

# OpenGAN: Open-Set Recognition via Open Data Generation

Shu Kong\*, Deva Ramanan\*,<sup>†</sup>  
 \*Carnegie Mellon University    <sup>†</sup>Argo AI  
 {shuk, deva}@andrew.cmu.edu  
 [Github Repository]

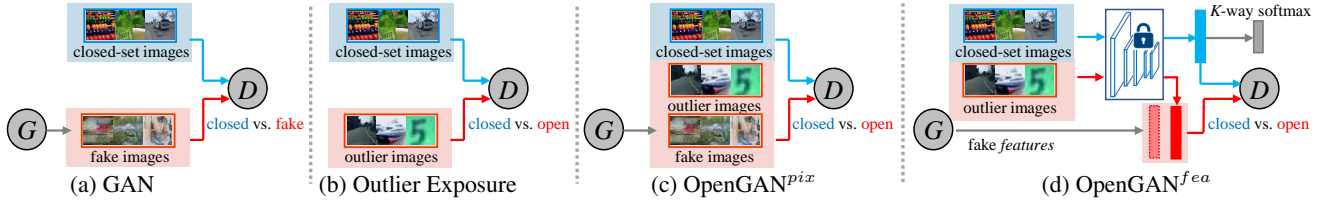


Figure 1: We explore open-set recognition, which requires the ability to discriminate open-set test examples outside  $K$  classes of interest. (a) Past work has suggested that GAN discriminators can serve as open-set likelihood functions, but this does not work well due to instable training of GANs [56, 53, 48, 66, 37]. (b) Outlier Exposure [31] exploits some outlier data to learn a binary discriminator  $D$  for open-set discrimination. Because outliers observed during training will not exhaustively span the open-world, the discriminator  $D$  tends to generalize poorly to diverse open data [57]. (c) We introduce OpenGAN, which augments training outliers with *fake* open data synthesized by a generator  $G$  trained to fool the discriminator  $D$ . Importantly, we find that a small number of outliers stabilizes training by enabling effective model selection of the discriminator  $D$ . (d) Because we are concerned with accurate discrimination rather than realistic pixel generation, we find it more efficient to generate (and discriminate) *features* from the off-the-shelf  $K$ -way classification network. This allows OpenGAN to be implemented via a lightweight discriminator head built on top of an existing  $K$ -way network, enabling closed-world systems to be readily modified for open-set recognition.

## Abstract

Real-world machine learning systems need to analyze test data that may differ from training data. In  $K$ -way classification, this is crisply formulated as open-set recognition, core to which is the ability to discriminate open-set data outside the  $K$  closed-set classes. Two conceptually elegant ideas for open-set discrimination are: 1) discriminatively learning an open-vs-closed binary discriminator by exploiting some outlier data as the open-set, and 2) unsupervised learning the closed-set data distribution with a GAN, using its discriminator as the open-set likelihood function. However, the former generalizes poorly to diverse open test data due to overfitting to the training outliers, which are unlikely to exhaustively span the open-world. The latter does not work well, presumably due to the instable training of GANs. Motivated by the above, we propose OpenGAN, which addresses the limitation of each approach by combining them with several technical insights. First, we show that a carefully selected GAN-discriminator on some real outlier data already achieves the state-of-the-art. Second, we augment the available set of real open training examples with adversarially synthesized “fake” data. Third and most impor-

tantly, we build the discriminator over the features computed by the closed-world  $K$ -way networks. This allows OpenGAN to be implemented via a lightweight discriminator head built on top of an existing  $K$ -way network. Extensive experiments show that OpenGAN significantly outperforms prior open-set methods.

## 1. Introduction

Machine learning systems that operate in the real open-world invariably encounter test-time data that is unlike training examples, such as anomalies or rare objects that were insufficiently (or never) observed during training. Fig. 2 illustrates two cases in which a state-of-the-art semantic segmentation network misclassifies a “stroller”/“street-market” — a rare occurrence in either training or testing — as a “motorcycle”/“building”. This failure could be catastrophic for an autonomous vehicle.

Addressing the open-world has been explored through anomaly detection [69, 31] and out-of-distribution detection [36]. In  $K$ -way classification, this task can be crisply formulated as open-set recognition, which requires discriminating open-set data that belongs to a  $(K+1)^{th}$  “other”

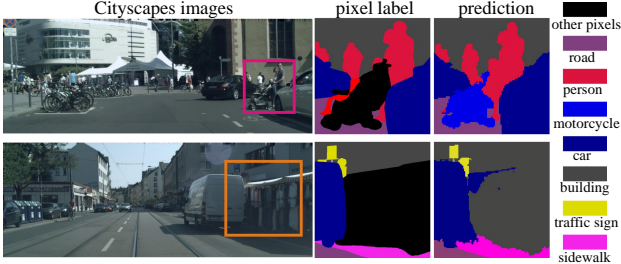


Figure 2: We motivate open-set recognition with safety concerns in autonomous vehicles (AVs). Contemporary benchmarks such as Cityscapes [14] focus on  $K$  classes of interest for evaluation, ignoring a sizeable set of “other” pixels that include vulnerable objects like wheelchairs and strollers (upper row). As a result, most state-of-the-art segmentators [61] also ignore these pixels during training, resulting in a **stroller** mislabeled as a “motorcycle” (top) and a **street-market** mislabeled as a “building”. Such misclassifications may be critical for AVs because these objects may require different plans for obstacle avoidance (e.g., “yield” or “slow-down”). Fig. 4 shows our approach, which explicitly augments state-of-the-art segmentors with open-set reasoning.

class, outside the  $K$  closed-set classes [54]. Typically, open-set discrimination assumes no examples from the “other” class available during training [6, 64, 44]. In this setup, one elegant approach is to learn the closed-set data distribution with a GAN, making use of the GAN-discriminator as the open-set likelihood function (Fig. 1a) [56, 53, 48, 66, 37]. However, this does not work well due to instable training of GANs. Recent work has shown that outlier exposure (Fig. 1b), or the ability to train on *some* outlier data as open-training examples, can work surprisingly well via the training of a simple open-vs-closed binary discriminator [18, 31]. However, such discriminators fail to generalize to diverse open-set data [57] because they overfit to the available set of training outliers, which are often biased and fail to exhaustively span the open-world. Motivated by above, we introduce **OpenGAN**, a simple approach that dramatically improves open-set classification accuracy by incorporating several key insights. First, we show that using outlier data as a valset to select the “right” GAN-discriminator *does* achieve the state-of-the-art on open-set discrimination. Second, with outlier exposure, we augment the available set of open-training data by adversarially generating *fake* open examples that fool the binary discriminator (Fig. 1c). Third and most importantly, rather than defining discriminators on pixels, we define them on off-the-shelf (OTS) features computed by the closed-world  $K$ -way classification network (Fig. 1d). We find such discriminators generalize much better.

Our formulation differs in three ways from other open-set approaches that employ GANs. (1) Our goal is *not* to generate realistic pixel images, but rather to learn a robust open-vs-closed discriminator that naturally serves as an open-set likelihood function. Because of this, our ap-

proach might be better characterized as a discriminative adversarial network! (2) We train the discriminator with both *fake* data (synthesized from the generator) and *real* open-training examples (cf. outlier exposure [31]). (3) We train GANs on OTS features rather than RGB pixels. We show that OpenGAN significantly outperforms prior work for open-set recognition across a variety of tasks including image classification and pixel segmentation. Moreover, we demonstrate that our technical insights improve the accuracy of other GAN-based open-set methods: training them on OTS features and selecting their discriminators via validation as the open-set likelihood function.

## 2. Related Work

**Open-Set Recognition.** There are multiple lines of work addressing open-set discrimination, such as anomaly detection [12, 36, 69], outlier detection [53, 48], and open-set recognition [54, 22]. The typical setup for these problems assumes that one does not have access to training examples of open-set data. As a result, many approaches first train a closed-world  $K$ -way classification network on the closed-set [30, 6] and then exploit the trained network for open-set discrimination [54, 35, 44]. Some others train “ground-up” models for both closed-world  $K$ -way classification and open-set discrimination by synthesizing fake open data during training, oftentimes sacrificing the classification accuracy on the closed-set [21, 42, 64, 59]. To recognize open-set examples, they resort to post-hoc functions like density estimation [69, 67], uncertainty modeling [20, 36], and input image reconstruction [48, 23, 17, 59]. We also explore open-set recognition through  $K$ -way classification networks, but we show OpenGAN, a simple and direct method of training an open-vs-closed classifier on adversarial data, performs significantly better than prior work.

**Open-Set Recognition with GANs.** As GANs can learn data distributions [25], conceptually, a GAN-discriminator trained on the closed-set naturally serves as an open-set likelihood function. However, this does not work well [56, 53, 48, 66, 37], presumably due to instable training of GANs. As a result, previous GAN-based methods focus on 1) generating fake open-set data to augment the training set, and 2) relying on the reconstruction error for open-set recognition [62, 56, 53, 1, 16]. With OpenGAN, we show that GAN-discriminator can achieve the state-of-the-art for open-set discrimination *once* we perform model selection on a valset of outlier examples. Therefore, unlike prior approaches, OpenGAN directly uses the discriminator as the open-set likelihood function. Moreover, our final version of OpenGAN generates features rather than pixel images.

**Open-Set Recognition with Outlier Exposure.** [18, 31, 52] reformulate the problem with the concept of “outlier exposure” which allows methods to access *some* outlier data as open-training examples. In this setting, simply train-

ing a binary open-vs-closed classifier works surprisingly well. However, such classifiers easily overfit to the available set of open-training data and generalize poorly, e.g., in a “cross-dataset” setting where open-set testing data differs from open-training data [57]. It appears fundamentally challenging to collect outlier data to curate an exhaustive training set of open-set examples. Our approach, OpenGAN, attempts to address this issue by augmenting the training set with adversarial *fake* open-training examples.

### 3. OpenGAN for Open-Set Recognition

Generally, solutions to open-set recognition contain two steps: (1) open-set discrimination that classifies testing examples into closed and open sets based on the open-set likelihoods, and (2)  $K$ -way classification on closed-set from step (1) [54, 5, 44]. The core problem to open-set recognition is the first step, i.e., open-set discrimination. Typically, open-set discrimination assumes that open-set examples are not available during training [42, 44]. However, [18, 31] demonstrate that outlier exposure, or the ability to train on *some* outlier examples, can greatly improve open-set discrimination via the training of a simple open-vs-closed binary classifier (Fig. 1b). Because it is challenging to construct a training set that exhaustively spans the open-world, such a classifier may overfit to the outlier data and not sufficiently generalize [57]. We demonstrate that OpenGAN alleviates this challenge by generating fake open-set training examples using a generator that is adversarially trained to fool the classifier. Importantly, with model selection on a valset, OpenGAN is also effective under the classic setup which assumes no availability of open-training data.

#### 3.1. Methodology

Let  $x$  be a data example, which can be an RGB image or its feature representation. We will show that using the latter performs better. Let  $\mathcal{D}_{closed}(x)$  be the closed-world distribution over  $x$  — that is, closed-set data from the  $K$  closed-set classes. Let  $\mathcal{D}_{open}(x)$  be the open-set data distribution of examples which do not belong to the closed-set.

**Binary Classifier.** We train a binary classifier  $D$  from both closed- and open-set data:

$$\max_D \mathbb{E}_{x \sim \mathcal{D}_{closed}} [\log D(x)] + \lambda_o \cdot \mathbb{E}_{x \sim \mathcal{D}_{open}} [\log (1 - D(x))]$$

where  $D(x) = p(y=\text{“closed-set”}|x)$ . Intuitively, we tune  $\lambda_o$  to balance the closed- and open-set training examples. This simple method is effective when the open-training examples are sufficiently representative of testing-time open-set data [31], but underperforms when they fail to span the open-world [57].

**Synthetic Open Data.** One solution to the above is to exploit synthetic training data, which might improve the generalizability of classifier  $D$ . Assume we have a generator network  $G(z)$  that produces synthetic images given

(Gaussian normal) random noise inputs  $z \sim \mathcal{N}$ . We can naively add them to the pool of negative or open-set examples that  $D$  should not fire on. But these synthetic images might be too easy for  $D$  to categorize as open-set data [41, 13]. A natural solution is to adversarially train the generator  $G$  to produce difficult examples that fool the classifier  $D$  using a GAN loss:

$$\min_G \mathbb{E}_{z \sim \mathcal{N}} [\log (1 - D(G(z)))] \quad (1)$$

Because a *perfectly* trained generator  $G$  would generate realistic closed-set images, eventually making the discriminator  $D$  inapplicable for open-set discrimination. We find that the following two techniques easily resolve this issue.

**OpenGAN** trains with both the *real* open&closed-set data and the *fake* open-data into a single (GAN-like) minimax optimization over  $D$  and  $G$ :

$$\begin{aligned} \max_D \min_G \mathbb{E}_{x \sim \mathcal{D}_{closed}} [\log D(x)] \\ + \lambda_o \cdot \mathbb{E}_{\bar{x} \sim \mathcal{D}_{open}} [\log (1 - D(\bar{x}))] \\ + \lambda_G \cdot \mathbb{E}_{z \sim \mathcal{N}} [\log (1 - D(G(z)))] \end{aligned} \quad (2)$$

where  $\lambda_G$  controls the contribution of generated fake open-data by  $G$ . When there are no open training examples (i.e.,  $\lambda_o=0$ ), the above minimax optimization can still train a discriminator  $D$  for open-set classification. In this case, training an OpenGAN is equivalent to training a normal GAN and using its discriminator as the open-set likelihood function. While past work suggests that GAN-discriminators do not work well as open-set likelihood functions, we show they *do* achieve the state-of-the-art *once* selected with a valset (detailed below). To distinguish our contribution on the crucial step of model selection via validation, we call this method OpenGAN-0.

**Open Validation.** Model selection is challenging for GANs. Typically, one resorts to visual inspection of generated images from different model checkpoints to select the generator  $G$  [25]. In our case, we must carefully select the discriminator  $D$ . We experimented with many approaches such as using the last model checkpoint or selecting the one with minimum training error, but neither works, because adversarial training will eventually lead to a discriminator  $D$  that is incapable of discriminating closed-set data and fake open-set data generated by  $G$  (details in the supplemental). We find it crucial to use a validation set of real outlier data to select  $D$ , when  $D$  achieves the best open-vs-closed classification accuracy on the valset. We find the performance to be quite robust to the val-set of outlier examples, even when they are drawn from a different distribution from those encountered at test-time (Table 3 and 4).

#### 3.2. Comparison to Prior GAN Methods

We compare OpenGAN to numerous prior art that use GANs for open-set discrimination.

**Discriminator vs. Generator.** GANs mostly aim at generating realistic images [2, 9]. As a result, prior work in open-set recognition has focused on using GANs to generate realistic open-training images [21, 34, 42]. These additional images are used to augment the training set for learning an open-set model, which oftentimes is designed for both the closed-world  $K$ -way classification and open-set discrimination [21, 34, 42]. In our case, we do not learn a separate open-set model but directly use the already-trained discriminator as the open-set likelihood function.

**Features vs. Pixels.** GANs are typically used to generate realistic pixel images. As a result, many GAN-based open-set methods focus on generating realistic images to augment the closed-set training data [48, 66, 31]. However, generating high-dimensional realistic images is challenging per se [2, 9] and may not be necessary to open-set recognition [48]. As such, we build GANs over OTS feature embeddings learned by the closed-world  $K$ -way classification networks, e.g., over pre-logit features from the penultimate layer. This allows for exploiting an enormous amount of engineering effort “for free” (e.g., network design).

**Classification vs. Reconstruction.** We note that most, if not all, GAN-based methods largely rely on the reconstruction error for open-set discrimination [56, 53, 48, 66, 44]. The underlying assumption is that closed-set data produces lower reconstruction error than the open-set. While this seems reasonable, it is challenging to reconstruct complex, high-resolution images [2, 9], like the Cityscapes images shown in Fig. 2. On the contrary, we directly use the discriminator as the open-set likelihood function. Crucially, such a baseline has been shown to perform poorly in large body of prior work [56, 53, 48, 66, 37]. To the best of our knowledge, ours is the first result to demonstrate the strong performance of GAN-discriminators, thanks in large part to model selection via open validation (Section 3.1).

## 4. Experiment

We conduct extensive experiments to validate OpenGAN under various setups, and justify the advantage of exploiting OTS features and using the GAN-discriminator as the open-set likelihood function. We first briefly introduce three experimental setups below (details in later sections).

- *Setup-I* open-set discrimination splits a *single dataset* into open and closed sets w.r.t class labels, e.g., define MNIST digits 0-5 as the closed-set for training, and digits 6-9 as the open-set in testing. Although small-scale, this is a common experimental protocol for open-set discrimination that classifies open-vs-closed test examples [42, 44, 47, 67].
- *Setup-II* open-set recognition requires both  $K$ -way classification on the closed-set and open-set discrimination. We follow a “less biased” protocol [57] that constructs the open train&test-sets with *cross-dataset* images [60].

- *Setup-III* examines the open-set discrimination at pixel level in semantic segmentation, which evaluates pixel-level open-vs-closed classification accuracy [7, 29].

**Implementation.** We describe how to train the closed-world  $K$ -way classification networks which compute OTS features used for training OpenGAN<sup>fea</sup> (Fig. 1d) and other methods (e.g., OpenMax [6] and C2AE [44]). For training  $K$ -way networks under *Setup-I* and *II*, we train a ResNet18 model [28] exclusively on the closed-train-set (with  $K$ -way cross-entropy loss). Under *Setup-III*, we use HRNet [61] as an OTS network, which is a top ranked model for semantic segmentation on Cityscapes [14]. We choose the penultimate/pre-logit layer of each  $K$ -way network to extract OTS features. Other layers also apply but we do not explore them in this work. Over the features, we train OpenGAN<sup>fea</sup> discriminator (2MB), as well as the generator (2MB), with a multi-layer perceptron architecture. For comparison, we also train a ground-up OpenGAN<sup>pix</sup> over pixels with a CNN architecture ( $\sim 14$ MB) [68]. We train our OpenGAN models using GAN techniques [49]. Compared to the segmentation network HRNet (250MB), OpenGAN<sup>fea</sup> is quite lightweight that induces minimal compute overhead. We conduct experiments with PyTorch [45] on a single Titan X GPU. Code is available at <https://github.com/aimerykong/OpenGAN>

**Evaluation Metric.** To evaluate open-set discrimination that measures the open-vs-closed binary classification performance, we follow the literature [35, 44] and use the area under ROC curve (AUROC) [15]. AUROC is a calibration-free and threshold-less metric, simplifying comparisons between methods and reliable in large open-closed imbalance situation. For open-set recognition that measures  $(K+1)$ -way classification accuracy ( $K$  closed-set classes plus the  $(K+1)^{th}$  open-set class), we report the macro average F1-score over all the  $(K+1)$  classes on the valsets [54, 6].

### 4.1. Compared Methods

We compare the following representative baselines and state-of-the-art methods for open-set recognition.

**Baselines.** First, we explore classic generative models learned on the closed-train-set, including Nearest Neighbors (NNs) [50] and Gaussian Mixture Models (GMMs) which were found to perform quite well over L2-normalized OTS features [33]. We refer the reader to the supplemental for details of GMMs as they are strong yet underexplored baseline in the literature. Both models can be used for open-set discrimination by thresholding NN distances or likelihoods. We further examine the idea of outlier exposure [31] that learns an open-vs-closed binary classifier (CLS). Lastly, following classic work in semantic segmentation [19], we evaluate a  $(K+1)$ -way classifier trained with outlier exposure, in which we use the softmax score corresponding to the  $(K+1)^{th}$  “other” class as the open-set like-

Table 1: **Open-set discrimination (Setup-I)** measured by area under ROC curve (AUROC) $\uparrow$ . We report methods marked by \* with their best reported numbers in the compared papers. Recall that OpenGAN-0 does not train on outlier data (i.e.,  $\lambda_0=0$  in Eq. 2) and only selects discriminator checkpoints on the validation set. OpenGAN-0<sup>fea</sup> clearly performs the best. Defined on the off-the-shelf (OTS) features of closed-world  $K$ -way networks, NN<sup>fea</sup> and OpenGAN-0<sup>fea</sup> work much better than their pixel version (NN<sup>pix</sup> and OpenGAN-0<sup>pix</sup>).

Dataset	MSP	MSP <sub>c</sub>	MCdrop	GDM	OpenMax	GOpenMax	OSRCI	C2AE	CROSR	RPL	Hybrid	GDFR	NN <sup>pix</sup>	NN <sup>fea</sup>	OpenGAN	OpenGAN
	[30]	[36]	[20]	[35]	[6]	[21]*	[42]*	[44]*	[64]*	[13]*	[67]*	[46]*	[50]	[50]	-0 <sup>pix</sup>	-0 <sup>fea</sup>
MNIST	.977	.985	.984	.989	.981	.984	.988	.989	.991	.996	.995	—	.931	.981	.987	<b>.999</b>
SVHN	.886	.891	.884	.866	.894	.896	.910	.922	.899	.968	.947	.935	.534	.888	.881	<b>.988</b>
CIFAR	.757	.808	.732	.752	.811	.675	.699	.895	.883	.901	.950	.807	.544	.801	.971	<b>.973</b>
TinyImgNet	.577	.713	.675	.712	.576	.580	.586	.748	.589	.809	.793	.608	.528	.692	.795	<b>.907</b>

likelihood.

**Likelihoods.** Many methods compute open-set likelihood on OTS features, including Max Softmax Probability (MSP) [30] and Entropy [58] (derived from softmax probabilities), and calibrated MSP (MSP<sub>c</sub>) [36]. OpenMax [6] fits logits to Weibull distributions [55] that recalibrate softmax outputs for open-set recognition. C2AE [44] learns an additional  $K$ -way classifier over the OTS features using reconstruction errors, which are then used as the open-set likelihoods. GDM [35] learns a Gaussian Discriminant Model on OTS features and computes open-set likelihood based on Mahalanobis distance.

**Bayesian Networks.** Bayesian neural networks estimate uncertainties via Monte Carlo estimates (MCdrop) [20, 38]. The estimated uncertainties are used as open-set likelihoods. We implement MCdrop via 500 samples.

**GANs.** GOpenMax [21] and OSRCI [42] train GANs to generate fake images to augment closed-set data for open-set recognition. Other types of GANs can also be used for open-set recognition, such as BiGANs [66], on which we show our technical insights (e.g., training on OTS features and directly using the discriminator) also apply (Table 2).

When possible, we train the methods using their open-source code. We implement NN, CLS and OpenGAN on both RGB images (marked with <sup>pix</sup>) and OTS features (marked with <sup>fea</sup>) for comparison. For fair comparison, we tune all the models for all methods on the same val-sets.

## 4.2. Setup-I: Open-Set Discrimination

**Datasets.** MNIST/CIFAR/SVHN/TinyImageNet are widely used in the open-set literature, and we follow the literature to experiment with these datasets [42, 44]. For each of the first three datasets that have ten classes, we randomly split 6 (4) classes of train/val-sets as the closed (open) train/val-sets respectively. For TinyImageNet that has 200 classes, we randomly split 20 (180) classes of train/val-sets as the closed (open) train/val-set. On each dataset and for each method, we repeat five times with different random splits and report the average AUROC on the val-set [42, 44]. As all methods have small standard deviations in their performance ( $<0.02$ ), we omit them for brevity.

**Results.** As this setup assumes no open training data, we cannot train discriminative classifiers like CLS. But we can still train OpenGAN-0 that uses GAN-discriminator (with

Table 2: Our technical insights apply to other GAN-based open-set discrimination methods: 1) using BiGAN-discriminator as the open likelihood function works better than using reconstruction errors (BiGAN<sub>d</sub><sup>fea</sup> vs. BiGAN<sub>r</sub><sup>fea</sup>), and 2) learning BiGANs on OTS features works much better than pixels (BiGAN<sub>d</sub><sup>fea</sup> vs. BiGAN<sub>d</sub><sup>pix</sup>). The results are comparable to Table 7.

dataset	BiGAN <sub>r</sub> <sup>pix</sup>	BiGAN <sub>r</sub> <sup>fea</sup>	BiGAN <sub>d</sub> <sup>pix</sup>	BiGAN <sub>d</sub> <sup>fea</sup>
MNIST	.976	.998	.986	.999
SVHN	.822	.976	.880	.993
CIFAR	.924	.967	.968	.973

model selection) as the open likelihood function. We have two salient conclusions from the results in Table 7. (1) Methods (e.g., NN and OpenGAN) work better on OTS features than pixels, suggesting that OTS features computed by the underlying  $K$ -way network are already good representations for open-set recognition. (2) OpenGAN-0<sup>fea</sup> performs the best and OpenGAN-0<sup>pix</sup> is competitive with prior methods such as GDM and GMM, suggesting that the GAN-discriminator is a powerful open likelihood function.

**Further Analysis.** There are many other GAN-based open-set methods, such as training BiGANs [56, 65, 66] or adversarial autoencoders [48, 53] on raw images, and using the reconstruction error as open-set likelihood [56, 53, 66, 1, 16]. We show our technical insights apply to different GAN architectures for open-set recognition: (1) using GAN-discriminator as the open-set likelihood function instead of pixel reconstruction errors, and (2) training them on OTS features rather than raw pixels. We hereby analyze a typical BiGAN-based method [66], which learns a BiGAN with both the reconstruction error and the GAN-discriminator. We compare BiGAN’s performance by either using the reconstruction error (BiGAN<sub>r</sub>) or its discriminator (BiGAN<sub>d</sub>) for open-set recognition. We also compare building BiGANs on either pixels (BiGAN<sup>pix</sup>) or features (BiGAN<sup>fea</sup>). Table 2 lists detailed comparisons under Setup-I (all models are selected on the val-sets). Clearly, our conclusions hold regardless of the base GAN architecture: 1) using OTS features rather than pixels (cf. BiGAN<sup>fea</sup> vs BiGAN<sup>pix</sup>), and 2) more importantly, using discriminators instead of reconstruction errors (cf. BiGAN<sub>d</sub> vs. BiGAN<sub>r</sub>).

## 4.3. Setup-II: Cross-Dataset Open-Set Recognition

Using cross-dataset examples as the open-set is another established protocol [36, 35, 31, 18]. We follow the “less bi-

Table 3: **Open-set recognition (Setup-II)** measured by AUROC $\uparrow$ , and macro-averaged F1-score $\uparrow$  over all  $(K+1)$  classes. We use TinyImageNet ( $K=200$ ) as the closed-set, and four different datasets as the open-sets. To report a method on a specific open-test-set out of four (first column), we perform four runs in which we use one of the four datasets as a validation set for training/tuning, and then average the performance measures over the four runs with a **superscript** marking the standard deviation. Methods such as Nearest Neighbor (NN) do not need tuning and hence have zero deviations. We provide a summary number in the bottom macro row by averaging the results over all open-test-sets. Detailed results in Table 4. Clearly, a binary classifier trained on features ( $\text{CLS}^{fea}$ ) already outperforms prior methods. However, when trained on pixels,  $\text{CLS}^{pix}$  works poorly in AUROC due to overfitting to high-dimensional raw images, but performs decently in F1. To note, without handling the open-set, the  $K$ -way model (trained only on the closed-set TinyImageNet) achieves 0.553 F1-score over  $(K+1)$  classes, suggesting that, when  $K$  is large ( $K=200$  here), F1-score can hardly reflect open-set discrimination performance which is better measured by AUROC. While largely underexplored in the literature, training a  $(K+1)$ -way model works quite well. Clearly,  $\text{OpenGAN}^{fea}$  works the best in both AUROC and F1-score. Please refer to Fig. 3(f-i) for ROC curves, and F1-scores vs. thresholds on the open-set likelihood.

open-test	metric	MSP [30]	OpenMax [6]	NN <sup>fea</sup> [50]	GMM [33]	C2AE [44]	MSP <sub>c</sub> [36]	MCdrop [20]	GDM [35]	CLS <sup>pix</sup> (K+1)	CLS <sup>fea</sup> Open	Open GAN <sup>pix</sup>	Open GAN <sup>fea</sup>
CIFAR	AUROC	.769 <sup>.000</sup>	.669 <sup>.011</sup>	.927 <sup>.000</sup>	.961 <sup>.013</sup>	.767 <sup>.020</sup>	.791 <sup>.007</sup>	.809 <sup>.005</sup>	.961 <sup>.007</sup>	.754 <sup>.367</sup>	.880 <sup>.091</sup>	.928 <sup>.113</sup>	.981 <sup>.027</sup>
	F1	.548 <sup>.002</sup>	.507 <sup>.001</sup>	.525 <sup>.000</sup>	.544 <sup>.002</sup>	.564 <sup>.002</sup>	.553 <sup>.003</sup>	.564 <sup>.001</sup>	.519 <sup>.003</sup>	.545 <sup>.032</sup>	.558 <sup>.017</sup>	.555 <sup>.027</sup>	.563 <sup>.035</sup>
SVHN	AUROC	.695 <sup>.000</sup>	.691 <sup>.014</sup>	.994 <sup>.000</sup>	.990 <sup>.016</sup>	.657 <sup>.018</sup>	.863 <sup>.013</sup>	.783 <sup>.009</sup>	.999 <sup>.006</sup>	.701 <sup>.224</sup>	.948 <sup>.068</sup>	.955 <sup>.052</sup>	.980 <sup>.014</sup>
	F1	.567 <sup>.002</sup>	.551 <sup>.002</sup>	.545 <sup>.000</sup>	.574 <sup>.002</sup>	.565 <sup>.001</sup>	.572 <sup>.002</sup>	.572 <sup>.001</sup>	.575 <sup>.002</sup>	.572 <sup>.027</sup>	.564 <sup>.015</sup>	.578 <sup>.014</sup>	.574 <sup>.009</sup>
MNIST	AUROC	.764 <sup>.000</sup>	.690 <sup>.019</sup>	.901 <sup>.000</sup>	.964 <sup>.021</sup>	.755 <sup>.008</sup>	.832 <sup>.017</sup>	.801 <sup>.009</sup>	.957 <sup>.007</sup>	.986 <sup>.327</sup>	.944 <sup>.015</sup>	.961 <sup>.083</sup>	.983 <sup>.068</sup>
	F1	.559 <sup>.001</sup>	.536 <sup>.013</sup>	.553 <sup>.000</sup>	.547 <sup>.008</sup>	.575 <sup>.001</sup>	.564 <sup>.001</sup>	.563 <sup>.001</sup>	.552 <sup>.002</sup>	.565 <sup>.020</sup>	.586 <sup>.021</sup>	.583 <sup>.010</sup>	.569 <sup>.016</sup>
Citysc.	AUROC	.789 <sup>.000</sup>	.693 <sup>.021</sup>	.715 <sup>.000</sup>	.867 <sup>.016</sup>	.814 <sup>.010</sup>	.851 <sup>.003</sup>	.868 <sup>.003</sup>	.513 <sup>.005</sup>	.646 <sup>.332</sup>	.971 <sup>.050</sup>	.828 <sup>.032</sup>	.933 <sup>.026</sup>
	F1	.579 <sup>.002</sup>	.514 <sup>.002</sup>	.583 <sup>.000</sup>	.572 <sup>.003</sup>	.589 <sup>.002</sup>	.583 <sup>.001</sup>	.571 <sup>.001</sup>	.546 <sup>.003</sup>	.589 <sup>.007</sup>	.561 <sup>.029</sup>	.587 <sup>.006</sup>	.588 <sup>.007</sup>
average	AUROC	.754	.686	.884	.945	.748	.834	.815	.857	.772	.936	.918	.969
	F1	.560	.527	.552	.559	.569	.568	.567	.548	.568	.565	.576	.573

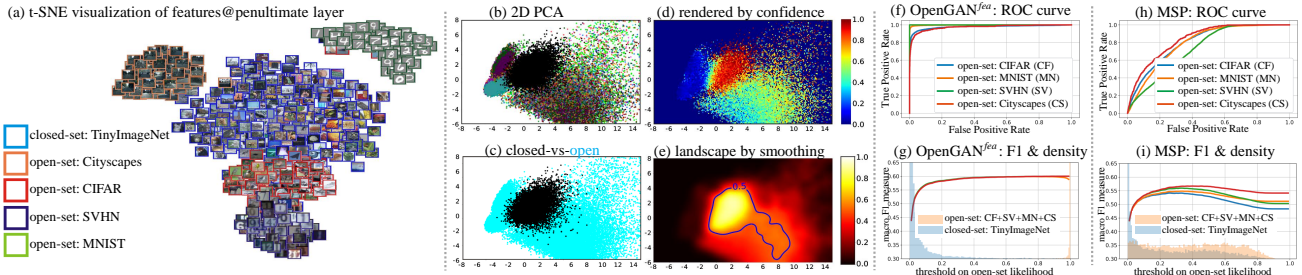


Figure 3: (a) We use t-SNE to visualize the embedding space through the OTS features computed by the  $K$ -way network trained on TinyImageNet train-set. Images from the other datasets are open-set examples. Clearly, closed and open examples are well separated in the feature space. We further visualize the “landscape” of the  $\text{OpenGAN}^{fea}$  open-set discriminator, by (b) projecting the OTS features into 2D using PCA; (c) coloring them with their closed/open labels; (d) rendering them with their open-set likelihoods computed by  $\text{OpenGAN}^{fea}$ ; (e) smoothing with Gaussian Filtering overlaid with the  $\text{OpenGAN}$ ’s decision boundary. We further compare  $\text{OpenGAN}^{fea}$  (tuned on SVHN) and MSP in (f-g) for open-set discrimination by ROC curves, and in (h-i) for open-set recognition by curves of the F1-score vs. thresholds of open likelihood. We render the density of open and closed testing data using shadows in (g) and (i). In these plots, we use each of the four cross-dataset open-test-sets (unseen in training) as an independent open-set to draw the curves. The curves clearly show that  $\text{OpenGAN}$  significantly outperforms MSP on open-set discrimination (AUROC) and open-set recognition (F1).

ased” protocol introduced in [57], which uses three datasets for benchmarking that reduces dataset-level bias [60]. This protocol tests the generalization of open-set methods to diverse open testing examples.

**Datasets.** We use TinyImageNet as the closed-set for  $K$ -way classification ( $K=200$ ). Images of each class are split into 500/50/50 images as the train/val/test sets. Following [57], we construct open train/val and test sets using different datasets [60], including MNIST (MN), SVHN (SV), CIFAR (CF) and Cityscapes (CS). For example, we use MNIST *train*-set to tune/train a model, and test it on CIFAR *test*-set as open-test set. This allows for analyzing how open-set methods generalize to diverse open testing examples (cf. Table 4). We use bilinear interpolation to resize all

images to 64x64 to match TinyImageNet image resolution.

**Results.** Table 3 shows detailed results. First, methods perform much better on features than pixels (e.g.,  $\text{CLS}^{fea}$  vs.  $\text{CLS}^{pix}$ ); and our  $\text{OpenGAN}$  performs the best. Perhaps surprisingly,  $\text{OpenMax}$ , a classic open-set, does not work well in this setup. This is consistent with the results in [18, 57]. We conjecture that  $\text{OpenMax}$  cannot effectively recognize cross-dataset open-set examples represented by logit features (computed by the  $K$ -way network) which are too invariant to be adequate for open-set recognition. Moreover, the  $(K+1)$ -way classifier also works quite well, even outperforming the open-vs-closed binary classifiers (CLS) in AUROC. Next we analyze why the binary classifier CLS (as widely done since [31]) are less effective.

Table 4: **Diagnostic analysis for cross-dataset open-set discrimination** measured by AUROC $\uparrow$ . In this setup, the TinyImageNet train/val/test sets serve as the closed train/val/test sets, and open train/test sets are the other two different datasets. Following outlier exposure [30], we train/tune CLS and OpenGAN on a cross-dataset as the open train-set. Recall that we do not train OpenGAN-0 on *any* open examples, although we tune it on the respective cross-dataset open train-set. CLS and OpenGAN use their last-epoch checkpoints to report performance. For better comparison, we report the **average** AUROC performance across all open-val-sets in the last column. We color the entries that have **AUROC  $< 0.9$**  with **red**, implying these models overfit to the open-train-set and generalize poorly on the other open-test-set. OpenGAN<sup>fea</sup> clearly performs the best; while CLS (esp. CLS<sup>pix</sup> which operates on pixels) generalizes poorly. Perhaps surprisingly, OpenGAN-0 performs equally well although it does not train on open training data.

<i>open-val-set</i>	CIFAR10 (CF)				SVHN (SV)				MNIST (MN)				Cityscapes (CS)				<i>avg.</i>
<i>open-test-set</i>	CF	SV	MN	CS	CF	SV	MN	CS	CF	SV	MN	CS	CF	SV	MN	CS	
CLS <sup>pix</sup>	.999	.999	.101	.895	.935	.999	.453	.972	.411	.340	.999	.113	.317	.512	.100	.999	.634
OpenGAN-0 <sup>pix</sup>	.999	.998	.550	.999	.999	.999	.993	.999	.999	.968	.999	.911	.999	.999	.915	.999	.958
OpenGAN <sup>pix</sup>	.999	.999	.989	.933	.974	.999	.997	.967	.976	.998	.999	.835	.967	.928	.950	.999	.969
CLS <sup>fea</sup>	.999	.933	.916	.699	.940	.999	.979	.863	.893	.961	.999	.781	.881	.926	.949	.968	.918
OpenGAN-0 <sup>fea</sup>	.999	.998	.997	.999	.964	.996	.996	.946	.952	.992	.994	.934	.994	.995	.992	.997	.984
OpenGAN <sup>fea</sup>	.999	.999	.990	.973	.974	.999	.996	.971	.976	.998	.999	.967	.973	.968	.970	.999	.984

**Further Analysis.** Table 4 lists detailed results of OpenGAN, CLS ( $\lambda_G=0$  in Eq. 2) and OpenGAN-0 ( $\lambda_o=0$  in Eq. 2), when trained/tuned and tested on different cross-dataset open-set examples. All methods perform better on OTS features than pixels (cf. CLS<sup>fea</sup> vs. CLS<sup>pix</sup>); and work almost perfectly when trained and tested with the same open-set dataset, e.g., column-cf under “CIFAR-train (cf)” where we use CIFAR images as the open-set data. However, when tested on a different dataset of open-set examples, CLS performs quite poorly (especially when built on pixels) because it overfits easily to high-dimensional pixel images [57]. In contrast, with *fake* open-data generated adversarially, OpenGAN and its special form OpenGAN-0 perform and generalize much better. Nevertheless, this implies a failure mode of OpenGAN, because the open-set data used in training could be quite different from those in testing, potentially leading to an OpenGAN that perform poorly in the real open world. Perhaps surprisingly, OpenGAN-0<sup>fea</sup> performs as well as OpenGAN<sup>fea</sup>, although it does not train on open-set data. This further shows the merit of generating *fake* open examples to augment heavily-biased open-set training data, and our technique insights (as previously analyzed under *Setup-I*): 1) using GAN-discriminator as the likelihood function, and 2) training GANs on OTS features rather than pixels.

**Visualization.** Fig. 1 shows some synthesized images by GAN<sup>pix</sup>, and we visualize more in the supplement. To intuitively illustrate how the synthesized images help better span the open-world, we analyze why a simple discriminator works so well when trained on the OTS features. We visualize the features in Fig. 3 (a) and “decision landscape” in Fig. 3 (b-e), demonstrating that the closed- and open-set images are clearly separated in the feature space.

#### 4.4. Setup-III: Open-Set Semantic Segmentation

Open-set semantic segmentation has been explored in recent work [7, 29], which creates synthetic open-set pix-

els by pasting virtual objects (e.g., cropped from Pascal VOC masks [19]) on Cityscapes images. In this work, we do not generate synthetic pixels but instead repurpose “other” pixels (outside the set of  $K$  classes) that already exist in Cityscapes. Interestingly, classic semantic segmentation benchmarks evaluate these “other” pixels as a separate background class [19], but Cityscapes ignores them in its evaluation (as do many other contemporary datasets [10, 4, 51, 43]). The historically-ignored pixels include vulnerable objects (e.g., strollers in Fig. 2), and can be naturally evaluated as open-set examples.

**Datasets.** Cityscapes [14] contains 1024x2048 high-resolution urban scene images with 19 class labels for semantic segmentation. We construct our train- and val-sets from its 2,975 training images, in which we use the last 10 images as val-set and the rest as train-set. We use its official 500 validation images as our test-set. The “other” pixels (cf. Fig. 2) are the open-set examples in this setup. We refer readers to the supplemental for details, such as model architecture, batch construction, weight tuning, etc.

**Pixel Generation.** As Cityscapes has high-resolution images (1024x2048), it is nontrivial to train OpenGAN<sup>pix</sup>, especially its special form OpenGAN-0<sup>pix</sup>, which must learn to generate high-resolution images. We find the successful training of OpenGAN-0<sup>pix</sup> depends on the resolution of images to be generated: we train OpenGAN-0<sup>pix</sup> by generating patches (64x64), not full-resolution images.

**Results.** Table 5 lists quantitative results. As we train OpenGAN and CLS with open pixels, we diagnose in Fig. 5 the open-set performance by varying the number of training images that provide the open-training pixels, along with closed-training pixels from all training images. First, these results show that OpenGAN<sup>fea</sup> substantially outperforms all other methods. Generally speaking, the methods that process features outperform those that process pixels (e.g., OpenGAN and CLS in Fig. 5). This suggests that OTS features (from the segmentation network) serve as a pow-

Table 5: **Comparison in open-set semantic segmentation** on Cityscapes (AUROC  $\uparrow$ ). All methods are implemented on top of the segmentation network HRNet [61] except the ones operating on pixels (as marked by  $^{pix}$ ). Our approach OpenGAN $^{fea}$  clearly performs the best. Fig. 5 analyzes OpenGAN trained with varied number of open-set pixels, when built on either pixels or OTS features.

MSP [30]	Entropy [58]	OpenMax [6]	C2AE [44]	MSP $_c$ [36]	MCdrop [20]	GDM [35]	GMM [33]	HRNet-(K+1)	OpenGAN-0 $^{fea}$	CLS $^{fea}$	OpenGAN $^{fea}$
.721	.697	.751	.722	.755	.767	.743	.765	.755	.709	.861	.885

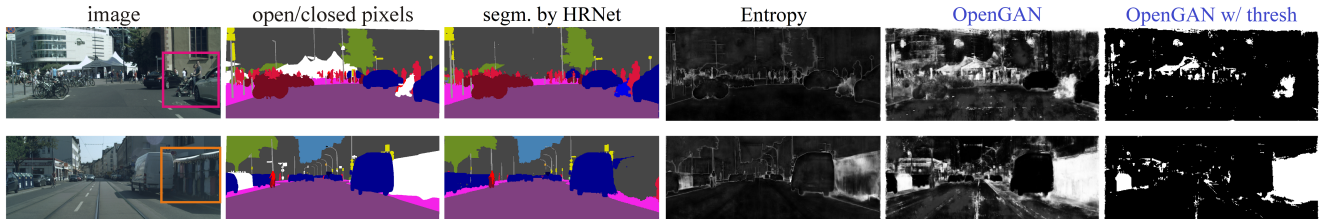


Figure 4: Qualitative results of two testing images, on which a state-of-the-art network (HRNet) misclassifies the *unknown* categories *stroller/street-shop* as motorcycle/building. From left to right of each row: the input image, its per-pixel semantic labels (in which white regions are open-set pixels), the semantic segmentation result by HRNet, open-set likelihoods by Entropy, our OpenGAN $^{fea}$ , and its thresholded open-pixel map (threshold=0.7). OpenGAN clearly captures most open-set pixels (the white ones). Note that the *street-shop* is a real open-set example because Cityscapes train-set does not have another *street-shop* like this size and content (i.e., selling clothes).

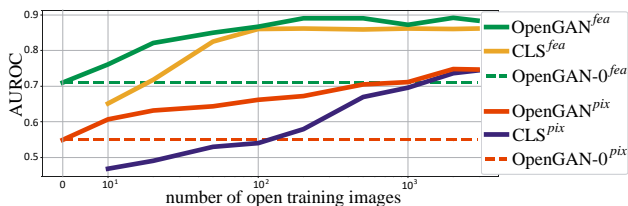


Figure 5: Diagnostic study w.r.t AUROC vs. number of open images which provide open-set training pixels. Our methods perform better on OTS features than pixels. Recall OpenGAN-0 is equivalent to training a normal GAN (without open training data) and using its discriminator as open-set likelihood. With some open training data (e.g., 100 open images), CLS outperforms OpenGAN-0; but OpenGAN consistently performs the best.

erful representation for open-set pixel recognition. The curves in Fig. 5 imply that methods with enough data on pixels should work (e.g., achieving similar performance as on features). This is consistent with evidence from semantic segmentation works. However, methods saturate more quickly on OTS features than pixels, suggesting the benefit of using OTS features for open-set recognition. Moreover, OpenGAN-0 performs better than CLS when trained on fewer open-training images (e.g., 10). But with modest number of open training images (e.g., 50), CLS outperforms OpenGAN-0 and other classic methods (e.g., OpenMax and C2AE in Table 5) which assume no open training data. This confirms the effectiveness of Outlier Exposure, even with a modest amount of outliers [31].

Bayesian networks (MCdrop and MSP $_c$ ) outperform the baseline MSP, showing that uncertainties can be reasonably used for open-set recognition. Lastly, we train a “ground-up”  $(K+1)$ -way HRNet model that treats “other” pixels as the  $(K+1)^{th}$  background class [19], shown by HRNet-(K+1) in Table 5. It performs better than other typical open-set methods but much lower than the simple open-vs-closed binary classifier CLS $^{fea}$ , presumably because the  $(K+1)$ -



Figure 6: Visuals of Cityscapes real image patches (left), synthesized patches by OpenGAN-0 $^{pix}$  (mid) and OpenGAN-0 $^{fea}$  (right). As OpenGAN-0 $^{fea}$  generates features instead of pixel patches, we “synthesize” the patch for a generated feature by finding the closest pixel feature on the training set and returning its surrounding image patch. We can see OpenGAN-0 $^{pix}$  synthesizes realistic patches w.r.t color and tone, but it (0.549 AUROC) notably underperforms OpenGAN-0 $^{fea}$  (0.709 AUROC) for open-set segmentation. The “synthesized” patches by OpenGAN-0 $^{fea}$  capture many open-set objects, such as bridge, back-of-traffic-sign and unknown-static-objects, none of which belong to any of the 19 closed-set classes in the Cityscapes benchmark. This intuitively shows why methods work better on OTS features than pixels.

way model has to strike a balance over all the  $(K+1)$  classes while the binary CLS benefits from training on more balanced batches of closed/open pixels.

**Visualization.** Fig. 4 qualitatively compares OpenGAN and the entropy method (more visual results are in the supplemental). The visualization shows OpenGAN sufficiently recognize open-set pixels. It also implies failure happens when OpenGAN misclassifies open-vs-closed pixels. Fig. 16 compares some generated patches by OpenGAN-0 $^{fea}$  and OpenGAN-0 $^{fea}$ , intuitively showing why using OTS features leads to better performance for open-set recognition.

## 5. Conclusion

We propose **OpenGAN** for open-set recognition by incorporating two technical insights, 1) training an open-vs-closed classifier on OTS features rather than pixels, and 2) adversarially synthesizing *fake* open data to augment the

set of open-training data. With OpenGAN, we show using GAN-discriminator *does* achieve the state-of-the-art on open-set discrimination, once being selected using a val-set of real outlier examples. This is effective even when the outlier validation examples are sparsely sampled or strongly biased. OpenGAN significantly outperforms prior art on both open-set image recognition and semantic segmentation.

## Acknowledgement

This work was supported by the CMU Argo AI Center for Autonomous Vehicle Research.

## References

- [1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision (ACCV)*, 2018. 2, 5
- [2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 4, 14
- [3] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *ICML*, 2017. 14
- [4] Jens Behley, Martin Garbade, Andres Milioto, Jan Quen- zel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*, 2019. 7
- [5] Abhijit Bendale and Terrance Boulton. Towards open world recognition. In *CVPR*, 2015. 3
- [6] Abhijit Bendale and Terrance E Boulton. Towards open set deep networks. In *CVPR*, 2016. 2, 4, 5, 6, 8, 16, 18
- [7] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *arXiv preprint arXiv:1904.03215*, 2019. 4, 7, 12
- [8] Oren Boiman, Eli Shechtman, and Michal Irani. In defense of nearest-neighbor based image classification. In *CVPR*. IEEE, 2008. 17
- [9] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019. 4
- [10] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Gi- ancarlo Baldan, and Oscar Beijbom. nuscenes: A multi- modal dataset for autonomous driving. In *CVPR*, 2020. 7
- [11] Yue Cao, Mingsheng Long, Jianmin Wang, Han Zhu, and Qingfu Wen. Deep quantization network for efficient image retrieval. In *AAAI*, 2016. 17
- [12] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009. 2
- [13] Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. In *ECCV*, 2020. 3, 5, 17, 18
- [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 *CVPR*, 2016. 2, 4, 7
- [15] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *ICML*, 2006. 4
- [16] Lucas Deecke, Robert Vandermeulen, Lukas Ruff, Stephan Mandt, and Marius Kloft. Image anomaly detection with generative adversarial networks. In *Joint european conference on machine learning and knowledge discovery in databases*. Springer, 2018. 2, 5
- [17] David Dehaene, Oriel Frigo, Sébastien Combreselle, and Pierre Eline. Iterative energy-based projection on a normal data manifold for anomaly localization. In *ICLR*, 2020. 2
- [18] Akshay Raj Dhamija, Manuel Günther, and Terrance Boulton. Reducing network agnostophobia. In *NeurIPS*, 2018. 2, 3, 5, 6
- [19] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christo- pher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 4, 7, 8, 12
- [20] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 2, 5, 6, 8, 18
- [21] ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multi-class open set clas- sification. In *British Machine Vision Conference (BMVC)*, 2017. 2, 4, 5, 18
- [22] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Re- cent advances in open set recognition: A survey. *IEEE Trans- actions on Pattern Analysis and Machine Intelligence*, 2020. 2
- [23] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *ICCV*, 2019. 2
- [24] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolu- tional activation features. In *ECCV*, 2014. 16
- [25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2, 3, 14
- [26] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Lar- lus. End-to-end learning of deep visual representations for image retrieval. *IJCV*, 2017. 16
- [27] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *ICLR*, 2019. 16, 17
- [28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 16
- [29] Dan Hendrycks, Steven Basart, Mantas Mazeika, Moham- madreza Mostajabi, Jacob Steinhardt, and Dawn Song.

- A benchmark for anomaly segmentation. *arXiv preprint arXiv:1911.11132*, 2019. 4, 7, 12
- [30] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 2, 5, 6, 7, 8, 16, 18
- [31] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019. 1, 2, 3, 4, 5, 6, 8, 14
- [32] Pedro R Mendes Júnior, Roberto M De Souza, Rafael de O Werneck, Bernardo V Stein, Daniel V Pazinato, Waldir R de Almeida, Otávio AB Penatti, Ricardo da S Torres, and Anderson Rocha. Nearest neighbors distance ratio open-set classifier. *Machine Learning*, 106(3):359–386, 2017. 17
- [33] Shu Kong and Deva Ramanan. An empirical exploration of open-set recognition via lightweight statistical pipelines, 2021. 4, 6, 8, 14, 16, 17, 18
- [34] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *ICLR*, 2018. 4
- [35] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018. 2, 4, 5, 6, 8, 16, 17, 18
- [36] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018. 1, 2, 5, 6, 8, 18
- [37] Yezheng Liu, Zhe Li, Chong Zhou, Yuanchun Jiang, Jian-shan Sun, Meng Wang, and Xiangnan He. Generative adversarial active learning for unsupervised outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1517–1528, 2019. 1, 2, 4
- [38] Antonio Loquercio, Mattia Segu, and Davide Scaramuzza. A general framework for uncertainty estimation in deep learning. *IEEE Robotics and Automation Letters*, 2020. 5
- [39] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. 17
- [40] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *ECCV*, 2012. 17
- [41] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don’t know? *arXiv preprint arXiv:1810.09136*, 2018. 3
- [42] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *ECCV*, 2018. 2, 3, 4, 5, 18
- [43] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Buló, and Peter Kotschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 7
- [44] Poojan Oza and Vishal M. Patel. C2AE: class conditioned auto-encoder for open-set recognition. In *CVPR*, 2019. 2, 3, 4, 5, 6, 8, 16, 17, 18
- [45] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 4
- [46] Pramuditha Perera, Vlad I Morariu, Rajiv Jain, Varun Manjunatha, Curtis Wigington, Vicente Ordonez, and Vishal M Patel. Generative-discriminative feature representations for open-set recognition. In *CVPR*, 2020. 5
- [47] Pramuditha Perera and Vishal M Patel. Deep transfer learning for multiple class novelty detection. In *CVPR*, 2019. 4
- [48] Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. In *NeurIPS*, 2018. 1, 2, 4, 5
- [49] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 4
- [50] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, 2000. 4, 5, 6
- [51] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 7
- [52] Lukas Ruff, Robert A Vandermeulen, Nico Görmitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *ICLR*, 2020. 2
- [53] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *CVPR*, 2018. 1, 2, 4, 5
- [54] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boulton. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2012. 2, 3, 4
- [55] Walter J Scheirer, Anderson Rocha, Ross J Micheals, and Terrance E Boulton. Meta-recognition: The theory and practice of recognition score analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011. 5
- [56] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017. 1, 2, 4, 5
- [57] Alireza Shafaei, Mark Schmidt, and James J. Little. A less biased evaluation of out-of-distribution sample detectors. In *British Machine Vision Conference (BMVC)*, 2019. 1, 2, 3, 4, 6, 7
- [58] Jacob Steinhardt and Percy S Liang. Unsupervised risk estimation using only conditional independence structure. In *NeurIPS*, 2016. 5, 8, 17, 18
- [59] Xin Sun, Zhenning Yang, Chi Zhang, Keck-Voon Ling, and Guohao Peng. Conditional gaussian distribution learning for open set recognition. In *CVPR*, 2020. 2, 17, 18
- [60] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528, 2011. 4, 6
- [61] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep

- high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2, 4, 8, 13
- [62] Yan Xia, Xudong Cao, Fang Wen, Gang Hua, and Jian Sun. Learning discriminative reconstructions for unsupervised outlier removal. In *ICCV*, pages 1511–1519, 2015. 2
  - [63] Songfan Yang and Deva Ramanan. Multi-scale recognition with dag-cnns. In *CVPR*, pages 1215–1223, 2015. 16
  - [64] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *CVPR*, 2019. 2, 5
  - [65] Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar. Efficient gan-based anomaly detection. *arXiv preprint arXiv:1802.06222*, 2018. 5
  - [66] Houssam Zenati, Manon Romain, Chuan-Sheng Foo, Bruno Lecouat, and Vijay Chandrasekhar. Adversarially learned anomaly detection. In *IEEE International Conference on Data Mining (ICDM)*, 2018. 1, 2, 4, 5
  - [67] Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. Hybrid models for open set recognition. In *ECCV*, 2020. 2, 4, 5, 17, 18
  - [68] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 4, 12
  - [69] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *ICLR*, 2018. 1, 2

## Outline

As elaborated in the main paper, our proposed OpenGAN trains an open-vs-closed binary classifier for open-set recognition. Our three major technical insights are (1) model selection of a GAN-discriminator as the open-set likelihood function via validation, (2) augment the available set of real open training examples with adversarially synthesized “fake” data, and (3) training OpenGAN on off-the-shelf (OTS) features rather than pixel images. We expand on the techniques of OpenGAN in the appendix, including architecture design, model selection and additional details for training. We also provide additional comparisons to recently published methods and qualitative results. Below is the outline.

**Section 6:** Model architectures for both OpenGAN<sup>fea</sup> and OpenGAN<sup>pix</sup>.

**Section 7:** Detailed setup for open-set semantic segmentation, such as data statistics (e.g., the number of open-set pixels in the testing set) and batch construction during training.

**Section 8:** Model selection that is performed on a validation set.

**Section 9:** Hyper-parameter tuning which is performed on a validation set.

**Section 10:** Statistical methods for open-set recognition that learn generative models (e.g., Gaussian Mixture) over off-the-shelf deep features.

**Section 11:** More quantitative comparisons to several approaches published recently.

**Section 12:** Visuals of synthesized images generated by OpenGAN-0<sup>pix</sup> and OpenGAN-0<sup>fea</sup>, intuitively demonstrating their effectiveness and limitations.

**Section 13:** Visual results of open-set semantic segmentation.

**Section 14:** Failure Cases and Limitations.

## 6. Model Architecture

We describe the network architectures of OpenGAN. Because our final version OpenGAN<sup>fea</sup> operates on off-the-shelf (OTS) features, we use multi-layer perceptron (MLP) networks for the generator and discriminator. Because OpenGAN<sup>pix</sup> operates on pixels, we make use of convolutional neural network (CNN) architectures. We begin with the former.

### 6.1. OpenGAN<sup>fea</sup> architecture

OpenGAN<sup>fea</sup> consists of a generator and a discriminator. OpenGAN<sup>fea</sup> is compact in terms of model size ( $\sim 2$ MB), because it adopts MLP network over OTS features

which are low-dimensional (e.g., 512-dim vectors) compared to pixel images. The MLP architectures are described below

- The MLP discriminator in OpenGAN<sup>fea</sup> takes a  $D$ -dimensional feature as the input. Its architecture has a set of fully-connected layers (fc marked with input-dimension and output-dimension), Batch Normalization layers (BN) and LeakyReLU layers (hyper-parameter as 0.2): fc ( $D \rightarrow 64 \times 8$ ), BN, LeakyReLU, fc ( $64 \times 8 \rightarrow 64 \times 4$ ), BN, LeakyReLU, fc ( $64 \times 4 \rightarrow 64 \times 2$ ), BN, LeakyReLU, fc ( $64 \times 2 \rightarrow 64 \times 1$ ), BN, LeakyReLU, fc ( $64 \times 1 \rightarrow 1$ ), Sigmoid.
- The MLP generator synthesizes a  $D$ -dimensional feature given a 64-dimensional random vector: fc ( $64 \rightarrow 64 \times 8$ ), BN, LeakyReLU, fc ( $64 \times 8 \rightarrow 64 \times 4$ ), BN, LeakyReLU, fc ( $64 \times 4 \rightarrow 64 \times 2$ ), BN, LeakyReLU, fc ( $64 \times 2 \rightarrow 64 \times 4$ ), BN, LeakyReLU, fc ( $64 \times 4 \rightarrow D$ ), Tanh.

For open-set image classification, the image features have dimension  $D = 512$  from ResNet18 (the  $K$ -way classification networks under *Setup-I* and *II*). For open-set segmentation, the per-pixel features have dimension  $D = 720$  at the penultimate layer of HRnet (a top-ranked semantic segmentation model used in this work under *Setup-III*).

### 6.2. OpenGAN<sup>pix</sup> architecture

OpenGAN<sup>pix</sup>’s generator and discriminator follow the CycleGAN architecture [68]. We change the stride size in the convolution layers to adapt the networks to specific image resolution (e.g., CIFAR 32x32 and TinyImageNet 64x64). The generator and discriminator in OpenGAN<sup>pix</sup> have model sizes as  $\sim 14$ MB and  $\sim 11$ MB, respectively. We find it important to ensure that OpenGAN<sup>pix</sup> has a larger capacity than OpenGAN<sup>fea</sup> to generate high-dimensional RGB raw images.

## 7. Setup for Open-Set Semantic Segmentation

In this work, we use Cityscapes to study open-set semantic segmentation. Prior work suggests pasting virtual objects (e.g., cropped from PASCAL VOC masks [19]) on Cityscapes images as open-set pixels [7, 29]. We notice that Cityscapes ignores a sizeable portion of pixels in its benchmark, as demonstrated by Figure 7. As a result, many methods also ignore them in training. Therefore, instead of introducing artificial open-set examples, we use the historically-ignored pixels in Cityscapes as the real open-set examples. We hereby describe in detail our configuration for open-set semantic segmentation setup and experiments on Cityscapes.

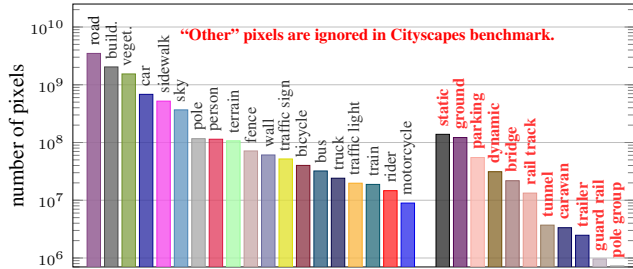


Figure 7: Cityscapes annotates a sizeable portion of pixels that do not belong to one of the  $K$  closed-set classes on which the Cityscapes benchmark evaluates. As a result, many methods also ignore them during training [61]. We repurpose these historically-ignored pixels as open-set examples that are from the  $(K+1)^{th}$  “other” class, allowing for a large-scale exploration of open-set recognition via semantic segmentation.

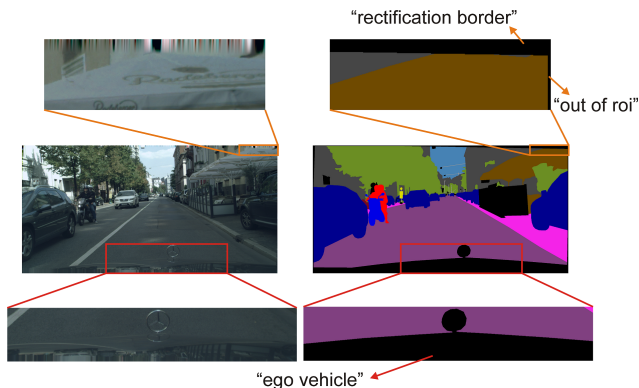


Figure 8: **Void pixels in Cityscapes** that are not from the closed-set classes nor open-set. We highlight these pixels over an image (left) and its semantic annotations (right). Cityscapes labels these pixels as *rectification-border* (artifacts at the image borders caused by stereo rectification), *ego-vehicle* (a part of the car body at the bottom of the image including car logo and hood) and *out-of-roi* (narrow strip of 5 pixels along the image borders). These noise-pixels can be easily identified without machine-learned methods. Therefore, we do not evaluate on these pixels.

**Data Setup.** Cityscapes training set has 2,975 images. We use the first 2,965 images for training, and hold out the last 10 as validation set for model selection. We use the 500 Cityscapes validation images as our test set. Here are the statistics for the full train/val/test sets.

- train-set for closed-pixels: 2,965 images providing 334M closed-set pixels.
- train-set for open-pixels: 2,965 images providing 44M open-set pixels.
- val-set for closed-pixels: 10 images providing 1M closed-set pixels.
- val-set for open-pixels: 10 images providing 0.2M

open-set pixels.

- test-set for closed-pixels: 500 images providing 56M pixels.
- test-set for open-pixels: 500 images providing 2M pixels.

Note that, we exclude the pixels labeled with *rectification-border* (artifacts at the image borders caused by stereo rectification), *ego-vehicle* (a part of the car body at the bottom of the image including car logo and hood) and *out-of-roi* (narrow strip of 5 pixels along the image borders). These pixels can be easily localized using camera information. We demonstrate such pixels in Figure 8. In this sense, these pixels are not *unknown* open-set pixels but *known* noises caused by sensors and viewpoint. Therefore, we do not include them for open-set evaluation.

#### Feature setup.

- $M^{pix}$ , where  $M \in \{\text{CLS}, \text{OpenGAN}\}$ , corresponds to a model defined on raw pixels.
- $M^{fea}$  corresponds to a model defined on embedding features at the penultimate layer of underlying semantic segmentation network (i.e., HRNet as introduced below).
- HRNet [61] is a top-ranked semantic segmentation model on Cityscapes. It has a multiscale pyramid head that produce high-resolution segmentation prediction. We extract embedding features at its penultimate layer (720-dimensional before the 19-way classifier). We also tried other layers but we did not observe significant difference in their performance.

**Batch Construction.** To fully shuffle open- and closed-set training pixels, we cache all the open-set training pixel features extracted from HRNet. We construct a batch consisting of 10,000 pixels for training  $\text{OpenGAN}^{fea}$ . To do so, we

- randomly sample a real image, run HRNet over it and randomly extract 5,000 closed-set training pixel features;
- randomly sample 2,500 open-set training features from cache;
- run the  $\text{OpenGAN}^{fea}$  generator (being trained on-the-fly) to synthesize 2,500 “fake” open-set pixel features.

Similarly, to train  $\text{OpenGAN}^{pix}$  which is fully-convolutional, we construct a batch of 10,000 pixels as below.

- We feed a random real image to the  $\text{OpenGAN}^{pix}$  discriminator, and penalize predictions on 5000 random closed pixels and 2500 random open pixels.
- We run the  $\text{OpenGAN}^{pix}$  generator (being trained on-the-fly) to synthesize a “fake” image. We feed this “fake” image to the discriminator along with *open-set* labels. We penalize 2500 random “fake” pixels.

## 8. Model Selection

Due to the unstable training of GANs [2], model selection is crucial and challenging. GANs are typically used for generating realistic images, so model selection for GANs focuses on selecting generators. To do so, one relies heavily on manual inspection of visual results over the generated images from different model epochs [25]. In contrast, we must select the discriminator, rather than the generator, because we use the discriminator as an open-set likelihood function for open-set recognition. It is important to note that, in theory, a perfectly trained discriminator would not be capable of recognizing fake open-set data because of the equilibrium in the discriminator/generator game [3]. Although such an equilibrium hardly exist in practice, we find it crucial to select GAN discriminators to be used as open-set likelihood function. For model selection, we further find it crucial to use a validation set that consists of both real open and closed data. We present this study below.

**Model selection is crucial.** In Figure 9, we plot the open-set classification performance as a function of training epochs. We study both  $\text{OpenGAN-0}^{fea}$  and  $\text{OpenGAN-0}^{pix}$  on the three datasets as typically used in open-set recognition (under *Setup-I*). Recall that  $\text{OpenGAN-0}$  is to train a normal GAN and use its discriminator as open-set likelihood function for open-set recognition. Clearly, we can see that long training time does *not* necessarily improve open-set classification performance. We posit that this is due to the unstable training of GANs. This motivates robust model selection using a validation set.

**Synthesized data are not sufficient for model selection.** To study how each checkpoint models perform in training (fake-vs-real classification) and testing (open-vs-closed classification), we scatter-plot Figure 10, where we render the dots with colors to indicate the model epoch (blue→red dots represent model epoch-0→50, respectively). For the scatter plot, the ideal case is that the train-time and test-time performance is linearly correlated, i.e., all dots appear in the diagonal line (from origin to top right). But their performances on the two sets are not correlated, suggesting that using the synthesized data for model selection is not sufficient. Instead, we find it crucial to use a validation set of real open examples to select the open-set discriminator. Our observation is consistent to what reported in [31]. It is worth noting that the models selected on

the validation set *do* generalize to test sets. This has been demonstrated in Table 3 and 4 in the main paper.

## 9. Hyper-Parameter Tuning

Strictly following the practice of machine learning, we tune hyper-parameters on the same validation set. We now study parameter tuning through open-set semantic segmentation (*Setup-III*). We select the best  $\text{OpenGAN}$  model according to the performance on the validation set (10 images).

In training  $\text{OpenGAN}$ , a training batch contains both real closed- and open-set pixels, and synthesized *fake open* pixels. Correspondingly, our loss function has three terms (refer to Eq. 2 in the main paper). Therefore, we tune the hyper-parameter  $\lambda_o$  and  $\lambda_G$  as below to balance the terms in the loss function that exploits real open data and generated data:

- The term exploiting real open data has a weight  $\lambda_o = 1$ . We do not tune this as we presume the sparsely sampled open-set examples are equally important as the real closed-set examples.
- The term using the generated “fake” data has varied parameter  $\lambda_G \in [0.05, 0.10, 0.15, 0.20, \dots, 0.80, 0.85, 0.90]$ . We mainly focus on tuning  $\lambda_G$  to study how the synthesized data help training.

In Table 6, we show the performance on the test set of  $\text{OpenGAN}^{pix}$  and  $\text{OpenGAN}^{fea}$  with varied open training images. For each selected model, we mark the corresponding  $\lambda_G$  that yields the best performance (on validation set). Roughly speaking, it is preferable to set a lower weight  $\lambda_G$  when we have more real open-set training data. However, we do not see a clear correlation between the weight  $\lambda_G$  and test-time performance. We believe this is due to the random initialization which affects adversarial learning.

We also study how models trained with different  $\lambda_G$  perform on validation set and test set, and if the model selected on the validation set can reliably perform well on the testing set. Figure 11 plots the performance as a function of  $\lambda_G$  on validation set and test set. Hereby we choose the  $\text{OpenGAN}^{fea}$  model trained with 1000 open training images. We can see the performance on the validation set reliably reflects the performance on the test set. This confirms that model selection on the validation set is reliable.

## 10. Statistical Models for Open-Set

Our previous work introduces a lightweight statistical pipeline that repurposes off-the-shelf (OTS) deep features for open-set recognition [33]. For the completeness of this paper, we briefly introduce this pipeline: (1) extracting OTS features (with appropriate processing detailed below) of closed-set training examples using the underlying

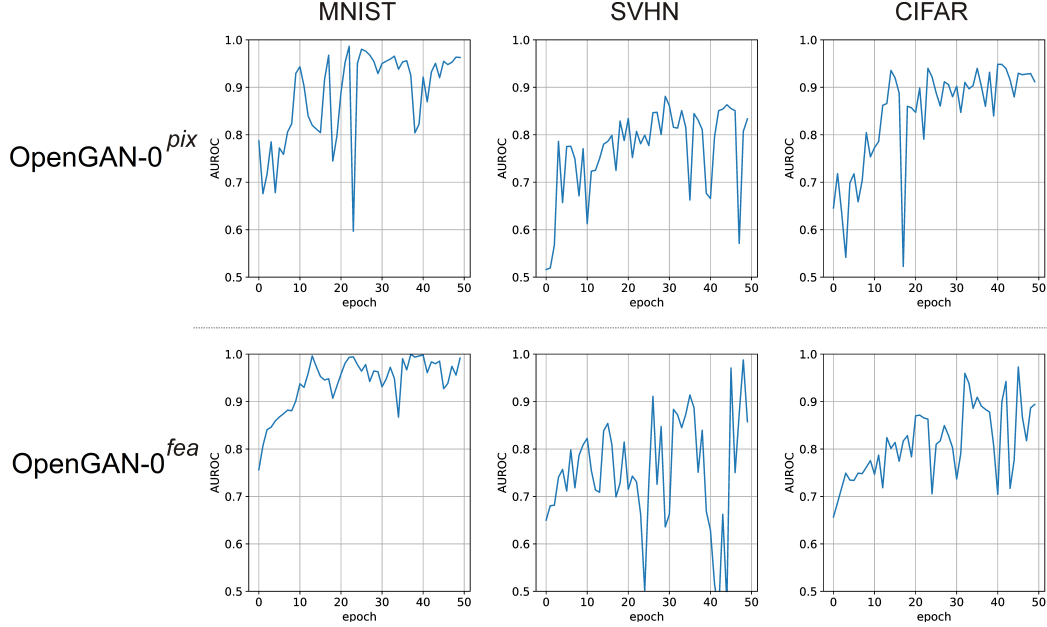


Figure 9: **Open-set image recognition performance vs. training epochs.** We show the performance (AUROC) by OpenGAN-0<sup>*pix*</sup> and OpenGAN-0<sup>*fea*</sup> on the val-sets of the three datasets which are widely studied in the open-set recognition literature (*Setup-I*). Recall that OpenGAN-0 is to train a normal GAN and use its discriminator as open-set likelihood function for open-set recognition. We can see that best open-set discrimination performance is achieved by intermediate checkpoints of GAN discriminators, and longer training does *not* necessarily improve performance. This is due to the unstable training of GANs with the min-max game. This motivates the need for robust model selection.

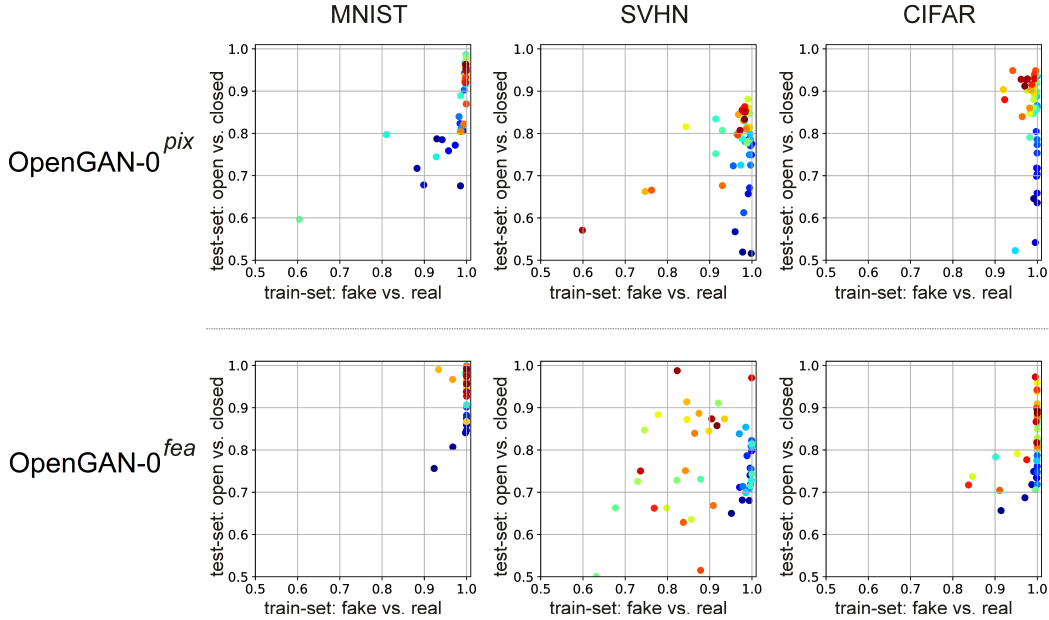


Figure 10: **Scatter plot of training performance (fake-vs-real classification) and testing performance (open-vs-closed classification).** We color the dots from blue  $\rightarrow$  red to marks models saved at epoch-0  $\rightarrow$  50, respectively. We use the three datasets (under *Setup-I*) following the typical setup of open-set recognition. The ideal correlation is that all the dots lie in the diagonal from bottom-left to top-right. However, there is no correlation between training (fake-vs-real) and validation (open-vs-closed) performance. Moreover, because the dots appear to be on the right part in the plots, this means that fake-vs-real classification (as denoted by the  $x$ -axis) is much easier than open-vs-closed classification (as denoted by the  $y$ -axis). These scatter plots demonstrate that (1) intermediate discriminators can perform quite well in open-set discrimination (i.e., on the validation set consisting of real open and closed-set images), and (2) synthesized data are insufficient to be used for model selection.

Table 6: **Hyper-parameter tuning for open-set semantic segmentation on Cityscapes.** Given a fixed number of open training images, we vary the hyper-parameter  $\lambda_G$  to train OpenGAN models. Recall that  $\lambda_G$  controls the contribution of synthesized data in the loss function. We conduct model selection on the val-set (10 images), and report here the performance (AUROC $\uparrow$ ) on the test set (500 images). We also mark the  $\lambda_G$  for each of the selected models. It seems to be preferable to set a lower weight  $\lambda_G$  (for the term exploiting synthesized data examples) when we have more real open-set data, but we do not see a tight correlation between  $\lambda_G$  and test-time performance. We believe this is because of the (random) initialization of model weights that has a non-trivial impact on training GANs and their final performance.

#images <sub>train</sub> <sup>open</sup>	10	20	50	100	200	500	1000	2000	2900
OpenGAN <sup>fea</sup>	.761	.821	.849	.866	.891	.890	.873	.891	.885
$\lambda_G$	0.20	0.20	0.05	0.10	0.05	0.05	0.20	0.05	0.05
OpenGAN <sup>pix</sup>	.607	.632	.643	.661	.672	.705	.711	.748	.746
$\lambda_G$	0.60	0.40	0.80	0.90	0.70	0.60	0.60	0.60	0.70

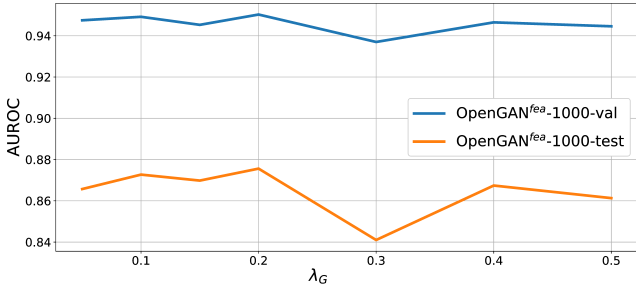


Figure 11: **Tuning hyper-parameter  $\lambda_G$ .** We plot the open-set discrimination performance (AUROC) as a function of  $\lambda_G$ , which controls the contribution of generated data examples in the loss function. The model we report here is OpenGAN<sup>fea</sup>-1000 that is trained with 1000 open-set training images. The validation set and test set contain 10 and 500 images. Although the validation set has much fewer images than the test set, the open-set classification performances align well on the two sets.

$K$ -way classification model, (2) learning statistical models over the OTS features. There are many statistical methods one can choose, e.g., nearest class centroids, Nearest Neighbors, and (class-conditional) Gaussian Mixture Models (GMMs). During testing, we extract the OTS features of the given example and resort to the learned statistical models to compute an open-set likelihood, e.g., based on the (inverse) closed-set probability from GMM. By thresholding the open-set likelihood, we decide whether it is an open-set example or one of the  $K$  closed-set classes, with the latter we report the predicted class label.

**Feature extraction.** OTS features generated at different layers of the trained  $K$ -way classification network can be repurposed for open-set recognition. Most methods leverage softmax [30] and logits [6, 27, 44] which can be thought of as features extracted at top layers. Similar to [35], we find it crucial to analyze features from intermediate layers for open-set recognition, because logits and softmax may be too invariant to be effective for open-set recognition (Fig. 13). One immediate challenge to extract features from an intermediate layer is their high dimensionality, e.g.



Figure 12: **street-shop as open-set.** Figure 4 in the main paper shows open-set pixel recognition results on a street-shop on a testing image (top-row). We verify if such a street-shop appears in the training set. We manually search for a similar street-shop in the training set, and find the one (bottom-row) most similar to the testing example in terms of size. Importantly, we did not find any other street-shops in the training set that sell clothes like the testing example shown in the top row. In this sense, the testing image in the top row does contain a real open-set example (i.e., the street-shop) in terms of not only size, but also novel content.

of size 512x7x7 from ResNet18 [28]. To reduce feature dimension, we simply (max or average) pool the feature activations spatially into a 512-dim feature vectors [63]. We can further reduce dimension by applying PCA, which can reduce dimensionality by 10 $\times$  (from 512-dimensional to 50 dimensional) without sacrificing performance. We find this dimensionality particularly important for learning second-order covariance statistics as in GMM, described below. Finally, following [24, 26], we find it crucial to L2-normalize extracted features (Fig. 13). We refer the reader to [33] for quantitative results. Note in this paper, we do not L2-normalize features for training OpenGANs.

**Statistical models.** Given the above extracted features, we use various generative statistical methods to learn the confidence/probability that a test example belongs to the closed-set classes. Such statistical methods include simple

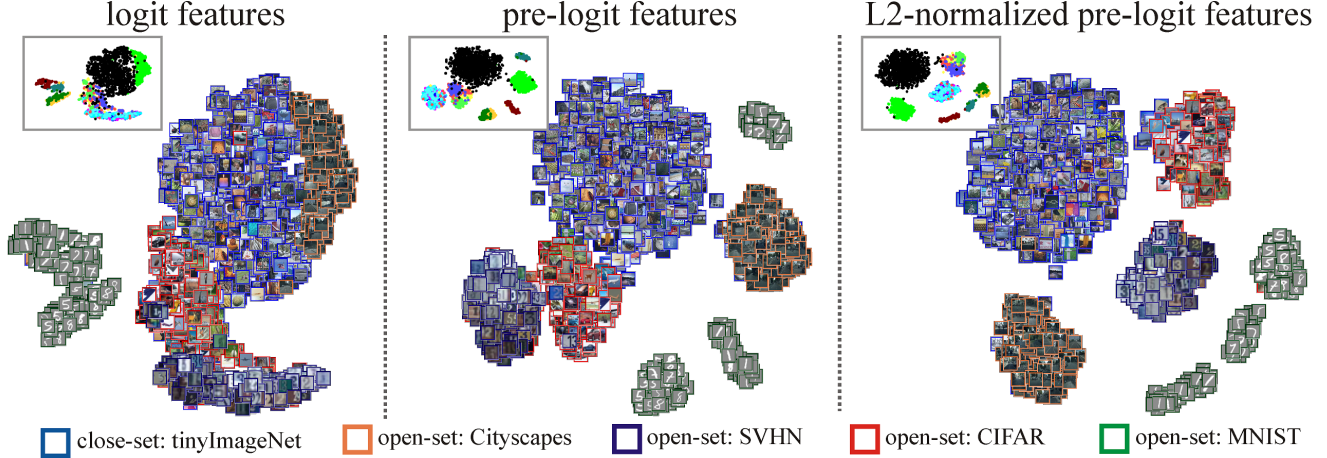


Figure 13: t-SNE plots [39] of open vs closed-set testing data, as encoded by different features from a ResNet18 network trained from scratch on the TinyImageNet dataset for 200-way classification. To better view the clustering results, we show zoom-out scatter plots in which we show closed-set data in black, and color open-set examples using their class labels provided by the respective datasets. **Left:** Logit features mix open and closed data, suggesting that methods based on them (Entropy, SoftMax and OpenMax) may struggle in open-set discrimination. **Mid:** Pre-logit features at the penultimate layer show better separation between closed- and open-set data. **Right:** Normalizing the pre-logits features separates them even better. These plots intuitively demonstrate the benefit of L2-normalization and using OTS features rather than the highly-invariant logits.

parametric models such as class centroids [40] and class-conditional Gaussian models [35, 27], non-parametric models such as NN [8, 32], and mixture models such as (class-conditional) GMMs and k-means [11]. A statistical model labels a test example as open-set when the inverse probability (e.g., of the most-likely class-conditional GMM) or distance (e.g., to the closest class centroid) is above a threshold. One benefit of such simple statistical models is that they are interpretable and relatively easier to diagnose failures. For example, one failure mode is an open-set sample being misclassified as a closed-set class. This happens when open-set data lie close to a class-centroid or Gaussian component mean (see Fig. 13). Note that a single statistical model may have several hyperparameters – GMM can have multiple Gaussian components and different structures of second-order covariance, e.g., either a single scalar, a diagonal matrix or a general covariance per component. We make use of validation set to determine these hyperparameters, as opposed to prior works that conduct model selection either unrealistically on the test-set [44] or on large-scale val-set which could be arguably used for training [35]. We refer the reader to [33] for detailed analysis.

**Lightweight Pipeline.** We re-iterate that the above feature extraction and statistical models result in a lightweight pipeline for open-set recognition. To understand this, we analyze the number of parameters involved in the pipeline. Assume we learn a GMM over  $512 \times 7 \times 7$  feature activations, and specify a general covariance and five Gaussian components. If we learn the GMM directly on the feature activations, the number of parameters from the second-order

covariance alone is at the scale of  $(512 * 7 * 7)^2$ . With the help of our feature extraction (including spatial pooling and PCA), we have 50-dim feature vectors, and the number of parameters in the covariance matrices is now at the scale of  $50^2$ . This means a huge reduction ( $10^5 \times$ ) in space usage! We count the total number of parameters in this GMM:  $3.3 \times 10^4$  32-bit float parameters including PCA and GMM’s five components, amounting to 128KB storage space. Moreover, given that PCA just runs once for all classes, even when we learn such GMMs for each of 19 classes (such as defined in Cityscapes), it merely requires 594KB storage space! Compared to the modern networks such as HRNet ( $>250\text{MB}$ ), our statistical pipeline for open-set recognition adds a negligible (0.2%) amount of compute, making it quite practical for implementation on autonomy stacks.

## 11. Further Quantitative Results

While in the main paper we compare OpenGAN to many methods and cannot include more due to space issues, we list a few more in this appendix including Entropy [58], GMM [33], CGDL [59], OpenHybrid [67], and RPL++ [13]. Except for Entropy which is a classic method, the rest were published recently. Table 7 lists the comparisons under *Setup-I*. Please refer to Section 4.2 of the main paper for the detailed setup. Numbers are comparable to Table 1 in the main paper. In summary, our OpenGAN outperforms all these prior methods under this setup, achieving the state-of-the-art.

Table 7: **Open-set discrimination (Setup-I)** measured by area under ROC curve (AUROC) $\uparrow$ . Numbers are comparable to Table 1 in the main paper. Recall that OpenGAN-0 does not train on outlier data (i.e.,  $\lambda_0=0$  in Eq. 2) and only selects discriminator checkpoints on the validation set. OpenGAN-0<sup>fea</sup> clearly performs the best, achieving the state-of-the-art.

	MSP Entropy		OpenMax	MSP <sub>c</sub>	GOpenMax	OSRCI	MCdrop	GDM	GMM	C2AE	CGDL	RPL-WRN	OpenHybrid	OpenGAN-0 <sup>fea</sup>
Dataset	[30]	[58]	[6]	[36]	[21]	[42]	[20]	[35]	[33]	[44]	[59]	[13]	[67]	(ours)
MNIST	.977	.988	.981	.985	.984	.988	.984	.989	.993	.989	.994	.996	.995	<b>.999</b>
SVHN	.886	.895	.894	.891	.896	.910	.884	.866	.914	.922	.935	.968	.947	<b>.988</b>
CIFAR	.757	.788	.811	.808	.675	.699	.732	.752	.817	.895	.903	.901	.950	<b>.973</b>

## 12. Visualization of Generated Images

In this section, we visualize some synthesized examples for intuitive demonstration.

**Generating Small Images.** Recall that OpenGAN-0<sup>pix</sup> trains a normal GAN and uses its discriminator as the open-set likelihood function. As demonstrated in the main paper, OpenGAN-0<sup>pix</sup> performs surprisingly well under *Setup-I* (i.e., using CIFAR10, MNIST and SVHN datasets) and *Setup-II* (using TinyImageNet as the closed-set and other datasets as the open-set). OpenGAN-0<sup>pix</sup> also enables us to generate visual results for intuitive inspection. In Figure 14, we display real and synthesized “fake” images under *Setup-I* on each of the three datasets. In Figure 15, we display real and fake images under *Setup-II* by using tinyImageNet as the closed-set and other datasets as the open-set. We can see the generated images look realistic in terms of color and tone. But they are not strictly open-set images as they contain synthesized *known* contents (e.g., the digits in the synthesized images are closed-set digits). This intuitively demonstrates that a perfectly trained discriminator will not be capable of discriminating open and closed-sets due to the nature of the min-max game in training GANs. However, from the low confidence scores of classifying the generated fake data as closed-set shown in Figure 14, we can see the discriminator almost naively recognizes these synthesized examples as “fake” data. This shows the synthesized data are insufficient to be used for model selection. Moreover, from the classification confidence scores on the closed-testing and open-testing images in each datasets, we can see the discriminator is not calibrated. In other words, we cannot naively set threshold as 0.5 for open-vs-closed classification. This is largely hidden by AUROC metric which is calibration-free. This implies a potential limitation and suggests future work to calibrate the open-set discriminators.

**Generating Cityscapes Patches.** In the main paper (Fig. 6), we have shown some generated patches. In the appendix, we provide more in Figure 16. As OpenGAN-0<sup>fea</sup> generates features instead of pixel patches, we “synthesize” the patches analytically – for a generated feature, from training pixels represented as OTS features, we find the nearest-neighbor pixel feature (w.r.t L1 distance), and use the RGB patch centered at that pixel as the “synthesized” patch. We can see OpenGAN-0<sup>pix</sup> synthesizes re-

alistic patches w.r.t color and tone, but it (0.549 AUROC) significantly underperforms OpenGAN-0<sup>fea</sup> (0.709 AUROC) for open-set segmentation. The “synthesized” patches by OpenGAN-0<sup>fea</sup> capture many *open-set* objects, such as bridges, vehicle logo and back of traffic sign, all of which are outside the 19 classes defined in Cityscapes. This intuitively explains why OpenGAN-0<sup>fea</sup> (0.709 AUROC) works much better than OpenGAN-0<sup>pix</sup> (0.549 AUROC).

## 13. Visual Results of Open-Set Segmentation

On the task of open-set semantic segmentation, first, we show in Figure 12 that our testing set contains real open-set examples never-before-seen in training; please refer to the caption for details. Then, we show more visual results in figures from 17 through 23. From these figures, we can see OpenGAN<sup>fea</sup> captures most open-set pixels, outperforming the other methods notably.

## 14. Failure Cases and Limitations

As we use an discriminator as the open-set likelihood function, straightforwardly, failure cases happen when the classification is not correct, as shown by the marked confidence scores in Figure 14, as well as the thresholded per-pixel predictions in figures from 17 through 23.

Hereby we point out other failure cases and limitations. First, as we have explained in the main paper, the GAN-discriminator will eventually become incapable of discriminating closed-set and fake/open-set images due to the nature of GANs that strikes an equilibrium between the discriminator and generator. Although we empirically show superior performance by model selection on a validation set, there surely exists risks that the validation set is biased in an unknown way which could catastrophically hurt the final open-set recognition performance. This is also true even with outlier training examples. Therefore, in the real open-world practitioners should be aware of such a bias, and exploit prior knowledge in constructing “reliable” training and validation sets in training OpenGANs. Second, as we adopt adversarial training for OpenGANs, it is straightforward to ask if OpenGAN is robust to adversarial perturbations on the input images. We have not investigated this point yet, and we leave it as future work.



Figure 14: **Demonstration of visuals along with the classification confidence scores as probabilities of being recognized as closed-set data.** On each of the three datasets, we show some random images that are closed-training image (for training GANs), synthesized “fake” images, closed-testing images (from known classes) and open-testing images (from unknown classes). We also mark the probability for each image of being classified as closed-set by the discriminator. We can see the synthesized images look realistic in terms of color, tone and shape. But the discriminator can easily recognize these fake images (as indicated by the low probability). Moreover, although the discriminator achieves good open-vs-closed classification performance measured by AUROC (which is calibration-free), the confidence scores (probability) are not calibrated well. This implies that the discriminator may need to be calibrated for real-world application.

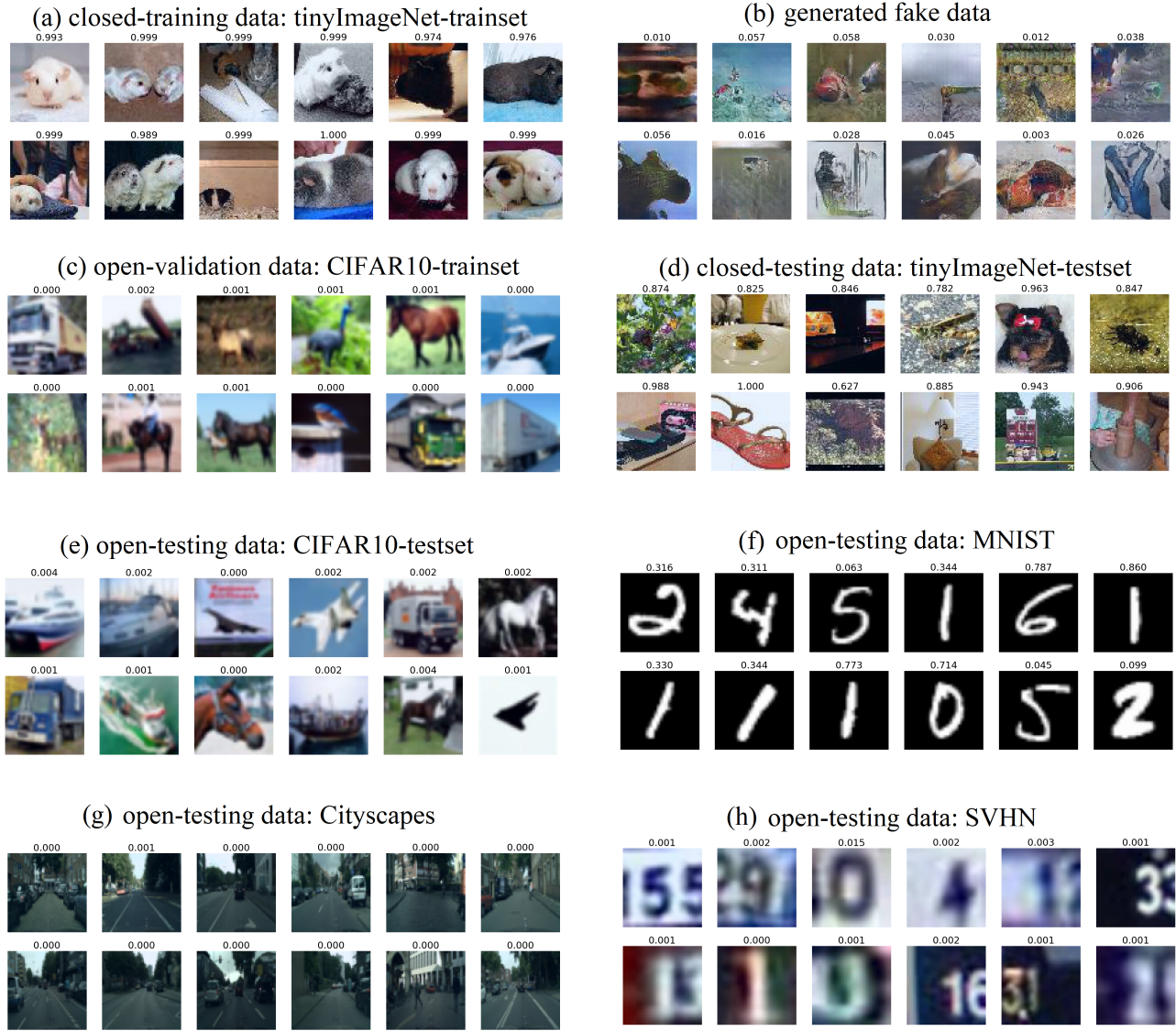


Figure 15: **Demonstration of visuals along with the OpenGAN-0<sup>pix</sup> classification confidence scores as probabilities of being recognized as closed-set data.** These visual results are generated under *Setup-II*, where the TinyImageNet is the closed-set for 200-way classification, and other datasets are treated as the open-set. The discriminator of the OpenGAN-0<sup>pix</sup> is selected over the CIFAR train-set. (a) The discriminator recognizes the closed-set training examples with a high confidence score. (b) OpenGAN-0<sup>pix</sup> synthesizes fake images that look realistic in terms of color, tone and shape, but not content. The discriminator can easily recognize these fake images (as indicated by the low probability). The discriminator generalizes well in terms of recognizing closed-set examples from the validation and test sets as shown in (c) and (d), and open-set examples from other datasets as shown in (e), (f), and (h).



Figure 16: **Visuals of real Cityscapes image patches (left), synthesized patches by OpenGAN-0<sup>pix</sup> (mid) and OpenGAN-0<sup>fea</sup> (right).** As OpenGAN-0<sup>fea</sup> generates feature vectors instead of RGB patches, we “synthesize” the patches in an analytical way – for a generated feature, we find the nearest-neighbor per-pixel feature (w.r.t L1 distance) from the training images, and then find the RGB patch centered at the associated pixel with the per-pixel feature. The real patch is our “synthesized” patch for that generated feature. The synthesized patches by OpenGAN-0<sup>pix</sup> do look realistic in terms of color and tone, but OpenGAN-0<sup>pix</sup> (0.549 AUROC) does not work as well as OpenGAN-0<sup>fea</sup> (0.709 AUROC). The “synthesized” patches by OpenGAN-0<sup>fea</sup> do capture some *unknown open-set* objects, such as bridge, back of traffic sign and unknown static objects, none of which belong to any of the 19 classes defined in Cityscapes for semantic segmentation (cf. Figure 7).

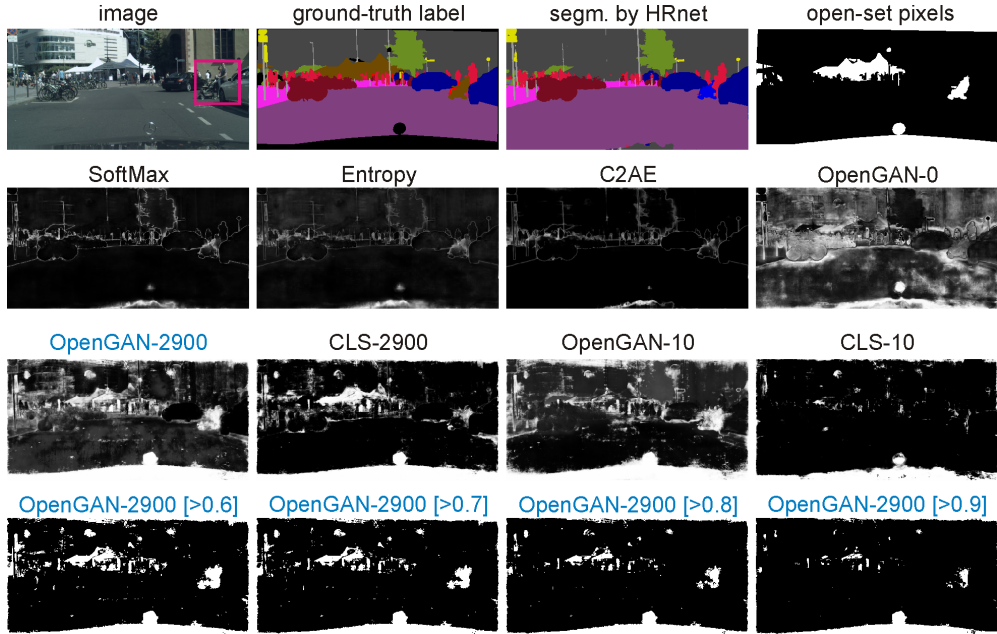


Figure 17: **Qualitative results of a testing image from Cityscapes.** [1<sup>st</sup> row] the input image, its per-pixel semantic labels, the semantic segmentation result by HRnet and open-set pixels colored by white. [2<sup>nd</sup> row] visual results as per-pixel scores of being classified as open-set pixel by SoftMax, Entropy, C2AE and our OpenGAN-0<sup>fea</sup>. [3<sup>rd</sup> row] visual results by our OpenGAN<sup>fea</sup> and CLS, trained with 2900 or 10 open training images, respectively. [4<sup>th</sup> row] visual results by thresholding OpenGAN-2900 with 0.6, 0.7, 0.8 and 0.9 respectively. OpenGAN clearly captures most open-set pixels (cf. the white pixels in top-right open-set map).

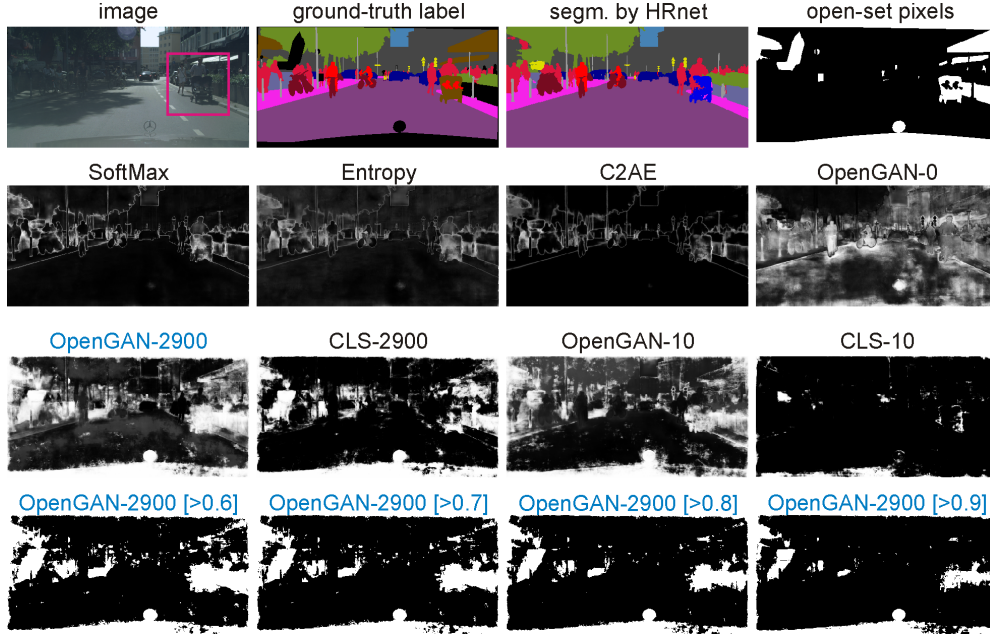


Figure 18: **Qualitative results of a testing image from Cityscapes.** [ $1^{st}$  row] the input image, its per-pixel semantic labels, the semantic segmentation result by HRnet and open-set pixels colored by white. [ $2^{nd}$  row] visual results as per-pixel scores of being classified as open-set pixel by SoftMax, Entropy, C2AE and our OpenGAN-0<sup>fea</sup>. [ $3^{rd}$  row] visual results by our OpenGAN<sup>fea</sup> and CLS, trained with 2900 or 10 open training images, respectively. [ $4^{th}$  row] visual results by thresholding OpenGAN-2900 with 0.6, 0.7, 0.8 and 0.9 respectively. OpenGAN clearly captures most open-set pixels (cf. the white pixels in top-right open-set map).

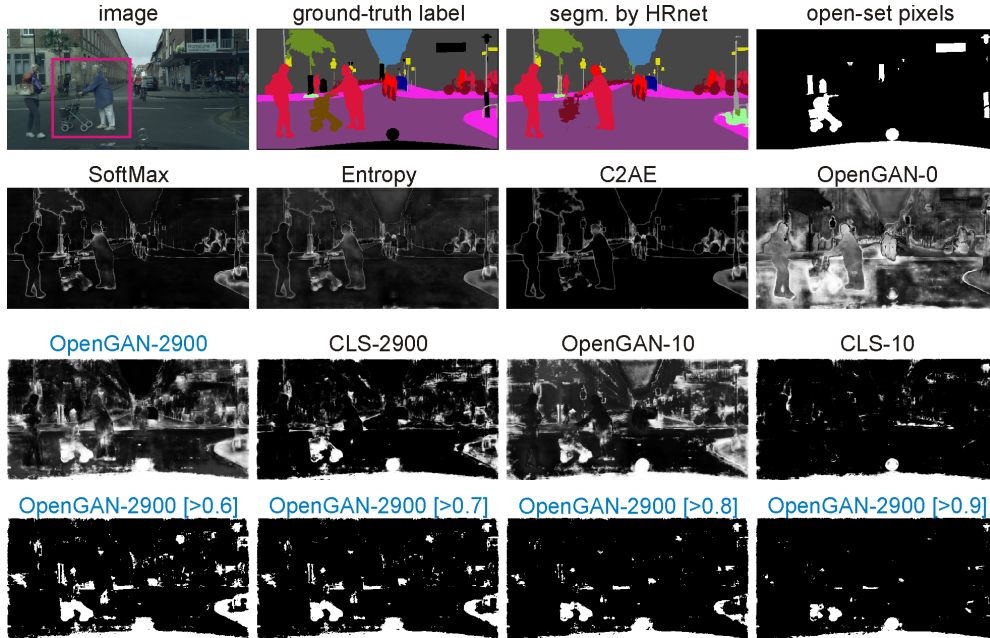


Figure 19: **Qualitative results of a testing image from Cityscapes.** [ $1^{st}$  row] the input image, its per-pixel semantic labels, the semantic segmentation result by HRnet and open-set pixels colored by white. [ $2^{nd}$  row] visual results as per-pixel scores of being classified as open-set pixel by SoftMax, Entropy, C2AE and our OpenGAN-0<sup>fea</sup>. [ $3^{rd}$  row] visual results by our OpenGAN<sup>fea</sup> and CLS, trained with 2900 or 10 open training images, respectively. [ $4^{th}$  row] visual results by thresholding OpenGAN-2900 with 0.6, 0.7, 0.8 and 0.9 respectively. OpenGAN clearly captures most open-set pixels (cf. the white pixels in top-right open-set map).

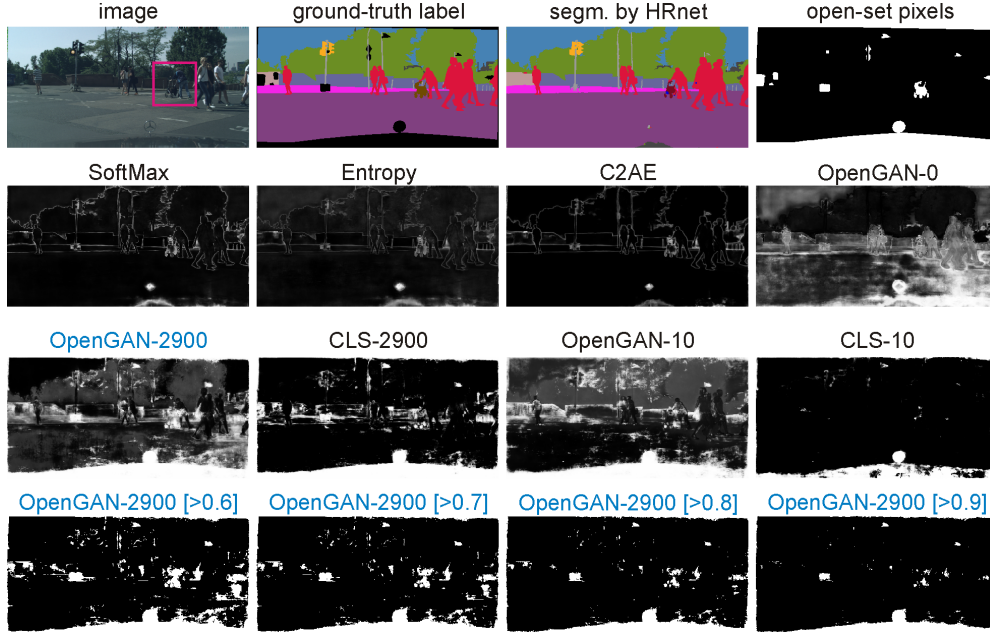


Figure 20: **Qualitative results of a testing image from Cityscapes.** [1<sup>st</sup> row] the input image, its per-pixel semantic labels, the semantic segmentation result by HRnet and open-set pixels colored by white. [2<sup>nd</sup> row] visual results as per-pixel scores of being classified as open-set pixel by SoftMax, Entropy, C2AE and our OpenGAN-0<sup>fea</sup>. [3<sup>rd</sup> row] visual results by our OpenGAN<sup>fea</sup> and CLS, trained with 2900 or 10 open training images, respectively. [4<sup>th</sup> row] visual results by thresholding OpenGAN-2900 with 0.6, 0.7, 0.8 and 0.9 respectively. OpenGAN clearly captures most open-set pixels (cf. the white pixels in top-right open-set map).

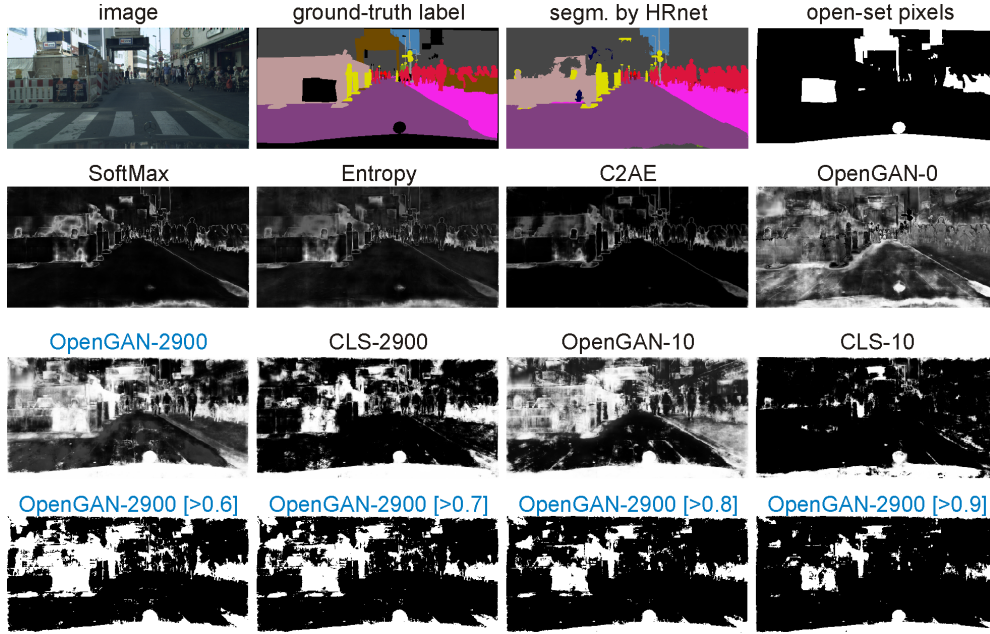


Figure 21: **Qualitative results of a testing image from Cityscapes.** [1<sup>st</sup> row] the input image, its per-pixel semantic labels, the semantic segmentation result by HRnet and open-set pixels colored by white. [2<sup>nd</sup> row] visual results as per-pixel scores of being classified as open-set pixel by SoftMax, Entropy, C2AE and our OpenGAN-0<sup>fea</sup>. [3<sup>rd</sup> row] visual results by our OpenGAN<sup>fea</sup> and CLS, trained with 2900 or 10 open training images, respectively. [4<sup>th</sup> row] visual results by thresholding OpenGAN-2900 with 0.6, 0.7, 0.8 and 0.9 respectively. OpenGAN clearly captures most open-set pixels (cf. the white pixels in top-right open-set map).

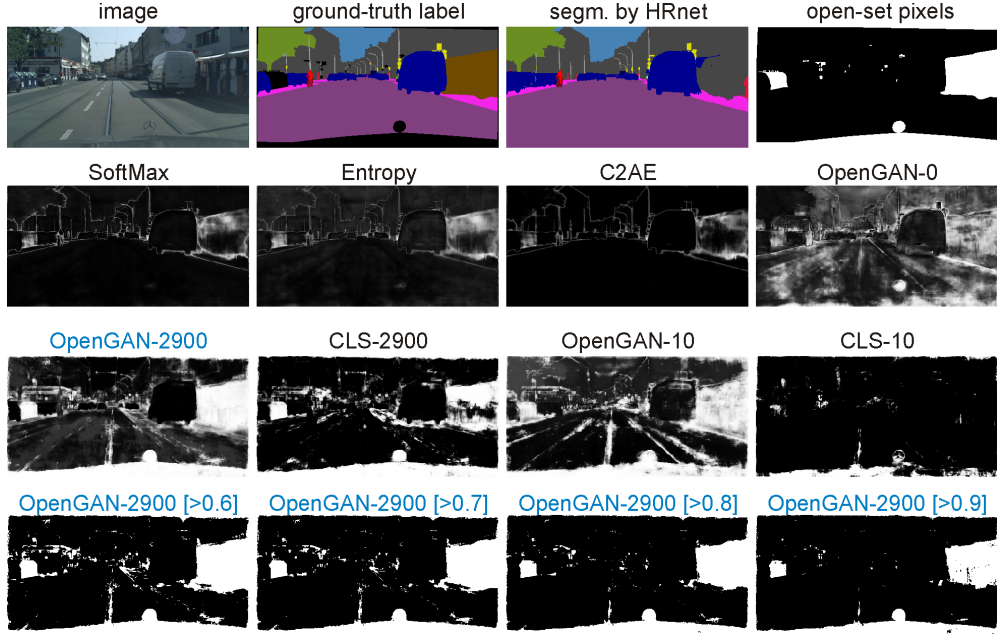


Figure 22: **Qualitative results of a testing image from Cityscapes.** [1<sup>st</sup> row] the input image, its per-pixel semantic labels, the semantic segmentation result by HRnet and open-set pixels colored by white. [2<sup>nd</sup> row] visual results as per-pixel scores of being classified as open-set pixel by SoftMax, Entropy, C2AE and our OpenGAN-0<sup>fea</sup>. [3<sup>rd</sup> row] visual results by our OpenGAN<sup>fea</sup> and CLS, trained with 2900 or 10 open training images, respectively. [4<sup>th</sup> row] visual results by thresholding OpenGAN-2900 with 0.6, 0.7, 0.8 and 0.9 respectively. OpenGAN clearly captures most open-set pixels (cf. the white pixels in top-right open-set map).

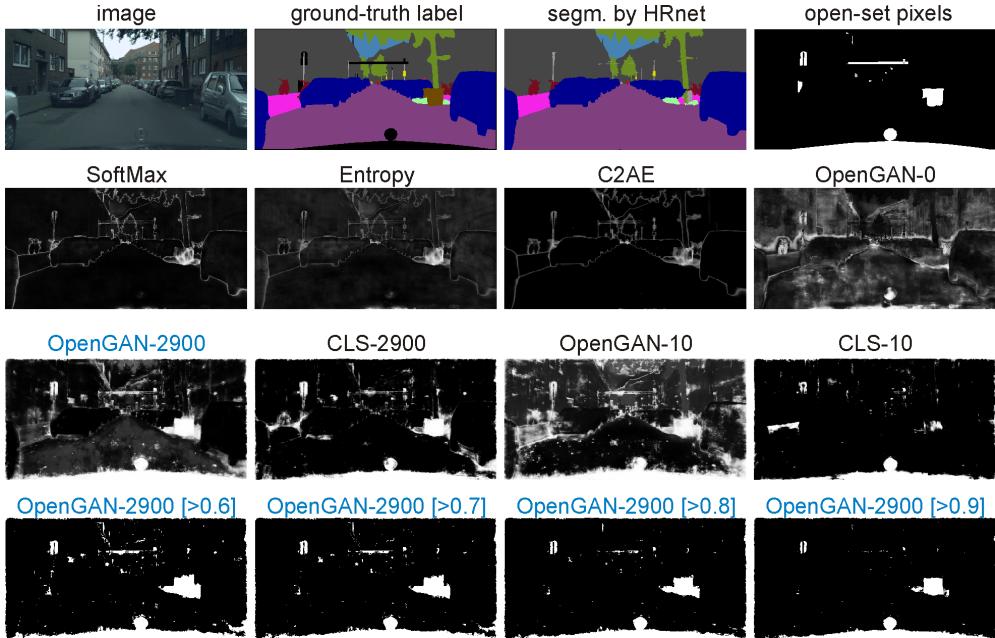


Figure 23: **Qualitative results of a testing image from Cityscapes.** [1<sup>st</sup> row] the input image, its per-pixel semantic labels, the semantic segmentation result by HRnet and open-set pixels colored by white. [2<sup>nd</sup> row] visual results as per-pixel scores of being classified as open-set pixel by SoftMax, Entropy, C2AE and our OpenGAN-0<sup>fea</sup>. [3<sup>rd</sup> row] visual results by our OpenGAN<sup>fea</sup> and CLS, trained with 2900 or 10 open training images, respectively. [4<sup>th</sup> row] visual results by thresholding OpenGAN-2900 with 0.6, 0.7, 0.8 and 0.9 respectively. OpenGAN clearly captures most open-set pixels (cf. the white pixels in top-right open-set map).