Simple Adaptive Projection with Pretrained Features for Anomaly Detection

Xintai Gui
University of Electronic Science and Technology of China
tabgui@std.uestc.edu.cn

Abstract
Deep anomaly detection aims to separate anomaly from normal samples with high-quality representations. Pretrained features bring effective representation and promising anomaly detection performance. However, with one-class training data, adapting the pretrained features is a thorny problem. Specifically, the existing optimization objectives with global target often lead to pattern collapse, i.e. all inputs are mapped to the same. In this paper, we propose a novel adaptation framework including simple linear transformation and self-attention. Such adaptation is applied on a specific input, and its k nearest representations of normal samples in pretrained feature space and the inner-relationship between similar one-class semantic features are mined. Furthermore, based on such framework, we propose an effective constraint term to avoid learning trivial solution. Our simple adaptive projection with pretrained features (SAP2) yields a novel anomaly detection criterion which is more accurate and robust to pattern collapse. Our method achieves state-of-the-art anomaly detection performance on semantic anomaly detection and sensory anomaly detection benchmarks including 96.5% AU-ROC on CIFAR-100 dataset, 97.0% AUROC on CIFAR-10 dataset and 88.1% AUROC on MvTec dataset.

Keywords Anomlay Detection · Pretrained Features · Simple Adaptive Projection · Self-attention

1 Introduction
Anomaly detection (AD) is a specific task designed to detect anomalies through data-based models or algorithms and an anomalous is an observation that deviates considerably from the concept of normality [1]. Such deviation can happen due to either covariate shift or semantic shift, while assuming the other distribution shift do not exist and these shifts leads to two sub-tasks: sensory AD and semantic AD, respectively [2].

In standard AD settings, labeled anomalous data are often nonexistent and only normal data are accessible which means all given training samples are normal samples. In this case, self-supervised method based on auxiliary tasks, or unsupervised method like autoencoder are widely used in AD. However, for self-supervised methods, the features trained on auxiliary domains may not generalize well to the target domain and for autoencoder-based methods, their good generalization performance can reconstruct the abnormal inputs well and lead to a misjudgment on anomalies.

Due to the restriction of one-class training data, the learned representation is indistinguishable to some extent resulting in the limitation of anomaly detection performance. Recently, anomaly detection based on pretrained features have been widely studied. Some works [3, 4] consider leveraging knowledge distillation to only transfer the pretrained features of anomaly-free data to the student. However, they do not adapt the features to the target data set. PANDA [5] proposes an AD baseline based on pretrained network, and implements finetune referring to Deep-SVDD[6]. But such finetune method will cause pattern collapse which means all features will shrink to the center point. PANDA leverages elastic weight consolidation to restrict the weight change in fine-tune. However, it still needs to further pretrained on an auxiliary task to obtain a Fisher information matrix.

In this work, we introduce a novel framework for adapting the pretrained feature to target data and anomaly detection scenario. In order to avoid potential pattern collapse, the framework abandons the traditional optimization with global parameter such as “c” in Deep-SVDD and PANDA but adopts a strategy of paying attention to normal patterns partly.
Simple Adaptive Projection with Pretrained Features for Anomaly Detection

For a specific input images, the top-k nearest normal representations in pretrained feature space are traced. Instead of finetuning layers in pretrained network, we propose a simple adaptive projection head which can play a role in adapting pretrained features properly. It maps the input and its top-k nearest normal pretrained features into an adaptive space and self-attention [7] is leveraged to learn the weights of projected normal features whose weighted average is regarded as a normal representation of the input. The similarity between the normal representation and projected feature of the input is optimized in training phase and such similarity is regarded as a criterion of anomalies.

We summarize our contributions as follows: i) we design a simple adaptive projection with pretrained feature (SAP2) framework for anomaly detection. Instead of finetuning the large pretrained network, with a simple projection head, the pretrained features are adapted for anomaly detection. ii) we propose the similarity between and projected feature of a specific input and its normal representation obtained by SAP2, as a criterion for anomaly detection. Consequently, without comparing global parameter, it can reduce the risk of pattern collapse in optimization. iii) extensive experiments and visualization results validate the effectiveness of our proposed framework for both sensory AD and semantic AD.

2 Related work

Unsupervised Anomaly Detection  In practical application scenarios, abnormal situations rarely occur which will lead to our network can only be trained through normal samples, so anomaly detection is widely described as unsupervised learning problem. Among the traditional anomaly detection methods, KNN method[8] and K-means[9] method are often used to detect outliers; such as OC-SVM[10] to seek a hyperplane to divide the normal points and outliers and SVDD[11] adopts the method of training a hypersphere some one-classification methods are used to solve this problem; based on statistical methods, such as Gaussian Mixture Model[12], a suitable probability models are established to fit the characteristic distribution of data. However, it is difficult for traditional unsupervised learning methods to extract effective data features to identify abnormal samples when facing high-dimensional data samples, such as image anomaly detection tasks.

Transferring Pretrained Representations  Pretrained Representations: In the development of deep learning, the performance of the model often matches the depth of the network and the size of the training data. Data annotation is an extremely resource-consuming task, so transfer learning[13] was proposed in 2010 to solve this problem. The pre-trained networks performed better in all areas of work than unsupervised learning methods. Huh et al.[14] demonstrated that depth feature representations pre-trained on ImageNet can improve metrics on target tasks, even though the data sets in these tasks are largely irrelevant to ImageNet. Nazare et al.[15] attempted to use pre-trained CNN features to detect abnormal behaviors in videos. Cohen et al.[16] proved that the performance of many self-supervised learning methods applied in image anomaly detection in recent years is far inferior to the simple method based on KNN with pretrained strong feature extractor. Salehi et al.[17] designed an adaptive anomaly location method based on gradient, which utilized the different behaviors of the clone network and the original expert network to detect and locate anomalies. Reiss et al.[5] proposed an approach to adapt pre-training features and suggested some training methods to overcome catastrophic collapse.

3 Our Approach

Our SAP2 framework designs a simple adaptive projection for adapting pretrained features on target data for anomaly detection. The whole framework is illustrated in Figure[1] which includes a KNN module, an adaptive project head module and self-attention module.

3.1 kNN

For a specific input, SAP2 firstly traces k normal pretrained representations that most similar to its pretrained feature and k-nearest-neighbor(kNN) is an effective method. In the specific implementation, instead of exhaustively computing pretrained representations every time, a memory bank is used to store the pretrained features of all normal training samples. The memory items are obtained through a pretrained large network $G_P$ (e.g. ResNet152) which is regarded as an encoder. A memory addressing module is leveraged to find k most similar items to a specific input.

Pretrained Memory Bank  The memory bank $M \in \mathbb{R}^{N \times D}$ is a real-valued matrix containing $N$ items of fixed dimension $D$ and $m_i$ represents a memory item. The memory items $m_i$ is obtained from the encoder $G_P$, $m_i = G_P(x_i^{train})$ where $x_i^{train}$ is a training normal image.
Simple Adaptive Projection with Pretrained Features for Anomaly Detection

Figure 1: The architecture of proposed SAP2 framework.

**Top-K Memory Addressing** For a specific image $x_i$, we first compute its pretrained feature $z_i = G_p(x_i)$, and then compare it against all the pretrained memory items. With a cosine similarity $s_{j,i} = \cos(m_j, z_i)$, the $k$ nearest neighbors in pretrained space are recorded in a subset $M_i \in \mathbb{R}^{K \times D}$ and would then be used to construct a normal representation. In the training phase, it should be noted that $M$ includes pretrained encodes of all training samples and the most nearest item of each $z_i$ is the corresponding $m_i$. Therefore, $M_i$ should contain top-k nearest memory items except $m_i$.

### 3.2 Simple Adaptive Projection Head

Instead of finetuning some layers of the pretrained large network, we add an adaptive projection head $G_\theta$ to obtain representation adapted to the target data. [18, 19] proposed that a function including linear transformation and nonlinear activation function can effectively improve the representation ability. In SAP2 model, a simple adaptive projection head including MLP block is leveraged to finetune $z_i$ and $M_i$ simultaneously and get their corresponding mapping

$$\hat{z}_i = G_\theta(z_i); \hat{M}_i = G_\theta(M_i)$$

MLP block includes linear transformation and various nonlinear activation functions. We have found and proved experimentally that in anomaly detection task, a simple one-layer mapping without activation function is more effective than multi-layer mapping. Moreover, for better adapting to target data, the structure of adaptive projection head containing only one linear layer needs to be further considered. These contents will be described in detail later.

### 3.3 Self-attention

For K nearest neighbors, the general way to calculate its corresponding feature is with mean value or calculates the weight through cosine similarity with a query. However, the relationship between these items is not considered. In SAP2 model, beyond the relationship with the input, we hope that the nearest k adapted representations in $M_i$ can learn the weight considering their inner-relationship at the same time. [21] proposed self-attention to learn the relationship in sequences paralleled and [20] applies a batch-attention module to capture the discriminative information from similar objects. We consider that the subset $M_i$ contains the information that can infer $\hat{z}_i$, naturally, we reform the self-attention mechanism to adapt to anomaly detection scenario.

$$A(M_i; W_Q, W_K) = \text{Softmax}(\frac{(\hat{M}_i W_Q)(\hat{M}_i W_K)^T}{\sqrt{D}})\hat{M}_i$$

$W_Q, W_K \in \mathbb{R}^{D \times D}$ are linear transforms and add sufficient expressive power for normal features in adaptive space.

In addition to the such architecture, a residual connection and average operation are inserted and the output of such
We make a hypothesis that the nearest k items in the pretrained space are still the nearest ones in the adapted space. Thus, we look upon $\hat{z}_i^n$ as the normal representation corresponding to $\hat{z}_i$. Naturally, like Figure II right, for a normal sample, $\hat{z}_i$ would be similar to $\hat{z}_i^n$ while for an anomaly, their dissimilarity is large. In the training phase, since all accessible samples are normal, we take the similarity between $\hat{z}_i$ and $\hat{z}_i^n$ as optimization target.

$$L_S = \frac{1}{n} \sum_{i=1}^{n} \left(1 - \frac{\hat{z}_i \cdot \hat{z}_i^n}{||\hat{z}_i|| ||\hat{z}_i^n||} \right)$$

Unlike the loss function mentioned in Deep-SVDD and PANDA where reduces the distance between training data and a global target, such loss function optimizes the similarity between a specific input and its pretrained kNN features in the adaptation space. For such loss function, SAP2 may tend to learn another trivial solution. However, for SAP2 framework, a simple and practicable constraint is proposed to effectively to avoid such potential risks. Here we theoretically demonstrate the potential trivial solution in SAP2 optimization if only consider $L_S$ and the feasible method.

**Proposition** Let $G_\theta$ be the all-zero network weights. For this choice of parameters, the adaptive projection head maps any input to the same output, i.e. $G_\theta(z_i) = 0$.

Since $G_\theta$ is a simple linear projection, $G_\theta = 0$ is a potential solution. As the output of the all-zero-weights projection is zero for every input, $L_S$ will become meaningless, however, $L_s$ would optimize $G_\theta$ into a sparse matrix tending to zero which greatly reduces the representation capacity of model, since with such solution, the loss error tends to be zero.

To avoid such trivial solution, we design $G_\theta$ as a square matrix ($G_\theta \in \mathbb{R}^{D \times D}$) and propose a constraint term:

$$\Omega = \frac{1}{n} \sum_{i=1}^{n} \left(1 - \frac{z_i \cdot \hat{z}_i}{||z_i|| ||\hat{z}_i||} + ||z_i - \hat{z}_i||^2_2 \right)$$

Where we consider the cosine similarity and Euclidean distance of $z$ and $\hat{z}_i$ simultaneously. This encourages $\hat{z}_i$ be not only close to the $z$ in terms of Euclidean distance but also be in the same direction. Such constraint, on the one hand, prevents the solution from approaching zero matrix, on the other hand, retain the pretrained information to a certain extent. Using the aforementioned loss and constraint, $L_{total}$ is formulated as

$$L_{total} = L_s + \lambda \Omega$$

For a given test image, we can naturally define an anomaly score by the similarity between its adaptive mapping $\hat{z}_i$ and normal representation $\hat{z}_i^n$

$$s(x_i) = 1 - \frac{\hat{z}_i \cdot \hat{z}_i^n}{||\hat{z}_i|| ||\hat{z}_i^n||}$$

### 4 Experiments

In our experiments, all input images are resized to 224×224 pixels to fit pretrained network, and intensity values are normalized to [-1,1]. ResNet152 pretrained on ImageNet is leveraged as the pretrained network which is same as PANDA[5]. Four datasets including CIFAR10, CIFAR100, FMNIST and MvTec are used as the benchmarks. The first three datasets is semantic AD, and in training phase, for a specific dataset, only one class is visible which is regarded as normal and in testing phase, all classes are accessible. MvTec is sensory AD and the images without anomaly region are trained.
4.1 Comparison with State-of-the-art

In this section, we present anomaly detection results of our SAP2 framework compared to the state-of-the-art deep learning based method: Deep-SVDD [6], MHRot [21], Distillation [4], DN2 and PANDA [5]. Deep-SVDD is the widely accepted deep-learning based method, MHRot is a high-performance self-supervised method, Distillation, DN2 and PANDA are pretrain based methods. All the results of others that are available in the original papers are copied exactly. To verify the effectiveness of SAP2, we additionally compare the results without adaptation whose anomaly score is calculated with the similarity between the feature of a specific input and mean value of its kNN features in pretrained space.

| Method         | CIFAR10 | CIFAR100 | FMNIST | MvTec |
|----------------|---------|----------|--------|-------|
| Deep-SVDD [6]  | 64.8    | 67.0     | 84.8   | 77.9  |
| MHRot [21]     | 90.1    | 80.1     | 93.2   | 65.5  |
| Distillation [4]| 87.2    | -        | 94.5   | 87.7  |
| DN2 [5]        | 92.5    | 94.1     | 94.5   | 86.5  |
| PANDA [5]      | 96.2    | 94.1     | **95.6** | 86.5 |
| SAP2 (no adaption) | 96.6    | 96.0    | 95.1   | 86.3  |
| SAP2           | **97.0** | **96.5** | 95.5   | **88.1** |

The performances of these methods in different dataset are reported in Table 1. We can make the following observations from this table: i) Proposed anomaly detection criterion achieve a significantly improvement. Specifically, without adaptation, the anomaly detection performance is better than DN2 which is also a baseline pretrained based method without adaptation. ii) Feature adaptation of a simple adaptive projection head improves the performance. For sensory AD datasets, the detection results are about 0.4% higher and for semantic AD dataset, it is 1.8% higher than method without adaptation. iii) Our SAP2 framework prominently outperforms the state-of-the-art pretrained based method PANDA (achieving around a 0.8%, 2.4% and 1.6% higher on the CIFAR10, CIFAR100 and MvTec datasets) while for FMNIST, SAP2 perform basically the same. Therefore, it demonstrates the effectiveness of the proposed adaptation strategy and anomaly criterion both on semantic AD and sensory AD.

4.2 Design of Adaptive Projection Head

For the structural design of adaptive projection head, we refer to the simple MLP block. Concretely, The layer number and nonlinear function would greatly affect the performance. We compare performance with different structure on a semantic AD dataset and a sensory AD dataset.

| Dataset | L+ReLU | L+ReLU | L+ReLU | L+ReLU | L+ReLU | L+ReLU |
|---------|--------|--------|--------|--------|--------|--------|
| CIFAR10 | ✓      | ✓      | ✓      | ✓      | ✓      | 96.4   |
|         |        |        |        |        | ✓      | 96.8   |
|         |        |        |        |        | ✓      | **97.0** |
|         |        |        |        |        | ✓      | 96.8   |
| MvTec   | ✓      | ✓      | ✓      | ✓      | ✓      | 87.4   |
|         |        |        |        |        | ✓      | 87.8   |
|         |        |        |        |        | ✓      | **88.1** |
|         |        |        |        |        | ✓      | 87.6   |

The performances of different head designs are reported in Table 2. The result shows that for anomaly detection scenario, the simpler the structure of projection head is, the more effective the adaptation would be which is contrary to general cognition. In the traditional deep learning task, deep model has a certain generalization capability, while for anomaly detection, generalization is disadvantageous where we hope feature adaptation only works in normal data. Thus, in all subsequent experiments, the one-layer linear mapping is leveraged as the adaptation function.

4.3 Effect of Self-attention

Self-attention mechanism aims to make the normal representation takes inner-relationship into account. We visualize the effect of self-attention module in Figure 2 where a specific input image, its kNN images in pretrained space and
the weights with or without self-attention are displayed. Moreover, performance with self-attention or not are shown in Table 2.

![Figure 2: The visualization of the effect of self-attention.](image)

Traditionally, the weights of k nearest features are often obtained by calculating the similarity with the input. However, we believe that kNN contains important inner-relationship in one-class setting. Like words in a sentence, input feature of a specific image should be inferred through the inner-relationship of its kNN features. The first weight bar illustrates that weights without self-attention are approximately same while the second bar shows the weights with self-attention have obvious discrimination. Moreover, these weight implies that models have consider semantic inner-information of the k nearest images. Specifically, The two items with larger weights are semantically closest to the input (black cars), while the items with smaller weights are very different (cars of other colors).

Combined with anomaly detection performance shown in Table 2, it can be concluded that the considering the inner-relationship of kNN semantic features is reasonable and the self-attention module is effective.

### 4.4 Benefit of Constraint Term

In training phase, we have mentioned a potential trivial solution and proposed a constraint term $\Omega$ to avoid such situation during optimization. To verify the effectiveness of the constraint, firstly we compare the performance under different weights of $\Omega$ on CIFAR10 which is shown in Table 3.

| $\lambda$ | 0   | 0.1 | 1   | 2   | 10  | 100 | no ada |
|-----------|-----|-----|-----|-----|-----|-----|--------|
| AUROC     | 95.1| 95.6| 96.9| 97.0| 97.0| 97.0| 96.6   |

Comparing with the performance without adaptation, when $\lambda$ is relatively small (e.g. $\lambda = 0$ or $\lambda = 0.1$), the performance has decreased greatly. The reason for this is that small constraint cannot avoid the model tend to learn trivial solutions resulting in the similarity between adapted feature $\hat{z}_i$ and normal representation $\hat{z}_n^i$ of a specific input is large no matter it is normal or not. While for larger $\lambda$, the model can adapt well, and we find that when the constraint weight reaches a certain value, the result tends to be stable. Based on this, we recommend that when adapting SAP2 to a specific dataset, the constraint weight should be adjusted with a larger one.

To show the benefit of constraint more intuitively, we show performance when $\lambda = 2$ and $\lambda = 0$ in Figure 3 including the trends of AUROC, and anomaly score of normal and anomaly sample. For $\lambda = 2$, it is shown that the gap of anomaly score between normal and anomaly is gradually growing larger. Specifically, The anomaly score of normal is basically unchanged, while the anomaly score of anomaly have increased, which reflects that the model does not generalize the adaptation to abnormal data. While for $\lambda = 0$, the anomaly score of all samples are basically equal to 0 which implies the similarity between $\hat{z}_i$ and $\hat{z}_n^i$ is equal to 1. In this case, combined with the proposed optimization target, we deem that the model would indeed tend to map all samples to 0 when there are no constraint.

### 4.5 Visualization

To show the mechanism of SAP2 more intuitively, we visualized result refer to class activation mapping(CAM)[22]. In CAM, the product of a global average pooling feature $z$ and the last layer feature map $f$ reveals that through which pixels does the model know that the picture belongs to a category. In SAP2, $z_i$ is the global average pooling, thus we
Simple Adaptive Projection with Pretrained Features for Anomaly Detection

Figure 3: CIFAR10 Class 1: left AUROC and anomaly score with $\lambda = 2$. right AUROC and anomaly score with $\lambda = 0$.

compute the element-wise cosine similarity between $z_i$ and the last layer 7*7 feature map. Beyond that, since $\hat{z}_i^n$ is the normal representation, we also compute the element-wise cosine similarity between $\hat{z}_i^n$ and the same feature map. And the two calculation results are upsampled to 224 * 224. Naturally, the absolute difference between the two can reflect the anomaly region. We show the visualization results in Figure 4 and Figure 5 which display the performance on semantic AD and sensory AD separately. The dark yellow area represents the anomaly region SAP2 infers.

**Semantic AD** Semantic AD only focuses on the semantic shift. To visualize how SAP2 works, we show the anomaly region on the normal class and anomaly classes. In the normal class, the anomaly region is some edge pixels and for anomaly class, the anomaly region is accurately located on the semantic target. This result illustrates that i) pretrained feature contains important semantic information. ii) adapted normal representation obtained by SAP2 contains the semantic information of the normal class. Meanwhile, this result also shows that the proposed anomaly detection criterion is reasonable and convincing.

Figure 4: Visualization of CIFAR10(Semantic AD). The inferred anomaly region of normal image is in the unimportant area and that of anomaly image is on the semantic target.

**Sensory AD** Sensory AD only focuses on objects with the same or similar semantics, and identifies the observational differences on their surface. We select some classes in MvTec and show the calculated anomaly region and the ground truth in Figure 5 where left side are anomaly with a large anomaly region while the right side are anomaly with the smaller anomaly region. The visualization results show that even if only the difference between the two semantic features $z_i$ and $\hat{z}_i^n$, is considered, the anomaly region can be located. For anomaly with large region, the localization performance is better than that with smaller region, and the reason for this is that the anomaly region is located in the 7*7 feature map and the upsampling is through interpolation, thus the inferred anomaly region will be larger than authentic region. However, these anomaly region can all be located more or less which implies SAP2 is an effective adaptation strategy for sensory AD task.

5 Conclusion

In this paper, we propose a simple adaptive projection head instead of finetuning layers in pretrained large network and leverage self-attention to mine the relationship between semantic features of a series of similar images. The proposed anomaly detection framework avoids the global optimization goal in the traditional one-class setting, and proposes a
simple constraint term to improve the representation ability of the adaptation strategy. We conducted experiments on multiple benchmarks to prove the effectiveness of SAP2, and proved the rationality of each module in the framework through a large number of ablation experiments.

References

[1] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE, 2021.

[2] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. arXiv preprint arXiv:2110.11334, 2021.

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4183–4192, 2020.

[4] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Basilzadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14902–14912, 2021.

[5] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2806–2814, 2021.

[6] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In International conference on machine learning, pages 4393–4402. PMLR, 2018.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

[8] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Sal Stolfo. A geometric framework for unsupervised anomaly detection. In Applications of data mining in computer security, pages 77–101. Springer, 2002.

[9] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. Journal of the royal statistical society. series c (applied statistics), 28(1):100–108, 1979.

[10] Yunqiang Chen, Xiang Sean Zhou, and Thomas S Huang. One-class svm for learning in image retrieval. In Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205), volume 1, pages 34–37. IEEE, 2001.
[11] David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54(1):45–66, 2004.

[12] Michael Glodek, Martin Schels, and Friedhelm Schwenker. Ensemble gaussian mixture models for probability density estimation. *Computational statistics*, 28(1):127–138, 2013.

[13] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[14] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

[15] Tiago S Nazare, Rodrigo F de Mello, and Moacir A Ponti. Are pre-trained cnns good feature extractors for anomaly detection in surveillance videos? *arXiv preprint arXiv:1811.08495*, 2018.

[16] Liron Bergman, Niv Cohen, and Yedid Hoshen. Deep nearest neighbor anomaly detection. *arXiv preprint arXiv:2002.10445*, 2020.

[17] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14902–14912, 2021.

[18] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[19] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Peter Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[20] Yanzhou Su, Yongjian Wu, Min Wang, Feng Wang, and Jian Cheng. Semantic segmentation of high resolution remote sensing image based on batch-attention mechanism. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 3856–3859. IEEE, 2019.

[21] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *arXiv preprint arXiv:1906.12340*, 2019.

[22] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.