Holistic Segmentation

Stefano Gasperini\textsuperscript{1,2,\textdegree} Alvaro Marcos-Ramiro\textsuperscript{2} Michael Schmidt\textsuperscript{2} Nassir Navab\textsuperscript{1,3} Benjamin Busam\textsuperscript{1} Federico Tombari\textsuperscript{1,4}

\textsuperscript{1} Technical University of Munich \textsuperscript{2} BMW Group \textsuperscript{3} Johns Hopkins University \textsuperscript{4} Google

Abstract

Panoptic segmentation methods assign a known class to each pixel given in input. Even for state-of-the-art approaches, this inherently enforces decisions that systematically lead to wrong predictions for unknown objects that are not part of the training categories. However, in safety-critical settings, robustness against out-of-distribution samples and corner cases is crucial to avoid dangerous consequences. Since real-world datasets cannot contain enough data points to properly sample the long tail of the underlying distribution, models must be able to deal with unknown and unseen scenarios as well. Previous methods targeted this issue by re-identifying already seen unlabeled objects. In this work, we propose the necessary step to extend segmentation with a new task which we term holistic segmentation. The aim of holistic segmentation is 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 not trained with unknown categories, reducing assumptions and leaving the settings as unconstrained as in real-life scenarios. Extensive experiments on publicly available data from Cityscapes\textsuperscript{[14]} and Lost&Found\textsuperscript{[52]} demonstrate the effectiveness of U3HS for the new challenging task of holistic segmentation.

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\textsuperscript{[27]}. Recently, several methods have been proposed to improve robustness and generalization\textsuperscript{[6, 64]}, by addressing corner cases and out-of-distribution data, via domain adaptation\textsuperscript{[45, 68]}, adversarial augmentations\textsuperscript{[43]}, simulations\textsuperscript{[1]}, sensor fusion\textsuperscript{[24]}, and uncertainty estimation\textsuperscript{[46, 54]}.

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\textsuperscript{[8, 33]}. In safety-critical applications, considering them is of utmost importance, as OOD samples could cause severe damage if not properly taken into account during development and deployment\textsuperscript{[3, 43]}.

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\textsuperscript{[27, 57]}. In safety-critical settings, reliable confidence together with interpretability techniques\textsuperscript{[20, 76]} increases trust for downstream tasks\textsuperscript{[27]}, e.g., trajectory prediction and path planning. Towards this end, es-

\degree Contact author: Stefano Gasperini (stefano.gasperini@tum.de).
timating the uncertainty of a model’s output is commonly considered a key enabler for its safe applicability [27,39].

While several works addressed some of these problems for image classification [44,54,57,60] and object detection [23,48], they remain mostly unexplored for dense tasks such as semantic and panoptic segmentation [5,65]. In dense tasks, a model needs to provide a prediction for every unit of the input (e.g., each pixel). Therefore, as shown in Figure 1, unseen objects (i.e., of new, unseen categories) will be systematically and wrongly assigned to one of the limited number of known classes (closed-set). This has led researchers towards designing new methods that work not only with the available data distribution, but also with OOD samples which are not available (open-set), thereby improving robustness against unseen scenarios [5,38,43,44].

Open-set panoptic segmentation [65] segments instances of unlabeled objects in addition to the standard panoptic segmentation, which is the combination of semantic and instance segmentation [41]. Unlike OOD segmentation [38], identifying unknown instances enables tracking and trajectory prediction. Prior works [36,65,70] tackled open-set panoptic segmentation by relying on seeing unlabeled categories already during training. They learned these categories through the \textit{void} class (i.e., unlabeled) and assumed unknowns to be confined within ground truth \textit{void} regions at training time and inside \textit{void} predictions at test time. By doing so, unknown objects are transformed into learned unlabeled instances (i.e., essentially known objects) [36], hence constraining the open-set task. Mainly intended to segment already seen unlabeled objects [36,70], current approaches cannot deal with the wide variability of unknown objects and corner cases outside the available training data.

In this paper, we propose the necessary next step for panoptic segmentation to include categories of objects outside the training data (i.e., unseen unknowns). We term the new task holistic segmentation. The aim of holistic segmentation is to identify and separate unseen unknowns into instances, while segmenting known classes in a panoptic fashion, without any external nor prior knowledge about unknown objects. Unseen categories pose different challenges compared to already seen unlabeled ones [36,70], requiring new solutions. Estimating the uncertainty is a key step towards holistic segmentation, to identify the knowledge boundaries of a model, leave the problem unconstrained and reduce assumptions on the training data. Therefore, we propose U3HS: \underline{U}nseen \underline{U}nknowns via \underline{U}ncertainty estimation for \underline{H}olistic \underline{S}egmentation. The main contributions of this paper can be summarized as follows:

- We introduce the task of holistic segmentation, which highlights the importance of not using prior knowledge about unknown objects, and leaves the settings as unconstrained as in real-life scenarios.
- We tackle this novel task with U3HS: the first panoptic framework to deal with unseen unknown object categories, capable to segment and separate them.
- We provide uncertainty measures for the output of U3HS to further improve its applicability in safety-critical settings, such as autonomous driving.

2. Related Work

\textbf{Closed-set panoptic segmentation} By combining semantic and instance segmentation, Kirillov et al. [41] proposed panoptic segmentation. They distinguished between \textit{thing} and \textit{stuff} classes, as countable objects and amorphous regions respectively. Most existing methods can be divided into two categories: bottom-up and top-down. The vast majority are top-down [35,40,50,53,59,69], which are two-stage methods exploiting box proposals and instance masks of \textit{things} from Mask R-CNN [31], and filling up \textit{stuff} regions with a semantic branch. Bottom-up methods, e.g., Panoptic-DeepLab [13], are proposal-free. They perform semantic segmentation, then cluster instances within \textit{thing} segments [13,63]. A different line of work proposed end-to-end solutions [25,62] where both instance and semantic segments are delivered directly by treating instance segmentation as a class-agnostic classification problem. Others explored panoptic segmentation with self-attention [35], videos [11,47,66], scene graphs [67,71], multi-task learning [16,29], neural fields [42], or text descriptions [18]. Our U3HS framework deals with unknowns and extends [13] to form instances via instance-aware embeddings.

\textbf{Zero-shot learning} Zero-shot learning aims to predict unseen classes that are not part of the training set [4,74,77] with the help of external knowledge [10,28], e.g., a language model [75], used to build semantic spaces common between seen and unseen classes [72]. While zero-shot methods detect only unseen classes at inference time, generalized zero-shot approaches [9,56] detect also seen ones, similarly to the proposed holistic segmentation. However, as shown in Figure 2, holistic segmentation does not use any external support, such that objects of unseen categories can be found and segmented solely via known ones.

\textbf{Uncertainty estimation} There are two types of uncertainty: epistemic caused by the model itself, and aleatoric due to the input [27,39]. OOD data typically results in high epistemic uncertainty, due to a knowledge gap. Methods can be categorized in sampling-based (e.g., MC Dropout [22]) and sampling-free [55], depending on whether they aggregate multiple predictions or not. The latter better suits real-time applications and includes single deterministic approaches, which provide both prediction and uncertainty estimates with the same model [27]. Among these, DUQ [60] learns class representatives and compares them with input features to find OOD data, while
unknown objects into instances and segments known areas, the classes distributions. Instead, our work separates unseen logits as post-processing, standardizing the max known classes [30, 34, 38]. Among those not learning from semantic segmentation, telling apart unknown areas from boxes and masks. Recently, several works tackled open-set practically to known and unknown instances, exploiting edges, clustering. Open world recognition [2,7] labels detected un-

SNGP [44] improves the model’s awareness to domain shifts via weight normalization and a Gaussian process. DPN [57] predicts the parameters of a Dirichlet distribution, instead of a categorical one, and uses a Dirichlet density function, representing each probability assignment and its uncertainty. Various works estimated uncertainty for object detection [23,48,49], and segmentation [5,55,58], improving robustness and generalization. While most compute uncertainty only to provide it as extra output [27,58], our U3HS integrates it to find unknown objects, which are then separated into instances for holistic segmentation. Exemplified with DPN [57], which we extended and modified for this task, our U3HS supports any of the above techniques.

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 [74] and inference is based only on what was learned from the training data. Therefore, uncertainty estimation becomes relevant, helping 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] proposed a Bayesian approach that groups 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 [30,34,38]. 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 classes distributions. Instead, our work separates unseen unknown objects into instances and segments known areas, addressing the new task of holistic segmentation.

Open-set panoptic segmentation The pioneering OSIS [65] was the first in this direction: applied on LiDAR data, it exploits 3D locations to cluster unlabeled points into instances. Later with EOPSN, Hwang et al. [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, they associate a label to an unlabeled object and use it to learn to segment its instances. Instead, Xu et al. [70] used 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 [36,65,70]. 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 a dataset cannot include all possible object types [43], by requiring to learn from unknowns and the void class, existing methods are not designed to deal with any completely unseen object, solving only part of the problem. Instead, the proposed holistic segmentation differs from the way open-set panoptic segmentation has been tackled so far [36,65,70], as it uses no OOD data at training time and prevents from learning priors for unknowns. As shown in Figure 2, this 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 and assumptions-free problem we named holistic segmentation.

3. Proposed Task: Holistic Segmentation

As shown in Figure 2, the proposed task of holistic segmentation is a logical extension of open-set panoptic segmentation [36,65]. In order to make the task as unconstrained as real-life scenarios, holistic segmentation aims to identify and separate unseen unknown objects into instances, while segmenting known classes. Other tasks allow to include unknowns in the training data [36,65] (e.g., within the void class) or use information about them [77], and only re-identify already seen unlabeled objects [36]. Therefore, holistic segmentation is more challenging, makes no assumptions about the training data (e.g., the presence of unknowns), and leaves the problem unconstrained to any type of object. The formal definition and evaluation metric follow those of open-set panoptic segmentation [36,65]. What changes is the definition of unknowns, and the inability of learning from unknowns.

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-
4.1. Closed-set: Panoptic Segmentation for Knowns

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 logis to the last layers of the heads. The prototype head predicts a feature vector for every pixel. However, for things, these are considered meaningful instance prototypes only at the corresponding object centers found by the detection branch. This is inspired by the box dimensions heatmap in [78]. Such prototypes represent each instance, distinguishing them from one another. The head predicts the set of thing prototypes \( \Omega_{th} = \{ (\mu_o, \sigma^2_o) \in \mathbb{R}^F \times \mathbb{R}^+ : o \in I \} \), one for each object \( o \) of all detected instances \( I \). \( F \) is the embedding size, \( \mu \) and \( \sigma^2 \) are mean and variance. The same head also predicts a stuff prototype \( \Omega_{st} = \{ (\mu_k, \sigma^2_k) \in \mathbb{R}^F \times \mathbb{R}^+ : k \in K_{st} \} \) for each stuff class \( k \in K_{st} \) independently.

Similarly to [65], 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 \( \hat{y}_{(i,j),\omega} \) for each pixel \( (i,j) \) and prototype as:

\[
\hat{y}_{(i,j),\omega} = \frac{-||\phi_{(i,j)} - \mu_{\omega}||^2}{2\sigma^2_{\omega}}
\]

Compared to [65], we relax the problem by not including the term \( -\frac{F}{2} \log \sigma^2_{\omega} \), 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 \( \sigma^2_{\omega} \) strictly positive 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, while the ID is computed from the highest score in Eq. 1. For stuff regions, we follow [65], determining the semantic classes by associating the pixel embeddings to the prototypes \( \Omega_{st} \) via the highest scoring class from Eq. 1. This decoupling allows semantic-awareness throughout the model.

4.2. Holistic: Dealing with Unseen Unknowns

We find unknown segments by relying on uncertainty estimates, which can help identifying the knowledge boundaries of a model [44, 57]. Specifically, instead of predicting the void class and searching in it for unknowns as in [36, 65], 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 [37, 57], 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 pos-
itive activation function, while providing good uncertainty estimates on OOD data, without training on such data [57].

Following [57], we consider the evidence $e_k = \alpha_k - 1$ as a measure of the amount of hints given by data for a pixel to be assigned to a class $k \in K$ known classes, with $\alpha_k$ being the parameters of the Dirichlet distribution $\text{Dir}(\alpha)$. We compute the uncertainty as $u = K / \sum_{k=1}^{K} \alpha_k$. Given that the class probabilities $\mathbf{p} = \{p_k : k = [1, \ldots, K]\}$ 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:

$$D(\mathbf{p}|\alpha) = B(\alpha)^{-1} \prod_{k=1}^{K} p_k^{\alpha_k-1}$$

with:

$$B(\alpha) = \prod_{k=1}^{K} \frac{\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$$

(2)

where $\Gamma$ is the gamma function and $B(\alpha)$ is the $K$-dimensional multinomial beta function [57].

We apply this on semantic segmentation by predicting a concentration parameter $\alpha_{Y(i,j)}$ 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 $e_{Y(i,j)}$ in the Dirichlet distribution. We learn this distribution with the semantic loss $L_s$ by minimizing the negative expected log likelihood of the correct class $Y(i,j)$, for the random variable $X_{Y(i,j)} \sim \text{Dir}(\alpha_{Y(i,j)})$:

$$L_s(i,j) = -E[\ln X_{Y(i,j)}^{Y(i,j)}]$$

$$= \psi \left( \sum_{k=1}^{K} \alpha_{Y(i,j),k} \right) - \psi (\alpha_{Y(i,j)})$$

(3)

where $\psi$ is the digamma function (i.e., $\Gamma$’s logarithmic derivative) and $\alpha_{Y(i,j),k}$ is the output of the semantic branch. Due to the difficulty of modelling the target distribution in our holistic task, we omit the KL term used in [57], 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 $\mu$ and $\sigma^2$ being mean and variance of the uncertainties of all training pixels, and $t$ being a hyperparameter.

**Separating unknowns** After finding unknown segments as described above, we cluster their corresponding instance-aware embeddings, trained only on known objects, into individual unknowns using DBSCAN [19]. In particular, we find the DBSCAN hyperparameters on the training closed-set data. As the algorithm provides outliers as well, we re-assign these to their originally predicted semantic class, thus ignoring the uncertainty estimate in these few cases. Additional details can be found in the appendix.

**4.3. Learning to Find Knowns and Unknowns**

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

$$L_s = \frac{1}{W \cdot H} \sum_{i,j} -E[\ln X_{Y(i,j)}^{Y(i,j)}]$$

$$= \frac{1}{W \cdot H} \sum_{i,j} \psi \left( \sum_{k=1}^{K} \alpha_{Y(i,j),k} \right) - \psi (\alpha_{Y(i,j)})$$

(4)

As in [13], the detection of $\hat{C}$ and ground truth $C$ center heatmaps:

$$L_o = \frac{1}{W \cdot H} \sum_{i,j} \left( \hat{C}(i,j) - C(i,j) \right)^2$$

(5)

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

$$L_p = \frac{1}{W \cdot H} \sum_{i,j} - \log(\tilde{z}(i,j),\omega(i,j))$$

(6)

Following [65], we learn significant embeddings $\phi(x,y)$ with a discriminative loss [15] $L_d$. Details can be found in the Supplementary Material. The training objective is:

$$L = \lambda_1 L_s + \lambda_2 L_o + \lambda_3 L_p + \lambda_4 L_d$$

(7)

**5. Experiments and Results**

**5.1. Experimental Setup**

**Datasets** We conducted our experiments on two public autonomous driving datasets, namely Cityscapes [14] and Lost&Found [52]. Cityscapes is a popular benchmark for outdoor segmentation tasks such as semantic and panoptic segmentation. 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.

**Evaluation metrics** We evaluated the holistic segmentation capability of a model, by means of the panoptic quality (PQ) metric [41] separately for known classes and unknown objects. Also the standard recognition (RQ) and segmentation (SQ) qualities were computed. We report PQ on 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 were set to detect both knowns and unknowns, while in closed settings, the same models predicted only known classes, which in practice meant ignoring the uncertainty estimates. Analyzing both cases allows to explore the trade-off between the ability of detecting unknowns (open) and the in-domain performance (closed). Furthermore, for uncertainty estimation, we evaluated the ability to identify unknowns reporting the AP on the unknown class [52], as well as the false positive rate at the recall 95 (FPR95). For semantic segmentation on Cityscapes, we computed the mIoU.

Network architecture All our models share the structure with Panoptic-DeepLab [13], using a ResNet50 [32] 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 allowing to quantify the uncertainty. The other branches follow Panoptic-DeepLab for the centers detection, and DeepLabV3+ for the embeddings, with two heads.

Implementation details Our models took in input RGB images sized 1024×512 (resized from 2048×1024) and were trained with batch size 16. Semantic models (Tables 3 and 4) used crops sized 512×256. We used the Adam optimizer until convergence, with an initial learning rate of 0.001, 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. The backbone was pretrained on ImageNet [17]. The losses were weighted \( \lambda_1 = \lambda_3 = \lambda_4 = 1 \) and \( \lambda_2 = 200 \) [13].

Prior works For a fair comparison, all methods were retrained with the setup described above. We compared our U3HS with open-set panoptic works: OSIS [65], which we adapted from LiDAR point clouds to images, and EOPSN [36]. Instead of training them directly on unknowns (as in [36]), we trained them on Cityscapes including the void class as fallback (as in [36, 65]), and applied them to Lost&Found. This is the same setup used for all other methods, except for learning void (ignored by our U3HS). We compared our uncertainty estimates with a variety of prior works, namely DPN [57], DUQ [60], SNGP [44], DML [5], SML [38], softmax and MC Dropout with 25 runs [22]. To do so, we repurposed and extended DPN, DUQ, and SNGP from image classification to semantic segmentation (appendix). Then, we extended them further to holistic segmentation, by incorporating them in our U3HS framework.

5.2. Quantitative Results

Table 1 compares our U3HS with prior approaches when segmenting instances of unseen unknowns. OSIS [65] was the first to address the more constrained open-set panoptic segmentation task, followed by EOPSN [36]. However, when applied to holistic segmentation, OSIS performance fell short on PQ, proving the severe limitations of relying on unknowns at training time. By learning void, OSIS achieved the highest SQ, which considers only matched segments. Instead, despite numerous attempts, EOPSN [36] did not work with the proposed holistic setup: it diverged as soon as the exemplars were mined. We attribute this to the inconsistent similarities within the void class of Cityscapes, compared to those across existing classes treated as void (e.g., car in their setup). This prevented EOPSN to form meaningful clusters from the proposal features during training [36]. Despite the similar setup to EOPSN [36], OSIS [65] could converge, since it does not rely on associating unknowns across images. In Table 1, we also compare various uncertainty estimations paired to our U3HS framework. While DUQ [60] and softmax underperformed compared to OSIS [65], DPN [57] and SNGP [44] achieved a higher PQ. Nevertheless, our improved DPN paired with our framework outperformed prior methods by a substantial margin, with a PQ 5.5 times higher than OSIS [65]. With respect to 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 completely new objects, such as those found in unconstrained settings (e.g., this transfer to Lost&Found).

| Method        | Framework | Uncertainty | Lost&Found | open Cityscapes | closed Cityscapes |
|---------------|-----------|-------------|------------|-----------------|-------------------|
|               |           |             | PQ | RQ | SQ    | PQ | RQ | SQ    | PQ | RQ | SQ    |
| OSIS [65]     | -         | -           | 1.45 | 2.23 | 65.11 | 39.42 | 50.20 | 78.53 | 39.42 | 50.20 | 78.53 |
| U3HS [ours]   | softmax   | -           | 0.10 | 0.20 | 51.45 | 39.70 | 50.77 | 78.20 | 45.12 | 56.83 | 79.40 |
| U3HS [ours]   | DUQ [60]  | -           | 0.56 | 0.89 | 62.56 | 41.68 | 53.14 | 78.42 | 45.90 | 58.17 | 78.90 |
| U3HS [ours]   | DPN [57]  | -           | 2.09 | 3.30 | 63.43 | 38.90 | 49.56 | 78.49 | 44.91 | 56.95 | 78.85 |
| U3HS [ours]   | SNGP [44] | -           | 4.65 | 7.57 | 61.49 | 41.02 | 51.98 | 78.91 | 46.23 | 58.56 | 78.95 |
| U3HS [ours]   | [ours]    |             | 7.94 | 12.37 | 64.24 | 41.21 | 51.67 | 79.77 | 46.53 | 58.99 | 78.87 |

Table 1. Holistic segmentation comparison of models trained on Cityscapes [14] and transferred to the Lost&Found [52] test set, without fine-tuning. All used the same backbone. The Uncertainty column reports different uncertainty estimators within our U3HS framework.
Table 2. Holistic and panoptic segmentation comparison of models trained on Cityscapes [14] and transferred to the test set of Lost&Found [52]. *: as in all other experiments, all methods were trained with the same constraints (e.g., ResNet50 [32], small batch and image sizes). Additionally, an ablation study (A1-A6) shows the impact of the main components of U3HS, with A3 being our full approach.

| Method* | Lost&Found | open Cityscapes | closed Cityscapes |
|---------|------------|-----------------|------------------|
|         | PQ RQ SQ   | PQ RQ SQ        | PQ RQ SQ         |
| -       |            |                 |                 |
| Panoptic-DeepLab [13] | 0.49 0.82 60.16 | 35.02 44.83 78.10 | 35.97 46.12 78.00 |
| OSIS [65] | 3.64 5.27 69.09 | 42.14 53.46 78.83 | 43.99 55.98 78.59 |
| A1 [ours] baseline: semantic uncertainty | 7.94 12.37 64.24 | 41.21 51.67 79.77 | 46.53 59.89 78.87 |
| A2 A1 + relaxed embedding association | 7.85 12.25 64.11 | 39.84 49.97 79.75 | 46.53 59.89 78.87 |
| A3 A2 + prototype head = U3HS | 7.85 12.25 64.11 | 39.84 49.97 79.75 | 46.53 59.89 78.87 |
| A4 A3 - reassigning outliers | 7.85 12.25 64.11 | 39.84 49.97 79.75 | 46.53 59.89 78.87 |
| A5 A4 - majority voting | 3.33 3.48 67.01 | 35.16 43.34 81.13 | 35.92 44.30 81.07 |
| A6 A4 - semantic embeddings | 3.33 3.48 67.01 | 35.16 43.34 81.13 | 35.92 44.30 81.07 |

Table 3. Comparison of open-set semantic segmentation on Lost&Found [52] test set of uncertainty estimators based on DeepLabV3+ [12] and trained only on Cityscapes (CS) [14].

| Uncertainty method | Lost&Found | open CS mIoU |
|--------------------|------------|--------------|
| softmax            | 16.72      | 22.88        |
| MC Dropout [22]    | 11.22      | 13.94        |
| DML [5]            | 3.14       | 83.04        |
| DUQ [60]           | 5.43       | 26.64        |
| DPN [57]           | 5.43       | 19.79        |
| SML [38]           | 16.91      | 51.67        |
| SNGP [44]          | 22.70      | 12.02        |
| improved DPN [ours]| 25.44      | 19.10        |

Knowns in panoptic segmentation Table 1 reports also the performances in-domain, under open and closed settings (Section 5.1). Ideally, an uncertainty estimator would suffer from no decrease in PQ between the two settings, meaning that its estimates are aligned with the distribution shift. OSIS [65] does not use uncertainty estimation, so it does not have these two operating modes. This results 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 [60] had the smallest gap, which could be attributed to its underestimated uncertainty, as supported by its low scores on Lost&Found (also in Table 3).

In-domain panoptic segmentation comparison In Table 2, we compare our U3HS with Panoptic-DeepLab [13]. For a fair comparison, both approaches, as well as all methods in this work, 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, under the same setting, our framework achieved a higher PQ on Cityscapes. 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. Experiments with improved training resources are out of the scope of this work.

Ablation on holistic and panoptic segmentation Table 2 reports an ablation of the main components of our U3HS, showing their benefits targeting holistic segmentation. Compared to the open-set panoptic OSIS [65], we reduced assumptions not learning the `void` class and mainly added a semantic branch with uncertainty for unknowns (A1), 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]. By transforming unknown predictions in standard in-domain outputs, this 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 had a major impact on 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 the model outputs, decreasing RQ and PQ.

Unknowns in semantic segmentation In Table 3, we compare the ability of a wide variety of uncertainty estimators to find unknowns in a semantic setting on Lost&Found [52], after training on Cityscapes [14]. This
meant retraining all methods under the same conditions, while also extending DPN [57], DUQ [60], and SNGP [44] to semantic segmentation. As seen in Table 1, DUQ [60] and DPN [57] performed worse than SNGP [44]. MC Dropout [22] underperformed softmax, probably due to the contrasting opinions from 25 forward passes. Our method was the best at finding unknowns (AP), with high quality uncertainty estimates (FPR$_{95}$). Table 3 reports also the mIoU on Cityscapes (CS), showing that all methods introduce a trade-off between OOD and in-domain outputs, as overestimating the uncertainty decreases the in-domain mIoU. Balancing these two complementary aspects is not trivial, with our approach and SNGP managing it best.

**Ablation on uncertainty estimation** Table 4 compares the DPN [57] we adapted from image classification to semantic segmentation with our extension. Our improvements were oriented to simplify the training process and help convergence. First we applied the softplus activation function to the last semantic layer, instead of exp as in DPN [57]. We chose softplus because it grows slower than exp and it is smooth, differentiable everywhere, and monotonic. This significantly improved the training stability at the cost of a reduced quality of the uncertainty estimates. Finally, due to the complexity of modeling the target distribution in our task, omitting the KL term used by DPN [57] further stabilized training and boosted the performance on all metrics.

### 5.3. Qualitative Results

Figure 4 shows example predictions of the proposed U3HS. The images illustrate the task difficulty, while providing examples of the OOD objects of Lost&Found. 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 properly 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 rather formed using the learned embeddings. As shown in Figure 4, the embeddings are closely coupled with the uncertainty estimates and the outputs.

**Considerations on Lost&Found and limitations**

Lost&Found [52] introduces a large domain shift from Cityscapes [14]. 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 known areas, due to the unseen configurations (e.g., fences). As we do not use any OOD data, nothing constrains high uncertainty to unknown segments alone, decreasing PQ. Nevertheless, results show that uncertainty is highly valuable in this holistic context, especially in safety-critical scenarios, allowing to leave the settings unconstrained. U3HS performance would mainly benefit from improvements in: uncertainty estimates, embeddings descriptiveness, and their clustering. So, learning-based clustering [25, 26] could be advantageous.

We refer to the **Supplementary Material** for more details on U3HS and the baselines, as well as additional results, including the trade-off between in-domain and OOD performances, the impact of embedding size, architecture

---

**Table 4. Ablation study on uncertainty estimates for open-set semantic segmentation. Models trained only on Cityscapes [14].**

| Configuration | Lost&Found | open CS |
|---------------|------------|---------|
|               | AP         | FPR$_{95}$ ↓ | mIoU    |
| [57] exp yes  | 5.43       | 19.89    | 66.99   |
| [ours] softplus yes | 3.43 | 25.97    | 64.36   |
| [ours] softplus no   | **25.44**  | **19.10**| **70.10**|

---
6. Conclusion

In this paper we introduced holistic segmentation: a new task addressing unseen unknown objects in unconstrained settings. Additionally, we presented U3HS: the first solution for the new task of holistic segmentation. Thanks to its uncertainty estimation and instance-aware learned embeddings, U3HS identifies and separates instances of completely unseen unknowns while segmenting known regions. Extensive experiments showed the effectiveness of U3HS for the challenging task of holistic segmentation.

A. Supplementary Material

In this appendix we include further details and results. Specifically, Sections A.1 and A.2 provide deeper insights on the method and the experimental setup respectively, while Sections A.3 and A.4 contain more results, both quantitative and qualitative.

A.1. Additional Details on the Method

Loss functions As described in Section 4.3, the proposed method is trained with a combination of losses: a semantic loss $L_s$, an object detection loss $L_o$, a prototype loss $L_p$, and a discriminative loss $L_d$. The discriminative loss is aimed at learning meaningful embeddings. It is composed of three different terms [15], namely variance $L_{va}$ to attract elements towards the mean, distance $L_{di}$ to push away different groups, and regularization $L_{re}$ to prevent the divergence of clusters from the origin:

$$L_d = \lambda_{41} L_{va} + \lambda_{42} L_{di} + \lambda_{43} L_{re}$$

$$L_{va} = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \sum_{a=1}^{N_\omega} [||\mu_{\omega} - \phi_a|| - \delta_a]_+^2$$

$$L_{di} = \frac{1}{|\Omega|(|\Omega|-1)} \sum_{\omega_A \in \Omega} \sum_{\omega_B \in \Omega} [2\delta_{\omega_A \omega_B} (\mu_{\omega_A} - \mu_{\omega_B})]^2$$

$$L_{re} = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} ||\mu_{\omega}||$$

where: $|\Omega|$ is the number of prototypes, $N_\omega$ is the number of embeddings associated to the prototype $\omega$, $\mu_{\omega}$ is the mean embedding of the cluster related to $\omega$, $||\cdot||$ is the L2 distance, $[x]_+ = \max(0, x)$ is the hinge (i.e., until which threshold the terms are active [15]), $\omega_A \neq \omega_B$, and we follow [15] for the hyperparameters, e.g., $\lambda_{41} = \lambda_{42} = 1$ and $\lambda_{43} = 0.001$.

Clustering unseen unknowns As described in Section 4.2, we use DBSCAN [19] to cluster the embeddings of unknown regions into individual unknown objects. Specifically, DBSCAN has multiple advantages: it does not need the number of clusters as input (which is unknown in our case), it is effective and very fast, has a low memory footprint, and distinguishes outliers (Table 2 shows the impact of this feature with A3-A4). Although other traditional clustering methods (e.g., Mean Shift, Affinity Propagation, Birch), are theoretically applicable in our setting, they come with drawbacks (e.g., have high memory requirements, do not output outliers, are significantly slower, or tend to deliver sub-optimal results). On the other hand, popular approaches that require the number of clusters as input cannot be applied in our settings (e.g., K-Means). Hence, DBSCAN was selected.

A.2. Additional Details on the Experimental Setup

Clustering parameters DBSCAN requires two parameters: $minPts$ (number of points in a neighborhood to count as core point) and $\epsilon$ (maximum neighborhood size). In order to find such parameters, we trained a model on Cityscapes [14], then selected ($minPts$, $\epsilon$) with a simple grid search maximizing PQ on a random subset of Cityscapes. Towards this end, we formed instances as follows: ignoring the detection output (i.e., using only the embeddings) and determining their class via majority voting from the semantic output. In particular, when finding ($minPts$, $\epsilon$), this means treating the embeddings of knowns as if they were unknowns (apart from their semantic class), assuming that the model treats them similarly. It is important to consider that the parameters were selected on the known objects of Cityscapes, despite DBSCAN was used only for separating unknowns in the Lost&Found [52] dataset. We did this to maintain the unknowns completely unseen (i.e., only as test set), as in real scenarios.

Comparisons with previous works As described in Section 5.1, we compared our uncertainty-based solution with prior works learning the void class and also tested a variety of uncertainty estimators within our proposed frame-
Table 5. Different embedding dimensions $F$ on closed-set panoptic segmentation on the validation set of Cityscapes [14]. The first column indicates the depth of the ResNet [32] backbone used (i.e., 18 for ResNet18).

| ResNet depth | $F$ | PQ   | RQ   | SQ   |
|--------------|-----|------|------|------|
| 18           | 2   | 33.0 | 42.3 | 77.9 |
| 18           | 4   | 38.9 | 49.8 | 78.0 |
| 18           | 8   | 41.3 | 52.7 | 78.3 |
| 18           | 16  | 42.3 | 53.8 | 78.7 |
| 18           | 32  | 42.1 | 53.6 | 78.5 |
| 50           | 8   | 47.7 | 60.4 | 79.0 |

Table 6. Transfer from Cityscapes [14] to Lost&Found-300 [52] test set (i.e., on the first 300 samples, see Section A.3). DBSCAN (our choice) is compared to Mean Shift to cluster the embeddings of unknown areas.

| Method         | Clustering | PQ   | RQ   | SQ   |
|----------------|------------|------|------|------|
| U3HS [ours]    | Mean Shift | 2.71 | 3.91 | **69.22** |
| U3HS [ours]    | DBSCAN     | **9.36** | **14.83** | 63.14 |

Ablation study for holistic segmentation With reference to Table 2, A1 is our baseline, which was built upon OSIS [65]. As OSIS, A1 included learned instance-aware embeddings, but unlike OSIS, it featured a semantic decoder delivering semantic segmentation and uncertainty estimates based on the semantic output via our improved DPN. Moreover, as for all our models, A1 did not learn the void class (unlike OSIS). A2 featured the relaxed score for the embedding association (described in Section 4.1), which lets the variance be indirectly controlled by the final task (i.e., the loss $L_p$, Section 4.3). Unlike A1 and A2, which had a shared head between embeddings and prototypes (i.e., as in OSIS), A3 introduced a dedicated prototype head. In practice, this meant having more layers fully dedicated for the embeddings and for the prototypes, separately, instead of sharing the computation until a later stage. Therefore, this allowed for more expressive and purposed features. A4 did not reassign to the known Cityscapes classes the outliers obtained from clustering unknowns via DBSCAN. Therefore, these pixels were kept as unknown and they shared the same instance ID. A5 did not perform majority voting (Section 4.1). This meant directly assigning to all known instance pixels the semantic classes predicted by the semantic branch, instead of enforcing coherence within an instance. This caused the instances to be fragmented according to how many semantic classes they contained, which in turn decreased RQ. Finally, A6 predicted the semantic classes for stuff areas directly from the semantic prediction branch, instead of matching the embeddings with stuff prototypes as in A1-A5 (Section 4.1).

A.3. Additional Quantitative Results

Trade-off between known and unknown Figure 5 shows the trade-off between the performance on known and unknown for our framework, both with SNGP [44] and our improved DPN, compared to that of OSIS [65]. The different data points were extracted by evaluating the outputs at different thresholds $t$, namely $[2, ..., 5]$, as well as ignoring the uncertainty estimates entirely (i.e., closed-set, reported where PQ Lost&Found is 0). The hyperparameter $t$ directly affects how high the uncertainty estimates need to be in order for their associated pixels to be considered unknown. This has an impact on the open-set performance on both Cityscapes [14] and Lost&Found [52], since changing in...
output what is considered unknown alters what is regarded as in-domain (i.e., known) as well. OSIS [65] does not have such hyperparameter as it considers unknown everything predicted as void. Overall, it can be seen that our proposed framework offers a better trade-off in both configurations (red and blue) than that of OSIS [65] (yellow). Furthermore, using our full approach (i.e., our framework with our improved DPN) typically gave the best trade-off between known and unknown without compromising the metrics too much (blue).

**Impact of embedding size and architecture** Table 5 shows the effect of different embeddings dimensions $F$ on a smaller ResNet18 [32]. In the rest of this work, all experiments used $F = 8$ and ResNet50 as in the last line of
Figure 7. Additional example predictions of the proposed U3HS on unknown categories from the test set of Lost&Found [52]. The model was trained on Cityscapes [14] and transferred to Lost&Found without any fine-tuning. White arrows mark labeled OOD objects.

the table, due to constrained training resources. The embedding dimension directly affects the learning capability of the model. Since the instance-aware embeddings are a critical part of the output, a smaller \( F \) is linked to inexpensive embeddings that cannot be as discriminative as those from a larger \( F \). Therefore, increasing \( F \) improved all metrics, except for the larger \( F = 32 \). This can be attributed to the small ResNet18 backbone being already saturated at \( F = 16 \), unable to extract rich and detailed features for the larger embeddings to exploit. With a larger model (e.g., ResNet101), even higher embedding dimensions \( F \) might be beneficial. Overall, Table 5 shows that our proposed approach, given less constrained resources, could deliver better results when using an embedding dimension higher than the \( F = 8 \) employed across this work. The table also shows the comparison between ResNet18 and ResNet50, with the latter delivering over 15% higher PQ at the same \( F = 8 \). This gives an idea of how our proposed approach would perform with a larger backbone.

**Impact of clustering method** In Table 6 we compare two popular clustering methods within our U3HS framework, namely DBSCAN and Mean Shift [21]. Due to the very high computation effort and memory required by Mean Shift, for this experiment, we opted for the following setup. First, instead of a standard CPU implementation, we used a parallelized CUDA version of the algorithm [73]. Then, due to the still very high memory requirements, certain samples of Lost&Found caused memory issues. Therefore, we reduced the size of the test set of Lost&Found [52] to its first 300 samples (Lost&Found-300), which were not problematic. These 300 samples are sufficient to indicate the effect of using Mean Shift instead of DBSCAN. Table 6 shows the superiority of DBSCAN for this task, with a 3.5x higher PQ and 3.8x better RQ. In particular, RQ should be the focus as we compare instance segmentation of unknowns.

**A.4. Additional Qualitative Results**

**Qualitative comparison** Figure 6 shows a comparison of the open-set panoptic segmentation predictions of the proposed U3HS with the prior work OSIS [65], as well as the regions each predicted as unknown. In particular, ours found unknowns as segments estimated as highly uncertain and OSIS found them as the pixels predicted to be part of the learned void class.

From the images, it can be seen how for the most part OSIS managed to learn a relatively good class boundary around the void class, as it was typically able to predict the OOD objects as unknown via void. This is interesting as it shows how OSIS can potentially work with challenging unseen unknowns. However, the same figure shows also the strong limitations of learning and predicting void, due to the assumptions about the data distributions that this entails. In the first image, OSIS completely ignored the unknown object, assigning it to the road class, while in the fifth image, it detected the toy as car. In contrast, in the last image, OSIS predicted almost everything as unknown. This proves how the binary aspect introduced by predicting the void class (a pixel is either unknown, by being void, or known, if another class) does not cope well with the diversity and unpredictability of the scenes in unconstrained real-world settings. Specifically, predicting the void class severely relies on the closed-set training data, as the suc-
cess of such method is directly related to the diversity of the \textit{void} class seen during training, which is limited as it cannot properly sample the long tail of the data distribution [43].

Nevertheless, as shown already in Section 5, estimating the uncertainty allows to properly cope with unknown objects by adding an extra layer of prediction. Contrary to the idea of prior works (Section 2) of predicting unknowns via the \textit{void} class, which directly competes with the other semantic classes for being part of the output, uncertainty estimates go on top of the standard semantic predictions. Although this complicates dealing with multiple network outputs, it offers a wider spectrum and deeper insights, since the uncertainty could be ignored or considered with various thresholds depending on the situation (Figure 5), for the same trained model and output. Since estimating the uncertainty aims at smoothly quantifying the domain gap from the training data, we believe it is better suited to highly unpredictable unseen real-world scenarios as in holistic segmentation settings.

Furthermore, Figure 6 shows the capability of each method to identify instances of unknown OOD objects. For both approaches, this is related to the clustering of embeddings corresponding to those pixels predicted as unknown, via \textit{void} (OSIS) or as highly uncertain (ours). In particular, OSIS had a tendency to over-fragment unknown objects into several different small instances, as it can be seen in the fourth, sixth, eighth and last images. This proves again the effectiveness of our modifications when dealing with the embeddings, as described in Section 4 and evaluated in Table 2. Additionally, in the sixth image, OSIS could not distinguish the two neighboring OOD objects. Moreover, OSIS often improperly assigned large regions to the same unknown instance. Similarly to ours, OSIS considers every unknown segment as part of an instance. By learning and predicting the \textit{void} class, during training OSIS learned to precisely segment the bonnet of the ego car (labeled as \textit{void} in Cityscapes [14]). However, at test time on Lost&Found it could not tell the ego vehicle bonnet apart from a wide variety of pixels. This was the case for the unknown object in the seventh image, which was entirely assigned to the same instance as the bonnet, or also many other segments around knowns and unknowns (colored in black). In fact, the ego vehicle bonnet unknown instance (black) often surrounded other predicted unknown instances (e.g., in the second, fourth, sixth, eighth, and last images).

A benefit of estimating the uncertainty is the ability to account for a wide array of unusual regions. This is valuable for downstream tasks, e.g., trajectory prediction and path planning. Specifically, uncertainty estimates by the proposed U3HS were high on the stroller in Figure 1, as well as in Figure 4 on the walking assistance device on the left of the upper image and the cart pushed by the man waving on the right of the bottom image in Figure 7, none of which were labeled as unknown in the dataset [52], as they were not part of the objects manually placed by the authors. In Figure 6 this repeated from a different perspective on the stroller in the background of the second image, as well as the unusual van with the open doors in the fourth image, and the duffel bag in the sixth. By learning and predicting \textit{void}, OSIS ignored these unusual regions as it lacks the flexibility and granularity that our U3HS offers by estimating the uncertainty.

Additional results on unknowns Figure 7 shows additional qualitative outputs, similarly to those of Figure 4. Once again, it can be seen how challenging the proposed holistic segmentation task is. As in the predictions of Figure 4, the model can distinguish most of the unknown objects. It can be seen how specific areas of the images trigger higher uncertainty estimates. This is the case of the fences in the second and third images of Figure 7, as well as unknown objects not part of the OOD labels of Lost&Found, such as the cart on the right of the bottom image as previously mentioned. As previously seen, \textit{stuff} structures (e.g., fences) are assigned to a single coherent instance ID throughout the whole image, while unusual objects (e.g., cart in the last image) have their own dedicated ID. Figure 7 also provides some examples of unusual scenes present in the Lost&Found [52] dataset, posing major challenges compared to Cityscapes [14].

Considerations on estimating the uncertainty Figure 8
shows failure cases caused by the necessary filtering of the uncertainty estimates. While the uncertainty was triggered by a variety of unusual areas including the vast majority of unknown objects, its a priori filtering (based on closed-set training data, Section 4.2) sometimes caused the unknown object to result completely undetected. Although this filtering is aimed at removing low uncertainty areas which are probably in-domain (e.g., the fence in the upper image), it could inadvertently remove also proper OOD objects (e.g., those marked by the white arrows). This is related to the trade-off shown in Figure 5, so keeping more unknowns (i.e., lower threshold \( t \)) reduces the in-domain performance. Nevertheless, in the embeddings visualizations it can be seen that the model properly isolated the entire marked box in the lower image and precisely segmented the cardboard box in the upper one. However, the two unknown objects were not detected, due to the difficulty of merging multiple outputs and of interpreting uncertainty estimates without access to OOD data. It should be considered that the proposed U3HS does not distinguish between the uncertainty for unknown objects and that of unusual known classes. The difference might lie in the amount of uncertainty corresponding to these regions, hence the filtering via the threshold \( t \) to attempt telling apart completely unknown from unusual, which remains highly challenging without using any information about unknowns at training time, as in our setup.

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.

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

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

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

[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.

[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.

[7] Jun Cen, Peng Yun, Shiwei Zhang, Junhao Cai, Di Luan, Michael Yu Wang, Ming Liu, and Mingqian Tang. Open-world semantic segmentation for LIDAR point clouds. arXiv preprint arXiv:2207.01452, 2022.

[8] Robin Chan, Krzysztof Lis, Svenja Uhlmeyer, Hermann Blum, Sina Honari, Roland Siegwart, Mathieu Salzmann, Pascal Fua, and Matthias Rottmann. SegmentMeIfYouCan: A benchmark for anomaly segmentation. arXiv preprint arXiv:2104.14812, 2021.

[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 European conference on computer vision, pages 52–68. Springer, 2016.

[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 Thirtieth International Joint Conference on Artificial Intelligence, pages 4366–4373, 2021.

[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 European Conference on Computer Vision, pages 695–714. Springer, 2020.

[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 (ECCV), pages 801–818, 2018.

[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.

[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
[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, 9

[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 *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 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, 9

[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] Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory, 21*(1):32–40, 1975. 12

[22] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning*, pages 1050–1059. PMLR, 2016. 2, 6, 7, 8

[23] 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

[24] Stefano Gasperini, Patrick Koch, Vinzent 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

[25] 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

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

[27] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Hutnt, 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

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

[29] 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

[30] Matej Grčić, Petra Bevandić, and Siniša Šegvić. DenseHybrid: Hybrid anomaly detection for dense open-set recognition. *arXiv preprint arXiv:2207.02606*, 2022. 3

[31] 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

[32] 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, 10, 11

[33] 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

[34] Jie Hong, Weihao Li, Junlin Han, Jiyang Zheng, Pengfei Fang, Mehrtaish Harandi, and Lars Petersson. GOSS: Towards generalized open-set semantic segmentation. *arXiv preprint arXiv:2203.12116*, 2022. 3

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

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

[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.

[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.

[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.

[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.

[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.

[42] Abhijit Kundu, Kyle Genova, Xiaoxi 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.

[43] 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.

[44] 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.

[45] Adrian Lopez-Rodriguez, Benjamin Busam, and Krystian Mikolajczyk. Project to adapt: Domain adaptation for depth completion from noisy and sparse sensor data. In Proceedings of the Asian Conference on Computer Vision, 2020.

[46] Fabian Manhardt, Diego Martin Arroyo, Christian Rupprecht, Benjamin Busam, Tolga Birdal, Nassir Navab, and Federico Tombari. Explaining the ambiguity of object detection and 6D pose from visual data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6841–6850, 2019.

[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.

[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.

[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.

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

[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, 2018.

[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 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 1099–1106, 2016.

[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.
[54] Janis Postels, Hermann 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. 2, 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] 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, 8

[58] Kshitij Sirohi, Sajad Marvi, Daniel Büscher, and Wolfram Burgard. Uncertainty-aware panoptic segmentation. *arXiv preprint arXiv:2206.14554*, 2022. 3

[59] 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, 8, 10

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

[61] 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

[62] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. *arXiv preprint arXiv:2207.11247*, 2022. 2

[63] 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

[64] Yuxuan Luo, Xinlei Zhou, 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

[65] Kelvin Wong, Shenlong Wang, Mengye Ren, Ming Liang, and Raquel Urtasun. Identifying unknown instances for autonomous driving. In *Conference on Robot Learning (CoRL)*, pages 384–393. PMLR, 2020. 2, 3, 4, 5, 6, 7, 9, 10, 11, 12

[66] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the
[75] 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.

[76] 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.

[77] 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.

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