Anomaly Detection With Multiple-Hypotheses Predictions

Duc Tam Nguyen [1 2] **Zhongyu Lou** [2] **Michael Klar** [2] **Thomas Brox** [1]

Abstract

In one-class-learning tasks, only the normal case (foreground) can be modeled with data, whereas the variation of all possible anomalies is too erratic to be described by samples. Thus, due to the lack of representative data, the wide-spread discriminative approaches cannot cover such learning tasks, and rather generative models, which attempt to learn the input density of the foreground, are used. However, generative models suffer from a large input dimensionality (as in images) and are typically inefficient learners. We propose to learn the data distribution of the foreground more efficiently with a *multi-hypotheses autoencoder*. Moreover, the model is criticized by a *discriminator*, which prevents artificial data modes not supported by data, and enforces diversity across hypotheses. Our multiple-hypotheses-based anomaly detection framework allows the reliable identification of out-of-distribution samples. For anomaly detection on CIFAR-10, it yields up to 3.9% points improvement over previously reported results. On a real anomaly detection task, the approach reduces the error of the baseline models from 6.8% to 1.5%.

1. Introduction

Anomaly detection classifies a sample as normal or abnormal. In many applications, however, it must be treated as a one-class-learning problem, since the abnormal class cannot be defined sufficiently by samples. Samples of the abnormal class can be extremely rare, or they do not cover the full space of possible anomalies. For instance, in an autonomous driving system, we may have a test case with a bear or a kangaroo on the road. For defect detection in manufacturing, new, unknown production anomalies due to critical changes in the production environment can appear. In medical data analysis, there can be unknown deviations from the healthy state. In all these cases, the well-studied discriminative models, where decision boundaries of classifiers are learned from training samples of all classes, cannot be applied. The decision boundary learning of discriminative models will be dominated by the normal class, which will negatively influence the classification performance.

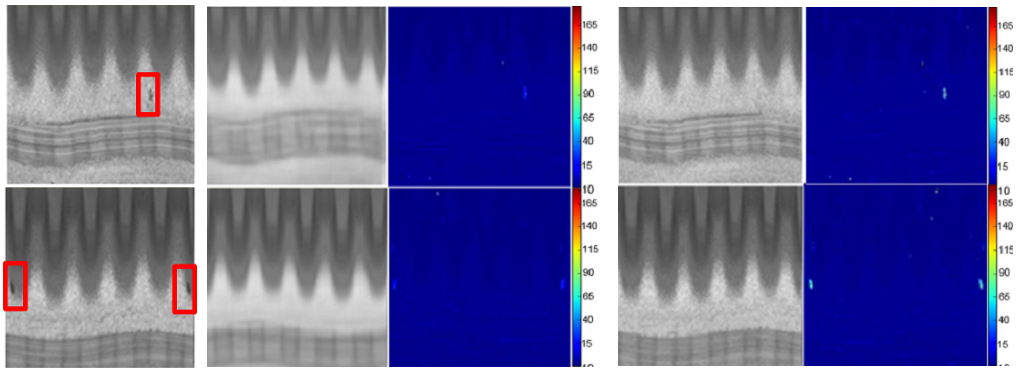
Anomaly detection as one-class learning is typically approached by generative, reconstruction-based methods (Zong et al., 2018). They approximate the input distribution of the normal cases by parametric models, which allow them to reconstruct input samples from this distribution. At test time, the data negative log-likelihood serves as an anomaly-score. In the case of high-dimensional inputs, such as images, learning a representative distribution model of the normal class is hard and requires many samples.

Autoencoder-based approaches, such as the variational autoencoder (Rezende et al., 2014; Kingma & Welling, 2013), mitigate the problem by learning a mapping to a lower-dimensional representation, where the actual distribution is modeled. In principle, the nonlinear mappings in the encoder and decoder allow the model to cover multi-modal distributions in the input space. However, in practice, autoencoders tend to yield blurry reconstructions, since they regress mostly the conditional mean rather than the actual multi-modal distribution (see Fig. 1 for an example on a metal anomaly dataset). Due to multiple modes in the actual distribution, the approximation with the mean predicts high probabilities in areas not supported by samples. The blurry reconstructions in Fig. 1 should have a low probability and be classified as anomalies, but instead they have the highest likelihood under the learned autoencoder. This is fatal for anomaly detection.

Alternatively, mixture density networks (Bishop, 1994) learn a conditional Gaussian *mixture distribution*. They directly estimate local densities that are coupled to a global density estimate via mixing coefficients. Anomaly scores for new points can be estimated using the data likelihood (see Appendix). However, global, multi-modal distribution es-

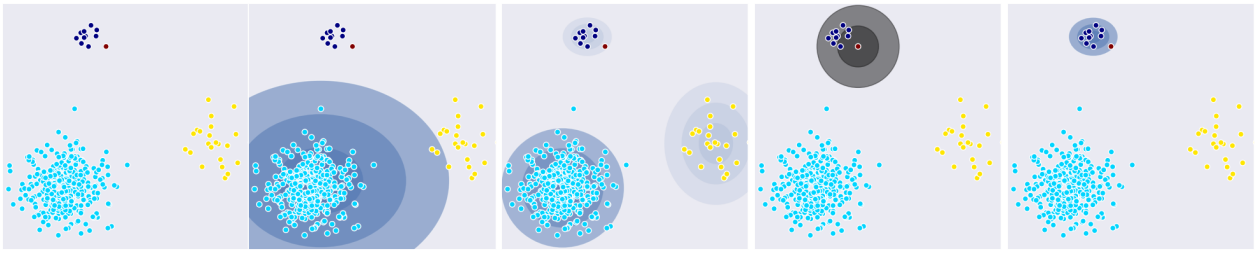
[1]Computer Vision Group, University of Freiburg, Freiburg, Germany [2]Corporate Research, Robert Bosch GmbH, Renningen, Germany. Correspondence to: Duc Tam Nguyen <Nguyen@informatik.uni-freiburg.de>, Zhongyu Lou <Zhongyu.Lou@de.bosch.com>, Michael Klar <Michael.Klar2@de.bosch.com>, Thomas Brox <Brox@informatik.uni-freiburg.de>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.



(a) Test images (b) Autonencoder reconstructions (c) ConAD reconstructions

Figure 1. Detection of anomalies on a Metal Anomaly dataset. (a) Test images showing anomalies (black spots). (b) An Autoencoder-based approach produces blurry reconstructions to express model uncertainty. The blurriness falsifies reconstruction errors (and hence anomaly scores)(c) Our model: Consistency-based anomaly detection (ConAD) gives the network more expressive power with a multi-headed decoder (also known as multiple-hypotheses networks). The resulting anomaly scores are hence much clearer in our framework ConAD.



(a) Cond. space (b) Autoencoder (c) MDN (d) LOF (e) Our model

Figure 2. Illustration of the different anomaly detection strategies. (a) In this example, two dimensions with details that are hard to capture in the conditional space are shown. The red dot is a new point. Dark blue indicates high likelihood, black indicates the neighborhood considered. The autoencoder (b) cannot deal with the multi-modal distribution. The mixture density network (c) in principle can do so, but recognition of the sample as a normal case is very brittle and will fail in case of mode collapse. Local-Outlier-Factor (d) makes a decision based on the data samples closest to the input sample. Our model (e) learns multiple local distributions and uses the data likelihood of the closest one as the anomaly score.

timation is a hard learning problem with many problems in practice. In particular, mixture density networks tend to suffer from mode collapse in high-dimensional data spaces, i.e., the relevant data modes needed to distinguish rare but normal data from anomalies will be missed.

Simple nearest neighbor analysis, such as the Local-outlier-factor (Breunig et al., 2000), operates in image-space directly without training. While this is a simple and sometimes effective baseline, such local analysis is inefficient in very high-dimensional spaces and is slow at test time. Fig. 2 illustrates these different strategies in a simple, two-dimensional example.

In this work, we propose the use of multiple-hypotheses networks (Rupprecht et al., 2016; Chen & Koltun, 2017; Ilg et al., 2018; Bhattacharyya et al., 2018) for anomaly detection to provide a more fine-grained description of the data distribution than with a single-headed network. In conjunction with a variational autoencoder, the multiple hypotheses can be realized with a multi-headed decoder. Concretely, *each network head* may predict a Gaussian density estimate. Hypotheses form clusters in the data space and can capture model uncertainty not encoded by the latent code.

Multiple-hypotheses networks have not yet been applied to anomaly detection due to several difficulties in training these networks to produce a multi-modal distribution consistent with the training distribution. The loosely coupled hypotheses branches are typically learned with a winner-takes-all loss, where all learning signal is transferred to one single best branch. Hence, bad hypotheses branches are not penalized and may support non-existing data regions. These artificial data modes cannot be distinguished from normal data. This is an undesired property for anomaly detection and becomes more severe with an increasing number of hypotheses.

We mitigate the problem of artificial data modes by com-

bining multiple-hypotheses learning with a discriminator D as a critic. The discriminator ensures the consistency of estimated data modes with the real data distribution. Fig. 3 shows the scheme of the framework.

This approach combines ideas from all three previous paradigms: the latent code of a variational autoencoder yields a way to efficiently realize a generative model that can act in a rather low-dimensional space; the multiple hypotheses are related to the mixture density of mixture density networks, yet without the global component, which leads to mode collapse.

We evaluate the anomaly detection performance of our approach on CIFAR-10 and a real anomaly image dataset, the *Metal Anomaly dataset* with images showing a structured metal surface, where anomalies in the form of scratches, dents or texture differences are to be detected. We show that anomaly detection performance with multiple-hypotheses networks is significantly better compared to single-hypotheses networks. On CIFAR-10, our proposed ConAD framework (consistency-based anomaly detection) improves on previously published results. Furthermore, we show a large performance gap between ConAD and mixture density networks. This indicates that anomaly score estimation based on the global neighborhood (or data likelihood) is inferior to local neighborhood consideration.

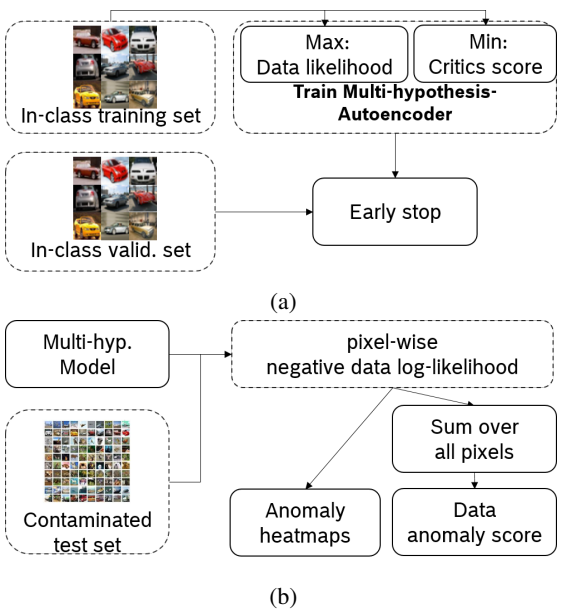


Figure 3. Training and testing overview of the proposed anomaly detection framework. (a) shows training the model to capture the normal data distribution. For the distribution learning, we use a multiple-hypotheses variational autoencoder (Fig. 4) with discriminator training (Fig. 5). During training, only data from the normal case are used. (b) At test time, the data likelihood is used for detecting anomalies. A low likelihood indicates an out-of-distribution sample, i.e., an anomaly.

2. Anomaly detection with multi-hypotheses variational autoencoders

2.1. Training and testing for anomaly detection

Fig. 3 shows the training and testing within our framework. The multiple-hypothesis variational autoencoder (Fig. 4) uses the data from the normal case for distribution learning. The learning is performed with the maximum likelihood and critics minimizing objectives (Fig. 5).

At test time (Fig 3b), the test set is contaminated with samples from other classes (anomalies). For each sample, the data negative log-likelihood under the learned multi-hypothesis model is used as an anomaly score. The discriminator only acts as a critic during training and is not required at test time.

2.2. Multiple-hypotheses variational autoencoder

For fine-grained data description, we learn a distribution with a multiple-hypotheses autoencoder. Figure 4 shows our multiple-hypotheses variational autoencoder. The last layer (head) of the decoder is split into H branches to provide H different hypotheses. The outputs of each branch are the parameters of an independent Gaussian for each pixel.

In the basic training procedure without discriminator training, the multiple-hypotheses autoencoder is trained with the winner-takes-all (WTA) loss:

$$L_{WTA}(x_i|\theta_h) = E_{z_k\sim q_\phi(z|x)}\left[\log p_{\theta_h}(x_i|z_k)\right] \quad (1)$$
$$\textbf{s.t.}\ h = \arg\max_j E_{z_k\sim q_\phi(z|x)}\left[\log p_{\theta_j}(x_i|z_k)\right],$$

whereby θ_j is the parameter set of hypothesis branch j, θ_h the best hypothesis w.r.t. the data likelihood of sample x_i, z_k is the noise and q_ϕ the distribution after the encoder. Only the network head with the best-matching hypothesis concerning the training sample receives the learning signal.

2.3. Training with discriminator as a critic

When learning with the winner-takes-all loss, the non-optimal hypotheses are not penalized. Thus, they can support any artificial data regions without being informed via the learning signal; for a more formal discussion see the Appendix. We refer to this problem as the inconsistency of the model regarding the real underlying data distribution.

As a new alternative, we propose adding a discriminator D as a critic when training the multiple-hypotheses autoencoder G; see Fig. 5. D and G are optimized together on the

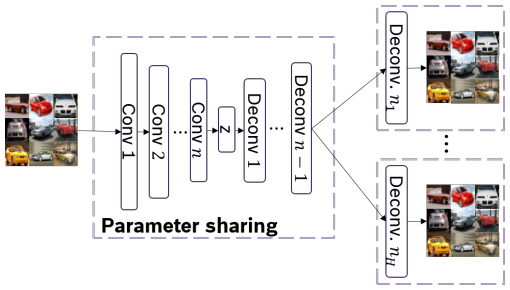


Figure 4. Multi-headed variational autoencoder. All heads share the same encoder, the same latent code, and large parts of the decoder, but the last layers create different hypotheses.

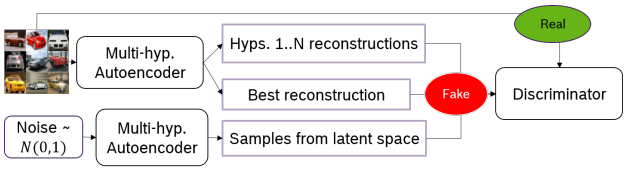


Figure 5. Discriminator training in the context of the multiple-hypotheses autoencoder. As in usual discriminator training, an image from the training set and a randomly sampled image are labeled as real and fake respectively. Additional fake samples are generated by the autoencoder.

minimax loss

$$\min_D \max_G L_D(x,z) = \min_D \max_G \underbrace{-\log(p_D(x_{real}))}_{L_{real}} + L_{fake}(x,z) \quad (2)$$

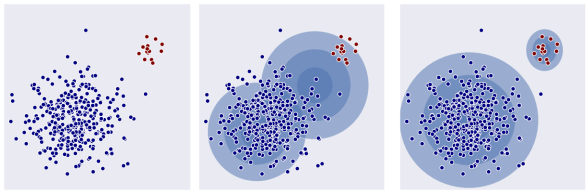
$$\text{with } L_{fake}(x,z) = \log(p_D(\hat{x}_{z\sim\mathcal{N}(0,1)})) + \log(p_D(\hat{x}_{z\sim\mathcal{N}(\mu_{z|x},\Sigma_{z|x})})) + \log(p_D(\hat{x}_{best-guess})) \quad (3)$$

Figure 5 illustrates how samples are fed into the discriminator. In contrast to a standard GAN, samples labeled as fake come from three different sources: randomly-sampled images $\hat{x}_{z\sim\mathcal{N}(0,1)}$, data reconstruction defined by individual hypotheses $\hat{x}_{z\sim\mathcal{N}(\mu_{z|x},\Sigma_{z|x})}$, the best combination of hypotheses according to the winner-takes-all loss $\hat{x}_{\text{best\_guess}}$.

Accordingly, the learning objective for the VAE generator becomes:

$$\min_G L_G = \min_G L_{WTA} + KL(q_\phi(z|x)||\mathcal{N}(0,1)) - L_D, \quad (4)$$

where KL denotes the symmetrized Kullback-Leibler divergence (Jensen-Shannon divergence). Intuitively, the discriminator enforces the generated hypotheses to remain in realistic data regions. The model is trained until the WTA-loss is minimized on the validation set.



(a) (b) (c)

Figure 6. (a) Modeling task with one extremely dominant data mode (dense region) and one under-represented mode. (b) shows how multiple-hypotheses predictions are used to cover data modes. Hypotheses tend to concentrate on the dominant mode, which leads to over-fitting in this region. (c) Increasing diversity across hypotheses (similar to maximizing inter-class variance) leads to better coverage of the underlying data.

2.4. Avoiding mode collapse

To avoid mode collapse of the discriminator training and hypotheses, we propose to employ hypotheses discrimination. This is inspired by minibatch discrimination (Salimans et al., 2016). Concretely, in each batch, the discriminator receives the pair-wise features-distance of generated hypotheses. Since batches of real images have large pair-wise distances, the generator has to generate diverse outputs to avoid being detected too easily. Training with hypotheses discrimination naturally leads to more diversity among hypotheses.

Fig. 6 shows a simple example of why more diversity among hypotheses is beneficial. The hypotheses correspond to cluster centers in the image-conditional space. Maximizing diversity among hypotheses is, hence, similar to the maximization of inter-class-variance in typical clustering algorithm such as Linear Discriminant Analysis (Mika et al., 1999).

2.5. Anomaly score estimation based on local neighborhood

Hypotheses are spread out to cover the data modes seen during training. Due to the loose coupling between hypotheses, the probability mass of each hypothesis is only distributed *within the respective cluster*. Compared to traditional likelihood learning, the conditional probability mass only sums up to 1 within each hypothesis branch, i.e., the combination of all hypotheses does not yield a proper density function as in mixture density networks. However, we can use the winner-takes-all loss as the pixel-wise sample anomaly score. Hence, each pixel likelihood is only evaluated based on the best-matching conditional hypothesis. We refer to this as anomaly detection based on local likelihood estimation.

Local likelihood is more effective for anomaly score estimation Fig. 2 provides an intuition, why the local neighborhood is more effective in anomaly detection. The red point represents a new normal point which is very close to one less dominant data mode. By using the global likelihood function (Fig. 2c), the anomaly score depends on all other points.

However, samples further away intuitively do not affect the anomaly score estimation. In Local-outlier-factor (Breunig et al., 2000), outlier score estimation only depends on samples close to the new point (fig. 2d). Similarly, our multi-hypotheses model considers only the next cluster (fig. 2e) and provides a more accurate anomaly score.

Further, learning local likelihood estimations is easier and more sample-efficient than learning from a global likelihood function, since the local model need not learn the global dependencies. During training, it is sufficient if samples are covered by at least one hypothesis.

In summary, we estimate the anomaly scores based on the consistency of new samples regarding the closest hypotheses. Accordingly, we refer to our framework as *consistency-based anomaly detection (ConAD)*.

3. Related works

In high-dimensional input domains such as images, modern generative models (Kingma & Welling, 2013; Goodfellow et al., 2014) are typically used to learn the data distribution for the normal data (Cong et al., 2011; Li et al., 2014; Ravanbakhsh et al., 2017). In many cases, anomaly detection might improve the models behavior in out-of-distribution cases (Nguyen et al., 2018).

For learning in uncertain tasks, Chen & Koltun (2017); Bhattacharyya et al. (2018); Rupprecht et al. (2016); Ilg et al. (2018) independently proposed multiple-hypotheses-predictions (MHP) networks. More details about theses works can be found in the Appendix.

In contrast to previous MHP-networks, we propose to utilize these networks for anomaly detection for the first time. To this end, we introduce a strategy to avoid the support of artificial data modes, namely via a discriminator as a critic. (Rupprecht et al., 2016) suggested a soft WTA-loss, where the non-optimal hypotheses receive a small fraction of the learning signal. Depending on the softening parameter ϵ, the model training results in a state between mean-regression (i.e., uni-modal learning) and large support of non-existing data modes (more details in the Appendix). Therefore, the soft-WTA-loss is a compromise of contradicting concepts and, thus, requires a good choice of the corresponding hyperparameter. In the case of anomaly detection, the hyperparameter search cannot be formalized, since there are not enough anomalous data points available.

Compared to previous reconstruction-based anomaly detection methods (using, e.g., Kingma & Welling (2013); Bishop (1994)), our framework evaluates anomaly score only based on the local instead of the global neighborhood. Further, the model learns from a relaxed version of likelihood maximizing, which results in better sample efficiency.

4. Experiments

In this section, we compare the proposed approach to previous deep learning and non-deep learning techniques for one-class learning tasks. Since true anomaly detection benchmarks are rare, we first tested on CIFAR-10, where one class is used as the normal case to be modeled, and the other 9 classes are considered as anomalies and are only available at test time. Besides, we tested on a true anomaly detection task on a metal anomaly dataset, where arbitrary deviations from the normal case can appear in the data.

4.1. Network architecture

The networks are following DCGAN (Radford et al., 2015) but were scaled down to support the low-resolution of CIFAR-10. Concretely, the decoder only uses a sequence of Dense-Deconv.-Conv.-Deconv. layers and on top, $2 * n$ Deconv. layer for n hypotheses branches. Each branch requires two layers since for each pixel position, the network predicts a μ andσ for the conditional distribution. Further, throughout the network, leaky-relu units are employed.

Hypotheses branches are represented as decoder networks heads. Each hypothesis predicts one Gaussian distribution with diagonal co-variance Σ and means μ. The winner-takes-all loss operates on the pixel-level, i.e., for each predicted pixel, there is a single winner across hypotheses. The best-combined-reconstructions is the combination of the winning hypotheses on pixel-level.

4.2. Training

For training with the discriminator in Fig. 5, samples are forwarded separately through the network. The batch-size n was set to 64 each on CIFAR-10, 32 on the Metal Anomaly dataset. Adam (Kingma & Ba, 2014) was used for training with a learning rate of 0.001. Per discriminator training, the generator is trained at most five epochs to balance both players. We use the validation set of samples from the normal class to early stop the training if no better model regarding the corresponding loss could be found.

4.3. Evaluation

Experiments details Quantitative evaluation is done on CIFAR-10 and the Metal Anomaly dataset (Tab.1). The typ-

Table 1. Dataset description. CIFAR-10 is transformed into 10 anomaly detection tasks, where one class is used as the normal class, and the remaining classes are treated as anomalies. The train & validation datasets contain only samples from the normal class. This scenario resembles the typical situation where anomalies are extremely rare and not available at training time, as in the Metal Anomaly dataset.

	TYPE	CIFAR-10	METAL ANOMALY
PROBLEM	-	1 VS. 9	1 VS. 1
TASKS	-	10	1
RESOLUTION	-	32X32	224X224
NORMAL DATA	TRAIN	4500	5408
	VALID	500	1352
	TEST	1000	1324
ANOMALY	TEST	9000	346

ical 10-way classification task in CIFAR-10 is transformed into 10 one vs. nine anomaly detection tasks. Each class is used as the normal class once; all remaining classes are treated as anomalies. During model training, only data from the normal data class is used, data from anomalous classes are abandoned. At test time, anomaly detection performance is measured in Area-Under-Curve of Receiver Operating Curve (AUROC) based on normalized negative log-likelihood scores given by the training objective.

In Tab. 2, we evaluated on CIFAR-10 variants of our multiple-hypotheses approaches including the following energy formulations: MDN (Bishop, 1994), MHP-WTA (Ilg et al., 2018), MHP (Rupprecht et al., 2016), ConAD, and MDN+GAN. We compare our methods against vanilla VAE (Kingma & Welling, 2013; Rezende et al., 2014) , VAEGAN (Larsen et al., 2015; Dosovitskiy & Brox, 2016), AnoGAN (Schlegl et al., 2017), AdGAN Deecke et al., 2018, OC-Deep-SVDD (Ruff et al., 2018). Traditional approaches considered are: Isolation Forest (Liu et al., 2008; 2012), OCSVM (Schölkopf et al., 2001). The performance of traditional methods suffers due to the curse of dimensionality (Zong et al., 2018).

Furthermore, on the high-dimensional Metal Anomaly dataset, we focus only on the evaluation of deep learning techniques. The GAN-techniques proposed by previous work AdGAN & AnoGAN heavily suffer from instability due to pure GAN-training on a small dataset. Hence, their training leads to random anomaly detection performance. Therefore, we only evaluate MHP-based approaches against their uni-modal counterparts (VAE, VAEGAN).

Anomaly detection on CIFAR-10 Tab. 3 and Tab. 4 show an extensive evaluation of different traditional and deep learning techniques. Results are adopted from (Deecke et al., 2018) in which the training and testing scenarios were similar. The average performance overall 10 anomaly detection tasks are summarized in Tab. 2. Traditional,

Table 2. Anomaly detection on CIFAR-10, performance measured in AUROC. Each class is considered as the normal class once with all other classes being considered as anomalies, resulting in 10 one-vs-nine classification tasks. Performance is averaged for all ten tasks and over three runs each (see Appendix for detailed performance). Our approach significantly outperforms previous non-Deep Learning and Deep Learning methods.

TYPE	MODELS			
NON-DL.	KDE-PCA	OC-SVM-PCA	IF	GMM
	59.0	61.0	55.8	58.5
DL	. ANOGAN	OC-D-SVDD	ADGAN	CONAD
	61.2	63.2	62.0	**67.1**

non-deep-learning methods only succeed to capture classes with a dominant homogeneous background such as ships, planes, frogs (backgrounds are water, sky, green nature respectively). This issue occurs due to preceding feature projection with PCA, which focuses on dominant axes with large variance. (Deecke et al., 2018) reported that even features from a pretrained AlexNet have no positive effect on anomaly detection performance.

Our approach ConAD outperforms previously reported results by 3.9% absolute improvement. Furthermore, compared to other multiple-hypotheses-approaches (MHP, MDN, MHP+WTA), our model could benefit from the increased capacity given by the additional hypotheses. The combination of discriminator training and a high number of hypotheses is crucial for high detection performance as indicated in our ablation study (Tab. 5).

Anomaly detection on Metal Anomaly dataset Fig. 7 shows a qualitative analysis of uni-modal learning with VAE (Kingma & Welling, 2013) compared to our framework ConAD. Due to the fine-grained learning with multiple-hypotheses, our maximum-likelihood reconstructions of samples are significantly closer to the input. Contrary, VAE training results in blurry reconstructions and hence falsified anomaly heatmaps, hence cannot separate possible anomaly from dataset details.

Tab. 6 shows an evaluation of MHP-methods against multi-modal density-learning methods such as MDN (Bishop, 1994), VAEGAN (Dosovitskiy & Brox, 2016; Larsen et al., 2015). Note that the VAE-GAN model corresponds to our ConAD with a single hypothesis. The VAE corresponds to a single hypothesis variant of MHP, MHP-WTA, and MDN.

Table 3. CIFAR-10 anomaly detection: AUROC-performance of different approaches. The column indicates which class was used as in-class data for distribution learning. Note that random performance is at 50% and higher scores are better. Top-2-methods are marked. Our ConAD approach outperforms traditional methods and vanilla MHP-approaches significantly and can benefit from an increasing number of hypotheses.

CIFAR-10	0	1	2	3	4	5	6	7	8	9	MEAN
VAE	77.1	46.7	68.4	53.8	71.	54.2	64.2	51.2	**76.5**	46.7	61.0
OC-D-SVDD	61.7	65.9	50.8	59.1	60.9	65.7	67.7	67.3	75.9	73.1	63.2
MDN-2	76.1	46.9	68.7	53.8	70.4	53.8	63.2	52.3	**76.8**	46.7	60.9
MDN-4	76.9	46.8	68.6	53.5	69.3	54.4	63.5	54.1	76.	46.9	61.0
MDN-8	76.2	46.9	68.6	53.3	70.4	54.7	63.3	53.	76.3	47.3	61.
MDN-16	76.2	47.9	68.2	52.8	70.1	54.	63.5	52.9	76.4	46.9	60.9
MHP-WTA-2	77.3	51.6	68.	55.2	69.5	54.3	64.3	55.5	76.	51.2	62.2
MHP-WTA-4	**77.8**	53.9	65.1	56.7	66.	54.2	63.5	56.3	75.2	54.1	62.2
MHP-WTA-8	76.1	56.	62.7	58.8	62.6	55.3	61.4	57.8	74.3	54.8	61.9
MHP-WTA-16	75.7	56.7	60.9	59.8	62.7	56.	61.	56.8	73.8	57.3	62.
MHP-2	75.5	49.9	67.6	54.6	69.3	54.3	63.6	57.7	76.4	50.8	61.9
MHP-4	75.2	51.	66.	56.8	67.7	55.1	64.4	56.	76.4	51.	61.9
MHP-8	75.7	54.	65.2	57.6	64.8	55.4	62.5	54.7	75.9	53.	61.8
MHP-16	75.8	53.9	64.1	58.5	64.6	55.2	62.3	54.5	75.9	53.2	61.7
MDN+GAN-2	74.6	48.9	68.6	52.1	71.1	52.5	66.8	57.7	76.5	48.1	61.6
MDN+GAN-4	76.2	50.4	69.	52.4	71.6	53.2	65.9	58.3	75.3	48.9	62.1
MDN+GAN-8	77.4	48.3	69.3	53.1	72.2	53.7	67.9	54.	76.	51.9	62.3
MDN+GAN-16	73.6	46.9	69.4	52.2	75.3	54.1	65.7	56.8	75.3	45.4	61.4
CONAD - 2 (OURS)	77.3	60.0	66.6	56.2	69.4	56.1	70.6	63.0	74.8	49.9	64.3
CONAD - 4 (OURS)	**77.6**	52.5	66.3	57.0	68.7	54.1	**80.1**	54.8	74.1	53.9	63.9
CONAD - 8 (OURS)	77.4	**65.2**	64.8	60.1	67.0	57.9	72.5	**66.2**	74.8	**66.0**	**67.1**
CONAD - 16 (OURS)	77.2	**63.1**	63.1	**61.5**	63.3	**58.8**	69.1	**64.0**	75.5	**63.7**	**65.9**

Table 4. Anomaly detection performance on CIFAR-10 dependent on multiple-hypotheses-predictions models and hypotheses number. Performance averaged over tasks and in multiple runs each.

	HYPOTHESES				
MODELS	1	2	4	8	16
MHP	61.0 =VAE	61.9	61.9	61.8	61.7
MHP+WTA		62.2	62.2	61.9	62.0
MDN		60.9	61.0	61.0	60.9
MDN+GAN	61.7 =VAEGAN	61.6	62.1	62.0	61.4
CONAD		**64.3**	**63.9**	**67.1**	**65.9**

Table 5. Ablation study of our approach ConAD on CIFAR-10, meausured in anomaly detection performance (AUROC-scores on unseen contaminated dataset).

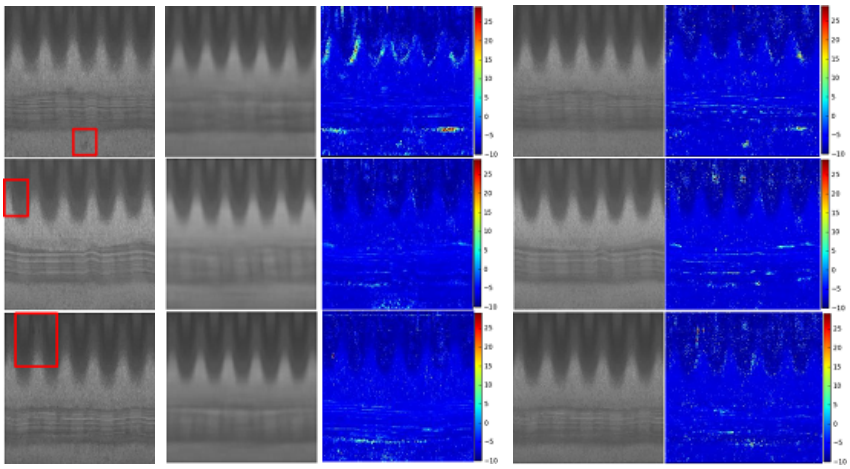
CONFIGURATION	AUROC
CONAD (8-HYPOTHESES)	**67.1**
- FEWER HYPOTHESES (2)	64.3
- DISCRIMINATOR	61.9
- WINNER-TAKES-ALL-LOSS (WTA)	61.8
- WTA & LOOSE HYP. COUPLING	61.0
- MULTIPLE-HYPOTHESES	61.7
- MULTIPLE-HYPOTHESES & DISCRIMINATOR	61.0

The significant improvement of up to 4.2% AUROC-score comes from the loose coupling of hypotheses in combination with a discriminator D as quality assurance. In a high-dimensional domain such as images, anomaly detection with MDN is worse than MHP approaches. This result from (1) typical mode collapse in MDN and (2) global neighborhood consideration for anomaly score estimation.

Using the MHP-technique, better performance is already achieved with two hypotheses. However, without the discriminator D, an increasing number of hypotheses rapidly leads to performance breakdown, due to the inconsistency property of generated hypotheses. Intuitively, additional non-optimal hypotheses are not strongly penalized during training, if they support artificial data regions.

With our framework ConAD, anomaly detection performance remains competitive or better even with an increasing number of hypotheses available. The discriminator D makes the framework adaptable to the new dataset and less sensitive to the number of hypotheses to be used.

When more hypotheses are used (8), the anomaly detection performance in all multiple-hypotheses models rapidly breaks down. The standard variance of performance of standard approaches remains high (up to $\pm$ 3.5). The reason might be the beneficial start for some hypotheses branches, which adversely affect non-optimal branches.



(a) (b) (c)

Figure 7. (a) anomalous samples on Metal Anomaly data-set. Anomalies are highlighted. (b) shows maximum-likelihood reconstructions under a Variational Autoencoder and the corresponding anomaly heatmaps based on negative-log-likelihood. (c) shows the reconstructions and anomaly maps for ConAD. In all cases, the maximum-likelihood expectation under the unimodal model is blurry and should itself be seen as an anomaly. Contrary, under our model, the maximum-likelihood expectation of the input is much closer to the input and more realistic. Due to the fine-grained learning, the anomaly heatmaps could reliably identify the location and strength of possible anomalies.

Table 6. Anomaly detection performance and their standard variance on the Metal Anomaly dataset. To reduce noisy residuals due to the high-dimensional input domain, only 10% of maximally abnormal pixels with the highest residuals are summed to form the total anomaly score. AUROC is computed on an unseen test set, a combination of normal and anomaly data. For more detailed results see Appendix. The anomaly detection performance of plain MHP rapidly breaks down with an increasing number of hypotheses.

	HYPOTHESES			
MODEL	1	2	4	8
MHP	94.2 (1.4)	98.0 (0.5)	97.0 (1.0)	95.0 (0.2)
MHP+WTA		98.0 (0.9)	**98.0** (0.1)	94.6 (3.3)
MDN		90.0 (1.1)	91.0 (1.9)	91.6 (3.5)
MDN+GAN	93.6 (0.7)	94.2 (1.6)	91.3 (1.9)	94.3 (1.1)
CONAD		**98.5** (0.1)	97.7 (0.5)	**96.5** (0.2)

This effect is less severe in our framework ConAD. The standard variance of our approaches is also significantly lower. We suggest that the noise is then learned too easily. Consider the extreme case when there are 255 hypotheses available. The winner-takes-all loss will encourage each hypothesis branch to predict a constant image with one value from [0,255]. In our framework, the discriminator as a critic attempts to alleviate this effect. That might be a reason why our ConAD has less severe performance breakdown. Our model ConAD is less sensitive to the choice of the hyper-parameter for the number of hypotheses. It enables better exploitation of the additional expressive power provided by the MHP-technique for new anomaly detection tasks.

Our method can detect more subtle anomalies due to the focus on extremely similar samples in the local neighborhood. However, the added capacity by the hypotheses branches makes the network more sensitive to large label noise in the datasets. Hence, robust anomaly detection under label noise is a possible future research direction.

5. Conclusion

In this work, we propose to employ multiple-hypotheses networks for learning data distributions for anomaly detection tasks. Hypotheses are meant to form clusters in the data space and can easily capture model uncertainty not encoded by the latent code. Multiple-hypotheses networks can provide a more fine-grained description of the data distribution and therefore enable also a more fine-grained anomaly detection. Furthermore, to reduce support of artificial data modes by hypotheses learning, we propose using a discriminator D as a critic. The combination of multiple-hypotheses learning with D aims to retain the consistency of estimated data modes w.r.t. the real data distribution. Further, D encourages diversity across hypotheses with hypotheses discrimination. Our framework allows the model to identify out-of-distribution samples reliably.

For the anomaly detection task on CIFAR-10, our proposed model results in up to 3.9% points improvement over previously reported results. On a real anomaly detection task, the approach reduces the error of the baseline models from 6.8% to 1.5%.

Acknowledgements

This research was supported by Robert Bosch GmbH. We thank our colleagues Oezgn Cicek, Thi-Hoai-Phuong Nguyen and the four anonymous reviewers who provided great feedback and their expertise to improve our work.

References

Bhattacharyya, A., Schiele, B., and Fritz, M. Accurate and diverse sampling of sequences based on a best of many sample objective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8485–8493, 2018.

Bishop, C. M. Mixture density networks. Technical report, Citeseer, 1994.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pp. 93–104. ACM, 2000.

Chen, Q. and Koltun, V. Photographic image synthesis with cascaded refinement networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 1520–1529, 2017.

Cong, Y., Yuan, J., and Liu, J. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, pp. 3449–3456. IEEE, 2011.

Deecke, L., Vandermeulen, R., Ruff, L., Mandt, S., and Kloft, M. Anomaly detection with generative adversarial networks. 2018.

Dosovitskiy, A. and Brox, T. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems*, pp. 658–666, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Ilg, E., Çiçek, Ö., Galesso, S., Klein, A., Makansi, O., Hutter, F., and Brox, T. Uncertainty Estimates with Multi-Hypotheses Networks for Optical Flow. In *European Conference on Computer Vision (ECCV)*, 2018. URL http://lmb.informatik.uni-freiburg.de/Publications/2018/ICKMB18. https://arxiv.org/abs/1802.07095.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

Li, W., Mahadevan, V., and Vasconcelos, N. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36 (1):18–32, 2014.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422. IEEE, 2008.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):3, 2012.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., and Mullers, K.-R. Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop.*, pp. 41–48. Ieee, 1999.

Nguyen, T. T., Spehr, J., Zug, S., and Kruse, R. Multisource fusion for robust road detection using online estimated reliabilities. *IEEE Transactions on Industrial Informatics*, 14(11):4927–4939, 2018.

Radford, A., Metz, L., and Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv:1511.06434 [cs]*, November 2015. URL http://arxiv.org/abs/1511.06434. arXiv: 1511.06434.

Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C., and Sebe, N. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 1577–1581. IEEE, 2017.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. Deep one-class classification. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4393–4402, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/ruff18a.html.

Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., and Hager, G. D. Learning in an Uncertain World: Representing Ambiguity Through Multiple Hypotheses. *arXiv:1612.00197 [cs]*, December 2016. URL

http://arxiv.org/abs/1612.00197. arXiv: 1612.00197.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pp. 146–157. Springer, 2017.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7): 1443–1471, 2001.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., and Chen, H. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. *International Conference on Learning Representations.*, 2018.