This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Segmenting Known Objects and Unseen Unknowns without Prior Knowledge

Stefano Gasperini[1,2,◦] Alvaro Marcos-Ramiro[2] Michael Schmidt[2]
Nassir Navab[1] Benjamin Busam[1] Federico Tombari[1,3]

[1] Technical University of Munich [2] BMW Group [3] Google

Abstract

Panoptic segmentation methods assign a known class to each pixel given in input. Even for state-of-the-art approaches, this inevitably enforces decisions that systematically lead to wrong predictions for objects outside the training categories. However, robustness against out-of-distribution samples and corner cases is crucial in safety-critical settings to avoid dangerous consequences. Since real-world datasets cannot contain enough data points to adequately sample the long tail of the underlying distribution, models must be able to deal with unseen and unknown scenarios as well. Previous methods targeted this by re-identifying already-seen unlabeled objects. In this work, we propose the necessary step to extend segmentation with a new setting which we term holistic segmentation. Holistic segmentation aims to identify and separate objects of unseen, unknown categories into instances without any prior knowledge about them while performing panoptic segmentation of known classes. We tackle this new problem with U3HS, which finds unknowns as highly uncertain regions and clusters their corresponding instance-aware embeddings into individual objects. By doing so, for the first time in panoptic segmentation with unknown objects, our U3HS is trained without unknown categories, reducing assumptions and leaving the settings as unconstrained as in real-life scenarios. Extensive experiments on public data from MS COCO, Cityscapes, and Lost&Found demonstrate the effectiveness of U3HS for this new, challenging, and assumptions-free setting called holistic segmentation. Project page: https://holisticseg.github.io.

1. Introduction

Since neural networks have achieved unprecedented performance in perception tasks (e.g., object detection and semantic segmentation), there has been a growing interest in ensuring their safe deployment, especially important for safety-critical scenarios, such as autonomous driving and robotics [25]. Recently, several works have been proposed to improve robustness and generalization by addressing corner cases and out-of-distribution data [6, 65, 22], via domain adaptation [70], adversarial augmentations [44], simulations [1], and uncertainty estimation [54].

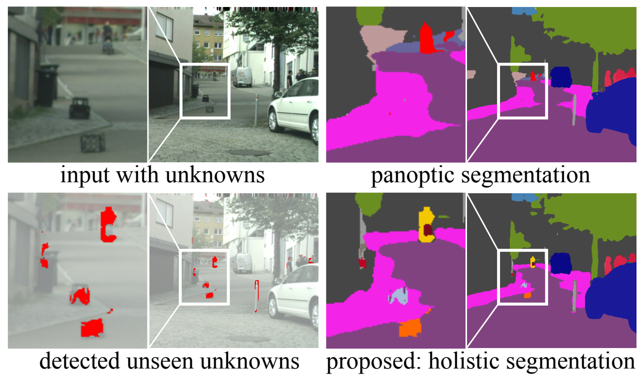

Figure 1. State-of-the-art panoptic segmentation methods [13] cannot deal with unseen classes from [52] (top right). Instead, our U3HS addresses the proposed holistic segmentation setting. U3HS finds unseen unknowns (bottom left) and separates them into instances (bottom right) without prior knowledge about unknowns.

Due to the difficulty of collecting corner cases from the long tail of the underlying data distribution, current datasets cannot fully represent the diversity of the world, leaving its vast majority as difficult out-of-distribution (OOD) samples [32, 8]. In safety-critical applications, considering them during development and deployment is of utmost importance [3, 44], or they could cause severe damage.

Furthermore, since the powerful and popular *softmax* highly promotes the probability of the highest *logit*, state-of-the-art methods tend to be overly confident even on wrong predictions [58, 25]. In safety-critical settings, reliable confidence together with interpretability techniques [20, 77] increases trust for downstream tasks [25], e.g., trajectory prediction and path planning. Towards this end, estimating the uncertainty of a model's output is commonly considered a key enabler for its safe applicability [39, 25].

◦ This work was conducted while working at BMW Group.
Contact author: Stefano Gasperini (*stefano.gasperini@tum.de*).

While several works addressed some of these problems for image classification [58, 46, 54, 61] and object detection [48, 21], they remain primarily unexplored for dense tasks such as semantic and panoptic segmentation [5, 66]. Compared to object detection, the problem is more severe for dense tasks, where a model needs to provide a prediction for every input unit (e.g., each pixel). So unseen objects (i.e., of new, unseen categories) are systematically and wrongly assigned to one of the limited number of known classes (closed-set), as shown in Figure 1. This has led researchers towards designing new methods that work not only with the available data distribution but also with OOD samples that are not available (open-set), thereby improving robustness against unseen scenarios [38, 46, 5, 44].

Open-set panoptic segmentation [66] segments instances of unlabeled objects in addition to panoptic segmentation of known areas, i.e., the combination of semantic and instance segmentation [41]. Unlike OOD segmentation [38], segmenting unknown instances enables tracking and trajectory prediction. Prior works [66, 36, 72] tackled this problem by relying on seeing unlabeled categories during training. They learned these categories through the *void* class (i.e., unlabeled) and assumed unknowns to be within ground truth *void* regions at training time and inside *void* predictions at test time. By doing so, unknowns are transformed into learned unlabeled instances (i.e., essentially known objects) [36], constraining the open-set task. Mainly intended to segment already-seen unlabeled objects [36, 72], current works cannot deal with the wide variability of unknowns and corner cases outside the training data.

In this paper, we propose the necessary next step for panoptic segmentation to include object categories outside the training data (i.e., unseen unknowns). We term the new setting holistic segmentation. The aim is to identify and segment unseen unknowns into instances while segmenting known classes in a panoptic fashion without any external nor prior knowledge about unknowns. Unseen categories pose new challenges compared to already-seen unlabeled ones [36], requiring new solutions. Estimating the uncertainty is a key step towards finding the knowledge boundaries of a model, leaving the problem unconstrained and reducing assumptions on the training data. Therefore, we propose U3HS: Unseen Unknowns via Uncertainty estimation for Holistic Segmentation. The main contributions of this paper can be summarized as follows:

- We introduce the setting of holistic segmentation, which highlights the importance of not using prior knowledge about unknown objects (e.g., text), and leaves the setup unconstrained as in real scenarios.
- We tackle this new setting with U3HS: the first panoptic framework to deal with unseen, unknown object categories, able to segment and separate them.
- We provide uncertainty measures for the output of U3HS to further improve its safe applicability.

2. Related Work

Closed-set panoptic segmentation Combining semantic and instance segmentation, panoptic segmentation [41] distinguishes *things* (countable classes) from *stuff* (amorphous). The vast majority of methods are top-down [71, 60, 53, 40, 34, 50]: two-stage exploiting box proposals and *thing* masks from Mask R-CNN [30], and filling up *stuff* areas with a semantic branch. Bottom-up are proposal-free, e.g., Panoptic-DeepLab [13]: they segment semantically and cluster instances within *thing* regions [13, 64]. A different line of work proposed end-to-end solutions [23, 63] where instance and semantic segments are delivered directly by treating instance segmentation as a class-agnostic classification task. Others explored self-attention [34], videos [67, 11, 47], scene graphs [69, 68], multi-task learning [16, 28], neural fields [43], or text descriptions [18]. Our U3HS framework deals with unseen unknowns and extends [13] via instance-aware embeddings.

Zero-shot learning aims to predict unseen classes outside the training set [4, 78, 75] with the help of external knowledge [10, 26], e.g., a language model [76], used to build semantic spaces common between seen and unseen classes [74]. While zero-shot methods detect only unseen classes at inference time, generalized zero-shot approaches also detect seen ones [9, 56], similarly to the proposed holistic segmentation. Also **open-vocabulary** methods are zero-shot [35, 27, 73, 18]. They exploit language models such as CLIP [57] to describe unknowns. CLIP has been trained on unknown classes and is treated as an oracle, as it is assumed to be able to describe every unknown, allowing open-vocabulary and zero-shot approaches to identify them. However, this implies that unknown classes are known, e.g., to CLIP. Moreover, CLIP is not immune to corner cases and long tail samples [57]. This limits the pool of objects that these methods can recognize. As shown in Figure 2, holistic segmentation segments unseen objects too, but unlike zero-shot and open-vocabulary, it does not use any external support (e.g., text descriptions of unknowns), such that objects of unseen categories are segmented solely by learning on known ones. More recent than this work, SAM [42] is a strong foundation model. Unlike SAM, ours does not use any prompts and outputs semantic classes.

Uncertainty estimation Epistemic uncertainty is caused by the model itself, while aleatoric is due to the input [39, 25]. OOD data typically results in high epistemic due to a knowledge gap. Single deterministic approaches are sampling-free and provide predictions and uncertainty estimates with the same model [25]. Among these, DUQ [61] learns class representatives and compares them with input features. SNGP [46] improves the awareness to domain

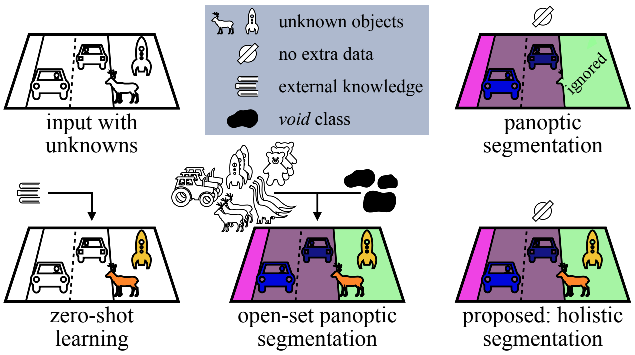

Figure 2. Comparison between closed-set (top right) [41] and open-set [36] panoptic segmentation, zero-shot learning [78], and the proposed holistic segmentation setting. While zero-shot and open-set panoptic methods commonly leverage knowledge about unknown objects, holistic segmentation does not use any priors.

shifts via weight normalization and a Gaussian process. DPN [58] predicts the parameters of a Dirichlet distribution, and uses a Dirichlet density function for each probability assignment and its uncertainty. Various works estimated uncertainty for object detection [49, 48, 21], and segmentation [55, 5, 59], improving robustness and generalization. While most compute uncertainty only to provide it as extra output [25, 59], our U3HS can be paired with any of the above techniques to find unknown objects, which it then separates into instances for holistic segmentation.

Open-set perception Open-set tasks are similar to generalized zero-shot learning [9], with the fundamental difference that here no external knowledge on the unseen classes is used [75]: inference is based only on what was learned from the training data. Uncertainty estimation helps to identify knowledge boundaries [49]. Bayesian SSD [49] uses Dropout sampling for open-set object detection. MLUC [6] tackles this for LiDAR point clouds via metric learning and unsupervised clustering. Open world recognition [2, 7] labels detected unknowns and adds them to the training set. Pham et al. [51] grouped regions perceptually to known and unknown instances, exploiting edges, boxes, and masks. Recently, several works tackled open-set semantic segmentation, telling apart unknown areas from known classes [38, 33, 29]. Among those not learning from OOD data, DML [5] uses metric learning and SML [38] acts as post-processing, standardizing the max *logits*, improving the class distributions. Instead, our work separates unseen, unknown objects into instances and segments known areas, addressing the proposed holistic segmentation.

Open-set panoptic segmentation The pioneering OSIS [66] was the first in this direction. Applied to LiDAR data, it exploits 3D locations to cluster unlabeled points into instances. Later, EOPSN [36] extended Panoptic FPN [40] and grouped its proposals into clusters. At training time, EOPSN clusters similar unlabeled objects across multiple inputs. When surrounded by known segments, it labels an unlabeled object and uses it to learn to segment its instances. Instead, DDOSP [72] uses a known-unknown class discriminator and class-agnostic proposals. However, as these approaches were intended to re-identify already-seen unlabeled objects, they all rely on seeing unknown data at training time [66, 36, 72]. They all cluster into instances what falls in the predicted *void* class, learned as a fallback (i.e., must be in the training samples) and assumed to contain all unknowns. Since datasets are limited [44], by requiring to learn from unknowns and the *void* class, existing works are not designed to deal with any completely unseen object, as their pool of identifiable unknowns is also limited [36, 72]. For these reasons, they solve only part of the problem. Instead, by using no OOD data at training time and preventing learning priors for unknowns, the proposed holistic segmentation differs from the way open-set panoptic segmentation has been tackled so far [66, 36, 72]. As shown in Figure 2, our setting allows to segment without constraints and separate even unseen unknown objects. In contrast to all existing approaches, the proposed U3HS neither relies on seeing unknowns at training time nor learns the *void* class. U3HS is the first method to solve this unconstrained, assumptions-free holistic segmentation setting.

3. Proposed Setting: Holistic Segmentation

As shown in Figure 2, the proposed setting of holistic segmentation is a logical extension of open-set panoptic segmentation [66, 36]. To make the setting unconstrained as real-life scenarios, we aim to identify and separate **any unseen, unknown object** into instances while segmenting known classes. Other settings allow to include unknowns in the training data [66, 36, 72] (e.g., within the *void* class) or use information about them [78], and only re-identify already-seen unlabeled objects [36]. Instead, we focus on the case where **no information is available about unknowns**. Therefore, holistic segmentation is more challenging, makes no assumptions about the training data (e.g., the presence of *void*, with unknowns in it), leaves the problem unconstrained to any object, and simplifies data collection as no unknowns need to be in the training data. Formal definition and metrics follow open-set panoptic segmentation [66, 36]. As shown in Figure 2, the outputs are comparable, but the definition of unknowns differs (here unseen), as well as the inability to learn from unknowns. Unknown instances can then be used for downstream tracking, trajectory prediction, path planning, or active learning.

4. Proposed Framework: U3HS

In Figure 3, we show a representation of U3HS, targeting holistic segmentation. U3HS outputs instances of unseen unknowns by clustering instance-aware embeddings corre-

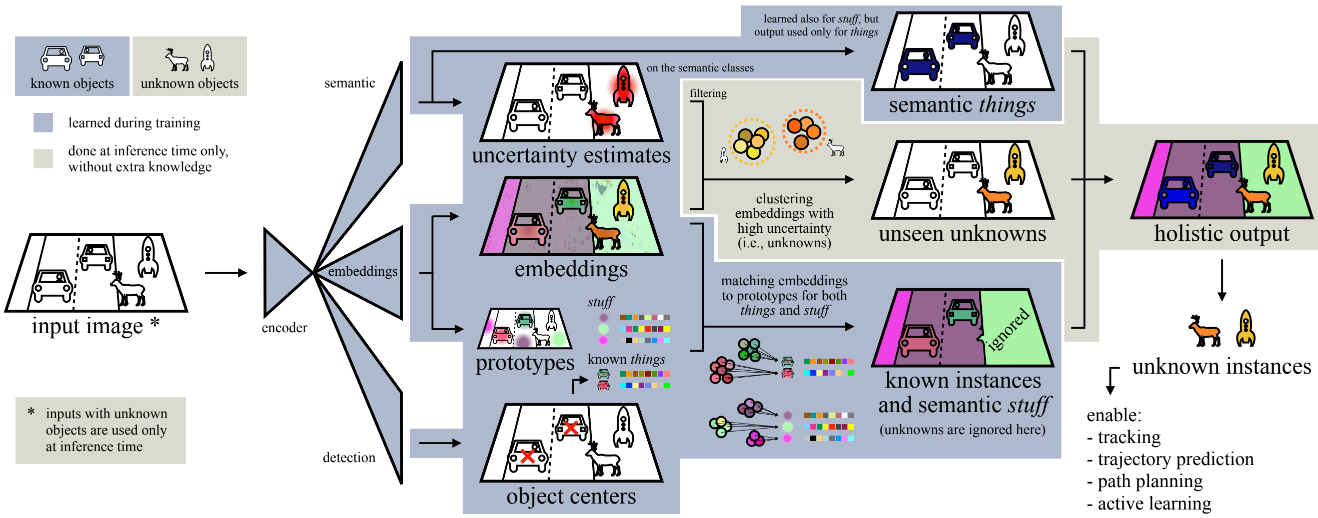

Figure 3. The proposed U3HS framework. Uncertainty is estimated in the semantic branch, and with the instance-aware embeddings, it determines unknown instances. Known instances are found via center regression and formed by grouping embeddings with their prototypes.

sponding to highly uncertain regions (Section 4.2). We use uncertainty estimation to distinguish known classes from unknowns, while embeddings are learned solely on known objects with panoptic segmentation (Section 4.1).

4.1. Panoptic Segmentation for Known Classes

Our approach for closed-set panoptic segmentation builds upon learning instance-aware embeddings. As shown in Figure 3, an encoder extracts features from an input image and propagates them to different decoders: 1) a semantic branch performing semantic segmentation and uncertainty estimation to identify unknown regions (Section 4.2); 2) a detection branch identifying object centers similarly to Panoptic-DeepLab [13]; and 3) an embeddings branch, with two separate heads, for prototypes and embeddings.

We make the embeddings instance-aware via discriminative loss functions (Section 4.3) and by concatenating the detection branch features to prototype and embeddings heads. Embeddings and detections are made also semantic-aware by concatenating the semantic *logits* to the last layers of the heads. Prototypes Ω are feature vectors the prototype head predicts to represent objects and *stuff* classes. While computed at each pixel, only the features at object centers C are considered *thing* prototypes, plus one for each *stuff* class. This is inspired by the size heatmap in [79]. *Thing* prototypes are $\Omega_{th} = \{(\mu_o, \sigma_o^2) \in \mathbb{R}^F \times \mathbb{R}^+ : o \in I\}$, one for each object o of all detected instances I. F is the embedding size, μ and σ^2 are mean and variance. *Stuff* prototypes are $\Omega_{st} = \{(\mu_k, \sigma_k^2) \in \mathbb{R}^F \times \mathbb{R}^+ : k \in K_{st}\}$, one for each *stuff* class $k \in K_{st}$ independently.

Simlarly to [66], the embeddings head predicts embeddings $\phi_{(i,j)} \in \mathbb{R}^F$ for each pixel (i, j), then matches them with their prototype $\omega \in \Omega$ as follows. We compute association scores $\hat{\mathbf{y}}_{(i,j),\omega}$ for each pixel (i, j) and prototype as:

$$\hat{\mathbf{y}}_{(i,j),\omega} = -||\phi_{(i,j)} - \mu_\omega||^2 / 2\sigma_\omega^2 \tag{1}$$

Compared to [66], we relax the problem by not including the term $-\frac{F}{2} \log \sigma_\omega^2$, and let the embedding variance be indirectly controlled by the final task, which naturally bounds it (shown empirically in Section 5.2). Then, we keep the prototype variance σ_ω^2 strictly positive by using *softplus*.

At inference time, for *things*, the semantic class of each instance is determined by majority voting of its semantic branch predictions, ensuring output consistency. Instead, the ID is computed from the highest score in Eq. 1. For *stuff* regions, we follow [66], determining the semantic classes by associating the pixel embeddings to the prototypes Ω_{st} via the highest scoring class from Eq. 1. This decoupling allows semantic awareness throughout the model.

4.2. Dealing with Unseen Unknown Objects

We find unknown segments by relying on uncertainty estimates, which can help identify the knowledge boundaries of a model [58, 46]. Specifically, instead of predicting the *void* class and searching in it for unknowns as in [66, 36], we estimate the uncertainty related to the semantic segmentation predictions and consider as unknown the areas with a high associated uncertainty. Although our framework can flexibly work with various uncertainty estimators (Section 5.2), here we exemplify it with DPN [58, 37], which we extended from image classification to semantic segmentation, and also improved its convergence in this context. We chose DPNs as they allow for minimal modifications at training time, i.e., replacing the *softmax* with a strictly positive activation function while providing good uncertainty estimates on OOD data without training on such data [58].

Following [58], we consider the evidence $e_k = \alpha_k - 1$ as a measure of the number of hints given by data for a pixel to be assigned to a class $k \in K$ known classes, with α_k being the parameters of the Dirichlet distribution $Dir(\alpha)$. We compute the uncertainty as $u = K/\sum_{k=1}^{K}\alpha_k$. Given that the class probabilities $\mathbf{p} = \{p_k : k = [1, ..., K]\}$ follow a simplex (i.e., are positive and sum to 1), the class assignment corresponds to a Dirichlet distribution parametrized over the evidence, as the probability density function:

$$\begin{aligned} D(\mathbf{p}|\boldsymbol{\alpha}) = &\ B(\boldsymbol{\alpha})^{-1}\prod_{k=1}^{K} p_k^{\alpha_k - 1} \\ \text{with: } B(\alpha) = &\ \prod_{k=1}^{K}\Gamma(\alpha_k)/\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) \end{aligned} \tag{2}$$

where Γ is the gamma function and $B(\alpha)$ is the K-dimensional multinomial beta function [58].

We apply this to semantic segmentation by predicting a concentration parameter $\alpha^{(i,j)}$ for each pixel (i, j), replacing the last layer with the smooth *softplus* activation function, thus converting the *logits* to a strictly positive vector, which we use as evidence $e^{(i,j)}$ in the Dirichlet distribution. We learn this distribution with the semantic loss $\mathcal{L}_s$ minimizing the negative expected log likelihood of the correct class $Y^{(i,j)}$, for the random variable $\mathcal{X}^{(i,j)} \sim \text{Dir}(\alpha^{(i,j)})$:

$$\begin{aligned} \mathcal{L}_s^{(i,j)} &= -E[\ln \mathcal{X}^{(i,j)}_{Y^{(i,j)}}] \\ &= \psi\left(\ \sum_{k=1}^{K}\alpha_{(i,j),k}\ \right) - \psi(\alpha_{Y^{(i,j)}}) \end{aligned} \tag{3}$$

where ψ is the digamma function (i.e., Γ's logarithmic derivative) and $\alpha_{(i,j),k}$ is the output of the semantic branch. Due to the difficulty of modeling the target distribution in our holistic setting, we omit the KL term used in [58], simplifying the loss design (Section 5.2). After training on the closed-set data, we consider all pixels (i, j) with an estimated uncertainty $u_{(i,j)} \geq \mu + t \cdot \sigma$ as unknown regions with μ and σ^2 being mean and variance of the uncertainties of all training pixels, and t being a hyperparameter.

Separating unknowns After finding the unknown segments, we cluster their instance-aware embeddings trained only on known objects into individual unknowns using DBSCAN [19]. We find the DBSCAN hyperparameters on the training closed-set data (Appendix). Finally, we re-assign the few DBSCAN's outliers to their originally predicted semantic class, thus ignoring their uncertainty estimates.

4.3. Learning to Find Knowns and Unknowns

We train our models with a combination of four losses. The semantic branch is optimized with $\mathcal{L}_s^{(i,j)}$ (Eq. 3) over the whole image sized $W \times H$ as:

$$\begin{aligned} \mathcal{L}_s = &\ \frac{1}{WH}\sum_{i,j} -E[\ln \mathcal{X}^{(i,j)}_{Y^{(i,j)}}] \\ = &\ \frac{1}{WH}\sum_{i,j}\psi\left(\sum_{k=1}^{K}\alpha_{(i,j),k}\right) - \psi(\alpha_{Y^{(i,j)}}) \end{aligned} \tag{4}$$

As in [13], the detection branch is trained with an L2 loss between predicted $\hat{C}$ and ground truth C center heatmaps:

$$\mathcal{L}_o = \ \frac{1}{WH}\sum_{i,j}\left(\hat{C}^{(i,j)} - C^{(i,j)}\right)^2 \tag{5}$$

For *stuff*, we use the predicted Ω_{st} as a pseudo label to learn the prototypes Ω. For *things*, the same is done with Ω_{th} at the true instance centers. The prototype loss $\mathcal{L}_p$ is the cross-entropy on the *softmax* of the association scores $\hat{\mathbf{y}}_{(i,j),\omega}$, as $\hat{\mathbf{z}}_{(i,j),\omega} = \exp(\hat{\mathbf{y}}_{(i,j),\omega})/\sum_{\omega'\in\Omega}\exp(\hat{\mathbf{y}}_{(i,j),\omega'})$, with $\omega_{(i,j)}$ being the pseudo label prototype:

$$\mathcal{L}_p = \ \frac{1}{WH}\sum_{i,j} -\log(\hat{\mathbf{z}}_{(i,j),\omega_{(i,j)}}) \tag{6}$$

We learn embeddings $\phi_{(x,y)}$ with a discriminative loss [15] $\mathcal{L}_d$ (Appendix). The overall training objective is:

$$\mathcal{L} = \lambda_1\mathcal{L}_s + \lambda_2\mathcal{L}_o + \lambda_3\mathcal{L}_p + \lambda_4\mathcal{L}_d \tag{7}$$

5. Experiments and Results

5.1. Experimental Setup

Datasets We conducted our experiments on three public datasets, namely Cityscapes [14], Lost&Found [52], and MS COCO [45]. **Cityscapes** is a popular outdoor benchmark. Recorded around 50 different cities, mainly in Germany, it contains 19 classes: 8 *things* and 11 *stuff*. We followed the standard split, with 2975 images for training and 500 as validation set, reporting all metrics on the latter. Also recorded in Germany, the **Lost&Found** dataset contains a variety of unusual OOD objects placed in the middle of the road. We selected it because: 1) it was recorded with the same sensor setup as Cityscapes, allowing seamless transfers and removing the need for fine-tuning; 2) it contains only real images; and 3) unlike similar datasets [3, 8], it provides instance annotations for unknowns. Therefore, it is a challenging complement to Cityscapes for holistic segmentation. We did not train on Lost&Found, but used it only to evaluate models trained on Cityscapes. We report all metrics on the *unknown* class of its 1202 test samples. **MS COCO** is a challenging large-scale benchmark for general image understanding, as it includes a variety of scenarios from indoor to outdoor. The 2017 panoptic split contains 80 *thing* categories, and 53 *stuff* classes. We followed EOPSN [36] by treating as unknown the least frequent 20% *thing* classes (e.g., *bear*, *frisbee*). However, instead of turning their segments into *void* and keeping their images in the training set as in [36], we removed their samples completely and regarded them as unseen unknowns. This reduced the training samples to 98112, with 117 classes to learn. We report on the 827 validation samples with unseen classes.

Evaluation metrics We evaluated the panoptic quality (**PQ**) metric [41] separately for known classes and unknowns, including recognition (RQ) and segmentation (SQ) qualities. We report PQ on the held-out classes of COCO [45], the unknown class of Lost&Found [52], as well as on the 19 known classes of Cityscapes [14] for both **open and closed** settings. Specifically, in open cases, models detect both knowns and unknowns, while in closed set-

Method	Assumptions	Lost&Found (*unseen*) PQ	RQ	SQ
EOPSN [36]	data, *void*	0^*	0^*	0^*
OSIS [66]	data, *void*	1.45	2.23	**65.11**
U3HS [ours]	**none**	**7.94**	**12.37**	64.24

Table 1. Segmentation of unseen unknown objects (*unknown* class) of **Lost&Found** [52] test set after training on Cityscapes [14] and transferring with no fine-tuning. *: EOPSN diverged (null TP).

Method	Assumptions	COCO (*unseen*) PQ	RQ	SQ
EOPSN [36]	data, *void*	0.40	0.50	<u>80.30</u>
void-train [36]	data, *void*	4.40	5.90	74.80
void-supp. [36]	data, *void*	4.50	6.00	75.90
DDOSP [72]	data, *void*	<u>9.30</u>	<u>11.20</u>	**82.50**
U3HS [ours]	**none**	**9.62**	**13.20**	72.84

Table 2. Segmentation of unseen, unknown objects (20% least frequent, held-out classes) on the validation set of **MS COCO** [45]. All others learn *void* and are taken from [72], added trailing zero.

tings, the same models predict only knowns, which in practice means ignoring the uncertainty estimates. By analyzing both, we explore the trade-off between detecting unknowns (open) and the in-domain performance (closed).

Network architecture All our models share the structure with Panoptic-DeepLab [13], using a ResNet50 [31] backbone and decoders following Deep-LabV3+ [12]. ResNet50 was chosen to increase reproducibility with limited resources. As described in Section 4.2, the only modification to the semantic decoder is applying the *softplus* activation to quantify the uncertainty. The other branches follow Panoptic-DeepLab for detecting centers and DeepLabV3+ for the embeddings, with two heads.

Implementation details For [14, 52], we used input images sized 1024×512 and batch size 16. For [45], we fed 8 images sized 640×480. We used the Adam optimizer until convergence, with an initial learning rate of 0.001, which was reduced by 2% at each epoch. We set $t = 3$ for the uncertainty threshold (i.e., 3 times the standard deviation) and $F = 8$ for the embedding size to keep the memory low. We adjusted to the different data distribution of COCO with $t = 1$. The backbone was pre-trained on ImageNet [17]. The losses were weighted $\lambda_1 = \lambda_3 = \lambda_4 = 1$ and $\lambda_2 = 200$ [13].

Prior works We compared our U3HS with open-set panoptic works: OSIS [66], which we adapted from LiDAR point clouds to images, EOPSN and its baselines [36], and DDOSP [72]. Instead of training them directly on the unknown categories being evaluated (as in [36]), we followed their setup [66, 36, 72] by training them with the *void* class as fallback, and applied them to unseen unknowns. All methods followed this setup, except that ours ignored *void*. On COCO, we facilitated other works following the K=5% setting of EOPSN [36], thus turning 4 classes into *void*, so they learned 4 classes less than ours. We repurposed and extended a variety of uncertainty estimators [58, 61, 46] from image classification to semantic segmentation (Appendix). We then extended them to holistic segmentation by incorporating them in our U3HS framework.

5.2. Quantitative Results

Unseen unknowns, L&F Table 1 compares our U3HS with prior approaches when segmenting instances of unseen unknowns from Lost&Found [52]. OSIS [66] was the first to address the more limited open-set panoptic segmentation setting, followed by EOPSN [36]. However, OSIS performance fell short on PQ for unseen unknowns, proving the severe limitation of relying on unknowns at training time. By learning *void*, OSIS achieved the highest SQ, which ignores wrong predictions [41]. Instead, despite numerous attempts, EOPSN [36] did not work: it diverged as soon as the exemplars were mined, obtaining 0 true positives (TP). We attribute this to the inconsistent similarities within the *void* class of Cityscapes, compared to those across existing major classes treated as *void* (e.g., *car* in their setup). This prevented EOPSN from forming meaningful clusters from the proposal features during training [36]. Despite the similar setup to EOPSN [36], OSIS [66] could converge since it does not rely on associating unknowns across images. Our U3HS outperformed OSIS by 5.5 times on PQ.

Unseen unknowns, COCO In Table 2, we show results on the 16 held-out categories of MS COCO [45]. In this case, other works were trained following the K=5% setup of EOPSN [36], where 4 classes were learned as *void* (*car*, *cow*, *pizza*, and *toilet*). While this allows them to learn more meaningful representations of unknowns (as unlabeled), it limits the number of classes they can distinguish semantically, e.g., they cannot identify cars. For the other works, the benefit of expanding the *void* distribution by enforcing the inclusion of a variety of recurring objects (e.g., pizza and cars) is evident as it allowed EOPSN to converge, although to a low PQ on the unseen objects. DDOSP [72] delivered a PQ similar to ours, albeit requiring to turn some knowns into *void*, learning unknowns via *void* and only 113 classes. Without altering the data nor making any assumption on the training samples (e.g., the presence of unknowns within *void* as in [36, 72, 66]), our U3HS performed the best on the unseen categories, especially on RQ (i.e., ability to form instances of unknowns), while learning the whole set of 117 classes, thereby distinguishing even more classes than the other works. Avoiding data assumptions with respect to unknowns made our U3HS effective at segmenting unseen unknowns across both datasets.

Known-unknown Table 3 reports the performances in-

ID	Method	Lost&Found (*unseen*) PQ	RQ	SQ	open Cityscapes PQ	RQ	SQ	closed Cityscapes PQ	RQ	SQ
-	Panoptic-DeepLab [13]	-	-	-	-	-	-	<u>45.82</u>	<u>57.66</u>	<u>79.46</u>
-	OSIS [66]	1.45	2.23	65.11	39.42	50.20	78.53	39.42	50.20	78.53
A1	[ours] baseline: semantic uncertainty	0.49	0.82	60.16	35.02	44.83	78.10	35.97	46.12	78.00
A2	A1 + relaxed embedding association	3.64	5.27	**69.09**	**42.14**	**53.46**	78.83	43.99	55.98	78.59
A3	A2 + prototype head = **U3HS**	**7.94**	**12.37**	64.24	<u>41.21</u>	<u>51.67</u>	<u>79.77</u>	**46.53**	**58.99**	78.87
A4	A3 – reassigning outliers	<u>7.85</u>	<u>12.25</u>	64.11	39.84	49.97	79.75	**46.53**	**58.99**	78.87
A5	A4 – majority voting	<u>7.85</u>	<u>12.25</u>	64.11	23.94	30.15	79.41	26.77	33.86	79.06
A6	A4 – semantic embeddings	2.33	3.48	<u>67.01</u>	35.16	43.34	**81.13**	35.92	44.30	**81.07**
U1	U3HS [ours] + *softmax* uncertainty	0.10	0.20	51.45	39.70	50.77	78.20	45.12	56.83	**79.40**
U2	U3HS [ours] + DUQ [61]	0.56	0.89	62.56	**41.68**	**53.14**	78.42	45.90	58.17	78.90
U3	U3HS [ours] + DPN [58]	2.09	3.30	63.43	38.90	49.56	78.49	44.91	56.95	78.85
U4	U3HS [ours] + SNGP [46]	<u>4.65</u>	<u>7.57</u>	61.49	41.02	<u>51.98</u>	<u>78.91</u>	<u>46.23</u>	<u>58.56</u>	<u>78.95</u>
A3	U3HS [ours] + improved DPN [ours]	**7.94**	**12.37**	<u>64.24</u>	<u>41.21</u>	51.67	**79.77**	**46.53**	**58.99**	78.87

Table 3. Segmentation comparison of models trained on Cityscapes [14] and transferred to the test set of Lost&Found [52] without fine-tuning. All were trained with the same constraints (e.g., ResNet50 [31], small batch, and image sizes). An ablation study (A1-A6) shows the impact of the main components of U3HS, with A3 being our full approach. A3 is paired with various uncertainty estimators (U1-U4).

domain, under open and closed settings (Section 5.1). Ideally, a method would suffer from no decrease in PQ between the two settings, meaning that its estimates are aligned with the distribution shift between knowns and unknowns. OSIS [66] does not use uncertainty estimation, so it does not have these two operating modes, resulting in identical open and closed-set outputs, as if it had only the open setting (via the prediction of *void*). Conversely, all others suffered from a reasonable decrease when extended to open-set. DUQ [61] had the smallest gap, which could be attributed to its underestimation of the uncertainty, as supported by its low scores on Lost&Found.

Closed-set In Table 3, we also compare our U3HS with Panoptic-DeepLab [13]. For a fair comparison, both approaches and all others were trained with the same backbone, image, and batch sizes (Section 5.1). As these were all smaller than those used in [13] due to the limited resources used, they resulted in a lower PQ than that reported in [13]. Nevertheless, our full approach (A3) achieved a slightly higher PQ on Cityscapes under the same setting. We attribute this to the effectiveness of the instance-aware discriminative embeddings learned by our approach, compared to the offset vectors and grouping used by Panoptic-DeepLab. As the focus is unknowns, experiments with improved training resources are out of the scope of this work.

Uncertainty In Table 3, we compare various uncertainty estimations paired to our U3HS framework (U1-U4, A3). While DUQ [61] and *softmax* underperformed compared to OSIS [66], DPN [58] and SNGP [46] achieved a higher PQ. Nevertheless, our improved DPN paired with our framework outperformed prior methods by a substantial margin (A3). For DPN and SNGP, this can be attributed to the superiority of our uncertainty estimates. Compared to OSIS and EOPSN, U3HS's combination of uncertainty estimation with instance-aware embeddings was more effective than learning *void* when encountering wholly new and unseen objects, such as those found in unconstrained settings (e.g., this transfer to Lost&Found).

Ablation study Table 3 reports an ablation of the main components of our U3HS, showing their benefits for holistic segmentation. Compared to the open-set panoptic OSIS [66], with A1, we reduced assumptions not learning the *void* class, and we added a semantic branch with uncertainty for unknowns, which by itself worsened the performance. However, combining this with a relaxed embedding association (Section 4.1) for *things* and *stuff* improved all metrics (A2). A dedicated prototype head (A3, i.e., full approach) increased them even further, more than doubling the PQ on unknowns (i.e., Lost&Found). Specifically, dedicated heads allow both prototypes and embeddings to be more meaningful and expressive without sacrificing the other. A4 shows the impact of reassigning outliers (Section 4.2). While its effect was limited on unknowns, it was more significant on Cityscapes [14]. Transforming unknown predictions in standard in-domain outputs is relevant only in open settings. A5 shows the effect of majority voting to enforce consistency between the outputs (Section 4.1). This did not affect unknowns since classes are not distinguished among them, but it significantly impacted RQ and PQ on Cityscapes. Finally, A6 shows the importance of learning the embeddings according to their semantic classes. In A6, predictions are made by the dedicated semantic branch without the model learning to distinguish the embeddings semantically. Although this increased SQ, it caused a discrepancy within

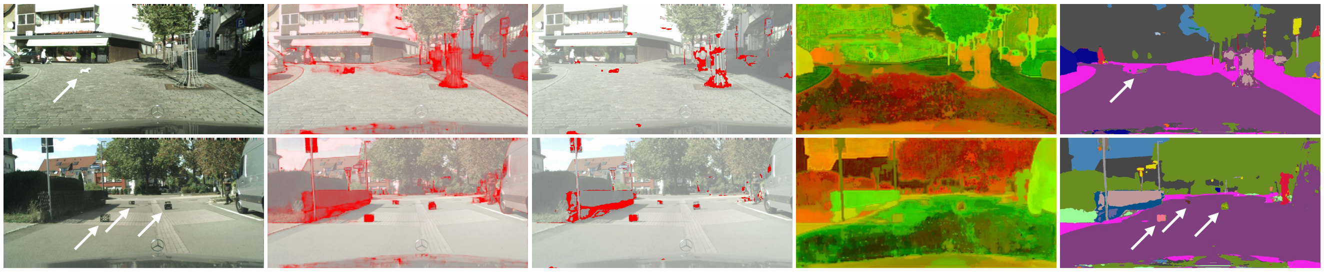

Figure 4. Example predictions of U3HS on OOD data from the Lost&Found [52] test set. The model was trained on Cityscapes [14] and transferred to Lost&Found without fine-tuning. Embeddings are projected to RGB via t-SNE [62]. White arrows mark labeled unknowns.

the model outputs, decreasing RQ and PQ.

5.3. Qualitative Results

Figure 4 shows example predictions of the proposed U3HS. The images illustrate the setting difficulty and provide examples of the OOD objects of **Lost&Found** [52]. These are often small and hard to see, hidden in the shade or far away from the camera. As seen in the quantitative results (Section 5.2), our U3HS could distinguish instances of unknowns (e.g., stroller in Figure 1), albeit leaving room for improvement. While unknowns correctly triggered high uncertainty estimates, their necessary filtering (third col.) sometimes left too few pixels, if any, on unknowns, leading to missed predictions. However, this is to be expected without any access to OOD data. Furthermore, without distinguishing between unknown *things* and unknown *stuff*, also structures (e.g., fence in the lower image) were given an ID. Nevertheless, thanks to our learned instance-aware embeddings, these were not further subdivided but formed a single large instance (e.g., blue in the lower output). Separate unusual *stuff* regions had the same effect, e.g., the structures around the trees in the upper image. This proves that instances are not simply created by separating disjoint OOD segments but are formed using the learned embeddings. As shown in Figure 4, the embeddings are closely coupled with the uncertainty estimates and the outputs.

Figure 5. Example predictions of U3HS on OOD data from the COCO [45] validation set. The model had never seen images containing *bear* or *frisbee* (part of the held-out classes), nor had any information about them. Colors represent the predicted instances.

Figure 5 reports predictions of U3HS on samples of **MS COCO** [45] containing two held-out classes (i.e., *bear* and *frisbee*). Remarkably, U3HS was able to separate both bears and frisbees into individual instances despite their high inter-class similarity and not having accessed any information about them. This is thanks to the uncertainty estimation and instance-aware embeddings of our U3HS.

Data considerations and limitations Lost&Found [52] introduces a significant domain shift from Cityscapes. By placing real OOD objects on the road, the authors had to choose unusual scenarios (Figure 4), causing the whole scenes to be OOD. This leads to high uncertainty estimates also on a few known areas. As we do not use any OOD data, nothing constrains high uncertainty to unknown segments, decreasing PQ. A similar issue occurs in COCO, albeit less severely, thanks to more training data. However, COCO has no dedicated unknown class, so it had to be extracted from the set of known ones. Nevertheless, results show that uncertainty is highly valuable, allowing to leave the settings unconstrained. U3HS would mainly benefit from improvements in uncertainty estimates, embeddings descriptiveness, and their clustering. So, learning-based clustering [23, 24] could be advantageous.

The **Supplementary Material** includes more details on the proposed holistic segmentation setting, U3HS and the baselines, as well as additional results, including the trade-off between in-domain and OOD performances, failure cases and qualitative comparisons.

6. Conclusion

In this paper we introduced holistic segmentation: a new setting addressing completely unseen unknown objects in unconstrained scenarios. Additionally, we presented U3HS: the first solution for this new problem. Thanks to its uncertainty estimation and instance-aware learned embeddings, U3HS identifies and separates instances of completely unseen unknowns without any information about them, while segmenting known regions. Extensive experiments on multiple datasets showed the effectiveness of U3HS.

References

[1] Sara Beery, Yang Liu, Dan Morris, Jim Piavis, Ashish Kapoor, Neel Joshi, Markus Meister, and Pietro Perona. Synthetic examples improve generalization for rare classes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 863–873, 2020. 1

[2] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1893–1902, 2015. 3

[3] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The Fishyscapes benchmark: Measuring blind spots in semantic segmentation. *Springer International Journal of Computer Vision*, 129(11):3119–3135, 2021. 1, 5

[4] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[5] Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. Deep metric learning for open world semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15333–15342, 2021. 2, 3

[6] Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. Open-set 3D object detection. In *International Conference on 3D Vision (3DV)*, pages 869–878. IEEE, 2021. 1, 3

[7] Jun Cen, Peng Yun, Shiwei Zhang, Junhao Cai, Di Luan, Mingqian Tang, Ming Liu, and Michael Yu Wang. Open-world semantic segmentation for LiDAR point clouds. In *Proceedings of the European Conference on Computer Vision*, pages 318–334. Springer, 2022. 3

[8] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. SegmentMeIfYouCan: A benchmark for anomaly segmentation. In *Neural Information Processing Systems - Datasets and Benchmarks Track*, 2021. 1, 5

[9] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *Proceedings of the European Conference on Computer Vision*, pages 52–68. Springer, 2016. 2, 3

[10] Jiaoyan Chen, Yuxia Geng, Zhuo Chen, Ian Horrocks, Jeff Z. Pan, and Huajun Chen. Knowledge-aware zero-shot learning: Survey and perspective. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 4366–4373, 2021. 2

[11] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D Collins, Ekin D Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 695–714. Springer, 2020. 2

[12] 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*, pages 801–818. Springer, 2018. 6

[13] 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, 2, 4, 5, 6, 7

[14] 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. 5, 6, 7, 8

[15] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*, 2017. 5

[16] Daan de Geus, Panagiotis Meletis, Chenyang Lu, Xiaoxiao Wen, and Gijs Dubbelman. Part-aware panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5485–5494, 2021. 2

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 6

[18] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation with MaskCLIP. *arXiv preprint arXiv:2208.08984*, 2022. 2

[19] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996. 5

[20] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2950–2958, 2019. 1

[21] Stefano Gasperini, Jan Haug, Mohammad-Ali Nikouei Mahani, Alvaro Marcos-Ramiro, Nassir Navab, Benjamin Busam, and Federico Tombari. CertainNet: Sampling-free uncertainty estimation for object detection. *IEEE Robotics and Automation Letters*, 7(2):698–705, 2021. 2, 3

[22] Stefano Gasperini, Patrick Koch, Vinzenz Dallabetta, Nassir Navab, Benjamin Busam, and Federico Tombari. R4Dyn: Exploring radar for self-supervised monocular depth estimation of dynamic scenes. In *International Conference on 3D Vision (3DV)*, pages 751–760. IEEE, 2021. 1

[23] Stefano Gasperini, Mohammad-Ali Nikouei Mahani, Alvaro Marcos-Ramiro, Nassir Navab, and Federico Tombari. Panoster: End-to-end panoptic segmentation of LiDAR point clouds. *IEEE Robotics and Automation Letters*, 6(2):3216–3223, 2021. 2, 8

[24] Stefano Gasperini, Magdalini Paschali, Carsten Hopke, David Wittmann, and Nassir Navab. Signal clustering with class-independent segmentation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3982–3986. IEEE, 2020. 8

[25] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021. 1, 2, 3

[26] Yuxia Geng, Jiaoyan Chen, Xiang Zhuang, Zhuo Chen, Jeff Z Pan, Juan Li, Zonggang Yuan, and Huajun Chen. Benchmarking knowledge-driven zero-shot learning. *Elsevir Journal of Web Semantics*, 75:100757, 2023. 2

[27] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Proceedings of the European Conference on Computer Vision*, pages 540–557. Springer, 2022. 2

[28] Kratarth Goel, Praveen Srinivasan, Sarah Tariq, and James Philbin. QuadroNet: Multi-task learning for real-time semantic depth aware instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 315–324, 2021. 2

[29] Matej Grcić, Petra Bevandić, and Siniša Šegvić. Densehybrid: Hybrid anomaly detection for dense open-set recognition. In *Proceedings of the European Conference on Computer Vision*, pages 500–517. Springer, 2022. 3

[30] 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

[31] 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. 6, 7

[32] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 1

[33] Jie Hong, Weihao Li, Junlin Han, Jiyang Zheng, Pengfei Fang, Mehrtash Harandi, and Lars Petersson. GOSS: Towards generalized open-set semantic segmentation. *Springer The Visual Computer*, pages 1–14, 2023. 3

[34] 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. 2

[35] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7031, 2022. 2

[36] Jaedong Hwang, Seoung Wug Oh, Joon-Young Lee, and Bohyung Han. Exemplar-based open-set panoptic segmentation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1175–1184, 2021. 2, 3, 4, 5, 6

[37] Taejong Joo, Uijung Chung, and Min-Gwan Seo. Being Bayesian about categorical probability. In *International Conference on Machine Learning*, pages 4950–4961. PMLR, 2020. 4

[38] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized Max Logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15425–15434, 2021. 2, 3

[39] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 2017. 1, 2

[40] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019. 2, 3

[41] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. 2, 3, 5, 6

[42] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[43] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic Neural Fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12871–12881, 2022. 2

[44] Alexander Lehner, Stefano Gasperini, Alvaro Marcos-Ramiro, Michael Schmidt, Mohammad-Ali Nikouei Mahani, Nassir Navab, Benjamin Busam, and Federico Tombari. 3D-VField: Adversarial augmentation of point clouds for domain generalization in 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17295–17304, 2022. 1, 2, 3

[45] 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 *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014. 5, 6, 8

[46] Jeremiah Z. Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Advances in Neural Information Processing Systems*, 2020. 2, 4, 6, 7

[47] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21033–21043, 2022. 2

[48] Dimity Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In *Proceedings of the International Conference on Robotics and Automation*, pages 2348–2354. IEEE, 2019. 2, 3

[49] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection

in open-set conditions. In *Proceedings of the International Conference on Robotics and Automation*, pages 3243–3249. IEEE, 2018. 3

[50] Rohit Mohan and Abhinav Valada. EfficientPS: Efficient panoptic segmentation. *Springer International Journal of Computer Vision*, 129(5):1551–1579, 2021. 2

[51] Trung Pham, Thanh-Toan Do, Gustavo Carneiro, Ian Reid, et al. Bayesian semantic instance segmentation in open set world. In *Proceedings of the European Conference on Computer Vision*, pages 3–18. Springer, 2018. 3

[52] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and Found: Detecting small road hazards for self-driving vehicles. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1099–1106, 2016. 1, 5, 6, 7, 8

[53] Lorenzo Porzi, Samuel Rota Bulo, Aleksander Colovic, and Peter Kontschieder. Seamless scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8277–8286, 2019. 2

[54] Janis Postels, Hermann Blum, Yannick Strümpler, Cesar Cadena, Roland Siegwart, Luc Van Gool, and Federico Tombari. The hidden uncertainty in a neural networks activations. *arXiv preprint arXiv:2012.03082*, 2020. 1, 2

[55] Janis Postels, Francesco Ferroni, Huseyin Coskun, Nassir Navab, and Federico Tombari. Sampling-free epistemic uncertainty estimation using approximated variance propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2931–2940, 2019. 3

[56] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[58] Murat Sensoy, Lance M. Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing*, pages 3183–3193, 2018. 1, 2, 3, 4, 5, 6, 7

[59] Kshitij Sirohi, Sajad Marvi, Daniel Büscher, and Wolfram Burgard. Uncertainty-aware panoptic segmentation. *IEEE Robotics and Automation Letters*, 8(5):2629–2636, 2023. 3

[60] Konstantin Sofiiuk, Olga Barinova, and Anton Konushin. AdaptIS: Adaptive instance selection network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7355–7363, 2019. 2

[61] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, pages 9690–9700. PMLR, 2020. 2, 6, 7

[62] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008. 8

[63] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. MaX-DeepLab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5463–5474, 2021. 2

[64] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 108–126. Springer, 2020. 2

[65] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 1

[66] Kelvin Wong, Shenlong Wang, Mengye Ren, Ming Liang, and Raquel Urtasun. Identifying unknown instances for autonomous driving. In *Proceedings of the Conference on Robot Learning*, pages 384–393. PMLR, 2020. 2, 3, 4, 6, 7

[67] Sanghyun Woo, Dahun Kim, Joon-Young Lee, and In So Kweon. Learning to associate every segment for video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2705–2714, 2021. 2

[68] Shun-Cheng Wu, Keisuke Tateno, Nassir Navab, and Federico Tombari. Incremental 3D semantic scene graph prediction from RGB sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5064–5074, 2023. 2

[69] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. SceneGraphFusion: Incremental 3D scene graph prediction from RGB-D sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7515–7525, 2021. 2

[70] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. DANNet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15769–15778, 2021. 1

[71] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 8818–8826, 2019. 2

[72] Hai-Ming Xu, Hao Chen, Lingqiao Liu, and Yufei Yin. Two-stage decision improves open-set panoptic segmentation. *arXiv preprint arXiv:2207.02504*, 2022. 2, 3, 6

[73] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *Proceedings of the European Conference on Computer Vision*, pages 736–753. Springer, 2022. 2

[74] Kai Yi, Xiaoqian Shen, Yunhao Gou, and Mohamed Elhoseiny. Exploring hierarchical graph representation for large-scale zero-shot image classification. In *Proceedings of the European Conference on Computer Vision*, pages 116–132. Springer, 2022. 2

[75] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15404–15414, 2021. 2, 3

[76] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 2

[77] Yang Zhang, Ashkan Khakzar, Yawei Li, Azade Farshad, Seong Tae Kim, and Nassir Navab. Fine-grained neural network explanation by identifying input features with predictive information. *Advances in Neural Information Processing Systems*, 34:20040–20051, 2021. 1

[78] Ye Zheng, Jiahong Wu, Yongqiang Qin, Faen Zhang, and Li Cui. Zero-shot instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2593–2602, 2021. 2, 3

[79] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 4