Hierarchical Lovász Embeddings for Proposal-free Panoptic Segmentation

Tommi Kerola[1*] Jie Li[2] Atsushi Kanehira[1] Yasunori Kudo[1] Alexis Vallet[1] Adrien Gaidon[2]
[1]Preferred Networks, Inc. [2]Toyota Research Institute (TRI)

Abstract

Panoptic segmentation brings together two separate tasks: instance and semantic segmentation. Although they are related, unifying them faces an apparent paradox: how to learn simultaneously instance-specific and category-specific (i.e. instance-agnostic) representations jointly. Hence, state-of-the-art panoptic segmentation methods use complex models with a distinct stream for each task. In contrast, we propose Hierarchical Lovász Embeddings, per pixel feature vectors that simultaneously encode instance- and category-level discriminative information. We use a hierarchical Lovász hinge loss to learn a low-dimensional embedding space structured into a unified semantic and instance hierarchy without requiring separate network branches or object proposals. Besides modeling instances precisely in a proposal-free manner, our Hierarchical Lovász Embeddings generalize to categories by using a simple Nearest-Class-Mean classifier, including for non-instance "stuff" classes where instance segmentation methods are not applicable. Our simple model achieves state-of-the-art results compared to existing proposal-free panoptic segmentation methods on Cityscapes, COCO, and Mapillary Vistas. Furthermore, our model demonstrates temporal stability between video frames.

1. Introduction

Holistic scene understanding is an important task in computer vision, where a model is trained to explain each pixel in an image, whether that pixel describes *stuff* – uncountable regions of similar texture such as grass, road or sky – or *thing* – a countable object with individually identifying characteristics, such as people or cars. While holistic scene understanding received some early attention [49, 55, 48], modern deep learning-based methods have mainly tackled the tasks of modeling stuff and things independently under the task names semantic segmentation and instance segmentation. Recently, Kirillov et al. proposed the panoptic quality (PQ) metric for unifying these two parallel tracks into the holistic task of panoptic segmentation [24]. Panoptic segmentation

*Correspondence to tommi@preferred.jp

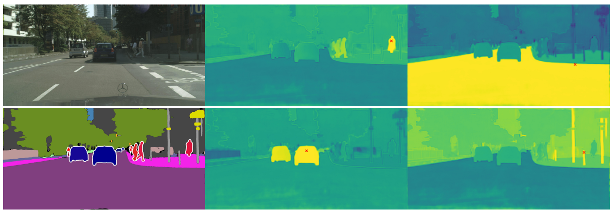

Figure 1. Example panoptic segmentation using our method (bottom left). Predictions are decoded from our Hierarchical Lovász Embeddings, which encode both instance and class information. The rightmost columns illustrate distances in embedding space between all pixels and a target pixel in red (warmer colors denote smaller distances). We can see the hierarchical structure: pixels on the same instance being closest, and pixels on other instances in the same category being closer than pixels from other categories.

is a key step for visual understanding, with applications in fields such as autonomous driving or robotics, where it is crucial to know both the locations of dynamically trackable things, as well as static stuff classes. For example, an autonomous car needs to be able to both avoid other cars with high precision, as well as understand the location of the road and sidewalk to stay on a desired path.

A strong but complex baseline for panoptic segmentation is to run independent methods for semantic segmentation and instance segmentation, and then fusing the results. To improve upon this, previous works combine both tasks in a joint model [52, 23, 27]. Early methods focus more on a joint model and the majority of them leverage two-stage instance detection models [43, 52, 23, 29, 13]; some recent works propose bottom-up [15, 9, 53], yet instance and semantic segmentation are still treated separately. Performing panoptic segmentation as a single task without duplicated information across sub-tasks remains an interesting question.

Intuitively, instances are contained in semantics, where semantic representations have higher variance in the embedding space to describe a general category, whereas instances have smaller variance to capture object-specific characteristics. This constitutes a natural hierarchical relationship between instances and semantics (cf. Figure 4).

In this work, we propose to model panoptic segmentation as a unified task via a novel formulation of the problem: learning hierarchical pixel embeddings. Creating a unified embedding for the task opens up the potential of leverag-

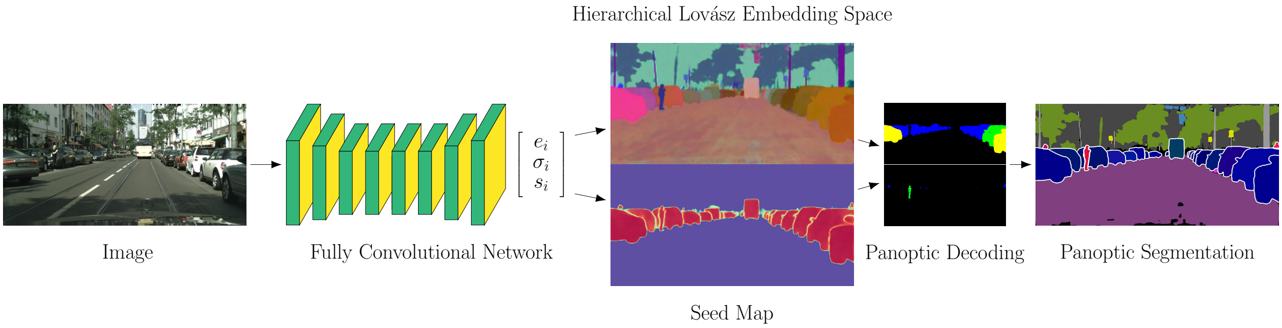

Figure 2. Overview of our method. We train a single-shot fully convolutional network to predict for each pixel i a hierarchical embedding $\mathbf{e}_i$ as well as an instance seed s_i and variance σ_i. The seed map represents probable instance locations, and the variance defines the margins of the hierarchical embedding space. These are used for panoptic decoding of the embedding space.

ing embedding space analysis as conducted in the natural language processing community [34, 37, 36, 35].

Leveraging hierarchical structure for representation learning is well studied [16, 14, 45]. However, previous work has not taken advantage of the semantic-instance visual hierarchy for end-to-end unified scene parsing. In this paper, we leverage advances in structural representation learning and encode "instance" and "category" features in a hierarchical embedding space. By doing so, we reduce the redundant information in the output space and optimize the information efficiency of network parameters for panoptic segmentation.

Our main contribution is a novel representation learning approach for panoptic segmentation that treats it as a unified task. We propose a simple architecture and loss to learn pixel-wise embeddings to represent instances, object categories and stuff classes, thus enabling **unified embedding-based single-shot panoptic segmentation**. In particular, we leverage the Lovász hinge loss to learn a structured Hierarchical Lovász Embedding space where categories can be represented with categories jointly. An overview of our method is shown in Figure 2.

Compared to conventional panoptic segmentation models, our model displays temporal stability between video frames, creating temporal smoothness in predictions that can be directly used in downstream applications such as object tracking and prediction in autonomous driving or mobile robotic systems, where data association is a key component (cf. Figure 8). Experiments on the Cityscapes [10], COCO [30] and Vistas [38] datasets show that our method establishes state-of-the-art results for proposal-free methods, and also yields competitive results compared with two-stage models.

2. Related Work

Deep learning-based dense prediction tasks have typically focused on either uncountable background or countable foreground objects. Semantic segmentation concerns the pixel-wise segmentation of semantics, treating each object class as uncountable. Instance segmentation, on the other hand, focuses explicitly on countable foreground classes, such as persons or cars. For the past few years, these tasks have evolved separately, with little interaction, leading to issues such as trouble with contextual clues in instance segmentation, or the confusion caused by the large variance of person classes in semantic segmentation. Recently, panoptic segmentation was proposed as a new task to bridge the gap between these methods and allow tackling them in a unified manner [24]. Embedding-based methods have recently become popular in the computer vision community for improving object detection [25, 57] and keypoint estimation [42]. In this section, we briefly review representative methods for each task.

Semantic segmentation. Following the seminal work of Long et al. [33], semantic segmentation is typically treated as a pixel-wise classification task (although exceptions exist [21]), where a fully convolutional network with an encoder-decoder architecture is trained to output a high softmax score for the ground-truth class. There have been several improvements since then. Namely, SegNet [3] introduces unpooling layers for more accurate upsampling in the decoder. Conversely, Deeplab [7] proposed to use a network with dilated convolution instead of a decoder, and leverages pooling to capture global information. PSPNet [56] improves global context by leveraging pyramid pooling and dilated convolution. DeeplabV3+ [8] combines the advantages of pooling and encoder-decoder architectures to better capture contextual information and sharper object boundaries.

Instance segmentation. Typical high-performing instance segmentation methods are variants of the Mask R-CNN [18] framework [32, 6]. These methods work in two stages, where the first stage computes regions of interest via a region proposal network, conducts non-maximum suppression on the proposed bounding boxes, and then runs a second stage via a head network on each proposal. While yielding high accuracy, these methods are typically too slow for real-time inference. Recently, there has been work addressing the creation of accurate single-shot (proposal-free) instance segmentation methods [2, 39, 15, 50]. In particular, Neven *et al.* [39] showed how to accurately predict a spatial embedding space for instance segmentation. Unlike previous methods, their network operates in a single stage and is able to produce an instance segmentation in a single shot, while having accuracy comparable to the more expensive Mask R-CNN. Their work inspires us to explore the possibility to learn a hierarchical embedding space for panoptic segmentation.

Panoptic segmentation. The current dominant methods for panoptic segmentation are two-stage frameworks. Kirillov et al. [24] run PSPNet and Mask R-CNN independently to obtain semantics and instance predictions. They subsequently combine these using heuristics. Subsequent work was done to combine the independent networks into one, by adding a semantic segmentation branch to Mask R-CNN [23, 43, 13, 29, 26, 27], but manual heuristics still remained. Other works aim to further remove manual merging heuristics of the semantics and instance predictions. E.g. UPSNet [52] proposes a panoptic head network for merging the predictions, and Liu et al. [31] leverages a spatial ranking module to conduct the merging between the two branches. Yang et al. [54] propose to resolve overlaps via instance relation reasoning. Recently, there has been some work adding a mask head on-top of a transformer to avoid heuristics [5].

Single-shot approaches, while less explored [53, 15, 9, 20], are an important complementary direction for panoptic segmentation. These methods demonstrate their potential in both network accuracy [9] and inference efficiency [20]. These existing methods feature separated feature streams for semantic segmentation- and instance-related representation, while in our approach, a unified embedding is used to model both semantic (category) and instance information. That is, our method has in fact *the same* downstream features showing that this pixel is both *a* car (instance-agnostic), and also *this* car (instance-specific), creating a natural and inherent feature representation for panoptic segmentation.

Embeddings for computer vision. Associative embeddings [41, 42] and variants have been popular for various vision tasks. The embedding space is learned by leveraging push and pull forces between embeddings in the image, depending on ground-truth annotation. The initial paper on associative embeddings [42] used 1-dimensional embeddings. The follow-up work [41] showed that increasing the embedding to be 8-dimensional improved convergence of the network. Embeddings have also been used for lane detection [40], instance segmentation [11, 39], and for instance and semantic segmentation of point clouds [51]. Further, embeddings have been used with success for improving object detection methods, using embedding space as a way to remove human-designed priors [25, 57]. Our proposed method is a variation of a deep nearest class mean (NCM) classifier [17], which has shown some promising results on image classification for replacing typical networks with softmax outputs directly after the last convolutional layer. To the best of our knowledge, our work is the first to leverage a deep NCM method for semantics in panoptic segmentation.

3. Panoptic Segmentation with Hierarchical Embeddings

In this section, we provide detailed discussion on how we learn the proposed embedding space and use the predicted embeddings for panoptic segmentation, based on the framework depicted in Figure 2. We start with problem formulation, and a discussion over the limitation of a current widely used loss in embedding learning for scene parsing. Then we propose the usage of a better alternative to the task with our novel loss for hierarchical embedding space learning. Finally, we describe panoptic decoding on the learned embedding space for creating a panoptic segmentation.

3.1. Problem Formulation

We formulate the task of panoptic segmentation as an embedding space learning problem, where we want to associate each pixel $\mathbf{x}_i$ with a latent embedding $\mathbf{e}_i$ such that we can decode both the object instance and semantic class correctly from *the same* single embedding (cf. Figure 4). Our goal is to learn a function f that maps a set of pixels $\mathcal{X}$, $|\mathcal{X}| = N$, into a set of embeddings $\mathcal{E}$ such that the embeddings can be partitioned into sets $\mathcal{S}$ and $\mathcal{I}$, where $\mathcal{S} = \{S_1, \ldots, S_{|\mathcal{C}|}\}$ are the embedding spaces defined by the set of semantic classes $\mathcal{C}$, and $\mathcal{I} = \{I_1, \ldots\}$ is the embedding subspace defined by the sets of instances such that we have the hierarchy $I_l \subseteq S_k \subseteq \mathcal{E}$ for any instance l with semantic class k. Further, the semantic classes $\mathcal{C}$ are assumed to be divided into thing classes $\mathcal{C}_{\text{thing}}$ and stuff classes $\mathcal{C}_{\text{stuff}}$. Note that we consider stuff classes to consist of a single instance.

3.2. Baseline – Associative Embeddings

A popular approach for learning such an embedding is the associative embedding (AE) loss proposed by Newell et al [42] (also known as discriminative loss [11]). The loss uses pull and push terms to attract or repel an embedding $\mathbf{e}_i$ from or to others $\mathbf{e}_j$ depending on the ground-truth pixel label of each embedding, given L instances:

$$\mathcal{L}_{\text{pull}} = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{|I_l|} \sum_{j \in \mathcal{I}_l} [\|\hat{\mathbf{e}}_l - \mathbf{e}_j\|_1 - \delta_{\text{pull}}]_+^2 \,, \tag{1}$$

$$\mathcal{L}_{\text{push}} = \frac{1}{L(L-1)} \sum_{l \neq k} [\delta_{\text{push}} - \|\hat{\mathbf{e}}_l - \hat{\mathbf{e}}_k\|_1]_+^2 \,, \tag{2}$$

where $\hat{\mathbf{e}}_l = \frac{1}{|I_l|} \sum_{j \in \mathcal{I}_l} \mathbf{e}_j$ is the instance mean embedding, and δ_{pull}, δ_{push} are problem-specific hinge hyperparameters and $[\cdot]_+$ is the ReLU function.

In theory, a set of two AE losses can be used to learn a hierarchical embedding space by enforcing the hinge hyperparameters to be larger for semantic ground-truth than for instance ground-truth. While this idea works for toy problems, the disadvantage is that we need to manually define the hinge hyperparameters which employ additional uncertainty and sensitivity in real-world datasets. We have found this difficult to do in practice for real-world datasets such as Cityscapes [10], as depicted in our ablation analysis in Section 4.4.

3.3. Lovász Hinge Loss

To decrease the requirement for additional hinge hyperparameters, we look into the Lovász hinge loss [4], which acts as a differentiable surrogate for the intersection over union (IoU), a common measure of mask overlap. In panoptic quality (PQ) [24] evaluation, IoU is a common base metric across thing and stuff classes.

First, we briefly review the Lovász hinge loss. Given a vector of binary ground-truth labels $\mathbf{t} \in \{0,1\}^N$ and predicted labels $\mathbf{y} \in \{0,1\}^N$, the IoU is defined as $\mathrm{IoU}(\mathbf{y},\mathbf{t}) = (|\{i : y_i = 1\} \cap \{i : t_i = 1\}|)/(|\{i : y_i = 1\} \cup \{i : t_i = 1\}|)$, which is a score between 0 and 1, where higher is better.

As the IoU is a discrete set function, it cannot be optimized via gradient descent, however the Lovász hinge provides a way of creating a continuous and differentiable surrogate function in terms of the prediction errors ξ_i for each pixel [4]. Given a continuous prediction vector $\mathbf{p} \in [0,1]^N$, the binary version of the Lovász hinge loss is defined as

$$\mathrm{LovászBinary}(\mathbf{p},\mathbf{t}) = \sum_i^N \xi_{\pi_i} \Delta_i^{\mathrm{IoU}} \,, \tag{3}$$

with prediction error ξ and IoU difference Δ^{IoU} defined as

$$\xi_i = \begin{cases} 1 - p_i & t_i = 1 \\ p_i & t_i = 0 \end{cases} \,, \tag{4}$$

$$\Delta_i^{\mathrm{IoU}} = \mathrm{IoU}(\pi_1, \ldots, \pi_{i-1}) - \mathrm{IoU}(\pi_1, \ldots, \pi_i) \,, \tag{5}$$

$$\pi = \text{indices that sort } \boldsymbol{\xi} \text{ in descending order.}$$

Here, with slight abuse of notation, $\mathrm{IoU}(\pi_1, \ldots, \pi_i)$ denotes calculating IoU given the i pixels with the largest prediction error ξ_{π_i}. For multi-class prediction, we apply the binary loss in an all-vs-one manner:

$$\mathrm{LovászSoftmax}(\mathbf{p},\mathbf{t}) = \tfrac{1}{|C|} \textstyle\sum_{c \in C} \mathrm{LovászBinary}(\mathbf{p}^c, \mathbf{t}^c) \,, \tag{6}$$

where $\mathbf{t}^c$ is 1 if the ground-truth class is c, and 0 otherwise.

3.4. Hierarchical Lovász Embeddings

We propose to leverage the Lovász hinge loss to learn a shared embedding space for both semantics and instances. Given an instance I_l, with mean $\boldsymbol{\mu}_l$ and variance σ_l, the score of an embedding $\mathbf{e}_i \in \mathbb{S}^{d-1}$ lying on the unit hypersphere belonging to instance I_l is

$$p_l(\mathbf{e}_i) = \exp\left(-\frac{d_{\cos}(\mathbf{e}_i, \boldsymbol{\mu}_l)}{2\sigma_l^2}\right) \,, \tag{7}$$

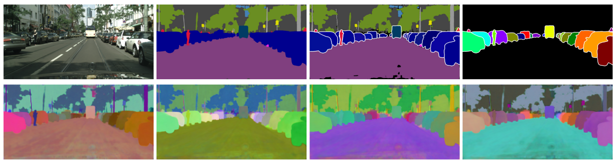

Figure 3. Our method predicts hierarchical embeddings that allow for a clear separation of semantics and instances. Top row from left: input image, semantic segmentation, panoptic segmentation, instance segmentation. Bottom row: our 12-dimensional embedding space visualized as RGB channels of 4 images, highlighting the hierarchical structure over categories and instances.

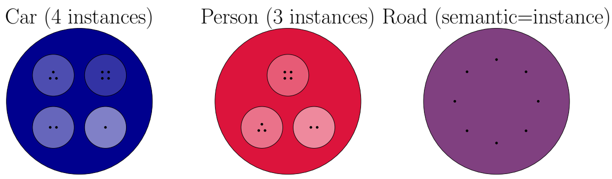

Figure 4. Our proposed loss groups pixel embeddings (black dots) into a hierarchy of jointly semantic- and instance-specific clusters. The border of each inner circle illustrates the decision boundary for the instance kernel ϕ. The outer circle shows the decision boundary for the semantic kernel ψ.

where $d_{\cos}(\mathbf{a},\mathbf{b}) = 1 - \mathbf{a}^\top \mathbf{b}$ denotes the cosine distance. Note that p_l plays both roles of the push and pull terms in the AE loss, as embeddings will be pulled together when the target score increases, and pushed apart otherwise. We use a unit hypersphere to enable the use of cosine distance, which is memory efficient. However, we emphasize that our method is compatible with any distance metric.

Instance and semantic losses via joint embedding. For instances, we use a spatial kernel to better separate far away objects as

$$\phi_l(\mathbf{e}_i) = p_l(\mathbf{e}_i) \exp\left(-\frac{\|\boldsymbol{\rho}_i - \boldsymbol{\rho}_l\|^2}{2\sigma_{l,\mathrm{spatial}}^2}\right) \,, \tag{8}$$

where $\boldsymbol{\rho}_i$ is the spatial position of embedding $\mathbf{e}_i$, and $\sigma_{l,\mathrm{spatial}}$ is a parameter learned by back propagation.

For handling the learning of semantics, we propose to associate with each semantic class S_k, a semantic mean $\boldsymbol{\mu}_k$ and semantic variance σ_k. We can then write the score of an embedding $\mathbf{e}_i$ belonging to this semantic class via the softmax function

$$\psi_k(\mathbf{e}_i) = \frac{p_k(\mathbf{e}_i)}{\sum_{c \in \mathcal{C}} p_c(\mathbf{e}_i)} \,. \tag{9}$$

While it is possible to simply use the Gaussian kernel p_k for semantics, preliminary experiments indicated that this is not optimal, as unlike instances, the number of semantic classes are known, and thus the decoding can be done by finding the closest embedding to each semantic mean, which motivates the softmax function.

We define the mean embedding $\boldsymbol{\mu}_k$ to be the mean embedding inside each image according to the ground-truth for instance indices k that exist in the current image, and the persistent estimate $\hat{\boldsymbol{\mu}}_k$ of the dataset mean otherwise:

$$\boldsymbol{\mu}_k = \begin{cases} \frac{1}{|I_k|} \sum_{j \in I_k} \mathbf{e}_j & \text{if } k \text{ is an instance} \\ \hat{\boldsymbol{\mu}}_k & \text{if } k \text{ is a semantic class} \end{cases} . \tag{10}$$

Given the predicted instance and semantic scores $\phi_l(\mathbf{e}_i)$ and $\psi_k(\mathbf{e}_i)$ for each embedding $\mathbf{e}_i$, we can minimize the Lovász hinge loss, to maximize the IoU metric on the training dataset. Our loss functions that act as the push and pull forces to form the hierarchical structure are defined as

$$\mathcal{L}_{\text{ins}} = \text{LovászBinary}(\phi_l(\mathbf{e}_i), \mathbf{t}_{\text{ins}}), \tag{11}$$
$$\mathcal{L}_{\text{seg}} = \text{LovászSoftmax}(\psi_k(\mathbf{e}_i), \mathbf{t}_{\text{seg}}) , \tag{12}$$

which are calculated for all embeddings $\mathbf{e}_i$, instances I_l and semantics S_k and the ground-truth vectors $\mathbf{t}$. This will cause the network to pull embeddings belonging to the same class together, and push apart embeddings that belong to different instances. The instance-semantic hierarchy is modeled by letting the network learn variance estimates that properly capture the hierarchy in the shared embedding space, with the semantic variance being larger than the instance variance.

Auxiliary losses. In addition to the embeddings for each pixel, our network also predicts a sigmoid output s_i for each pixel, estimating the value of the Gaussian kernel $\phi_l(\mathbf{e}_i)$. As shown in previous work, this "seed" value can be used at the decoding step to help finding an arbitrary number of instances [39]. For each pixel, we calculate

$$\mathcal{L}_{\text{seed}} = \begin{cases} \|s_i - \text{sg}[\phi_l(\mathbf{e}_i)]\|^2 & i \in I_l \\ \|s_i\|^2 & c_i \in \mathcal{C}_{\text{stuff}} \end{cases} , \tag{13}$$

where $\text{sg}[\cdot]$ is the stop gradient operator, e.g. as used in VQ-VAE [44]. Note that only instances are predicted to have a seed value, and stuff classes are regressed to zero.

We also add a loss term for each embedding $\mathbf{e}_i \in \mathcal{I}$ to encourage the predicted instance variances to be similar for the same object, to ease instance separation:

$$\mathcal{L}_{\text{ins-var}} = \gamma \|\sigma_i - \text{sg}[\hat{\sigma}_l]\|^2 , \tag{14}$$

where $\hat{\sigma}_l = \frac{1}{|I_l|} \sum_{j \in I_l} \sigma_j$, to encourage uniform predicted variance, and $\gamma = 10$.

Similar to previous work by Guerriero et al. [17], we found that the semantic means are difficult to learn via backpropagation, as the update is too slow to catch up with the other parts of the network. Therefore, similar to VQ-VAE [44], for semantic classes, we add an L2 regression loss term to store a persistent representation of each semantic class in the model

$$\mathcal{L}_{\text{seg-mean}} = \|\hat{\boldsymbol{\mu}}_k - \text{sg}[\boldsymbol{\mu}_k^{\text{batch}}]\|^2 , \tag{15}$$

where $\boldsymbol{\mu}_k^{\text{batch}}$ is the batch semantic mean gotten by averaging the predicted embeddings at the ground-truth locations of semantic class k. We illustrate the effect of this VQ-VAE style loss term in our ablation analysis in Section 4.4. The semantic variance σ_k is learned via backpropagation.

Proposed loss. Our proposed loss function is independent of the specific network architecture, and can be used with any standard fully convolutional neural network for semantic segmentation, such as DeeplabV3+ [8] or PSPNet [56]. The final form of our proposed loss function is

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{seg-mean}} + \mathcal{L}_{\text{ins}} + \mathcal{L}_{\text{ins-var}} + \mathcal{L}_{\text{seed}} , \tag{16}$$

averaged over all pixels $\mathcal{X}$. Specifically, our network predicts, for each pixel $\mathbf{x}_i$, a tuple $(\mathbf{e}_i, \sigma_i, s_i)$, where $\mathbf{e}_i$ is the predicted embedding, σ_i is the predicted instance variance, and s_i is the seed probability score.

Thomson initialization. To increase coverage of the embedding space, we can initialize the means to uniformly cover the unit hypersphere by solving the generalized Thomson problem [47], originally proposed in 1904 for the purpose of determining an electron configuration in physics. We initialize the means as

$$\text{argmin}_{\{\boldsymbol{\mu}_k\}\forall k} \sum_{i \neq j} \frac{1}{d_{\cos}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)} . \tag{17}$$

There is no known general solution to this problem, but we can find a local minimum via gradient descent. In our initial experiments, we found that Thomson initialization makes the training more stable.

3.5. Panoptic Decoding

For predicting semantics and instances from the learned hierarchical embedding space, we use a simple post-processing algorithm. We first assign the semantic class to each pixel via argmax of $\psi_k(\mathbf{e}_i)$. For thing classes, we first reduce the number of seed candidates s_i by 3×3 max pooling, similar to CornerNet [25]. Then we threshold the remaining seeds to get an initial set of candidate seeds. Each candidate seed s_i has an associated hierarchical embedding $\mathbf{e}_i$, so we can use the kernel $\Phi(\mathbf{e}_i, \mathbf{e}_j) = \exp(-d_{\cos}(\mathbf{e}_i, \mathbf{e}_j)/(2\sigma_i^2) - \|\boldsymbol{\rho}_i - \boldsymbol{\rho}_j\|^2/(2\sigma_{i,\text{spatial}}^2))$ to calculate the probability of two seeds (s_i, s_j) representing the same instance. We then merge seeds representing the same object based on the probability $\Phi(\mathbf{e}_i, \mathbf{e}_j)$. We now have one estimated seed s_i per instance. For each s_i, we then estimate

Method	Backbone	Pretrain.	PQ	PQ_{th}	PQ_{st}
Proposal-based					
Seamless [43]	ResNet50	ImageNet	**60.3**	**56.1**	63.3
UPSNet [52]	ResNet50	ImageNet	59.3	54.6	62.7
Real-time PS [20]	ResNet50	ImageNet	58.8	52.1	**63.7**
Pan.FPN [23]	ResNet50	ImageNet	57.7	51.6	62.2
Attn.-Guid. [29]	ResNet50	ImageNet	56.4	52.7	59.0
Li et al. [28]	ResNet101	ImageNet	47.3	39.6	52.9
DeGeus [13]	ResNet50	ImageNet	45.9	39.2	50.8
Proposal-free					
Pan. DeepL. [9]	ResNet50	ImageNet	59.7	-	-
SSAP [15]	ResNet50	ImageNet	57.6	50.4	-
DeeperLab [53]	Xception71	ImageNet	56.5	-	-
DeGeusFast [12]	ResNet50	ImageNet	55.1	48.3	60.1
HLE (Ours)	ResNet101	ImageNet	60.6	51.4	67.2
HLE (Ours)	ResNet50	ImageNet	**59.8**	**51.1**	**66.1**
HLE (Ours)	MobileNetV2	ImageNet	56.0	45.0	64.0

Table 1. Single-scale experimental results on the Cityscapes validation set. The best ResNet50 result in each category is highlighted.

the instance mask by thresholding $\Phi(\mathbf{e}_i, \mathbf{e}_k)$, for all remaining pixel embeddings $\mathbf{e}_k$, resolving mask disagreement by assigning each pixel to the instance with the highest probability among the estimated seeds. The semantic class of the instance is decided by the seed pixel. For stuff classes, we threshold the value of the semantic kernel $\psi_k(\mathbf{e}_i)$, to reduce the number of false positives. The panoptic decoding process can be easily optimized in favor of inference speed with little degradation of accuracy (details in the supplementary).

4. Experiments

4.1. Datasets

We conduct experiments on the Cityscapes [10], COCO [30], and Mapillary Vistas [38] datasets, to evaluate the performance of our model. Cityscapes is an image dataset for autonomous driving, depicting European street-level imagery at 1024×2048 resolution, labeled with 19 classes, out of which 8 are thing classes. COCO contains various indoor and outdoor images at varying resolutions, with 80 thing classes, and 53 stuff classes. Vistas contains a wide variety of street-level images, at varying resolution, with 37 thing classes and 28 stuff classes.

4.2. Experimental Details

We evaluate our method using the standard panoptic segmentation metric panoptic quality (PQ) [24]. In the supplementary material, we explain the PQ metric, and include its variations, $PQ^\dagger$ [43] and parsing covering (PC) [53], as well as mean Intersection-over-Union (mIoU) and Average Precision (AP).

On Cityscapes, we only use the fine annotations in training. Our models are trained on the training set ($2,975$ images) and evaluated on the validation set (500 images). For COCO, the training and validation sets have $118,287$ and $5,000$ images, respectively. Vistas has $18,000$ images in the training set, and $2,000$ in the validation set. We use the Adam optimizer [22] with learning rate 10^{-5}, polynomial learning rate decay, and test-time flip. We use ResNet50 as a backbone for DeepLabV3+ (pretrained on ImageNet) with an embedding dimension of 12 on Cityscapes, and 128 on COCO and Vistas. During training, we use crop size 1024×2048 on Cityscapes, and 512×512 on COCO, and crop around thing classes. The experiments on COCO and Vistas uses a max side length of 640 and 2048 pixels, respectively. We use Optuna [1] to find good hyperparameters for the decoder.

Method	Backbone	Pretrain.	PQ
Proposal-based			
Seamless [43]	ResNet50	ImageNet & Vistas	**62.6**
Proposal-free			
SSAP [15]	ResNet101	ImageNet	58.9
Dynam. inst. [2]	ResNet101	ImageNet	55.4
HLE (Ours)	ResNet101	ImageNet	**59.4**
HLE (Ours)	ResNet50	ImageNet	57.9

Table 2. Experimental results on the Cityscapes test set.

Method	Backbone	Pretrain.	PQ	PQ_{th}	PQ_{st}
Proposal-based					
UPSNet [52]	ResNet50	ImageNet	**42.5**	48.5	**33.4**
Attn.-Guid. [29]	ResNet50	ImageNet	39.6	**49.1**	25.2
Pan. FPN [23]	ResNet50	ImageNet	39.0	45.9	28.7
Real-time PS [20]	ResNet50	ImageNet	37.1	41.0	31.3
Proposal-free					
SSAP [15]	ResNet101	ImageNet	36.5	-	-
Pan. DeepL. [9]	ResNet50	ImageNet	35.1	-	-
DeeperLab [53]	Xception71	ImageNet	33.8	-	-
HLE (Ours)	ResNet101	ImageNet	38.1	42.8	31.0
HLE (Ours)	ResNet50	ImageNet	**37.1**	**41.1**	**30.9**

Table 3. Single-scale experimental results on the COCO validation set. The best ResNet50 result in each category is highlighted.

Network inference on a 1024×2048 image takes 91 ms running on single NVIDIA V100 Tensor Core GPU. Post-processing takes 113 ms. Alternatively, by running post-processing on a down-sampled embedding space, post-processing speed can be easily sped up to 8 ms at the expense of lowering Cityscapes validation set PQ to 58.4 (details are provided in the supplementary material). On a 800×1300 COCO image, inference takes 51 ms. Inference speed depends heavily on the backbone used. With a MobileNetV2 backbone, model inference takes 36 ms on Cityscapes.

4.3. Experimental Results

Our results on Cityscapes, COCO, and Vistas can be seen in Tables 1, 3 and 5, respectively. The best ResNet50

Method	Backbone	Pretrain.	PQ	PQ_{th}	PQ_{st}
Proposal-based					
UPSNet [52]	ResNet101	ImageNet	**46.6**	53.2	**36.7**
Attn.-Guid. [29]	ResNeXt-152	ImageNet	46.5	**55.8**	32.5
Pan. FPN [23]	ResNet50	ImageNet	40.9	48.3	29.7
Proposal-free					
SSAP [15]	ResNet101	ImageNet	36.9	40.1	32.0
DeeperLab [53]	Xception71	ImageNet	34.3	37.5	29.6
DeeperLab [53]	Wider MNV2	ImageNet	28.1	30.8	24.1
DeeperLab [53]	L. W. MNV2	ImageNet	24.5	26.9	20.9
HLE (Ours)	ResNet101	ImageNet	**39.9**	**45.0**	**32.2**
HLE (Ours)	ResNet50	ImageNet	38.2	42.7	31.4

Table 4. Experimental results on the COCO test-dev2019 set.

Method	Backbone	Pretrain.	PQ	PQ_{th}	PQ_{st}
Proposal-based					
Seamless [43]	ResNet50	ImageNet	**37.7**	**33.8**	**42.9**
Proposal-free					
Pan. DeepL. [9]	ResNet50	ImageNet	33.3	-	-
DeeperLab [53]	Xception71	ImageNet	32.0	-	-
HLE (Ours)	ResNet50	ImageNet	**34.6**	**27.8**	**43.5**

Table 5. Single-scale experimental results on the Vistas validation set. The best ResNet50 result in each category is highlighted.

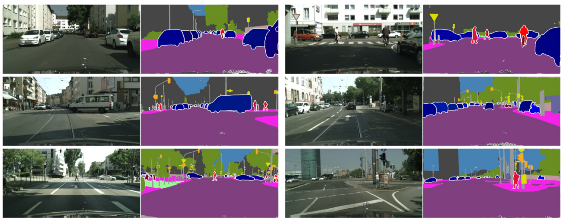

Figure 5. Visualization of our method on the Cityscapes validation set. Each pair: (left) input image, (right) panoptic segmentation.

Figure 6. Visualization of our method on the COCO validation set. Each pair: (left) input image, (right) panoptic segmentation.

result in each category is highlighted. Some examples of our model's outputs can be seen in Figures 5, 6 and 7. For a fair comparison, we report the results on ImageNet-trained models with a ResNet50 for previous works, if possible. We also report some results of our method with alternative backbones [19, 46], to illustrate the change in performance depending on the backbone network. Our method achieves state-of-the-art results among proposal-free methods on all three datasets with the same or even lighter backbone.

On Cityscapes, we can see that our method is competitive with heavier two-stage methods, such as PanopticFPN [23],

Figure 7. Visualization of our method on the Vistas validation set. Each pair: (left) input image, (right) panoptic segmentation.

Assoc. emb. loss	✓						
Cross-ent. loss		✓		✓			
Lovász hinge loss			✓	✓	✓	✓	✓
Sep. seg. branch.				✓			
Split emb. space			✓				
Hier. emb. space	✓	✓			✓	✓	✓
VQ-VAE style		✓	✓			✓	✓
Thomson init.		✓	✓	✓	✓		✓
PQ	38.8	45.0	50.2	57.0	57.0	58.6	**59.1**

Table 6. Ablative results on the Cityscapes validation set. Our proposed model is in the rightmost column.

and has higher PQ than previous proposal-based methods, such as Panoptic DeepLab [9]. Particularly, note that our method with a ResNet50 backbone network beats DeeperLab that uses a much heavier Xception71 backbone in terms of the PQ metric.

Notably, we achieve a high PQ_{st} score, indicating that using embeddings for semantic encoding is a promising direction, especially in panoptic segmentation. We can reason about the rationale that our model achieves top PQ_{st} results as follows. In our method, we learn both embeddings and variances for each semantic class. Each semantic class is treated in an adaptive manner based on distribution. For example, a larger volume in the embedding space will be assigned to semantic classes with higher variance, compared to direct regression where all classes are treated equally in feature space. An example of the learned hierarchical embedding space can be seen in Figure 3. Our proposed hierarchical embedding is able to encode both semantic and instance level information precisely, although a few mistaken predictions can also be seen.

We also submitted our model to the Cityscapes test set benchmark, where our method is competitive among both published and unpublished models, despite the big variety in network backbones and extra data included in the current benchmark. The results for peer-reviewed work can be seen in Table 2. The gap between our validation and test set PQ score is not large, indicating that our method generalizes well. While SSAP does not report any ResNet50 result, our model has higher PQ than theirs with ResNet101. In the supplementary material, we include Cityscapes results comparison on the metrics $PQ^{\dagger}$ (61.3), PC (76.6), $mIoU$ (77.3) and AP (23.9).

On COCO, we saw less proposal-free methods reported with similar network scale as shown in Table 3. We present COCO test set results in Table 4. Our method achieves the best performance among proposal-free models. Compared to Cityscapes, the gap between proposal-based and proposal-free models on COCO is much larger, indicating that there is still a demand for further research in order to close this gap. We find the same observation on the Vistas dataset as on COCO, where our proposed method is able to outperform other proposal-free methods. Notably, we find that bottom-up methods tend to under-perform on larger datasets [9, 20, 15], while our proposed hierarchical embedding space-based method is able to generalize better on these datasets.

4.4. Ablative Analysis

We conduct ablative experiments to better understand design choices we have made for our method. In Table 6, we provide different variations on each key design of our proposed method. For simplicity, we did not conduct hyper-parameter search in this experiment.

First of all, we compared our proposed hierarchical Lovász embeddings with conventional embeddings based on associative embedding or cross entropy loss. As can be seen in the table, Lovász embeddings have vastly better performance than associative embeddings or cross entropy losses.

Second, we compare with a model that uses a softmax classifier semantic segmentation branch with cross entropy loss, instead of our joint hierarchical embedding space, and still handling instance predictions via the embedding space. We can see that this yields a lower PQ of 57.0, which can be regarded as a naïve extension of previous instance segmentation work to panoptic segmentation [39]. Next, we illustrate the benefit of jointly learning the embedding space for both semantics and instances, compared to splitting the embedding space into two halves: one for semantics and the other for instance embeddings. We split the 12-dimensional embedding space of our model into 6 dimensions each for semantics and instances. It can be seen that jointly learning both semantics and instances in the same embedding space leads to higher performance. This interpretation is also strengthened by our method outperforming DeeperLab on the Cityscapes validation set – a method using a split embedding space.

Finally, we show that VQ-VAE style loss and Thomson initialization for learning the semantic means help improve the model performance.

4.5. Temporal Stability

As our embeddings are of low dimension, we are interested in seeing whether they exhibit temporal stability when run on a video. In Figure 8, we show the predicted embedding spaces and panoptic segmentation for a few consecutive frames in the Cityscapes demo video sequence. The figure qualitatively shows that matching pixels in different frames have similar embeddings. We believe this qualitative result indicates the potential of our method towards temporal downstream applications such as multi-object tracking. Temporal stability is important for applications such as tracking in autonomous driving, where flickering false positives can be fatal, and low-dimensional embeddings could be leveraged as a feature for cross-frame data association. We are not aware of other panoptic methods featuring this property.

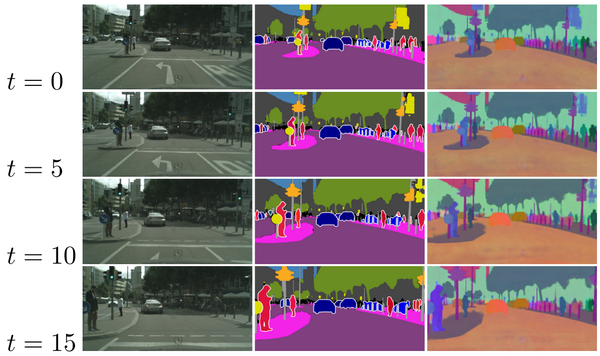

Figure 8. Our embeddings are temporally stable across video frames. Each row (video frame), from left to right: Input image, panoptic segmentation, first 3 dimensions of the embedding space. Matching pixels in consecutive frames have similar embeddings.

5. Conclusion

In this work, we presented a unified output representation for panoptic segmentation leveraging a hierarchically clustered embedding space. Via our proposed hierarchical Lovász hinge loss, we created a simple single-shot model for panoptic segmentation, predicting an embedding space structured into an instance-semantic hierarchy. Our method was shown to achieve state-of-the-art results among proposal-free panoptic segmentation methods, and be competitive with heavier two-stage methods on the Cityscapes, COCO and Vistas datasets. Results indicated that our hierarchical Lovász embeddings are a viable alternative for panoptic segmentation, for both thing and stuff classes.

We believe that it is an important and promising direction to explore a unified representation for panoptic segmentation. In the future, we will focus on leveraging the structure of the embedding space with self-adaptation to ontology distribution and its downstream applications.

Acknowledgments This work was supported by Woven Core, Inc. We thank Richard Calland, Aditya Ganeshan, Karim Hamzaoui, Daisuke Okanohara and Shunta Saito for discussions and feedback on the manuscript.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631. ACM, 2019. 6

[2] Anurag Arnab and Philip HS Torr. Pixelwise instance segmentation with a dynamically instantiated network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 441–450, 2017. 2, 6

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 2

[4] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4413–4421, 2018. 4

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 3

[6] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2018. 2

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2, 5

[9] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12485, 2020. 1, 3, 6, 7, 8

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 4, 6

[11] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*, 2017. 3

[12] Daan de Geus, Panagiotis Meletis, and Gijs Dubbelman. Fast panoptic segmentation network. *arXiv preprint arXiv:1910.03892*, 2019. 6

[13] Daan de Geus, Panagiotis Meletis, and Gijs Dubbelman. Single network panoptic segmentation for street scene understanding. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, 2019. 1, 3, 6

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[15] Naiyu Gao, Yanhu Shan, Yupei Wang, Xin Zhao, Yinan Yu, Ming Yang, and Kaiqi Huang. Ssap: Single-shot instance segmentation with affinity pyramid. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 642–651, 2019. 1, 2, 3, 6, 7, 8, 13

[16] Joshua Goodman. Classes for fast maximum entropy training. In *26th International Conference on Acoustics, Speech, and Signal Processing*, 2001. 2

[17] Samantha Guerriero, Barbara Caputo, and Thomas Mensink. Deepncm: Deep nearest class mean classifiers. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*, 2018. 3, 5

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[20] Rui Hou, Jie Li, Arjun Bhargava, Allan Raventos, Vitor Guizilini, Chao Fang, Jerome Lynch, and Adrien Gaidon. Real-time panoptic segmentation from dense detections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8523–8532, 2020. 3, 6, 8, 13

[21] Jyh-Jing Hwang, Stella X Yu, Jianbo Shi, Maxwell D Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7334–7344, 2019. 2

[22] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 6

[23] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019. 1, 3, 6, 7, 13

[24] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. 1, 2, 3, 4, 6, 12

[25] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018. 2, 3, 5

[26] Justin Lazarow, Kwonjoon Lee, and Zhuowen Tu. Learning instance occlusion for panoptic segmentation. *arXiv preprint arXiv:1906.05896*, 2019. 3

[27] Jie Li, Allan Raventos, Arjun Bhargava, Takaaki Tagawa, and Adrien Gaidon. Learning to fuse things and stuff. *arXiv preprint arXiv:1812.01192*, 2018. 1, 3

[28] Qizhu Li, Anurag Arnab, and Philip HS Torr. Weakly-and semi-supervised panoptic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 102–118, 2018. 6, 13

[29] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7026–7035, 2019. 1, 3, 6, 7, 13

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 6

[31] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. An end-to-end network for panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6181, 2019. 3

[32] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018. 2, 13

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[34] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 2

[35] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2

[36] Andriy Mnih and Geoffrey E Hinton. A scalable hierarchical distributed language model. In *Advances in neural information processing systems*, pages 1081–1088, 2009. 2

[37] Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In *Aistats*, volume 5, pages 246–252. Citeseer, 2005. 2

[38] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *International Conference on Computer Vision (ICCV)*, 2017. 2, 6

[39] Davy Neven, Bert De Brabandere, Marc Proesmans, and Luc Van Gool. Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8837–8845, 2019. 2, 3, 5, 8

[40] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Towards end-to-end lane detection: an instance segmentation approach. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 286–291. IEEE, 2018. 3

[41] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. In *Advances in neural information processing systems*, pages 2171–2180, 2017. 3

[42] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems*, pages 2277–2287, 2017. 2, 3

[43] Lorenzo Porzi, Samuel Rota Bulo, Aleksander Colovic, and Peter Kontschieder. Seamless scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8277–8286, 2019. 1, 3, 6, 7, 12, 13

[44] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *arXiv preprint arXiv:1906.00446*, 2019. 5

[45] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 2

[46] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 7

[47] JJ Thomson. On the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure. *Philos. Mag., Ser. 6*, 7:237–265, 1904. 5

[48] Joseph Tighe, Marc Niethammer, and Svetlana Lazebnik. Scene parsing with object instances and occlusion ordering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3748–3755, 2014. 1

[49] Zhuowen Tu, Xiangrong Chen, Alan L Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of computer vision*, 63(2):113–140, 2005. 1

[50] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. SOLO: Segmenting objects by locations. In *Proc. Eur. Conf. Computer Vision (ECCV)*, 2020. 2

[51] Xinlong Wang, Shu Liu, Xiaoyong Shen, Chunhua Shen, and Jiaya Jia. Associatively segmenting instances and semantics in point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4096–4105, 2019. 3

[52] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8818–8826, 2019. 1, 3, 6, 7, 13

[53] Tien-Ju Yang, Maxwell D Collins, Yukun Zhu, Jyh-Jing Hwang, Ting Liu, Xiao Zhang, Vivienne Sze, George Papandreou, and Liang-Chieh Chen. Deeperlab: Single-shot image parser. *arXiv preprint arXiv:1902.05093*, 2019. 1, 3, 6, 7, 12, 13

[54] Yibo Yang, Hongyang Li, Xia Li, Qijie Zhao, Jianlong Wu, and Zhouchen Lin. Sognet: Scene overlap graph network for panoptic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12637–12644, 2020. 3

[55] Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 702–709, 2012. 1

[56] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2, 5

[57] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2, 3