Synthesize then Compare: Detecting Failures and Anomalies for Semantic Segmentation

Yingda Xia*, Yi Zhang*, Fengze Liu, Wei Shen(✉), Alan L. Yuille

Johns Hopkins University

Abstract. The ability to detect failures and anomalies are fundamental requirements for building reliable systems for computer vision applications, especially safety-critical applications of semantic segmentation, such as autonomous driving and medical image analysis. In this paper, we systematically study failure and anomaly detection for semantic segmentation and propose a unified framework, consisting of two modules, to address these two related problems. The first module is an image synthesis module, which generates a synthesized image from a segmentation layout map, and the second is a comparison module, which computes the difference between the synthesized image and the input image. We validate our framework on three challenging datasets and improve the state-of-the-arts by large margins, *i.e.*, 6% AUPR-Error on Cityscapes, 7% Pearson correlation on pancreatic tumor segmentation in MSD and 20% AUPR on StreetHazards anomaly segmentation.

Keywords: failure detection, anomaly segmentation, semantic segmentation

1 Introduction

Deep neural networks [15, 19, 31, 45, 47] have achieved great success in various computer vision tasks. However, when they come to real world applications, such as autonomous driving [22], medical diagnoses [5] and nuclear power plant monitoring [36], the safety issue [1] raises tremendous concerns particularly in conditions where failure cases have severe consequences. As a result, it is of enormous value that a machine learning system is capable of detecting the failures, *i.e.*, wrong predictions, as well as identifying the anomalies, *i.e.*, out-of-distribution (OOD) cases, that may cause these failures.

Previous works on failure detection [7, 17, 24] and anomaly (OOD) detection [9, 18, 33–35] mainly focus on classifying small images. Although failure detection and anomaly detection for semantic segmentation have received little attention in the literature so far, they are more closely related to safety-critical applications, *e.g.*, autonomous driving and medical image analysis. The objective of failure detection for semantic segmentation is not only to determine whether there are failures in a segmentation result, but also to locate where the failures

* The first two authors equally contributed to the work. ✉ Corresponding Author.

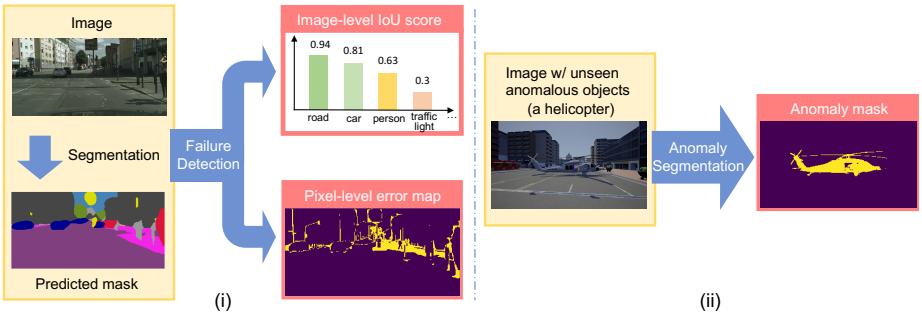

Fig. 1. We aim at addressing two tasks: (i) failure detection, *i.e.*, image-level per-class IoU prediction (top left) and pixel-level error map prediction (bottom left) (ii) anomaly segmentation i.e. segmenting anomalous objects (right middle).

are. Anomaly detection for semantic segmentation, *a.k.a* anomaly segmentation, is related to failure detection, and its objective is to segment anomalous objects or regions in a given image.

In this paper, our goal is to build a reliable alarm system to address failure detection for semantic segmentation (Fig. 1(i)) and anomaly segmentation (Fig. 1(ii)). Unlike image classification outputs only a single image label, semantic segmentation outputs a structured semantic layout. Thus, this requires that the system should be able to provide more detailed analysis than those for image classification, *i.e.*, pixel-level error/confidence maps. Some previous works [7, 16, 26] directly applied the failure/anomaly detection strategies for image classification pixel by pixel to estimate a pixel-level error map, but they lack the consideration of the structured semantic layout of a segmentation result.

We propose a unified framework to address failure detection and anomaly detection for semantic segmentation. This framework consists of two components: an image synthesis module, which synthesizes an image from a segmentation result to reconstruct its input image, *i.e.*, a reverse procedure of semantic segmentation, and a comparison module which computes the difference between the reconstructed image and the input image. Our framework is motivated by the fact that the quality of semantic image synthesis [21, 43, 48] can be evaluated by the performance of segmentation network. Presumably the converse is also true, the better is the segmentation result, the closer a synthesized image generated from the segmentation result is to the input image. If a failure occurs during segmentation, for example, if a person is mis-segmented as a pole, the synthesized image generated from the segmentation result does not look like a person and an obvious difference between the synthesized image and the input image should occur. Similarly, when an anomalous (OOD) object occurs in a test image, it would be classified as any possible in-distribution objects in a segmentation result, and then appear as in-distribution objects in the synthesized image generated from the segmentation result. Consequently, the anomalous object can be identified

by finding the differences between the test image and the synthesized image. We refer to our framework as SynthCP, for "synthesize then compare".

We model this synthesis procedure by a semantic-to-image conditional GAN (cGAN) [43], which is capable of modeling the mapping from the segmentation layout space to the image space. This cGAN is trained on label-image pairs. Given the segmentation result of an input image obtained by an semantic segmentation model, we apply the trained cGAN to the segmentation result to generate a reconstructed image. Then, the reconstructed image and the input image are fed into the comparison module to identify the failures/anomalies. The comparison module is designed task-specifically: For failure detection, the comparison module is modeled by a Siamese network, outputting both image-level confidences and pixel-level confidences; For anomaly segmentation, the comparison module is realized by computing the distance defined on the intermediate features extracted by the semantic segmentation model.

We validate SynthCP on the Cityscapes street scene dataset, a pancreatic tumor segmentation dataset in the Medical Segmentation Decathlon (MSD) challenge and the StreetHazards dataset, and show its superiority to other failure detection and anomaly segmentation methods. Specifically, we achieved improvements over the state-of-the-arts by approximately 6% AUPR-Error on Cityscapes pixel-level error prediction, 7% Pearson correlation on pancreatic tumor DSC prediction and 20% AUPR on StreetHazards anomaly segmentation.

We summarize our contribution as follows:

- To the best of our knowledge, we are the first to systematically study failure detection and anomaly detection for semantic segmentation
- We propose a unified framework, SynthCP, which enjoys the benefits of a semantic-to-image conditional GAN, to address both of the two tasks.
- SynthCP achieves state-of-the-art failure detection and anomaly segmentation results on three challenging datasets.

2 Related Work

In this section, we first review the topics closely related to failure detection and anomaly segmentation, such as uncertainty/confidence estimation, quality assessment and out-of-distribution (OOD) detection. Then, we review generative adversarial networks (GANs), which serves a key module in our framework.

Uncertainty estimation or confidence estimation has been a hot topic in the field of machine learning for years, and can be directly applied to the task of **failure detection**. Standard baselines was established in [17] for detecting failures in classification where maximum softmax probability (MSP) provides reasonable results. However, the main drawback of using MSP for confidence estimation is that deep networks tend to produce high confidence predictions [7]. Geifman *et al.* [12] controled the user specified risk-level by setting up thresholds on a pre-defined confidence function (e.g. MSP). Jiang *et al.* [24] measured the agreement between the classifier and a modified nearest-neighbor classifier on the test examples as a confidence score. A recent approach [7] proposed to direct

regress "true class probability" which improved over MSP for failure detection. Additionally, Bayesian approaches have drawn attention in this field of study. Dropout based approaches [10, 26] used Monte Carlo Dropout (MCDropout) for Bayesian approximation. Computing statistics such as entropy or variance is capable of indicating uncertainty. However, all these approaches mainly focus on small image classification tasks. When applied to semantic segmentation, they lack the information of semantic structures and contexts.

Segmentation quality assessment aims at estimating the overall quality of segmentation, without using ground-truth label, which is suitable to make alarms when model fails. Some approaches [25,32] utilize Bayesian CNNs to predict the segmentation quality of medical images. [29,44] regressed the segmentation quality from deep features computed from a pair of an image and its segmentation result. [20, 23] plugged an extra IoU regression head into object detection or instance segmentation. [4, 11] used unsupervised learning methods to estimate the segmentation quality using geometrical features. Recently, Liu *et al.* [38] proposed to use VAE [28] to capture a shape prior for segmentation quality assessment on 3D medical image. However, it is hardly applicable to natural images considering the complexity and large shape variance in 2D scenes and objects. Segmentation quality assessment will be referred to as image-level failure detection in the rest of the paper.

OOD detection aims at detecting out-of-distribution examples in testing data. Since the baseline MSP method [17] was brought up, many approaches have improved OOD detection from various aspects [9,18,33–35]. While these approaches mainly focus on image level OOD detection, *i.e.*, to determine whether an image is an OOD example, *e.g.*, [3, 30] targeted at detecting hazardous scenes in the Wilddash dataset [49]. On the contrary, we focus on **anomaly segmentation**, *i.e.*, a pixel-level OOD detection task that aims at segmenting anomalous regions from an image. Pixel-wise reconstruction loss [2, 14] with auto-encoders(AE) are the main stream approaches for anomaly segmentation. However, they can hardly model the complex street scenes in natural images and AEs can not guarantee to generate an in-distribution image from OOD regions. Recently, it was found that MSP surprisingly outperform AE and Bayesian network based approaches on a newly built larger scale street scene dataset StreetHazards [16] - with 250 types of anomalous objects and more than 6k high resolution images. Lis *et al.* [37] proposed to re-synthesize an image from the predicted semantic map to detect OOD objects in street scenes, which is the **pioneer** work for synthesis-based anomaly detection for semantic segmentation. SynthCP also follows this spirit, but we use a simple yet effective feature distance measure rather than a discrepancy network to find anomalies. In addition, we extend this idea to do a systematic study of failure detection.

Generative adversarial networks [13] generate realistic images by playing a "min-max" game between a generator and a discriminator. GANs effectively minimize a Jensen-Shannon divergence, thus generating in-distribution images. SynthCP utilizes conditional GANs [42] (cGANs) for image translation [21], *a.k.a* pixel-to-pixel translation. Approaches designed for semantic image syn-

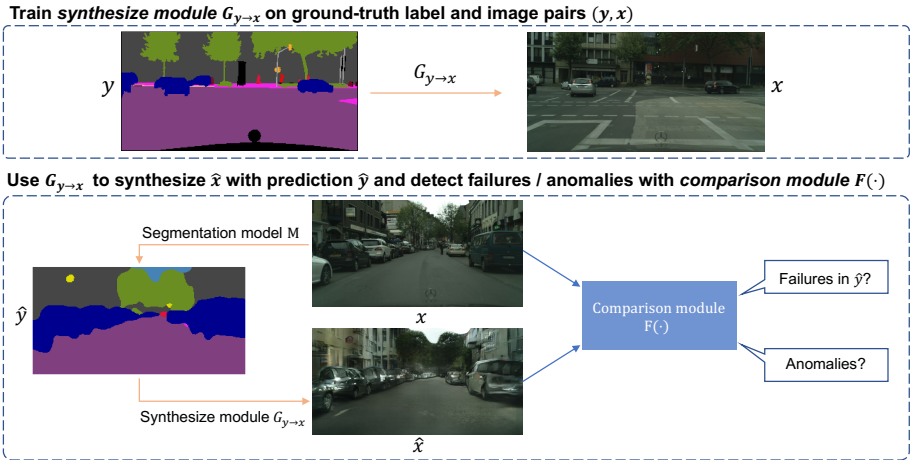

Fig. 2. We first train the synthesis module $G_{y\to x}$ on label-image pairs and then use this module to synthesize the image conditioning on the predicted segmentation mask $\hat{y}$. By comparing x and $\hat{x}$ with a comparison module $F(\cdot)$, we can detect failures as well as segment anomalous objects. $F(\cdot)$ is instantiated in Sec 3.2 and Sec 3.3.

thesis [40, 43, 48] improves pixel-to-pixel translation in synthesizing real images from semantic masks, which is the reverse procedure of semantic segmentation. Since semantic image synthesis is commonly evaluated by the performance of a segmentation model, reversely, we are motivated to use a semantic-to-image generator for failure detection for semantic segmentation.

3 Methodology

In this section, we introduce our framework, SynthCP, for failure detection and anomaly detection for semantic segmentation. SynthCP consists of two modules, an image synthesis module and a comparison module. We first introduce the general framework (shown in Fig. 2), then describe the details of the modules for failure detection and anomaly detection in Sec 3.2 and Sec 3.3, respectively. Unless otherwise specified, the notations in this paper follow this criterion: We use a lowercase letter, *e.g.*, x, to represent a tensor variable, such as a 1D array or a 2D map, and denote its i-th element as $x^{(i)}$; We use a capital letter, *e.g.*, F, to represent a function.

3.1 General Framework

Let x be an image with size of $w \times h$ and $\mathbb{L} = \{1, 2, \ldots, L\}$ be a set of integers representing the semantic labels. By feeding image x to a segmentation model M, we obtain its segmentation result, *i.e.*, a pixel-wise semantic label map $\hat{y} =$

$M(x) \in \mathbb{L}^{w \times h}$. Our goal is to identify and locate the failures in $\hat{y}$ or detect anomalies in x based on $\hat{y}$.

Image Synthesis Module We model this image synthesize module by a pixel-to-pixel translation conditional GAN (cGAN) [42], which is known for its excellent ability for semantic-to-image mapping. It consists of a generator G and a discriminator D.

Training. We train this translation conditional GAN on label-image pairs: (y, x), where y is a grouth-truth pixel-wise semantic label map and x is its corresponding image. The objective of the generator G is to translate semantic label maps to realistic-looking images, while the discriminator D aims to distinguish real images from the synthesized ones. This cGAN minimizes the conditional distribution of real images via the following min-max game:

$$\min_G \max_D \mathcal{L}_{GAN}(G, D), \tag{1}$$

where the objective function $\mathcal{L}_{GAN}(G, D)$ is defined as:

$$\mathbb{E}_{(y,x)}[\log D(y, x)] + \mathbb{E}_y[\log(1 - D(y, G(y))]. \tag{2}$$

Testing. After training, we fix the generator G. Given an image x and a segmentation model M, we feed the predicted segmentation mask $\hat{y} = M(x)$ into G, and obtain a synthesized (*i.e.*, reconstructed) image $\hat{x}$:

$$\hat{x} = G(\hat{y}). \tag{3}$$

$\hat{x}$ and x are then served as the input for the comparison module.

Comparison Module We detect failures and anomalies in $\hat{y}$ by comparing $\hat{x}$ with x. Our assumption is that, if $\hat{x}$ is more similar to x, then $\hat{y}$ is more similar to y. However, since the optimization of G does not guarantee that the synthesized image $\hat{x}$ has the same style as the original image x, simple similarity measurements such as ℓ_1 distance between x and $\hat{x}$ is not accurate. In order to address this issue, we model the comparison module by a task-specific function F which estimates a trustworthy task-specific confidence measure $\hat{c}$ between x and $\hat{x}$:

$$\hat{c} = F(x, \hat{x}) = F(x, G(\hat{y})). \tag{4}$$

For the task of failure detection, the confidence measure $\hat{c} = (\hat{c}_{iu}, \hat{c}_m)$ includes an image-level per-class intersection over union (IoU) array $\hat{c}_{iu} \in [0, 1]^{|\mathbb{L}|}$ and a pixel-level error map $\hat{c}_m \in [0, 1]^{w \times h}$; For the task of anomaly segmentation, the confidence measure $\hat{c}$ is a pixel-level confidence map $\hat{c}_n \in [0, 1]^{w \times h}$ for anomalous objects.

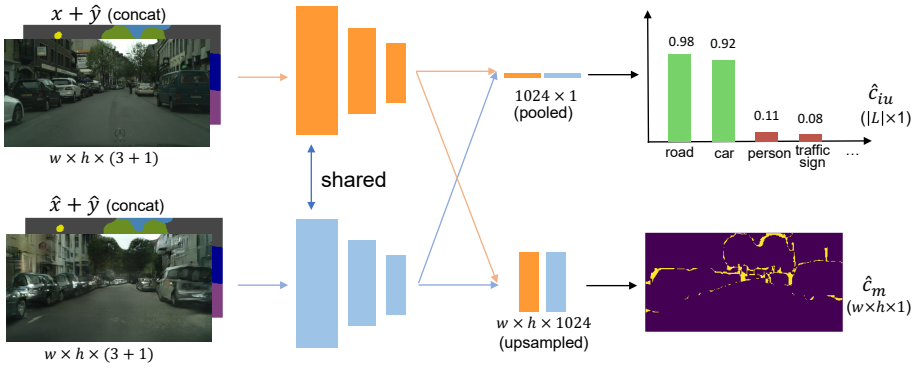

Fig. 3. We instantiate $F(\cdot)$ as a light-weighted siamese network $F(x, \hat{x}, \hat{y}; \theta)$ for joint image-level per-class IoU prediction and pixel-level error map prediction.

3.2 Failure Detection

Problem Definition Our failure detection contains two tasks: 1) a per-class IoU prediction $\hat{c}_{iu} \in [0,1]^{|\mathbb{L}|}$, which is useful to indicate whether there are failures in the segmentation result $\hat{y}$, and 2) to locate the failures in $\hat{y}$, which needs to compute a pixel-level error map $\hat{c}_m \in [0,1]^{w \times h}$.

Instantiation of Comparison Module We instantiate the comparison module $F(\cdot)$ as a light-weighted deep network. In practice, we use ResNet-18 [15] as the base network and follow a siamese-style design for learning the relationship between x and $\hat{x}$. As illustrated in Fig. 3, x and $\hat{x}$ are first concatenated with $\hat{y}$ and then separately encoded by a shared-weight siamese encoder. Then two heads are built upon the siamese encoder and output the image-level per-class IoU array $\hat{c}_{iu} \in [0,1]^{|\mathbb{L}|}$ and pixel-level error map $\hat{c}_m \in [0,1]^{w \times h}$, respectively. We rewrite the function F for failure detection as below:

$$\hat{c}_{iu}, \hat{c}_m = F(x, \hat{x}, \hat{y}; \theta) \tag{5}$$

where θ represents the network parameters.

In the training stage, the supervision of network training is obtained by computing the ground-truth confidence measure c from y and $\hat{y}$. For the ground-truth image-level per-class IoU array c_{iu}, we compute it by

$$c_{iu}^{(l)} = \frac{|\{i|\hat{y}^{(i)} = l\} \cap \{i|y^{(i)} = l\}|}{|\{i|\hat{y}^{(i)} = l\} \cup \{i|y^{(i)} = l\}|}, \tag{6}$$

where l is the l-th semantic class in label set $\mathbb{L}$. The ℓ_1 loss function $\mathcal{L}_{\ell_1}(c_{iu}^{(l)}, \hat{c}_{iu}^{(l)})$ is applied to learning this image-level per-class IoU prediction head. For the

ground-truth pixel-level error map, we compute it by

$$c_m^{(i)} = \begin{cases} 1 & \text{if } y^{(i)} \neq \hat{y}^{(i)} \\ 0 & \text{if } y^{(i)} = \hat{y}^{(i)} \end{cases}. \tag{7}$$

The binary cross-entropy loss $\mathcal{L}_{ce}(c_m^{(i)}, \hat{c}_m^{(i)})$ is applied to learning this pixel-level error map prediction head. The overall loss function of failure detection $\mathcal{L}$ is the sum of the above two:

$$\mathcal{L} = \frac{1}{|\mathbb{L}|} \sum_{l}^{|\mathbb{L}|} \mathcal{L}_{\ell_1}(c_{iu}^{(l)}, \hat{c}_{iu}^{(l)}) + \frac{1}{wh} \sum_{i}^{wh} \mathcal{L}_{ce}(c_m^{(i)}, \hat{c}_m^{(i)}). \tag{8}$$

3.3 Anomaly Segmentation

Problem Definition The goal of anomaly segmentation is segmenting anomalous objects in a test image which are unseen in the training images. Formally, given a test image x, an anomaly segmentation method should output a confidence score map $\hat{c}_n \in [0,1]^{w \times h}$ for the regions of the anomalous objects in the image, *i.e.*, $\hat{c}_n^{(i)} = 1$ and $\hat{c}_n^{(i)} = 0$ indicate the i^{th} pixel belongs to an anomalous object and an in-distribution object (the object is seen in the training images), respectively.

Instantiation of Comparison Module As the same as failure detection, we first train a cGAN generator G on the training images, which maps the in-distribution object labels to realistic images. Given a semantic segmentation model M, we feed its prediction $\hat{y} = M(x)$ into G and obtain $\hat{x} = G(\hat{y})$. Since $\hat{y}$ only contains in-distribution object labels, $\hat{x}$ also only contain in-distribution objects. Thus, we can compare x with $\hat{x}$ to find the anomalies. The pixel-wise semantic difference of x and $\hat{x}$ is a strong indicator of anomalous objects. Here, we simply instantiate the comparison function $F(\cdot)$ as the cosine distance defined on the intermediate features extracted by the segmentation model M:

$$\hat{c}_n^{(i)} = F(x, \hat{x}; M) = 1 - \langle \frac{\mathbf{f}_M^i(x)}{\|\mathbf{f}_M^i(x)\|_2}, \frac{\mathbf{f}_M^i(\hat{x})}{\|\mathbf{f}_M^i(\hat{x})\|_2} \rangle \tag{9}$$

where $\mathbf{f}_M^i$ is the feature vector at the i^{th} pixel position outputted by the last layer of segmentation model M and $\langle \cdot, \cdot \rangle$ is the inner product of the two vectors.
Post-processing with MSP. Due to the artifacts and generalized errors of GANs, our approach may mis-classify an in-distribution object into an anomalous object (false positives). We use a simple post-processing to address this issue. We refine the result by maximum softmax probability (MSP) [17], which is known as an effective uncertainty estimation strategy: $\hat{c}_n^{(i)} \leftarrow \hat{c}_n^{(i)} \cdot \mathbb{1}\{p^{(i)} \leq t\} + (1 - p^{(i)}) \cdot \mathbb{1}\{p^{(i)} > t\}$, where $p^{(i)}$ is the maximum soft-max probability at the i-th pixel outputted by the segmentation model M, $t \in [0,1]$ is a threshold and $\mathbb{1}\{\cdot\}$ is the indicator function.

3.4 Conceptual Explanation

We give conceptual explanations of SynthCP in Fig. 4, where $\mathcal{X}$ and $\mathcal{Y}$ correspond to image space and label space. $M_{x\to y}$ is the segmentation model and $G_{y\to x}$ is a semantic-to-image generator. The left image shows when $M_{x\to y}$ correctly maps an image to its corresponding segmentation mask, the synthesized image generated from $G_{y\to x}$ is close to the original image. However, when $M_{x\to y}$ makes a failure (middle) or encounters an OOD case (right), the synthesized image should be far away from the original image. As a result, the synthesized image serves as a strong indicator for either failure detection or OOD detection.

4 Experiments

4.1 Failure Detection

Evaluation Metrics Following [38], we evaluate the performance of image-level failure detection, *i.e.*, per-class IoU prediction, by four metrics: **MAE**, **STD**, **P.C** and **S.C**. MAE (mean absolute error) and its STD measure the average error between predicted IoUs and ground-truth IoUs. P.C (Pearson correlation) and S.C.(Spearman correlation) measures their correlation coefficients. For pixel-level failure detection, *i.e.*, pixel-level error map prediction, we use the metrics in literature [7,17]: **AUPR-Error**, **AUPR-Success**, **FPR at 95% TPR** and **AUROC**. Following [7], AUPR-Error is our main metric, which computes the area under the Precision-Recall curve using errors as the positive class.

The Cityscapes Dataset We validate SynthCP on the Cityscapes dataset [8], which contains 2975 high-resolution training images and 500 validation images. As far as we know, it is the largest one for failure detection for semantic segmentation.

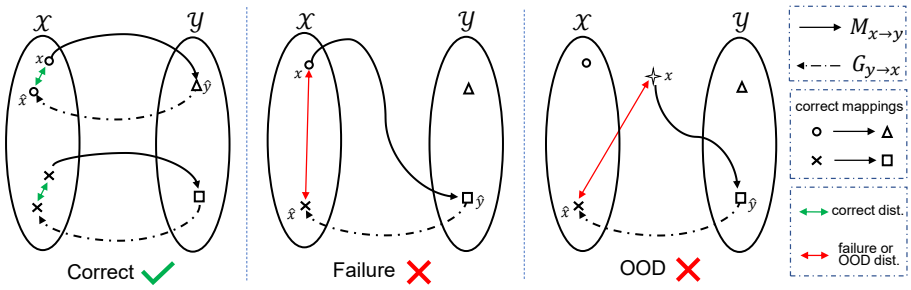

Fig. 4. An analysis of SynthCP. Left: $M_{x\to y}$ correctly maps x to $\hat{y}$, resulting in small distance between x and the synthesized $\hat{x}$. However, when there are failures in $\hat{y}$ (middle) or there are OOD examples in x (right), the distance between x and $\hat{x}$ is larger, given a reliable reverse mapping $G_{y\to x}$.

Baselines. We compare SynthCP to MCDropout [10], VAE alarm [38], MSP [17], TCP [7] and "Direct Prediction". MCDropout, MSP and TCP output pixel-level confidence maps, serving as standard baselines for pixel-level failure prediction. VAE alarm [38] is the state-of-the-art in image-level failure prediction method. Following [38], we also use MCDropout to predict image-level failures. Direct Prediction is a method that directly uses a network to predict both image-level and pixel-level failures, by taking an image and its segmentation result as input. Note that, Direct Prediction shares the same experimental settings (backbone and training strategies) with SynthCP, which can be seen as an ablation study on the effectiveness of the synthesized image $\hat{x}$.

Implementation details. We use the state-of-the-art semantic-to-image cGAN - SPADE [43] in SynthCP. We re-trained SPADE from scratch following the same hyper-parameters as in [43] with only semantic segmentation maps as the input (without the instance maps). The backbone of our comparison module is ResNet-18 [15] pretrained from Image-Net. We use ImageNet pre-trained model and train the network for 20k iterations using Adam optimizer [27] with initial learning 0.01 and $\beta = (0.9, 0.999)$, which takes about 6 hours on one single Nvidia Titan Xp GPU. Since we use a network to predict failures, we need to generate training data for this network. A straightforward strategy is to divide the original training set into a training subset and a validation subset, then train the segmentation model on the training subset and test it on the validation subset. The testing results on the validation subset can be used to train the failure predictor. We extend this strategy by doing 4-fold cross validation on the training set. Since the cross-validated results cover all samples in the original training set, we are able to generate sufficient training data to train our failure prediction network.

Table 1. Experiments on the Cityscapes dataset. We detect failures in the segmentation results of FCN-8 and Deeplab-v2. "SynthCP-separate" and "SynthCP-joint" mean training the image-level and pixel-level failure detection heads in our network separately and jointly, respectively.

	FCN-8				Deeplab-v2			
image-level	MAE↓	STD↓	P.C.↑	S.C.↑	MAE↓	STD↓	P.C.↑	S.C.↑
MCDropout [10]	17.28	13.33	3.62	5.97	19.31	12.86	4.55	1.37
VAE alarm [38]	16.28	11.88	21.82	18.26	16.78	12.21	17.92	19.63
Direct Prediction	13.25	11.96	58.34	59.74	14.45	12.20	60.94	62.01
SynthCP-separate	**11.58**	11.50	**64.63**	**65.63**	**13.60**	12.32	62.51	63.41
SynthCP-joint	12.69	**11.29**	62.52	61.23	13.68	**11.60**	**64.05**	**65.42**
	FCN-8				Deeplab-v2			
pixel-level	AP-Err↑	AP-Suc↑	AUC↑	FPR95↓	AP-Err↑	AP-Suc↑	AUC↑	FPR95↓
MSP [17]	50.31	99.02	91.54	25.34	48.46	99.24	92.26	24.41
MCDropout [10]	49.23	99.02	91.47	25.16	47.85	99.23	92.19	24.68
TCP [7]	48.54	98.82	90.29	32.21	45.57	98.84	89.14	36.98
Direct Prediction	52.17	99.15	92.55	**22.34**	48.76	99.34	92.94	**21.56**
SynthCP-separate	54.14	99.15	92.70	22.47	48.79	99.31	92.74	22.15
SynthCP-joint	**55.53**	**99.18**	**92.92**	22.47	**49.99**	**99.34**	**92.98**	21.69

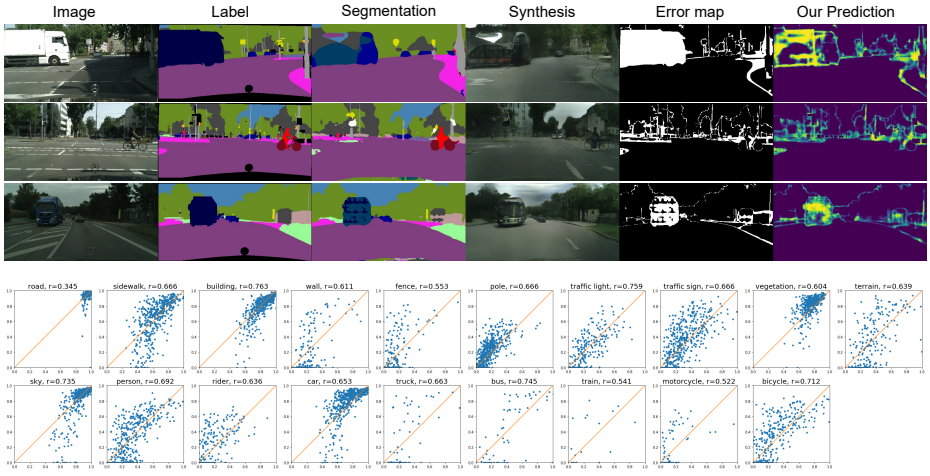

Fig. 5. Visualization on the Cityscapes dataset for pixel-level error map prediction (top) and image-level per-class IoU prediction (bottom). For each example from left to right (top), we show the original image, ground-truth label map, segmentation prediction, synthesized image conditioned on the segmentation prediction, (ground-truth) errors in the segmentation prediction and our pixel-level error prediction. The plots (bottom) show significant correlations between the ground-truth IoU and our predicted IoU on most of the classes.

Results. Experimental results are shown in Table 1 and visualizations are shown in Fig. 5. We use the well-known FCN8 [41] and Deeplab-v2 [6] as the segmentation models. For image-level failure detection, our approach consistently outperforms other methods on all metrics. Results are averaged over 19 classes for all four metrics (detailed results in supplementary). We find that VAE alarm does not perform well 2D images of street scenes, since small objects are easily missed in the VAE reconstruction. Without the synthesized images from the segmentation results, Direct Prediction performs worse than ours despite achieving better performance than the others.

For pixel-level failure detection, our approach achieves the state-of-the-art performance as well, especially for AP-Error metric where our approach outperforms other methods by a considerable margin. The comparison to Direct Prediction demonstrates that the improvements come from the image synthesis module in our framework. We hypothesis that TCP performs not as well because it is mainly designed for classification and it might be hard to fit the true class probability for dense predictions on large images in our settings. We find that our method produces slightly more false positives than "Direct Prediction" baseline (FPR95 is lower). We think the reason might be some correctly segmented regions are not synthesized well by the generative model.

We conducted another experiment to validate the **generalizability** on unseen segmentation models. We directly test our failure detection model, which is trained on Deeplab-v2 masks, on the segmentation masks produced by FCN8. We achieve an AUPR-Error of 53.12 for pixel-level error detection and MAE of 12.91 for image-level failure prediction. Full results are available in the supplementary material. The results are comparable to those obtained by our model trained on FCN8 segmentation model, as shown in table 1.

The Pancreatic Tumor Segmentation Dataset We also validate SynthCP on medical images. Following VAE alarm [38], we applied SynthCP to the challenging pancreatic tumor segmentation task of Medical Segmentation Decathlon [46], where we randomly split the 281 cases into 200 training and 81 testing. The VAE alarm system [38] is the main competitor on this dataset. Since their approach explored shape prior for accurate quality assessment and tumor shapes have large variance, we expect SynthCP can outperform shape-based models or be complementary to the VAE-based alarm model. We only compare image-level failure detection in this dataset, because the VAE alarm system [38] is targeted to this task and sets up standard baselines.

We use the state-of-the-art network 3D AH-Net [39] as the segmentation model. Instead of IoU, the segmentation performance is measured by Dice coefficient (DSC), a standard evaluation metric used for medical image segmentation. Moving into 3D is challenging for SynthCP, since training 3D GANs is extremely hard, considering the limited GPU memory and high computational costs. In practice, we modify SPADE into 3D. Results and visualizations are shown in Table 2 and Fig. 6 respectively. In terms of baselines, we re-implement the VAE alarm system for a fair comparison in our settings, while the results of other methods are quoted from [38]. SynthCP achieved comparable performances as VAE alarm system. When combined with VAE alarm (a simple ensemble of the predicted DSC), all of the four metric improves significantly (P.C. and S.C correlation coefficient both improves by approximately 7% and 10% respectively), illustrating SynthCP which captures label-to-image information is complementary to the shape-based VAE approach.

Table 2. Failure detection results on the pancreatic tumor segmentation dataset in MSD [46]

	tumor DSC prediction			
Method	MAE ↓	STD ↓	P.C. ↑	S.C. ↑
Direct Prediction	23.20	29.81	45.50	45.36
Jungo *et al.* [25]	26.57	29.78	-23.87	-20.23
Kwon *et al.* [32]	26.14	29.24	14.61	14.70
VAE alarm [38]	20.21	23.60	60.24	63.30
VAE (our imple.)	18.60	13.73	63.42	58.47
SynthCP	18.13	13.77	61.11	62.66
SynthCP + VAE	**15.19**	**13.37**	**67.97**	**71.35**

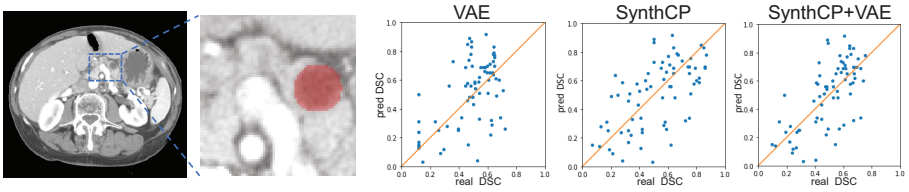

Fig. 6. Left two: an example of pancreatic tumor segmentation (in red). Right three: plots for tumor segmentation DSC score prediction by VAE alarm [38], SynthCP and the ensemble of SynthCP and VAE alarm.

4.2 Anomaly Segmentation

Evaluation metrics. We use the standard metrics for OOD detection and anomaly segmentation: area under the ROC curve (AUROC), false positive rate at 95% recall (FPR95), and area under the precision recall curve (AUPR).

The StreetHazards Dataset We validate SynthCP on the StreetHazards dataset of CAOS Benchmark [16]. This dataset contains 5125 training images, 1000 validation images and 1500 test images. 250 types of anomaly objects appears only in the testing images.

Baselines. Baseline approaches include MSP [17], MSP+CRF [16], Dropout [10] and an auto-encoder (AE) based approach [2]. Except for AE, all the other three approaches require a segmentation model to provide either softmax probability or uncertainty estimation. AE is the only approach that requires extra training of an auto-encoder for the images and computes pixel-wise ℓ_1 loss for anomaly segmentation.

Implementation details. Following [16], we use two network backbones as the segmentation models: ResNet-101 [15] and PSPNet [50]. The cGAN is also SPADE [43] trained with the same training strategy as in Sec 3.2. The post-processing threshold $t = 0.999$ is chosen for better AUPR and is discussed in detail in the following paragraph.

Results Experimental results are shown in Table 3. SynthCP improves the previous state-of-the-art approach MSP+CRF from 6.5% to 9.3% in terms of AUPR. Fig. 7 shows some anomaly segmentation examples.

Table 3. Anomaly segmentation results on StreetHazards dataset [16]

Method	FPR95↓	AUROC↑	AUPR↑
AE [2]	91.7	66.1	2.2
Dropout [10]	79.4	69.9	7.5
MSP [17]	33.7	87.7	6.6
MSP + CRF [16]	29.9	88.1	6.5
SynthCP	**28.4**	**88.5**	**9.3**

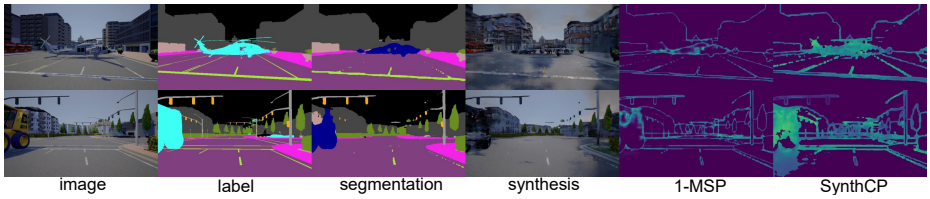

Fig. 7. Visualizations on the StreetHazards dataset. For each example, from left to right, we show the original image, ground-truth label map, segmentation prediction, synthesized image conditioned on segmentation prediction, MSP anomaly segmentation prediction and our anomaly segmentation prediction.

To study how much MSP post-processing contributes to SynthCP, we conduct experiments on different thresholds of t for post-processing. As shown in Table 4, without post-processing ($t = 1.0$), SynthCP achieves higher AUPR, but also produces more false positives, resulting in degrading FPR95 and AUROC. After pruning out false positives at high MSP positions ($p^{(i)} > 0.999$), we achieved the state-of-the-art performances under all three metrics.

Table 4. Performance change by varying post-processing threshold t

t	0.8	0.9	0.99	0.999	1.0
FPR95 ↓	**28.6**	28.5	28.2	28.4	46.0
AUROC ↑	88.3	88.4	**88.6**	88.5	81.9
AUPR ↑	7.4	7.7	8.8	**9.3**	8.1

5 Conclusions

We present a unified framework, SynthCP, to detect failures and anomalies for semantic segmentation, which consists of an image synthesize module and a comparison module. We model the image synthesize module with a semantic-to-image conditional GAN (cGAN) and train it on label-image pairs. We then use it to reconstruct the image based on the predicted segmentation mask. The synthesized image and the original image are fed forward to the comparison module and output either failure detection (both image-level and pixel-level) or the mask of anomalous objects, depending on the specific task. SynthCP achieved the state-of-the-art performances on three challenging datasets.

Acknowledgement This work was supported by NSF BCS-1827427, the Lustgarten Foundation for Pancreatic Cancer Research and NSFC No. 61672336. We also thank the constructive suggestions from Dr. Chenxi Liu, Qing Liu and Huiyu Wang.

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in ai safety. arXiv preprint arXiv:1606.06565 (2016)
2. Baur, C., Wiestler, B., Albarqouni, S., Navab, N.: Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In: MICCAI Brainlesion Workshop (2018)
3. Bevandić, P., Krešo, I., Oršić, M., Šegvić, S.: Discriminative out-of-distribution detection for semantic segmentation. arXiv preprint arXiv:1808.07703 (2018)
4. Chabrier, S., Emile, B., Rosenberger, C., Laurent, H.: Unsupervised performance evaluation of image segmentation. EURASIP Journal on Applied Signal Processing (2006)
5. Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., Tsaneva-Atanasova, K.: Artificial intelligence, bias and clinical safety. BMJ Quality and Safety (2019)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI (2018)
7. Corbière, C., Thome, N., Bar-Hen, A., Cord, M., Pérez, P.: Addressing failure prediction by learning model confidence. In: Advances in Neural Information Processing Systems (2019)
8. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2016)
9. DeVries, T., Taylor, G.W.: Learning confidence for out-of-distribution detection in neural networks. arXiv preprint arXiv:1802.04865 (2018)
10. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning, ICML (2016)
11. Gao, H., Tang, Y., Jing, L., Li, H., Ding, H.: A novel unsupervised segmentation quality evaluation method for remote sensing images. Sensors (2017)
12. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems (2017)
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (2014)
14. Haselmann, M., Gruber, D.P., Tabatabai, P.: Anomaly detection using deep learning based image completion. In: International Conference on Machine Learning and Applications, ICMLA. IEEE (2018)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2016)
16. Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., Song, D.: A benchmark for anomaly segmentation. arXiv preprint arXiv:1911.11132 (2019)
17. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. International Conference on Learning Representations, ICLR (2017)
18. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. International Conference on Learning Representations, ICLR (2019)

19. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2017)
20. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring r-cnn. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2019)
21. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2017)
22. Janai, J., Güney, F., Behl, A., Geiger, A.: Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. arXiv preprint arXiv:1704.05519 (2017)
23. Jiang, B., Luo, R., Mao, J., Xiao, T., Jiang, Y.: Acquisition of localization confidence for accurate object detection. In: Proceedings of the European Conference on Computer Vision, ECCV (2018)
24. Jiang, H., Kim, B., Guan, M., Gupta, M.: To trust or not to trust a classifier. In: Advances in Neural Information Processing Systems (2018)
25. Jungo, A., Meier, R., Ermis, E., Herrmann, E., Reyes, M.: Uncertainty-driven sanity check: Application to postoperative brain tumor cavity segmentation. arXiv preprint arXiv:1806.03106 (2018)
26. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems (2017)
27. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations, ICLR (2015)
28. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. International Conference on Learning Representations, ICLR (2014)
29. Kohlberger, T., Singh, V., Alvino, C., Bahlmann, C., Grady, L.: Evaluating segmentation error without ground truth. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI (2012)
30. Krešo, I., Oršić, M., Bevandić, P., Šegvić, S.: Robust semantic segmentation with ladder-densenet models. arXiv preprint arXiv:1806.03465 (2018)
31. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
32. Kwon, Y., Won, J.H., Kim, B.J., Paik, M.C.: Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. Computational Statistics and Data Analysis (2020)
33. Lee, K., Lee, H., Lee, K., Shin, J.: Training confidence-calibrated classifiers for detecting out-of-distribution samples. International Conference on Learning Representations, ICLR (2018)
34. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Advances in Neural Information Processing Systems (2018)
35. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. International Conference on Learning Representations, ICLR (2018)
36. Linda, O., Vollmer, T., Manic, M.: Neural network based intrusion detection system for critical infrastructures. In: International Joint Conference on Neural Networks, IJCNN (2009)
37. Lis, K., Nakka, K., Fua, P., Salzmann, M.: Detecting the unexpected via image resynthesis. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2152–2161 (2019)

38. Liu, F., Xia, Y., Yang, D., Yuille, A.L., Xu, D.: An alarm system for segmentation algorithm based on shape model. In: Proceedings of the IEEE International Conference on Computer Vision, ICCV (2019)
39. Liu, S., Xu, D., Zhou, S.K., Pauly, O., Grbic, S., Mertelmeier, T., Wicklein, J., Jerebko, A., Cai, W., Comaniciu, D.: 3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI (2018)
40. Liu, X., Yin, G., Shao, J., Wang, X., et al.: Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In: Advances in Neural Information Processing Systems (2019)
41. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2015)
42. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
43. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2019)
44. Robinson, R., Oktay, O., Bai, W., Valindria, V.V., Sanghvi, M.M., Aung, N., Paiva, J.M., Zemrak, F., Fung, K., Lukaschuk, E., et al.: Real-time prediction of segmentation quality. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI (2018)
45. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations, ICLR (2014)
46. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
47. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2015)
48. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2018)
49. Zendel, O., Honauer, K., Murschitz, M., Steininger, D., Fernandez Dominguez, G.: Wilddash-creating hazard-aware benchmarks. In: Proceedings of the European Conference on Computer Vision, ECCV (2018)
50. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2017)