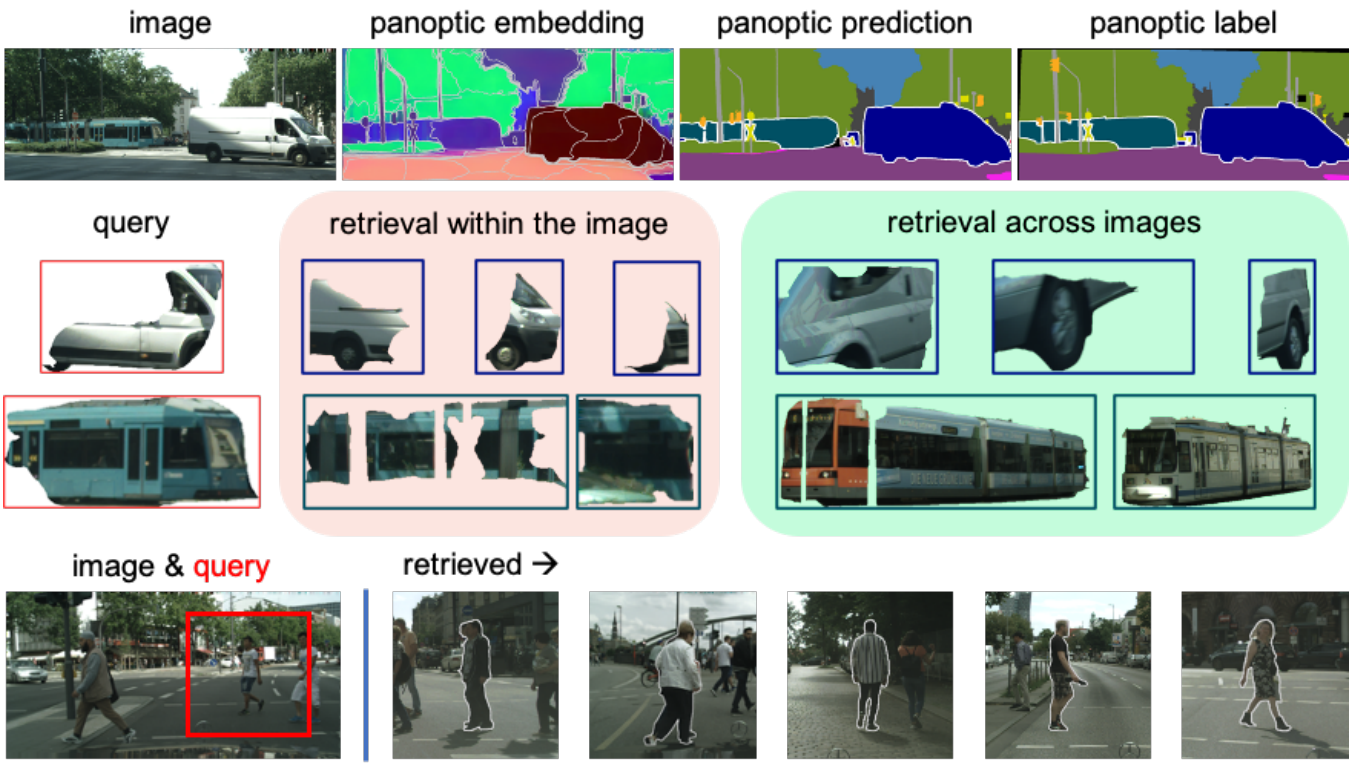

Figure 1: We approach contextual object recognition as a pixel-wise feature representation learning problem that accomplishes supervised panoptic segmentation while discovering and encoding visual context automatically. Top row from left to right) input image, panoptic embeddings, panoptic predictions, and panoptic labels. We overlay panoptic embeddings with the resultant over-segmentation boundaries. Middle row) After extracting panoptic embeddings from a CNN and the resultant over-segmentation, we use the segment prototype features to find nearest neighbors, within the image (middle) or across images (right), of each query segment (in red). These retrieval results probe what's learned in the embedding space. Bottom row) sample pedestrian retrieval results. We can retrieve a person *crossing a somewhat empty street* **without any such context labeling during training.**

ABSTRACT

Real-world visual recognition is far more complex than object recognition: There is *stuff* without distinctive shape or appearance, and the same object appearing in different contexts calls for different actions. While we need context-aware visual recognition, visual context is hard to describe and impossible to label manually.

We consider visual context as semantic correlations between objects and their surroundings that include both object instances and stuff categories. We approach contextual object recognition as a pixel-wise feature representation learning problem that accomplishes supervised panoptic segmentation while discovering and encoding visual context automatically.

Panoptic segmentation is a dense image parsing task that segments an image into regions with both semantic category and object instance labels. These two aspects could conflict each other, for two adjacent cars would have the same semantic label but different instance labels. Whereas most existing approaches handle the two labeling tasks separately and then fuse the results together, we propose a single pixel-wise feature learning approach that unifies both aspects of semantic segmentation and instance segmentation.

Our work takes the metric learning perspective of SegSort but extends it non-trivially to panoptic segmentation, as we must merge

*The first two authors contributed equally to this research.

MULL '21, October 24, 2021, Virtual Event, China.

ACM ISBN 978-1-4503-8681-4/21/10…$15.00
https://doi.org/10.1145/3476098.3485056

segments into proper instances and handle instances of various scales. Our most exciting result is the emergence of visual context in the feature space through contrastive learning between pixels and segments, such that we can retrieve a person *crossing a somewhat empty street* without any such context labeling.

Our experimental results on Cityscapes and PASCAL VOC demonstrate that, in terms of surround semantics distributions, our retrievals are much more consistent with the query than the state-of-the-art segmentation method, validating our pixel-wise representation learning approach for the unsupervised discovery and learning of visual context.

CCS CONCEPTS

• **Computing methodologies** → **Image segmentation**; **Structured outputs**.

KEYWORDS

contrastive learning, context encoding, context discovery, image parsing, panoptic segmentation

ACM Reference Format:
Jyh-Jing Hwang*, Tsung-Wei Ke*, Stella X. Yu. 2021. Contextual Image Parsing via Panoptic Segment Sorting. In *Proceedings of the 1st Workshop on Multimedia Understanding with Less Labeling (MULL '21), October 24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3476098.3485056

1 INTRODUCTION

Visual recognition is often modeled as identifying the semantic category of an object, based on its distinctive appearance or geometry. However, real-world visual recognition is far more complex: There are not only uncountable visual semantics (known as *stuff* as opposed to *things*) such as *shrubs*, *hedges*, and *rivers*, but the same object identified in different visual contexts could call for drastically different actions. For example, seeing a pedestrian in the middle of a road would prompt an autonomous driving system to quickly stop the car, whereas seeing a pedestrian on a side walk would require no special attention.

We need context-aware visual recognition, but visual context is hard to describe, as it could be anything related to given images and visual task. In practice, visual context has been modeled as a global image feature such as *scene gist* [59], statistically co-occurring object categories [16, 17, 37, 56, 64] and object instances [69], 2D spatially co-occurring instances [53], and more recently fine-grained interactions among objects [7, 25].

We consider visual context as semantic correlations between objects and their surroundings, which could include both object instances and stuff categories. We approach contextual object recognition as a pixel-wise feature representation learning problem that accomplishes supervised panoptic segmentation while discovering and encoding visual context automatically (Fig. 1).

Panoptic segmentation [34] is a dense image parsing task [72] that segments an image into regions with both semantic category and object instance labels. These two aspects could conflict each other, for two adjacent cars in an image would be labeled with the same semantic label (*car*) but different instance labels (*car #1 vs. car #2*). Existing approaches thus tackle them separately with two branches, each optimized for semantic segmentation and instance segmentation respectively [15, 33, 40, 78, 79]. However, additional modules are required to integrate semantic and instance predictions and resolve disagreements between the two branches.

We propose a single pixel-wise feature learning approach that unifies both aspects of semantic segmentation and instance segmentation. We map each pixel to a point in the latent feature space with a convolutional neural network (CNN). Our learning objective is to bring pixels belonging to the same instance or the same stuff closer, and to separate them far from other object instances and stuff regions, both within the same image and across different images. With subsequent feature clustering and classification, we can derive instance and semantic segmentation predictions from this common feature representation.

Unlike most methods that formulate image segmentation as a pixel-wise classification task that predicts the semantic category or object instance directly, our supervised panoptic segmentation method takes the metric learning perspective as SegSort [29], with a contrastive learning loss between individual pixels and segments, but extends segment sorting according to both semantic and instance labels. This extension is highly non-trivial, as we must merge segments into proper instances and handle instances of various scales. We dub our method Panoptic Segment Sorting (PSS).

Our most exciting result is the emergence of visual context in the feature space through contrastive learning between pixels and segments. Although the labels provided during training are only concerned with semantic instance categories not visual context, our learned feature is able to encode not only the object instance or stuff category that a pixel is part of, but also the visual context it is embedded into. That is, by design of supervised segmentation, pixels for different parts of a *bus* assume similar features (Fig. 1 middle), yet by contrastive learning of grouping relationships between pixels and segments, the feature of the pedestrian in the query image also encodes the visual context of *a person crossing a somewhat empty street*, and all the nearest neighbour retrievals from the image gallery also have such visual context (Fig. 1 bottom).

We propose a context metric that measures how well the semantic distribution in the surround of an object of interest in the query image is captured in the retrieval. Our experimental results on Cityscapes [18] and PASCAL VOC [21] demonstrate that the context of instances retrieved by our panoptic embedding is much more consistent with the query. Our pixel-wise representation learning points to a novel promising way to discover and learn visual context without any context supervision.

2 RELATED WORK

Image parsing and panoptic segmentation. The task of image parsing is first introduced in [72], where they formulate the solution in a Bayesian framework and construct a parsing graph as output. Since then, a lot of work has attempted to solve holistic scene understanding ([54, 64, 71, 80, 85]). Recently, [34] reintroduce image parsing in the context of deep learning with large-scale datasets and new evaluation metric, renaming the task as panoptic segmentation as to unify the well-developed semantic and instance segmentation. Many research efforts [14, 23, 33, 38–41, 45, 63, 73, 74, 78, 79] have followed quickly. The common approaches embrace the concept of

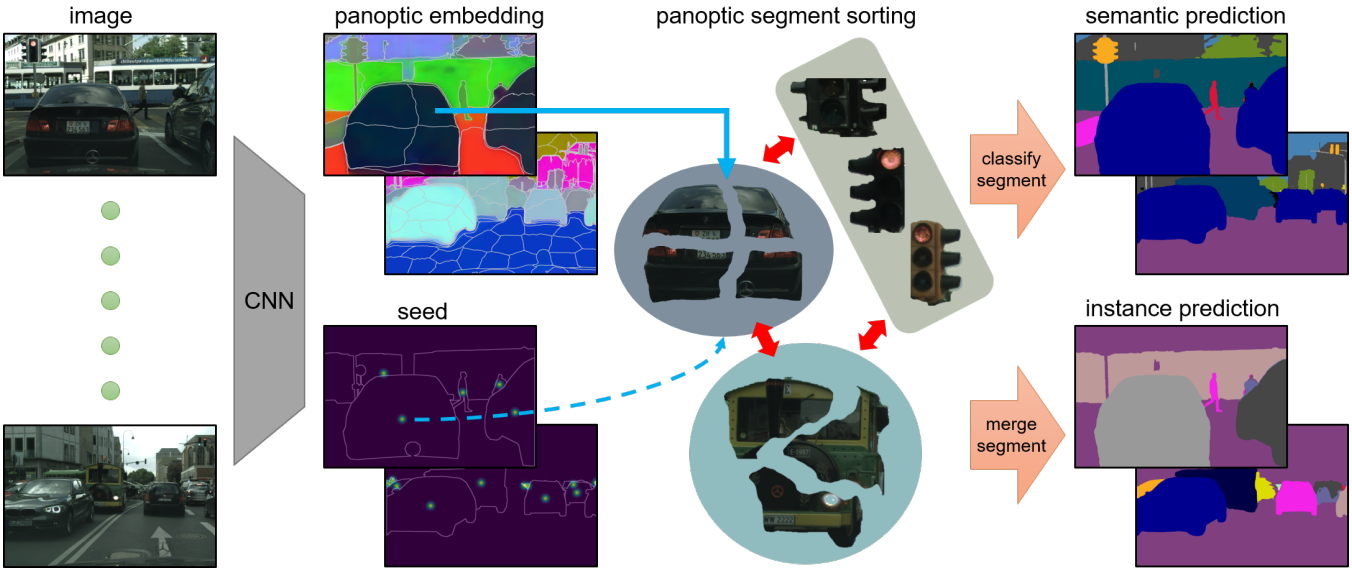

Figure 2: Our proposed Panoptic Segment Sorting (PSS) adopts feature learning with a pixel-to-segment contrastive loss, followed by segment merging (classification) for inferring instance (semantic) segmentation. We first over-segment an image with the pixel-wise embedding extracted from a CNN. Each segment is represented by a prototype feature (the average of pixel embeddings), which is then used for classifying segments (semantic predictions) and/or merging segments into an object instance (instance predictions). Our features automatically encode object-centric visual context. An extra center seeding branch can facilitate the merging process by designating seed segments. The overall losses include (1) the SegSort loss [29] for embeddings, (2) the cross-entropy softmax loss for classification, and (3) the regression loss for seed locations. PSS uses joint feature representations for both instance and semantic segmentation tasks.

unifying instance and semantic segmentation by integrating the time-tested object proposal and segmentation framework popularized by Mask R-CNN [26].

Instance segmentation. This task is generally approached by two camps of solutions: top-down or bottom-up. The top-down approaches [8, 19, 20, 26, 42, 48] adopt a two-stage framework where the bounding boxes are proposed by a detection network [65] and the segmentation masks are produced by an add-on head. The bottom-up approaches [1, 2, 6, 22, 32, 35, 47, 49, 58, 60–62, 84] predict and encode pair-wise relationships in various forms and segment the instance accordingly.

Instance context. Instance contexts and relationships are explored mainly to enhance the detection performance. Earlier work [53] models the appearances and 2D spatial context as a graph. Handcrafted features [37, 56], tree-based models [16, 17] are then developed to model co-occurring statistics and spatial configurations among object categories and object instances [69]. Recently, researchers integrate graphs [13] or spatial memory [12] into the deep learning framework. The distinction of our work is that our model does not explicitly model contexts yet is able to discovers novel contexts automatically.

Semantic segmentation. Current state-of-the-art semantic segmentation approaches develop from fully convolutional networks [9, 50], with various innovations. Incorporating contextual information [10, 11, 31, 66, 77, 81, 82], and encoding pair-wise relationships [4, 5, 28–30, 36, 46, 52, 55, 83] are the two major research lines.

Non-parametric segmentation. Prior to deep learning's emergence, non-parametric models [44, 67, 70] usually use hand-craft features with statistical models or graphical models to segment images with pixel-wise labels. Deep metric learning methods [22, 57] for instance segmentation emphasize the simplicity and fast computation. More recently, inspired by non-parametric models [75, 76] for image recognition, SegSort[29], upon which our work is built, captures pixel-to-segment relationships via pixel-wise embeddings, proposing the first deep non-parametric semantic segmentation in both supervised and unsupervised settings.

3 METHOD

We adapt the Segment Sorting approach [29] to panoptic segmentation by sorting segments according to both of its semantic and instance labels. With the learned feature representations, we classify segments into categories with a softmax classifier and merge them into instances by our proposed clustering algorithm. We also facilitate the merging process with a seeding branch that predicts the center of each instance.

Our end-to-end framework consists of a major SegSort branch and a seeding branch, both of which share one backbone network that generates multi-scale pixel-wise features. The SegSort branch outputs pixel-wise panoptic embeddings, which encode both semantic and instance information and are thus used to discover instance-centric context. The over-segmentations induced by the embeddings are then merged into instances and segments are classified by a softmax classifier. The seeding branch predicts the center of instances, which guide the merging process to reduce false positives. The overall framework is illustrated in Figure 2.

This section is organized as follows. We first briefly review the Segment Sorting framework for semantic segmentation in Sec. 3.1. We then describe how to extend it for panoptic segmentation in Sec. 3.2. In Sec. 3.3, we further develop a dynamic partitioning mechanism to alleviate the problem of varying scales of instances. Finally, we briefly describe the seeding branch in Sec. 3.4 that helps decide the ownership of boundaries.

3.1 Segment Sorting

We briefly review the Segment Sorting (SegSort) approach proposed by [29]. SegSort is an end-to-end optimization framework for non-parametric semantic segmentation. It produces pixel-wise semantic embeddings and their corresponding over-segmentation, each segment of which is then, during inference, assigned a semantic category via K-Nearest Neighbor search.

The basic idea of SegSort is assuming independent normal distributions (or von Mises-Fisher distributions for normalized embeddings) for individual segments, and seeking a maximum likelihood estimation of the feature mapping, so that the feature induced partitioning in the image and clustering across images provide maximum discrimination among segments. SegSort can be summarized as two components: spherical k-means clustering [3] and a maximum likelihood loss formulation with soft nighborhood assignments [24].

The spherical k-means clustering [3] alternates the expectation (E) and maximization (M) steps to partition the unit-length pixel-wise embeddings $\boldsymbol{v}$ of an image into K regions $(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_K)$. The M-step calculates the mean embedding direction of each region, or the *prototype* $\boldsymbol{\mu}_k = \frac{\sum_{i \in \boldsymbol{R}_k} \boldsymbol{v}_i}{||\sum_{i \in \boldsymbol{R}_k} \boldsymbol{v}_i||}$. The E-step assigns each pixel embedding $\boldsymbol{v}_i$ to a region $\boldsymbol{R}_k$ with nearest corresponding prototype $\boldsymbol{\mu}_k$, or $z_i = \arg\max_k \boldsymbol{\mu}_k^\top \boldsymbol{v}_i$, where z_i is the segment index that the pixel i is assigned. Note that the dot product on the right hand side is equivalent to cosine similarity as both $\boldsymbol{v}$ and $\boldsymbol{\mu}$ are of unit length. By alternating E- and M-steps, we over-segment an image.

After over-segmentation, one can derive a maximum likelihood loss with soft neighborhood assignments [24] to train the deep neural networks end-to-end. Interested readers are referred to the SegSort paper [29] for detailed derivation. The principle is to connect each pixel with one of its same-class segments, excluding its own segment, and to push away all the other segments in different classes. We define the corresponding probabilities given semantic segmentation ground truth labels as follows.

$$p(z_i = c^+ \mid \boldsymbol{v}_i, \Theta) = \frac{\exp(\kappa \boldsymbol{\mu}_{c^+}^\top \boldsymbol{v}_i)}{\sum_{l \neq c} \exp(\kappa \boldsymbol{\mu}_l^\top \boldsymbol{v}_i)}; \quad p(z_i = c \mid \boldsymbol{v}_i, \Theta) = 0, \quad (1)$$

where κ is the concentration (around μ) hyper-parameter in the von Mises-Fisher distributions, c denotes the segment index to which the pixel i is assigned, and c^+ denotes the segment index of any other same-class segment across all images in a batch. The final SegSort loss is therefore the negative log-likelihood of a pixel i selecting a same-class prototype as its neighbor:

$$L^i_{\text{SegSort}} = -\log \sum_{s \in C_i^+} p'_\phi(z_i = s \mid \boldsymbol{v}_i, \Theta) = -\log \frac{\sum_{s \in C_i^+} \exp(\kappa \boldsymbol{\mu}_s^\top \boldsymbol{v}_i)}{\sum_{l \neq c} \exp(\kappa \boldsymbol{\mu}_l^\top \boldsymbol{v}_i)}, \quad (2)$$

where C_i^+ denotes the set of c^+ segment indices w.r.t. the pixel i, which is selected by the semantic segmentation ground truth labels. Minimizing this loss is equivalent to maximizing the expected number of pixels correctly classified by voting of their nearest neighbor prototypes.

3.2 Panoptic Segment Sorting

Since the SegSort loss does not require a fixed number of classes as opposed to the conventional cross-entropy softmax loss, a way to extend it for instance discrimination is by changing the definition of ground truth labels and its corresponding selections of neighbor prototypes. In other words, we instead consider c^+ as the segment index of any other 'same-instance' segment. For stuff categories without instances, we consider all the segments in that class have the same instance label. With this modification, the SegSort loss in Eqn. 2 can be used to train panoptic embeddings.

Such trained embeddings, therefore, group each instance against all the other instances, regardless of their semantic categories. Still, since this loss pushes all the instances as far away as possible, visually similar instances are forced to stay closer on the hypersphere. We thus hypothesize two kinds of additional information are encoded: **(1)** The embeddings encode the semantic labels inherently as instances of the same class appear similar. To extract such information, we then stack two 1×1 convolutional layers on top of segment prototypes, followed by a softmax classifier to predict the semantic class of each segment. Note that no conflict between semantic and instance segmentations is introduced in this setting as they are built on the same over-segmentation. **(2)** The embeddings also encode object-centric context. This is endowed by the design of supervised semantic and instance segmentation with an unified representations. The feature of pedestrians walking across a road (on a sidewalk) encodes surrounding cars (buildings).

Given the panoptic embeddings and the resultant over-segmentations, the challenge is to group segments into instances correctly during inference. We need two criteria: 1) how to merge segments, and 2) when to stop the merging. To align with the formulation of the SegSort loss, we adopt a nearest neighbor clustering criterion [68] to greedily merge two segments $\boldsymbol{R}_m, \boldsymbol{R}_n$ with nearest prototypes, and stop the merging if the distance between two prototypes $\boldsymbol{\mu}_m, \boldsymbol{\mu}_n$ is greater than a threshold, or their dot product is less than a threshold T_P. The merging criteria can be summarized as:

$$\boldsymbol{R} = \{\boldsymbol{R}_m, \boldsymbol{R}_n\} \text{ if } \left(\mathcal{N}(\mu_m) = n \text{ or } \mathcal{N}(\mu_n) = m\right) \text{ and } \boldsymbol{\mu}_m^\top \boldsymbol{\mu}_n \geq T_P, \quad (3)$$

where $\{\cdot, \cdot\}$ denotes merging segments, $\mathcal{N}(\cdot)$ denotes the index of the nearest neighbor prototype. We sort all the pairs of distances (dot products) of the prototypes in an image and consider merging greedily from the closest pair. We also update the new prototype after merging.

3.3 Dynamic Partitioning for Hybrid Scale Exemplars

The vanilla SegSort partitions an image into a fixed number of regions regardlessly. For semantic segmentation, this setting is reasonable as the number and sizes of homogeneous regions do not vary a lot from an image to another. However in instance segmentation, scales of objects can change drastically from 100 to 100K pixels. Oftentimes cluttered small instances will fall into one single segment, or even worse be included in another big instance.

Figure 3: Our proposed metric Context Error (CE) evaluates the context similarity between two instances by measuring their semantic distributions of neighboring regions. From left to right: whole image, instance of interest, semantic distribution of the middle right extended region. We calculate the symmetric KL-divergence between semantic distributions from corresponding 8 extended regions as the context error. Our assumption is that visual context can be characterized by co-occurrence and spatial relationships among object semantics.

To alleviate this scale problem, we propose a hybrid scale setting for training and dynamic partitioning for inference accordingly. The illustrations can be found in Appendix.

During training, we consider regular embeddings $\boldsymbol{v}$ and their upscaled embeddings $\boldsymbol{v}^{(u)}$ by bilinear interpolation. The idea is to use the upscaled embeddings for small instances so that the gradient flows are finer. After the spherical k-mean clustering, we calculate segment prototypes using embeddings in different scales according to the instance sizes. Note that there is still only one prototype for each segment in the SegSort loss, be it either regular or upscaled.

During inference, the sizes of instances are unknown and have to be inferred. We notice that if a segment contains multiple small instances or multiple parts from a big instance, the corresponding pixel embeddings are usually noisy, resulting in a low concentration. Therefore, we define an approximated concentration $\tilde{\kappa}_k$ of a segment $\boldsymbol{R}_k$ as $\tilde{\kappa}_k = \frac{||\sum_{i \in \boldsymbol{R}_k} \boldsymbol{v}_i||}{|\boldsymbol{R}_k|} \in [0, 1]$, where $|\boldsymbol{R}_k|$ denotes the number of pixels in the segment. If this value for a segment falls below a certain threshold T_S, we again partition this segment using the same spherical k-means (here $k = 4$ usually).

3.4 Seeding Branch

We notice the boundaries between objects sometimes form their own segments, causing false positive instances. To remedy this issue, we build a second branch for predicting instance seeds, which are used for guiding the merging process, described in Sec. 3.2. We define seeds as the centers of instances and mark the segments that cover seeds as *seed segments*. For building this seeding branch, we follow closely the instance proposal branch in [26, 78] and use the centers of the predicted bounding boxes as the seeds.

Once we predict the seeds and the corresponding seed segments, we perform a seeding variant of merging. The only modification of the merging of $\{\boldsymbol{R}_m, \boldsymbol{R}_n\}$ (in Eqn. 3) is that the segments to merge $\boldsymbol{R}_m$&$\boldsymbol{R}_n$ are restricted to one seed segment and one non-seed segment; the merged segments are then marked as seed segments. Note that the merging only happens between same class segments. After this modification, all the boundary segments are then forced to be merged into one of the seed segments. The visualization of the merging processes can be found in the supplementary.

4 EXPERIMENTS

In this section, we demonstrate the efficacy of our framework through extensive experiments and analysis. We first describe the experimental setup in Sec. 4.1. We present the context specific instance retrieval results in Sec. 4.2. Finally in Sec. 4.3, we present the panoptic segmentation results. Hyper-parameters, ablation study, and more visual results such as panoptic predictions, context retrieval, and t-SNE [51] feature analysis, can be found in the Appendix.

4.1 Experimental Setup

Datasets. We carry out experiments mainly on two datasets: Cityscapes and PASCAL VOC 2012.

Cityscapes [18] is a dataset for semantic urban street scene understanding. 5,000 high quality pixel-level finely annotated images are divided into training, validation, and testing sets with 2,975 / 500 / 1,525 images, respectively. It defines 19 semantic categories

containing flat, human, vehicle, construction, object, nature, *etc.* , of which 8 categories have instance labels.

PASCAL VOC 2012 [21] segmentation dataset contains 20 object categories and one background class. The augmented dataset contains 10,582 (train) / 1,449 (val) / 1,456 (test) images. All the semantic classes, except for backgrounds, have instance labels.

Network architecture. We use the Feature Pyramid Networks (FPN) [43], with ResNet-50 [27] backbone pretrained on ImageNet, to provide the multi-scale pixel-wise features. For each of the seeding and panoptic embedding branch, we follow [78] by building three layers of deformable convolutional layers [20] (with shared weights across different scales) on top of each scale of FPN features. We then concatenate the multi-scale features, followed by a final fusion 1×1 convolutional layer. On top of the panoptic embeddings, we stack two 1×1 convolutional layers for the segment softmax classifier.

We consider UPSNet [78] as our baseline method on visual context retrieval. UPSNet achieves state-of-the-art performance on panoptic segmentation on Cityscapes dataset. It is in fact a good baseline method as it embraces two-branch models for tackling semantic and instance segmentation, respectively.

4.2 Context Specific Instance Retrieval

In this section, we experimentally verify our panoptic embeddings encode the object-centric context automatically.

Discovery of novel context. We retrieve the nearest neighbors of query instances on the Cityscapes validation set using their averaged embeddings. We notice that the retrieved instances are usually in similar context as the query. We showcase five interesting examples in the Appendix, *i.e.*, pedestrians crossing an intersection (also in Figure 1) or walking next to cars, riders riding bikes together or next to cars, and cluttered parked motorbikes. Note that these contexts are not given in the ground truth labels, yet our PSS can discover them unsupervisedly. We believe these examples are relevant in street scene understanding, especially for self-driving vehicles.

Quantitative evaluation. We wonder if such phenomena can be measured quantitatively. The challenge lies in the complicated scenarios and the lack of a complete label set. For example, crosswalks, which are labeled as roads, are visually similar as yet functionally different from roads. Furthermore, riding motorbikes next to cars is dangerous but difficult to describe precisely for annotating tasks.

We notice that the semantic category distribution of a larger patch captures some of such cases. For example, if there are multiple pedestrians nearby with cars around them, the chance of them walking on a crosswalk is higher. Based on these observations, we propose to evaluate the context similarity between query and the retrieval by comparing their semantic categories in 8 extended regions (Fig. 3).

To be specific, we denote the 8 neighbor regions (with the same size as the instance) as B_j for $j = 1, \ldots, 8$. We calculate the semantic distribution in each region by the occupancy ratio of each class and denote it as P_{B_j}. That is, for each class, given a semantic label mask S_P, then $P^c_{B_j} = \frac{1}{|B_j|} \sum_{B_j[S_P=c]}$ for each category c, where

method	person	rider	car	truck	bus	train	mbike	bike	mean CE
UPSNet [78]	1.15	1.21	0.88	1.20	1.08	1.33	1.23	1.21	1.16
PSS	0.96	1.01	0.65	1.12	1.04	1.27	1.11	1.05	1.02 (**-13.7%**)

Table 1: Our method is better at capturing visual context than our two-branch baseline, UPSNet [78]. We report results with *Context Errors* (CE) on the Cityscapes [18] validation set. We observe PSS performs better in every category and reduces 13.7% relative CE. Our PSS can retrieve object instances in more similar context.

$|B_j|$ denotes the area of region B_j. We then compare the semantic context distribution of the query $P^{(q)}_{B_j}$ against its i-th retrieval $P^{(r_i)}_{B_j}$ by calculating the symmetric KL divergence between the two, or

$$\text{CE} = \frac{1}{8K} \sum_{j=1}^{8} \sum_{i=1}^{K} \left(D_{\text{KL}}(P^{(q)}_{B_j}, P^{(r_i)}_{B_j}) + D_{\text{KL}}(P^{(r_i)}_{B_j}, P^{(q)}_{B_j}) \right), \quad (4)$$

where CE is our proposed metric, *Context Error*, and K is the number of retrievals per query instance. If the reference probability is 0, the KL divergence will be invalid; in this case, we use a small probability 0.1 instead. We set K to 20 nearest neighbors.

We compute Context Error (CE) for each instance category, *i.e.,* , we restrict both query and retrieval to be a certain instance category. The final CE is the average errors of all instance categories. We compare our PSS against state-of-the-art UPSNet [78] and summarize the results in Tab. 1. We observe PSS performs better in every category and reduces 13.7% relative context errors.

Visual Comparison. Next, we present the visual comparison in Fig 4 between our PSS and UPSNet using three query instances from the same validation image and display 3 retrieved instances for each network in the training set. We observe that our retrieved instances are usually in a similar context and are sometimes even from the same training image. It indicates that PSS encodes not only the appearances of an instance but also its nearby environment.

Visual Context Cluster Analysis We conduct visual context cluster analysis and visualize the results in Fig. 5. We first collect all the pedestrian prototypes in the Cityscapes training set. We plot their surrounding ground truth mask at their t-SNE feature locations and the aggregated density map. We observe interesting clusters such as pedestrians next to a car (center) and pedestrians alone on sidewalks (top left). We also notice some rare contexts on the middle left by examining the density map: a pedestrian is behind a clutter of a motorbike and a bike, which could lead to collision.

4.3 Panoptic Segmentation

Main results on Cityscapes. We summarize the main results on the Cityscapes validation set and compare with the state-of-the-art in Table 2. Our PSS achieves competitive performance in PQ (Panoptic Quality, explained in Appendix) and outperforms all the other methods in PQ^{St}. Notably, our framework performs particularly well in semantic segmentation related benchmarks.

Main results on PASCAL VOC. We summarize the main results on the PASCAL VOC validation set and compare with the state-of-the-art in Table 3. We show that PSS outperforms [39] by 2% PQ even with a weaker backbone (ResNet-50 vs 101).

Figure 4: PSS is better at encoding visual context in the learned representations. We compare our method with baseline for context specific instance retrieval. We show 3 query examples (left) and their top retrieval results by our PSS (middle) and UPSNet [78] (right), respectively. We observe that retrieved instances by PSS are usually in similar context or sometimes even from the same training image.

method	backbone	PQ	PQ^{Th}	PQ^{St}
[39]	ResNet-101	47.3	39.6	52.9
DeeperLab [79]	Xception-71	56.5	-	-
AUNet [41]	ResNet-50	56.4	52.7	59.0
SSAP [23]	ResNet-50	56.6	49.2	-
Panoptic FPN [33]	ResNet-50	57.7	51.6	62.2
UPSNet [78]	ResNet-50	59.3	54.6	62.7
UPSNet* [78]	ResNet-50	59.1	54.2	62.6
PSS	ResNet-50	58.7	51.7	**63.7**

Table 2: PSS achieves competitive performance on panoptic segmentation over the Cityscapes validation set. Our proposed framework PSS achieves comparable performance in PQ and outperforms all the other methods in PQ^{St}. * denotes retraining the model using released code; other results are copied from the published papers and '-' denotes missing metrics.

method	backbone	PQ
[39]	ResNet-101	62.7
PSS	ResNet-50	64.8

Table 3: PSS achieves competitive performance on panoptic segmentation over Pascal VOC 2012 validation set. Our PSS outperforms baseline [39] by large margin.

5 SUMMARY

We propose Panoptic Segment Sorting (PSS), a pixel-to-segment contrastive learning framework for contextual image parsing. Our method unifies semantic segmentation and instance segmentation in the pixel-wise panoptic embedding that also encodes and discovers visual context automatically.

We propose a context error metric that measures the distribution similarity in surround semantics between the query and the retrieval. Our experimental results demonstrate that PSS not only performs segmentation competitively with the state-of-the-art, but more importantly, its retrievals capture visual context much better, validating our pixel-wise representation learning approach for the unsupervised discovery and learning of visual context.

ACKNOWLEDGMENTS

This work was supported, in part, by Berkeley Deep Drive and Berkeley AI Research Commons with Facebook.

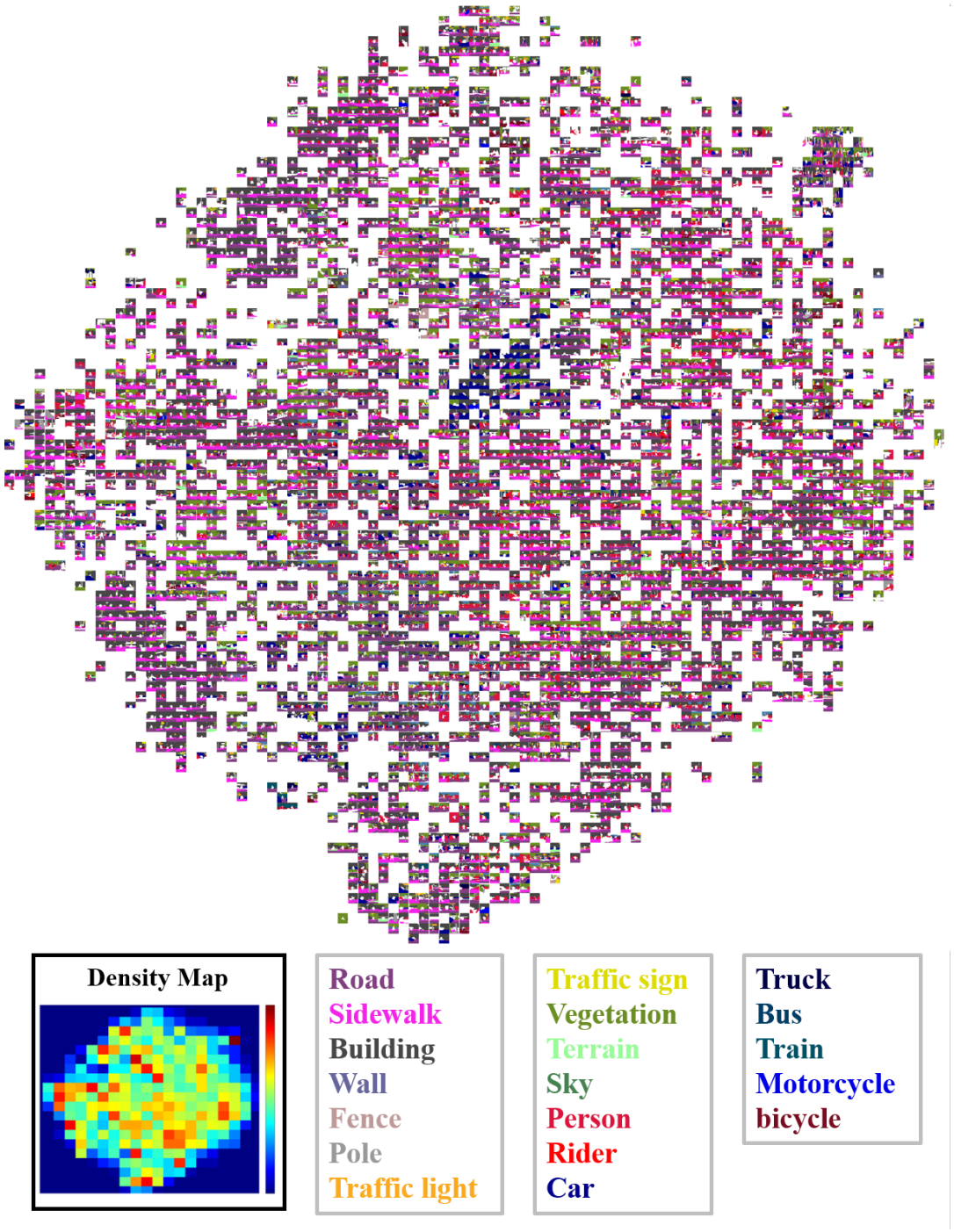

Figure 5: Clusterings of pedestrian prototypes in the latent feature space are corresponded to their visual contexts (best viewed with zoom-in). We first collect all the pedestrian prototypes in the Cityscapes training set. We plot their surrounding ground truth mask at their t-SNE feature locations and the aggregated density map (bottom left). We observe interesting clusters such as pedestrians next to a car (center) and pedestrians alone on sidewalks (top left). We also notice some rare contexts on the middle left by examining the density map: a pedestrian is behind a clutter of a motorbike and a bike, which could possibly lead to collision. We show that pedestrians in the similar context are grouped in the learned feature space.

REFERENCES

[1] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. 2014. Multiscale combinatorial grouping. In *CVPR*.

[2] Min Bai and Raquel Urtasun. 2017. Deep watershed transform for instance segmentation. In *CVPR*.

[3] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. 2005. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research* (2005).

[4] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. 2016. Semantic segmentation with boundary neural fields. In *CVPR*.

[5] Gedas Bertasius, Lorenzo Torresani, Stella X Yu, and Jianbo Shi. 2017. Convolutional Random Walk Networks for Semantic Image Segmentation. In *CVPR*.

[6] Joao Carreira and Cristian Sminchisescu. 2011. CPMC: Automatic object segmentation using constrained parametric min-cuts. *TPAMI* (2011).

[7] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. 2015. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*. 1017–1025.

[8] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. 2018. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *CVPR*.

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2016. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915* (2016).

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).

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

[12] Xinlei Chen and Abhinav Gupta. 2017. Spatial memory for context reasoning in object detection. In *ICCV*.

[13] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. 2018. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7239–7248.

[14] Yifeng Chen, Guangchen Lin, Songyuan Li, Omar Bourahla, Yiming Wu, Fangfang Wang, Junyi Feng, Mingliang Xu, and Xi Li. 2020. BANet: Bidirectional Aggregation Network with Occlusion Handling for Panoptic Segmentation. In *CVPR*.

[15] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. 2020. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR*.

[16] Myung Jin Choi, Joseph J Lim, Antonio Torralba, and Alan S Willsky. 2010. Exploiting hierarchical context on a large database of object categories. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 129–136.

[17] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. 2011. A tree-based context model for object recognition. *IEEE transactions on pattern analysis and machine intelligence* 34, 2 (2011), 240–252.

[18] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *CVPR*.

[19] Jifeng Dai, Kaiming He, and Jian Sun. 2016. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*.

[20] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable convolutional networks. In *ICCV*.

[21] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. [n.d.]. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[22] Alireza Fathi, Zbigniew Wojna, Vivek Rathod, Peng Wang, Hyun Oh Song, Sergio Guadarrama, and Kevin P Murphy. 2017. Semantic instance segmentation via deep metric learning. *arXiv preprint arXiv:1703.10277* (2017).

[23] Naiyu Gao, Yanhu Shan, Yupei Wang, Xin Zhao, Yinan Yu, Ming Yang, and Kaiqi Huang. 2019. SSAP: Single-Shot Instance Segmentation With Affinity Pyramid. In *ICCV*.

[24] Jacob Goldberger, Geoffrey E Hinton, Sam T Roweis, and Ruslan R Salakhutdinov. 2005. Neighbourhood components analysis. In *NIPS*.

[25] Saurabh Gupta and Jitendra Malik. 2015. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474* (2015).

[26] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *ICCV*.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

[28] Jyh-Jing Hwang, Tsung-Wei Ke, Jianbo Shi, and Stella X Yu. 2019. Adversarial Structure Matching for Structured Prediction Tasks. In *CVPR*.

[29] Jyh-Jing Hwang, Stella X Yu, Jianbo Shi, Maxwell D Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. 2019. SegSort: Segmentation by Discriminative Sorting of Segments. In *ICCV*.

[30] Tsung-Wei Ke, Jyh-Jing Hwang, Ziwei Liu, and Stella X Yu. 2018. Adaptive affinity fields for semantic segmentation. In *ECCV*.

[31] Tsung-Wei Ke, Jyh-Jing Hwang, and Stella X Yu. 2021. Universal Weakly Supervised Segmentation by Pixel-to-Segment Contrastive Learning. *ICLR* (2021).

[32] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*.

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

[34] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. 2019. Panoptic segmentation. In *CVPR*.

[35] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother. 2017. Instancecut: from edges to instances with multicut. In *CVPR*.

[36] Shu Kong and Charless Fowlkes. 2018. Recurrent pixel embedding for instance grouping. In *CVPR*.

[37] Yong Jae Lee and Kristen Grauman. 2010. Object-graphs for context-aware category discovery. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

[38] Jie Li, Allan Raventos, Arjun Bhargava, Takaaki Tagawa, and Adrien Gaidon. 2018. Learning to fuse things and stuff. *arXiv preprint arXiv:1812.01192* (2018).

[39] Qizhu Li, Anurag Arnab, and Philip HS Torr. 2018. Weakly- and semi-supervised panoptic segmentation. In *ECCV*.

[40] Qizhu Li, Xiaojuan Qi, and Philip HS Torr. 2020. Unifying training and inference for panoptic segmentation. In *CVPR*.

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

[42] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. 2017. Fully convolutional instance-aware semantic segmentation. In *CVPR*.

[43] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *CVPR*.

[44] Ce Liu, Jenny Yuen, and Antonio Torralba. 2011. Nonparametric scene parsing via label transfer. *PAMI* (2011).

[45] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. 2019. An end-to-end network for panoptic segmentation. In *CVPR*.

[46] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. 2017. Learning Affinity via Spatial Propagation Networks. In *NIPS*.

[47] Shu Liu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 2017. Sgn: Sequential grouping networks for instance segmentation. In *ICCV*.

[48] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. 2018. Path aggregation network for instance segmentation. In *CVPR*.

[49] Yiding Liu, Siyu Yang, Bin Li, Wengang Zhou, Jizheng Xu, Houqiang Li, and Yan Lu. 2018. Affinity derivation and graph merge for instance segmentation. In *ECCV*.

[50] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*.

[51] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* (2008).

[52] Michael Maire, Takuya Narihira, and Stella X Yu. 2016. Affinity CNN: Learning pixel-centric pairwise relations for figure/ground embedding. In *CVPR*.

[53] Tomasz Malisiewicz and Alyosha Efros. 2009. Beyond categories: The visual memex model for reasoning about object relationships. In *NIPS*.

[54] Tomasz Malisiewicz and Alexei A Efros. 2008. Recognition by association via learning per-exemplar distances. In *CVPR*.

[55] Mohammadreza Mostajabi, Michael Maire, and Gregory Shakhnarovich. 2018. Regularizing Deep Networks by Modeling and Predicting Label Structure. In *CVPR*.

[56] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 2014. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 891–898.

[57] Davy Neven, Bert De Brabandere, Marc Proesmans, and Luc Van Gool. 2019. Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. In *CVPR*.

[58] Alejandro Newell, Zhiao Huang, and Jia Deng. 2017. Associative embedding: End-to-end learning for joint detection and grouping. In *NeurIPS*.

[59] Aude Oliva and Antonio Torralba. 2007. The role of context in object recognition. *Trends in cognitive sciences* 11, 12 (2007), 520–527.

[60] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. 2018. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *ECCV*.

[61] Pedro O Pinheiro, Ronan Collobert, and Piotr Dollár. 2015. Learning to segment object candidates. In *NeurIPS*.

[62] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. 2016. Learning to refine object segments. In *ECCV*.

[63] Lorenzo Porzi, Samuel Rota Bulo, Aleksander Colovic, and Peter Kontschieder. 2019. Seamless Scene Segmentation. In *CVPR*.

[64] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge J Belongie. 2007. Objects in Context.. In *ICCV*.

[65] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*.

[66] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

[67] Bryan Russell, Alyosha Efros, Josef Sivic, Bill Freeman, and Andrew Zisserman. 2009. Segmenting scenes by matching image composites. In *NIPS*.

[68] M Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. 2019. Efficient Parameter-free Clustering Using First Neighbor Relations. In *CVPR*.

[69] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. 2019. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6619–6628.

[70] Joseph Tighe and Svetlana Lazebnik. 2010. Superparsing: scalable nonparametric image parsing with superpixels. In *ECCV*.

[71] Joseph Tighe and Svetlana Lazebnik. 2013. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*.

[72] Zhuowen Tu, Xiangrong Chen, Alan L Yuille, and Song-Chun Zhu. 2005. Image parsing: Unifying segmentation, detection, and recognition. *IJCV* (2005).

[73] Haochen Wang, Ruotian Luo, Michael Maire, and Greg Shakhnarovich. 2020. Pixel Consensus Voting for Panoptic Segmentation. In *CVPR*.

[74] Yangxin Wu, Gengwei Zhang, Yiming Gao, Xiajun Deng, Ke Gong, Xiaodan Liang, and Liang Lin. 2020. Bidirectional Graph Reasoning Network for Panoptic Segmentation. In *CVPR*.

[75] Zhirong Wu, Alexei A Efros, and Stella X Yu. 2018. Improving Generalization via Scalable Neighborhood Component Analysis. In *ECCV*.

[76] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. 2018. Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination. In *CVPR*.

[77] Saining Xie, Xun Huang, and Zhuowen Tu. 2016. Top-down learning for structured labeling with convolutional pseudoprior. In *ECCV*.

[78] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. 2019. UPSNet: A Unified Panoptic Segmentation Network. In *CVPR*.

[79] Tien-Ju Yang, Maxwell D Collins, Yukun Zhu, Jyh-Jing Hwang, Ting Liu, Xiao Zhang, Vivienne Sze, George Papandreou, and Liang-Chieh Chen. 2019. DeeperLab: Single-Shot Image Parser. *arXiv preprint arXiv:1902.05093* (2019).

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

[81] Fisher Yu and Vladlen Koltun. 2016. Multi-scale context aggregation by dilated convolutions. In *ICLR*.

[82] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *CVPR*.

[83] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. 2015. Conditional random fields as recurrent neural networks. In *ICCV*.

[84] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. 2019. Bottom-up object detection by grouping extreme and center points. In *CVPR*.

[85] Song-Chun Zhu and David Mumford. 2007. A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision* (2007).