CONSISTENCY-BASED ANOMALY DETECTION WITH ADAPTIVE MULTIPLE-HYPOTHESES PREDICTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

In out-of-distribution classification tasks, only some classes - the normal cases - can be modeled with data, whereas the variation of all possible anomalies is too large to be described sufficiently by samples. Thus, the wide-spread discriminative approaches cannot cover such learning tasks and rather generative models, which attempt to learn the input density of the ordinary cases, are used. However, generative models suffer under a large input dimensionality (as in images) and are typically inefficient learners. Motivated by the Local-Outlier-Factor (LOF) method, in this work, we propose to allow the network to directly estimate the local density functions since, for the detection of outliers, the local neighborhood is more important than the global one. At the same time, we retain consistency in the sense that the model must not support areas of the input space that are not covered by samples. Our method allows the model to identify out-of-distribution samples reliably. For the anomaly detection task on CIFAR-10, our ConAD model results in up to 5% points improvement over previously reported results.

1 INTRODUCTION

Anomaly detection tasks belong to the category of one-class-learning and are crucial in many applications, where a fixed set of classes cannot be defined, for instance, because a subset of classes is extremely rare or some classes are unknown at training time. For example, there might be a bear crossing the street as part of validation scenarios for automatic cars, unknown production anomalies due to critical change of the production environment, or unknown deviations from the healthy state in medical data. In all these cases, the well-established discriminative feature learning, where decision boundaries of the classifier are defined by samples from all classes, cannot be applied. The decision boundary is preferable to be set by the common classes, whereas the anomalies are defined indirectly as deviations from these common classes.

On high-dimensional inputs such as images, anomaly detection is typically solved by generative, reconstruction-based methods. They approximate the input distribution of the common cases by parametric models, which allow them to reconstruct inputs from this distribution. At test time, the data log-likelihood serves as a sample-conditional anomaly-score. However, the difficulty of the task is shifted to learning the meaningful and representative model. In the case of high-dimensional inputs, learning this distribution model is hard and requires many samples from the standard class.

In earlier works, Breunig et al. (2000) identified that the local neighborhood is more important for the detection of outliers than the global one. Their Local-Outlier-Factor (LOF) method determines outlierness of data points only relative to the density of the samples in the intermediate neighborhood. Motivated by this work, we propose to estimate the local density functions with neural networks by using the Multiple-Hypotheses-Prediction-technique (MHP) (Ilg et al., 2018), Rupprecht et al. (2016a), and Chen & Koltun (2017)).

For example, assume a network which predicts the density function with an isotropic Gaussian distribution parametrized by $\mu_{y|x}$ and $\Sigma_{y|x}$, with x, y being the input and output respectively. As one possible realization of MHP, the last layers of the networks could be split into n-branches to provide n-conditional hypotheses given a single input x. Typically, these hypotheses branches learn via a variant of the winner-takes-all loss, where only the best hypothesis receives a learning signal to finetune itself.

As a consequence of MHP-learning, each of these conditional hypothesis represents a local density estimation which covers one data mode. However, this potentially results in the following possible problems: (1) modecollapse among hypotheses and (2) support of non-realistic data regions.

On the one hand, mode collapse may occur since the data modes covered by hypotheses are not necessarily non-overlapping. In theory, the majority of hypotheses could concentrate on few dominant data modes and at the same time ignore less dense regions. On the other hand, due to the WTA objective, hypotheses may also support non-realistic data modes, i.e., which are not covered by real samples. This property is fatal for anomaly detection since incoming data points could be next to these artificial regions. For simplicity, imagine an anomaly detection system for images, in which a hypothesis always predicts a black image. Indeed, a new incoming black image is likely classified as typical by the system, although it does not belong to the real distribution.

In the ideal case, the space spanned by different conditional hypotheses should only cover realistic modes instead of supporting artificial, non-existent data regions. To preserve this consistency of hypotheses space w.r.t. the realistic distribution, we propose the framework: consistency-based anomaly detection (ConAD) which incorporates a discriminator D to penalize wrong hypotheses. Furthermore, since mode collapse is a well-studied problem, we propose to employ hypotheses discrimination (based on minibatch discrimination from Salimans et al. (2016b)) to foster data mode coverage through the diversity of hypotheses. Our framework naturally leads to meaningful and interpretable hypothesis generation without the need for specific prior knowledge in the form of other auxiliary tasks.

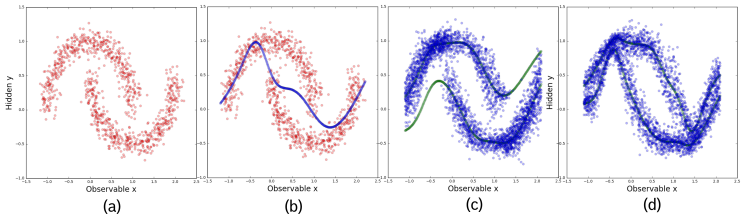

Figure 1: (a) Half-moon data-set: mapping from x to y is not unique, therefore the conditional output distribution is multi-modal (with two dense data regions at its maximum) where the unimodal regressor (b) fails. With MDN (c), local density functions support non-existing data regions while ConAD (d) forces all density functions to remain in realistic data regions. Note that the behavior is more complicated in high-dimensional inputs such as images. Compared to MDN, our ConAD is significantly less sensitive to outlier dimensions due to local density estimation in these domains.

In summary, our contributions are as follows:

- We relax the distribution modeling for reconstruction-based anomaly detection by leveraging MHP.
- We identify and address potential inconsistency property of MHP-techniques, which make them unsuitable for anomaly detection.
- We propose ConAD in combination with a discriminator as a general solution to avoid support of non-existent data regions and amplify the coverage of real data modes.

2 Related work

Popular traditional anomaly detection techniques comprise discriminative learning of normal data-containing hypersphere (Tax & Duin, 2004), estimation of decision boundary between data distribution and the origin (Schölkopf et al., 2001) or estimation of outlierness by the number of random splits needed to separate a data point in case of Isolation Forest (Liu et al., 2008; 2012). Furthermore, Local-Outlier-Factor (Breunig et al., 2000) measures the anomaly score of data points by the distance to their neighboring points, relative to the local density of the neighborhood. However, these methods typically suffer from the high-dimensional nature of inputs (as for images) and require a careful selection of low-dimensional features projections Zong et al. (2018).

Since traditional methods often underperform on high-dimensional input domain such as images (Zong et al., 2018), recent progress in deep generative modeling has enabled effective reconstruction-based anomaly detection. This paradigm includes a two-phase process: (1) estimation of the normal data distribution and (2) use the negative-log-likelihood or its variants as an anomaly score for unseen data points. The typical model of choice is the Variational Autoencoder (Rezende et al., 2014; Kingma & Welling, 2013) or Generative Adversarial Networks (Goodfellow et al., 2014). In this line of work, Schlegl et al. (2017) searches for best-matching candidates in learned noise space of a GAN for given input images, while Zong et al. (2018) estimates a GMM directly on reconstruction residuals and the corresponding latent code. Closely related to that, Kliger & Fleishman (2018) and Dai et al. (2017) attempt to use the discriminator D as an anomaly detector by enforcing the generator G to learn the low-density regions to provide D samples anomalies.

For the learning of complex multimodal distributions such as in Fig. 1, there is a long history of work. Bishop (1994) proposes neural networks which attempt to estimate all parameters Gaussian mixture Models conditioned on the inputs. This technique is closely related to Multi-Choice Learning (Dey et al., 2015; Lee et al., 2017; 2016). Motivated by these works, Prokudin et al. (2018) propose estimation of circular GMM to operate on a circular domain such as angles. Prokudin et al. (2018) performs uncertainty estimation for pose regression task by using the GMM. These works typically attempt to estimate a global density function, in which Gaussian Mixtures are connected via their corresponding mixing coefficients. In contrast to that, our framework ConAD foster learning of many decoupled local density estimators, represented by multiple-hypotheses-predictions (MHP), which have the advantage of (1) higher sample efficiency since no estimation of mixing coefficient is needed and (2) concentration on the local instead of the global neighborhood for outlier detection. Intuitively, the outlier-degree depends highly on the local distribution, samples further away almost do not influence this score.

The MHP-technique was introduced by Guzmán-Rivera et al. (2012) for SSVMs and recently applied to CNN by Rupprecht et al. (2016b) for future frame and pose estimation, Ilg et al. (2018) for uncertainty estimation in optical flow networks, Bhattacharyya et al. (2018) for image sequence predictions and Chen & Koltun (2017) for photo-realistic image synthesis. In these works, they consider the winner-takes-all loss, where only the best-generated hypothesis w.r.t. defined learning objective receives a penalty and therefore a learning signal. In order to make the network generate meaning and interpretable hypotheses, Rupprecht et al. (2016b), Bhattacharyya et al. (2018) and Chen & Koltun (2017) rely on the hypotheses-wise WTA definition, whereby Ilg et al. (2018) reports better performance with pixel-wise WTA-loss, when combined with a smoothing auxiliary task using total-variation. Compared to our ConAD-framework, these approaches were not developed for and applied on distribution learning itself, and there is no mechanism to avoid mode collapse among hypotheses. Furthermore, generated hypotheses could support non-existing data regions, which can be fatal for anomaly detection tasks.

Instead, our framework ConAD employs a discriminator D to assess the quality of generated hypotheses and avoid support of non-existent data modes. ConAD is based on but not restricted to the VAE for an efficient inference step and meaningful manifold learning. Compared to other works, which unify VAE and GAN into one learning framework (Dosovitskiy & Brox, 2016; Larsen et al., 2015), our framework focuses on the multi-hypotheses (i.e., multi-headed) architecture of the generator. Furthermore, ConAD contains the hypotheses discrimination mechanism to foster diversity among hypotheses and therefore increase mode coverage.

Our hypotheses discrimination bases on minibatch discrimination (Salimans et al., 2016a). To avoid the mode collapse problem among hypotheses, the discriminator also receives different hypotheses given the same data point during training. Therefore, a broad consensus among hypotheses is detected and penalized by the discriminator D.

The spirit of our work is similar to LOF, in which the local density is used for outlier detection. However, in contrast to LOF, ConAD requires no explicit dimensions reductions or hyper-parameters for the definition of the local neighborhood and its density estimation at test time. ConAD computes the local densities functions in a one-forward pass, which then can efficiently be used to determine the anomaly score of unseen samples.

3 INCONSISTENCY PROPERTY OF MULTIPLE HYPOTHESES TECHNIQUES

In case of a one-to-many-mapping task with the observable random variable x and multi-modal conditional hidden variable y, a distribution approximator needs the ability to address this multi-modality. A simple problem visualization is in the App. A.

In this context, recent MHP-techniques provides a potential solution. Rupprecht et al. (2016b) employed hypotheses learning with the energy function as described in Eq. 1.

$$E_{MHP}(\Theta) = -\sum_i \sum_h \log\left(p_{\theta_h}(y_i|x_i)\right) * \begin{cases} 1-\epsilon, p_{\theta_h}(y_i|x_i) \geq p_{\theta_k}(y_i|x_i), \forall k \\ \frac{\epsilon}{H-1}, \text{else} \end{cases} \tag{1}$$

Whereby x_i,y_i are corresponding input-output pair from the training dataset, $1 \leq h \leq H$ is a hypothesis branch, which is generated by a parametrized neural network with parameters set θ_h. Furthermore, ϵ is a hyperparameter used to distribute learning signal to the non-optimal hypothesis.

Intuitively, the authors propose to learn conditional hypotheses based on negative-log-likelihood with learning signal dispersion parameter ϵ. Setting $\epsilon = 0$ corresponds to a winner-takes-all (WTA) variant, where only the best-predicted hypotheses receive all learning signal such as proposed by Ilg et al. (2018). However, by learning hypotheses using the WTA-formulation, the generated hypotheses could enter artificial data modes, which are not consistent w.r.t. the underlying data distribution. The problem is derived formally in App. B. Intuitively, assume a perfect fitting of training data using n-hypotheses. By adding additional randomly initialized hypotheses branches to the model, the energy function most likely does not change, because the new hypotheses are unlikely better than the optimal hypotheses from previous training.

The introduction of ϵ parameter alleviate this problem. However, this only represents a trade-off between the inconsistencies by WTA-formulation and uni-hypotheses learning. Intuitively, if $\epsilon = (H-1)/H$, the learning signal is equally distributed to all hypotheses. This is equivalent to learning H-uni-modal-hypotheses independent of each other, which provides no solution for one-to-many-mapping tasks. See App. C for more detail. Therefore, this solution provides no clear guarantee for consistency problem and has to be chosen carefully. In anomaly detection, this is difficult since there is no anomalous data point contained in the training dataset.

Typical density approximation with Autoencoder structure correspond to a one-to-mapping from latent code to the output layer. Note, that given a sufficient network capacity and training time, the network could transform the problem into one-to-one mapping from latent code to output layer. In this case, the reconstructed training images are sharp and not blurry as usual. However, it is a common practice to use bottleneck layers and regulation techniques to enforce generalization, which only focuses on important data characteristics. Therefore, a point in the learned latent space corresponds to many points in image space, with slightly different unimportant nuances.

4 CONSISTENCY-BASED ANOMALY DETECTION (CONAD) FRAMEWORK

In our ConAD framework for anomaly detection, we propose to capture the normal data distribution with multiple-hypotheses-VAE, supported by a discriminator D, which avoids support of non-existing modes by generated hypotheses and at the same time foster mode coverage with hypotheses discrimination. The details are explained in the following.

4.1 MHP-LEARNING WITH VAE

As discussed previously, in this work, we consider density estimation and distribution learning with Autoencoder structure as a one-to-mapping regression problem. However, instead of relying on the estimation of the global density function, we propose to estimate a set of local density functions conditioned on the inputs, which are represented by multiple-hypotheses predictions (MHP). Each conditional hypothesis covers a data mode i.e. dense data region, in which there are many real samples. The learning of different hypotheses is performed based on winner-takes-all-objective as

given in Eq. 2.

$$L_{WTA}(x) = E_{z_i \sim q_\phi(z_i|x)} \left[\log p_{\theta_h}(x|z_i)\right] \textbf{ s.t. } h = \arg\max_j E_{z_i \sim q_\phi(z_i|x)} \left[\log p_{\theta_j}(x|z_i)\right] \quad (2)$$

Whereby L_{WTA} is the winner-takes-all energy function, $1 \leq j \leq H$ indicates the different hypotheses networks, z_i the respective latent code. To reduce free parameters, hypotheses networks with params θ_j share all layers but the last output layer. Intuitive, it means that only the best matching hypothesis receives all of the learning signals from the NLL-loss during training.

An efficient variant to realize MHP in neural networks is by using multi-headed-network, where only the last layer is split to provide different hypotheses. Every other layer are shared as shown in Fig. 2. Our framework is based on the Variational Autoencoder from Kingma & Welling (2013); Rezende et al. (2014) which provides an effective manifold learning and an efficient inference stage with a parameterized encoder q_ϕ.

4.2 Discriminator D to avoid non-existent mode coverage and mode collapse of hypotheses

As discussed earlier in chap. 3, hypotheses generated by the network could support non-existent data regions which are not covered by samples due to the WTA loss. To alleviate this, we propose to match the data density induced by the hypotheses to the real underlying density by learning from a symmetric variant of the Kullback-Leibler divergence (KLD). In detail, we employ the Jensen-Shannon divergence (JSD)-metric for learning by integration of a subsequent discriminator D for quality assessment of generated hypotheses. Fig. 2 illustrates a sample realization with VAE.

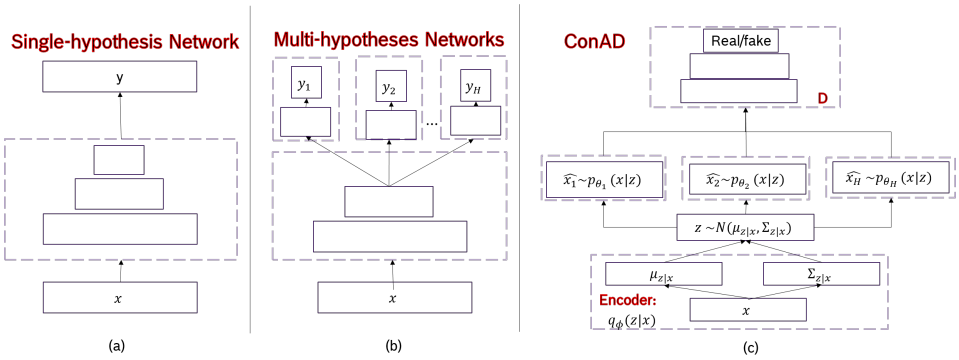

Figure 2: Comparision of single (a) and multi-hypothesis (b) network for a mapping task from input x to output y. Our ConAD framework (c), which integrates a discriminator D to avoid support of non-realistic data modes and foster higher mode coverage with the generated hypotheses.

More concretely, the D and G are in a mini-max game, in which D minimizes Eq. 3.

$$L_D(x, z) = \underbrace{\log(p_D(x_{real}))}_{L_{real}} + L_{fake}(x, z) \quad (3)$$

$$L_{fake}(x, z) = \log(p_D(\hat{x}_{z \sim \mathcal{N}(0,1)})) + \log(p_D(\hat{x}_{z \sim \mathcal{N}(\mu_{z|x}, \Sigma_{z|x})})) + \log(p_D(\hat{x}_{\text{best_guess}})) \quad (4)$$

In this energy formulation, the standard GAN loss is extended to assure quality of generated hypotheses. It comprises two parts: (1) L_{real} which forces D to classify real samples as such using the log-likelihood of real data x and (2) L_{fake} which focuses on identifying generated samples. L_{fake} itself consists of assessment for noise- ($\hat{x}_{z \sim \mathcal{N}(0,1)}$) and data-conditioned ($\hat{x}_{z \sim \mathcal{N}(\mu_{z|x}, \Sigma_{z|x})}$) hypotheses and the best guess given by the WTA objective.

Accordingly, the learning objective for the VAE generator becomes:

$$L_G = L_{Hyps} + KLD(q_\phi(z|x)||\mathcal{N}(0, 1)) - L_D \quad (5)$$

Whereby the right part $-L_D$ is reformulated according to (Goodfellow et al., 2014) for improved learning dynamics.

Name	Problem	Tasks	Train/Valid	Resolution	Test:normal data	Test: anomalies
CIFAR-10	1 vs. 9	10	4500/500	32x32	1000	9000
Metal anomaly	1 vs. 1	1	5408/1352	224x224	1324	346

Table 1: Dataset description.

To address the mode collapse problem of hypotheses, we propose to employ hypotheses discrimination, which is based on minibatch discrimination (Salimans et al., 2016a). In our method, a samples batch contains multiple hypotheses conditioned on an individual data point at the same time. In this way, D learns to distinguish and penalize similarity among hypotheses.

In summary, our framework ConAD proposes multiple-hypotheses learning with VAE, which is supported by a discriminator to avoid support of non-existing data modes and foster mode coverage. Anomaly detection is performed using local NLL (L_{WTA}) in Eq. 2.

5 EXPERIMENTS

5.1 EXPERIMENTS DESCRIPTIONS

The first section starts with the qualitative evaluation of MHP-approaches on a synthetic toy dataset to demonstrate the need for multimodal output capacity and effects of different MHP-techniques. Subsequently, we evaluate the proposed framework regarding anomaly detection performance on CIFAR-10 and a Metal Anomaly dataset provided by a metal parts manufacturer .

The toy dataset: inverted half-moon is a simple one-to-many-mapping task, where a hidden variable y is to be predicted given an observable variable x. This task shows a qualitative analysis of how different uni- and multimodal approaches such as MDN or MHP perform. We choose vanilla VAE (Kingma & Welling, 2013) with Gaussian output distribution as a basic model for evaluation of MHP-techniques. The network architecture follows principles from Radford et al. (2015) and Springenberg (2015). The framework can be easily extended to recent advances in deep generative modeling.

Further quantitative evaluation is done on CIFAR-10 and Metal Anomaly dataset. The typical 10-way classification task in CIFAR is transformed into 10 one vs. nine anomaly detection tasks. Detail can be found at Tab. 1. During model training, only data from the normal data class is used, data from anomalous classes are abandoned. At test time, anomaly detection is measured in Area-Under-Curve of Receiver Operating Curve (AUROC) based on normalized negative log likelihood scores given by the training objective.

On CIFAR-10, we evaluated variants of our multiple-hypotheses approaches including energy formulations: MDN (Bishop, 1994), MHP-WTA (Ilg et al., 2018), MHP (Rupprecht et al., 2016a), ConAD, and MDN+ConAD. We compare our methods against vanilla VAE (Kingma & Welling, 2013; Rezende et al., 2014) , VAEGAN (Larsen et al., 2015; Dosovitskiy & Brox, 2016) as well as results from traditional and deep learning based approaches from previous related works such as Isolation Forest (Liu et al., 2008; 2012), OCSVM (Schölkopf et al., 2001), AnoGAN (Schlegl et al., 2017), AdGAN Deecke et al. (2018). Since the performance of conventional methods suffers due to Curse of Dimensionality (Zong et al., 2018), on the high-dimensional dataset Metal anomaly, we focus only on the evaluation of deep learning techniques. The GAN-techniques proposed by previous work heavily suffer from instability, which leads to just random performance. Therefore, we evaluate MHP-based approaches against their multi-modal counterparts.

5.2 TOY DATASET: FLIPPED HALF-MOON

Fig. 4 shows the flipped half-moon to demonstrate MHP-learning in contrast to unimodal output distribution learning. In this section, Fig 3 shows a qualitative evaluation of different MHP-techniques. This task is a one-to-many mapping with a discontinuity at the point $x = 0$ and $x = 0.5$.

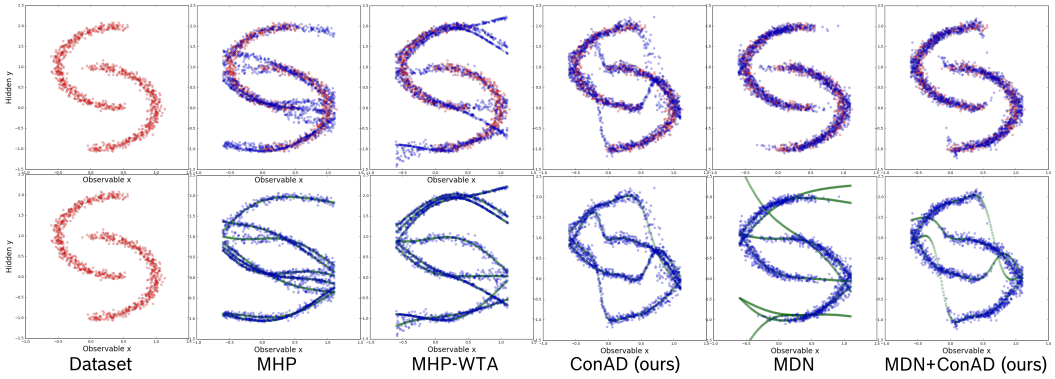

Figure 3: Flipped half-moon dataset: y prediction given observable x. Top: sampled points from learned models. Bottom: conditional μ_i of hypotheses i. Note that this one-to-many mapping is not continuous and has up to 4 modes at the point $x = 0$. When the local data density function abruptly ends, hypotheses from MHP-approaches continue and induce artificial data modes, while our ConAD-framework minimizes this effect. Note that the behaviour in high-dimensional domain is more complex.

Model	AUROC	Model	AUROC	Model	AUROC
KDE-PCA	.590	MHP-2	.619	ConAD - 2 (ours)	**.643**
OC-SVM-PCA	.610	MHP-4	.619	ConAD - 4 (ours)	**.639**
IF	.558	MHP-8	.618	ConAD - 8 (ours)	**.671**
GMM	.585	MHP-16	.617	ConAD - 16 (ours)	**.659**
AnoGAN	.612	MDN+ConAD-2	.616		
ADGAN	.620	MDN+ConAD-4	.621		
VAE	.610	MDN+ConAD-8	.623		
VAEGAN	.609	MDN+ConAD-16	.614		
MDN-2	.609	MHP-2-WTA	.622		
MDN-4	.610	MHP-4-WTA	.622		
MDN-8	.610	MHP-8-WTA	.619		
MDN-16	.609	MHP-16-WTA	.620		

Table 2: Anomaly detection on CIFAR-10 in AUROC. Performance is averaged for 10 times 1 vs. 9 classification and over 3 runs each. See the attachment for detailed performance For MHP-methods, $-n$ indicates the number of hypotheses. Our approach significantly outperforms previous traditional and deep learning methods.

When the local density function abruptly ends, MHP-techniques induce artificial data regions since they are not penalized for artificial modes by the objective function as discussed before. This issue even applies to Mixture-Density-Network, since the mixing coefficients of incorrect hypotheses rapidly approach to zero. In contrast to that, our prosposed consistency-based MHP-method ConAD minimizes the artificial data regions created by the hypotheses generation, so that even non-optimal hypotheses quickly reenter realistic data modes.

Note that on high-dimensional inputs with highly correlated features, these methods behave differently. While MDN assigns low likelihood to samples with extreme dimensions because of the global density function, MHP-approaches are less sensitive to outlier dimensions due to local density estimation.

5.3 CIFAR-10

Tab. 2 shows an extensive evaluation of different traditional and deep learning techniques. Results are adapted from Deecke et al. (2018) in which the training and testing scenarios were similar. Refer to App. D for more results. Traditional, non-deep-learning methods only succeed to capture classes with a dominant homogeneous background such as ships, planes, frogs (backgrounds are water, sky,

Model	AUROC	Model	AUROC	Model	AUROC
MDN-2	.900	MHP-2	.980	VAE	.942
MDN-4	.910	MHP-4	.970	VAEGAN	.936
MDN-8	.916	MHP-8	.950		
MDN+ConAD-2	.942	MHP-WTA-2	.980	ConAD-2	**.985**
MDN+ConAD-4	.913	MHP-WTA-4	**.980**	ConAD-4	.977
MDN+ConAD-8	.943	MHP-WTA-8	.946	ConAD-8	**.965**

Table 3: Anomaly detection performance on Metal Anomaly dataset. In order to avoid noisy residuals due to high-dimensional input domain, only 10% of maximally abnormal pixels are summed to form the NLL-score used for anomaly detection. For more detailed results, refer to attachment E. Anomaly detection performance rapidly breaks down with increasing number of hypotheses.

green nature respectively). This issue occurs due to preceding features projection of PCA, which focus on dominant axes with large variance. Deecke et al. (2018) reported that even discriminative features from pretrained AlexNet have no positive effect on anomaly detection performance.

In contrast to that, deep learning methods are performing significantly better, even without careful parameter tuning. When the MHP-technique is applied to this task, performance comparable to previously reported results is achieved. Note that having the multiple output distributions is not sufficient to meet high performance: MDNs are performing worse than the relaxation of density estimation provided by the MHP-technique. Nevertheless, the best performance is achieved in our ConAD- framework, by utilizing the flexibility of multiple hypotheses more effectively, leading to significantly higher detection performance of up to 5.1%.

5.4 Metal Anomaly dataset

Tab. 5.4 shows an evaluation of MHP-methods against density-learning methods such as VAE (Kingma & Welling, 2013), MDN (Bishop, 1994), VAEGAN (Dosovitskiy & Brox, 2016; Larsen et al., 2015) which corresponds to our ConAD with single hypothesis. The significant improvement of up to 3.8% AUROC-score comes from our relaxation of density estimation into local density estimation in the spirit of LOF (Breunig et al., 2000), i.e., each dense data region (mode) receive at least one hypothesis to cover the local density.

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 as discussed earlier. Intuitively, additional non-optimal hypotheses are not strongly penalized during training, if they induce artificial data regions which are not consistent w.r.t. the real underlying data distribution.

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 less sensitive to the number of hypotheses to be used. The hypotheses generated by the networks tend to remain in realistic regions due to the critics provided by the D.

6 Conclusion

In this work, we propose for anomaly detection a relaxation of density estimation with deep neural network by using multiple-hypotheses predictions (MHP)-technique. We identify the inconsistency property of MHP w.r.t. the real underlying distribution, i.e., that additional hypotheses can support non-existing data regions without being penalized during training. Therefore we propose a consistency-based anomaly detection-framework with MHP (ConAD), in which a discriminator D supports G to create realistic and diverse hypotheses with our newly adapted technique: hypotheses discrimination. This results in up to 5.1% points AUROC improvement on CIFAR-10 and alleviates performance breakdown when the number of hypotheses is increased.

REFERENCES

Apratim Bhattacharyya, Bernt Schiele, and Mario Fritz. 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.

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

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pp. 93–104. ACM, 2000.

Qifeng Chen and Vladlen Koltun. 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.

Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Ruslan R Salakhutdinov. Good semi-supervised learning that requires a bad gan. In *Advances in Neural Information Processing Systems*, pp. 6510–6520, 2017.

Lucas Deecke, Robert Vandermeulen, Lukas Ruff, Stephan Mandt, and Marius Kloft. Anomaly detection with generative adversarial networks. 2018.

Debadeepta Dey, Varun Ramakrishna, Martial Hebert, and J Andrew Bagnell. Predicting multiple structured visual interpretations. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2947–2955, 2015.

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Abner Guzmán-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25*, pp. 1799–1807. Curran Associates, Inc., 2012.

Eddy Ilg, Özgün Çiçek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. 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.

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

Mark Kliger and Shachar Fleishman. Novelty detection with gan. *arXiv preprint arXiv:1802.10560*, 2018.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

Kimin Lee, Changho Hwang, KyoungSoo Park, and Jinwoo Shin. Confident multiple choice learning. *arXiv preprint arXiv:1706.03475*, 2017.

Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. Stochastic multiple choice learning for training diverse deep ensembles. In *Advances in Neural Information Processing Systems*, pp. 2119–2127, 2016.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422. IEEE, 2008.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):3, 2012.

Sergey Prokudin, Peter Gehler, and Sebastian Nowozin. Deep directional statistics: Pose estimation with uncertainty quantification. *arXiv preprint arXiv:1805.03430*, 2018.

Alec Radford, Luke Metz, and Soumith Chintala. 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.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D. Hager. Learning in an Uncertain World: Representing Ambiguity Through Multiple Hypotheses. *arXiv:1612.00197 [cs]*, December 2016a. URL http://arxiv.org/abs/1612.00197. arXiv: 1612.00197.

Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D. Hager. Learning in an Uncertain World: Representing Ambiguity Through Multiple Hypotheses. *arXiv:1612.00197 [cs]*, December 2016b. URL http://arxiv.org/abs/1612.00197. arXiv: 1612.00197.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016a.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. *arXiv:1606.03498 [cs]*, June 2016b. URL http://arxiv.org/abs/1606.03498. arXiv: 1606.03498.

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*, pp. 146–157. Springer, 2017.

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

Jost Tobias Springenberg. Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks. *arXiv:1511.06390 [cs, stat]*, November 2015. URL http://arxiv.org/abs/1511.06390. arXiv: 1511.06390.

David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54(1): 45–66, 2004.

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. *International Conference on Learning Representations.*, 2018.

A ONE-TO-MANY MAPPING TASKS REQUIRE MULTI-MODALITY

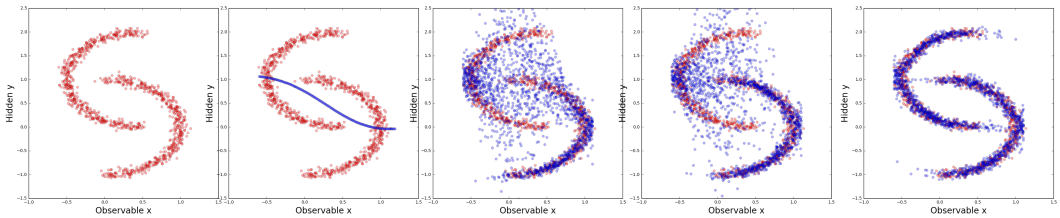

Figure 4: From left to right: (a) Flipped half-moon data-set: mapping from x to y is not unique. E.g. for $x = 0$, the conditional output distribution $p_{data}(y|x = 0)$ has 4 different modes, where points are concentrated b) data sampled from an unimodal neural regressor fails to capture the underlying data distribution (c)-(e) Multimodal regressors: increasing number of hypotheses up to 4 allows the network to fully capture the conditional data distribution

Consider a simple toy problem with an observable x and hidden y which is to be predicted and expressed by the conditional distribution $p_{true}(y|x)$ such as in Fig. 4. Since the data conditional is multi-modal for some x, an uni-modal output distribution cannot fully capture the underlying distribution. Instead, the bias-free solution for the MSE-minimizer is the empirical mean $\overline{y_{x_i}}$ of $p_{train}(y|x_i)$ on the training set. However, this learned conditional density does not comply with the underlying distribution; their sampled data point fall into the low-likelihood regions under $p_{true}(y|x)$. With increasing number of output hypotheses, the data modes could be gradually captured. For this task, the energy to be minimized is given by the Negative-log-likelihood of the Mixture Density Network (MDN) under a Gaussian Mixture with hypotheses h in Eq. 7 :

$$E_{MDN}(\Theta) = -\log L(\Theta|X;Y) = -\log p_{GMM}(Y|X,\Theta) = -\sum_i \sum_h \log \alpha_h p_{\theta_h}(y_i|x_i) \tag{6}$$

with

$$p_{\theta_h}(y_i|x_i,\theta) = \frac{1}{\sqrt{2\pi}\sigma_h} \exp -\frac{(y_i-\mu_h)^2}{2\sigma_h^2} \tag{7}$$

B LEMMA 4.1

Given a sufficient number of hypotheses H', an optimal solution Θ^* for $E_{WTA}(\Theta^*)$ is not unique (permutation is excluded). There exists a $\Theta^{'}$ with $E_{WTA}(\Theta^*) = E_{WTA}(\Theta^{'})$ which is not consistent w.r.t. the underlying output distribution $p_{train}(y_i|x_i)$.

Proof. : Suppose c is the maximal modes count of the dataset sampled from the real underlying conditional output distribution $p(y_i|x_i)$. Since $|\{(x_i, y_i)\}| < \infty \rightarrow c < \infty$.

Suppose $H = c$ then a trivial optimal solution for $E_{WTA}(\Theta_H)$ is found by center each hypothesis μ_{ik} at an different empirical data point k $y_{ik} \sim (y_i, x_i)$ and $\sigma_{ik} \mapsto 0$. In this case $\lim_{\sigma_{ik}\mapsto 0;\forall i,k} E_{WTA}(\widehat{\Theta}_H) = 0$.

Suppose $H' > c$, then a solution $\widehat{\Theta}_{H'}$ can be formulated s.t.: $E(\widehat{\Theta}_H) = E(\widehat{\Theta}_{H'})$.

Let $\widehat{\Theta}_{H'} = \widehat{\Theta}_H \cup \widehat{\Theta}_{H+1\ldots H'} = \widehat{\Theta}_H \cup \{\theta_{h+1} \ldots \theta_{h'}\}$ for some **random** $\widehat{\Theta}_{H+1\ldots H'}$. Due to randomness and without loss of generality, one can assume that $\forall (x_i, y_i), \forall \theta_i \in \Theta_{H+1\ldots H'}$, θ_i is not the optimal hypothesis for any training point $(x_i, y_i) \in D_{train}$.

In this case due to the winner-takes-all energy formulation we have:

$$E_{WTA}(\widehat{\theta}_{H'}) = -\sum_i \max_{1\leq h\leq H'} \log p_{\theta_h}(y_i|x_i) = -\sum_i \max_{1\leq h\leq H} \log p_{\theta_h}(y_i|x_i) = E_{WTA}(\widehat{\theta}_H) \tag{8}$$

So $\widehat{\Theta}_H$ and $\widehat{\Theta}_{H'}$ with $H' > H$ are both solutions to the loss formulation and share the same energy level. The extended hypotheses can support arbitrary artificial data regions without being penalized. □

C LEMMA 4.2

Lemma C.1. *Similar to Lemma B, minimizing on E_{MHP} in Eq. 1 might also lead to inconsistent approximation of real underlying output distribution.*

Proof. First, note that $0 \leq \epsilon \leq \frac{H-1}{H}$, since $\epsilon < 0$ would push away non-locally optimal hypotheses from the empirical solution, $\epsilon > \frac{H-1}{H}$ would penalize best hypothesis more than others. Both are undesired properties of MHP-Learning. First consider the case where $\epsilon \mapsto \frac{H-1}{H}$:

$$\lim_{\epsilon \mapsto \frac{H-1}{H}} E_{MHP}(\Theta) = \sum_i \sum_h \log\left(p_{\theta_h}(y_i|x_i)\right) * \frac{1}{H} \tag{9}$$

$$= \frac{1}{H} \sum_h \left(\sum_i \log\left(p_{\theta_h}(y_i|x_i)\right) \right)$$

$$= \frac{1}{H} \sum_h E_{\theta_h}$$

$\forall \theta_h$ and training data points (x_i, y_{ik}) the optimal least-square solution is the geometrical mean, therefore we have:

$$\theta_h^*(y_i|x_i) = E_{y_{ik \sim p(y|x_i)}[y_i]}$$

$$= \frac{1}{l} \sum_{i=1}^{l} y_i; y_{ik} \sim p(y_i|x_i)$$

In this case, all hypotheses are optimized independently and converge to the same solution similar to Single-hypothesis-approach. The resulting distribution is inconsistent w.r.t to real output distribution (see Fig. 4 for a sample).

Now consider $\epsilon \mapsto 1$:

$$\lim_{\epsilon \mapsto 1} E_{MHP}(\Theta) = -\sum_i \sum_h \log\left(p_{\theta_h}(y_i|x_i)\right) * \begin{cases} 1; \text{if } \theta_h \text{ is best hypothesis} \\ 0; \text{else} \end{cases}$$

$$= -\sum_i \max_{1 \leq h \leq H'} \log p_{\theta_h}(y_i|x_i)$$

$$= E_{WTA}(\Theta)$$

In this case E_{MHP} shares the same inconsistency property with E_{WTA}. Consequently, choosing $\epsilon \in [0, \frac{H-1}{H}]$ only smooths penalty on suboptimal hypothesis. The risk remains that distributions induced by non-optimal hypotheses are beyond the real modes of underlying distribution.

□

D CIFAR-RESULTS

E METAL ANOMALY DATASET

CIFAR-10	0	1	2	3	4	5	6	7	8	9	Mean
KDE-PCA	.705	.493	**.734**	.522	.691	.439	.771	.458	.595	.490	.590
KDE-Alexnet	.559	.487	.582	.531	.651	.551	.613	.593	.600	.529	.570
OC-SVM-PCA	.666	.473	.675	.530	.827	.438	**.787**	.532	.720	.453	.610
OC-SVM-Alexnet	.594	.540	.588	.575	.753	.558	.692	.547	.630	.530	.601
IF	.630	.379	.630	.408	**.764**	.514	.666	.480	.651	.459	.558
GMM	.709	.443	.697	.445	**.761**	.505	.766	.496	.646	.384	.585
AnoGAN	.610	.565	.648	.528	.670	.592	.625	.576	.723	.582	.612
ADGAN	.632	.529	.580	**.606**	.607	**.659**	.611	.630	.744	.644	.62
VAE	.771	.467	.684	.538	.71	.542	.642	.512	**.765**	.467	.610
VAEGAN	.762	.469	**.697**	.520	.756	.536	.588	.554	.754	.460	.609
MDN-2	.761	.469	.687	.538	.704	.538	.632	.523	**.768**	.467	.609
MDN-4	.769	.468	.686	.535	.693	.544	.635	.541	.76	.469	.610
MDN-8	.762	.469	.686	.533	.704	.547	.633	.53	.763	.473	.61
MDN-16	.762	.479	.682	.528	.701	.54	.635	.529	.764	.469	.609
MHP-WTA-2	.773	.516	.68	.552	.695	.543	.643	.555	.76	.512	.622
MHP-WTA-4	**.778**	.539	.651	.567	.66	.542	.635	.563	.752	.541	.622
MHP-WTA-8	.761	.56	.627	.588	.626	.553	.614	.578	.743	.548	.619
MHP-WTA-16	.757	.567	.609	.598	.627	.56	.61	.568	.738	.573	.62
MHP-2	.755	.499	.676	.546	.693	.543	.636	.577	.764	.508	.619
MHP-4	.752	.51	.66	.568	.677	.551	.644	.56	.764	.51	.619
MHP-8	.757	.54	.652	.576	.648	.554	.625	.547	.759	.53	.618
MHP-16	.758	.539	.641	.585	.646	.552	.623	.545	.759	.532	.617
MDN+ConAD-2	.746	.489	.686	.521	.711	.525	.668	.577	.765	.481	.616
MDN+ConAD-4	.762	.504	.69	.524	.716	.532	.659	.583	.753	.489	.621
MDN+ConAD-8	.774	.483	.693	.531	.722	.537	.679	.54	.76	.519	.623
MDN+ConAD-16	.736	.469	.694	.522	.753	.541	.657	.568	.753	.454	.614
ConAD - 2 (ours)	.773	.600	.666	.562	.694	.561	.706	.630	.748	.499	.643
ConAD - 4 (ours)	**.776**	.525	.663	.570	.687	.541	**.801**	.548	.741	.539	.639
ConAD - 8 (ours)	.774	**.652**	.648	.601	.670	.579	.725	**.662**	.748	**.660**	**.671**
ConAD - 16 (ours)	.772	**.631**	.631	**.615**	.633	**.588**	.691	**.640**	.755	**.637**	**.659**

Table 4: 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-1-methods are marked. Our ConAD approach outperforms traditional methods and vanilla MHP-approaches significantly and could benefit from increasing number of hypothesis. Furthermore, Mixture Density networks performs similarly to single-modal output distribution of VAE.

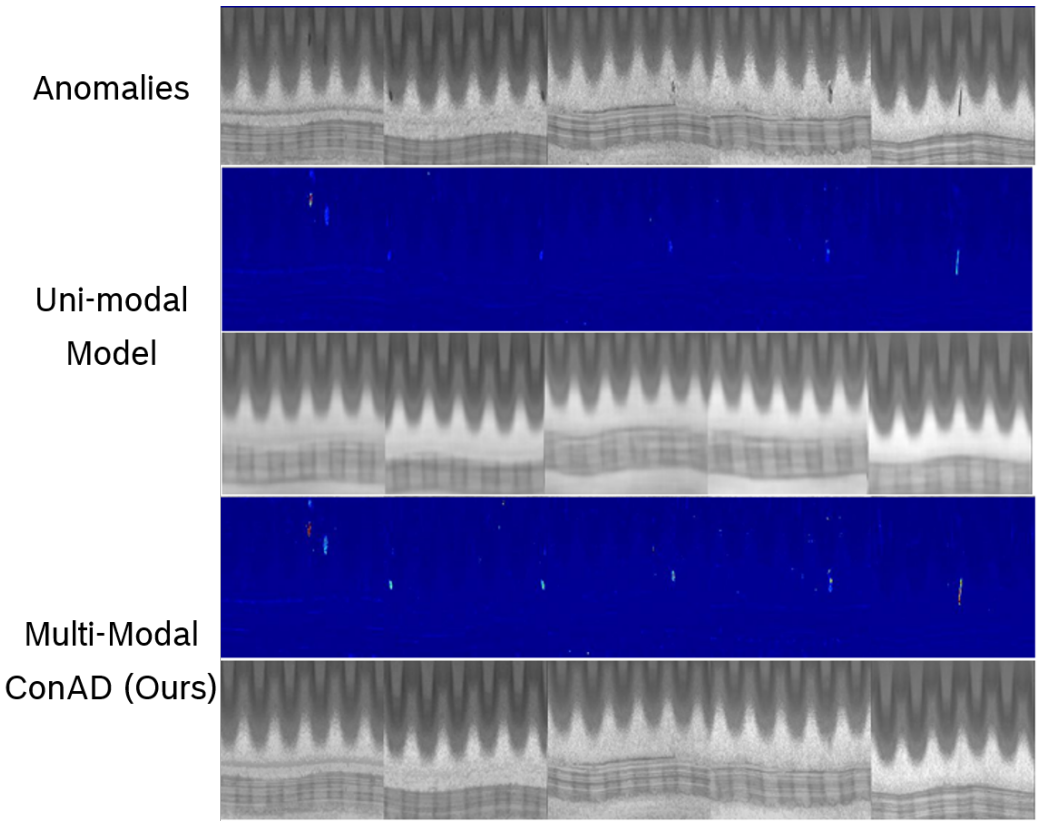

Figure 5: Image Reconstructions: Reconstructions from Uni-modal models are blurry at convergence. Using our ConAD-approach (last two rows), the maximally consistent reconstruction is closer to the original, capturing many details needed to differentiate between normal data noise and real anomalies, such as black spots or scratches. The likelihood maximizer in the hypotheses space is much closer to the original and also more realistic. The error residuals are also significantly clearer for our ConAD-method.

Model	NLL-All-pixels	10% abnormal pixels	1%-abnormal pixels
VAE	.795	.942	.977
VAEGAN (1-hyp)	.782	.936	.978
MDN-2	.746	.900	.970
MDN-4	.765	.910	.960
MDN-8	.743	.916	.975
MDN+ConAD-2 (ours)	.810	.942	.966
MDN+ConAD-4 (ours)	.781	.913	.951
MDN+ConAD-8 (ours)	.810	.943	.978
MHP-2	.876	.980	**.993**
MHP-4	.834	.970	.990
MHP-8	.793	.950	.984
MHP-WTA-2	.851	.980	.990
MHP-WTA-4	**.878**	.980	.990
MHP-WTA-8	.800	.946	.981
ConAD-2 (ours)	.867	**.985**	**.992**
ConAD-4 (ours)	.812	.977	.990
ConAD-8 (ours)	.817	.965	.987

Table 5: anomaly detection performance on data-set anomaly A in AUROC.anomaly detection on anomalyA measured in AUROC: How different Multiple Hypothesis approaches perform with increasing number of hypotheses. Vanilla single-hypothesis approach such as VAE and VAE+GAN under-perform on this task. Even with more sophisticated multi-modal output distribution capacity (MDN), the discriminability is not improved. The integration of MDN into the GAN-framework only slightly improves the results. On the other hand, all other MHP-approaches perform similarly well $> 99\%$ AUROC (at 1% of most abnormal pixels considered), which indicates that the task has become easily solvable for these methods.