MixedTeacher: Knowledge Distillation for Fast Inference Textural Anomaly Detection

Simon Thomine[1,2], Hichem Snoussi[1] and Mahmoud Soua[2]
[1]*University of Technology Troyes, Troyes, France*
[2]*AQUILAE, Troyes, France*

Keywords: Anomaly Detection, Texture, Knowledge Distillation, Layer Selection, Unsupervised.

Abstract: For a very long time, unsupervised learning for anomaly detection has been at the heart of image processing research and a stepping stone for high performance industrial automation process. With the emergence of CNN, several methods have been proposed such as Autoencoders, GAN, deep feature extraction, etc. In this paper, we propose a new method based on the promising concept of knowledge distillation which consists of training a network (the student) on normal samples while considering the output of a larger pretrained network (the teacher). The main contributions of this paper are twofold: First, a reduced student architecture with optimal layer selection is proposed, then a new Student-Teacher architecture with network bias reduction combining two teachers is proposed in order to jointly enhance the performance of anomaly detection and its localization accuracy. The proposed texture anomaly detector has an outstanding capability to detect defects in any texture and a fast inference time compared to the SOTA methods.

1 INTRODUCTION

Anomaly detection in industry is a vast topic since there is a lot of possible applications. For instance, defect detection aims at identifying specific anomaly classes and locations in industrial manufacturing processes (Kähler et al., 2022). This detection is crucial for ensuring the high quality of final products (Minhas and Zelek, 2019). A common property of defects is that their visual texture is inherently different from the defect-free surface. The specificity of textures is the pattern structure which, if known, allows the detection and the extraction of anomalies. However, the texture anomaly generally appears in a small region in few samples, which makes it difficult to build consistent normal and abnormal datasets to be used in supervised learning methods. Hence, unsupervised anomaly detection networks are very suitable for industrial scenarios as they represent the strong basis for building a detection model without any annotated samples (Huang et al., 2022). Several unsupervised anomaly detection methods have been introduced for texture anomaly detection. These methods could achieve high performance up to 99.6 AUROC. However, they suffer from complex networks and high latency.

In another context, knowledge distillation has been introduced with the purpose of reducing the network size while increasing performance. Knowledge distillation aims to train a smaller network (student) to imitate pretrained one or several larger ones (teachers) on normal samples. As the teacher is pretrained, it has the ability to generalize even if the sample contains an anomaly, whereas the student won't be able. Hence, by comparing the extracted features between the teacher and the student networks, an abnormal sample could be detected. According to some studies (Iglesias and Zseby, 2015), using too many features can significantly reduce the accuracy of anomaly detection. Recently, a Student-Teacher Feature Pyramid Matching (STPM) method has been proposed in (Wang et al., 2021), where the first three network layers are used in order to focus on edges, colors and shapes instead of context information. Even if using layer selection technique is an interesting approach, there is still a lack of explanation concerning the layer choice and the relevance of the relative information. Looking at the same layers for an object and for a texture reduces the relevance of the extracted information. For example, looking at context information in a texture is pointless and for an object, pure edge/color/texture information may not be the most interesting information.

Thomine, S., Snoussi, H. and Soua, M.
MixedTeacher: Knowledge Distillation for Fast Inference Textural Anomaly Detection.
DOI: 10.5220/0011633100003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 487-494
ISBN: 978-989-758-634-7; ISSN: 2184-4321

Another recurrent problem is the classifier bias. The best current methods use a pretrained classifier network on imageNet which is biased by the classes of imageNet and can have an impact on the localization and the detection of defects.

The main contributions of the paper are as follows:

- A new reduced student architecture for texture-specific object category.
- In order to reduce the classification bias, we propose a new architecture combining two teachers pretrained on imageNet but with different architectures (respectively ResNet-18 (He et al., 2015) and EfficientNet-b0 (Tan and Le, 2020)) and a single student network. This new mixed Teacher network structure outperforms competitive state-of-the-art methods both in inference time and SOTA scores, on anomaly datasets such as MVTEC AD textures and BTAD textures (Mishra et al., 2021). The proposed MixedTeacher model uses a score and anomaly localisation function based on each complementary teacher features with a careful feature selection.

The paper is organized as follows. In section 2, we review the related work, especially on MVTEC dataset and present the different approaches proposed in literature. In section 3, we compare the results of training with different architectures and different layer selection schemes and introduce our proposed texture-specific reduced student architecture. Section 4 is dedicated to describing a novel mixed Student-Teacher network. In section 5, we compare our results to the SOTA methods for both the reduced student architecture and the MixedTeacher in terms of AUROC, pixel-AUROC and inference time.

2 RELATED WORK

Anomaly detection is a problem that pops up in many areas and is often very difficult to deal with. Indeed, detecting the "abnormal" is a rather vague concept and is difficult to define according to the use cases, which makes research on this subject very specific.

For several years, the rise of deep learning has never ceased to impress with high quality results and interesting methods. Most of these methods are based on an unsupervised representation approach to discriminate outliers. Some specific work has been done for fabrics defect detection such as the multi-scale Convolutionnal denoising autoencoder (Mei et al., 2018). For unsupervised anomaly detection in general, we can also cite the GEE, a gradient based VAE (Nguyen et al., 2019) or the Gaussian mixture model VAE (Nguyen et al., 2019). Another common way to detect anomaly is to use generative adversarial networks (Goodfellow et al., 2014). Ano-GAN (Schlegl et al., 2019) was one of the first utilization of GAN for anomaly detection but since then a lot of approaches emerged such as G2D (Pourreza et al., 2021) and OCR-GAN (Liang et al., 2022). Other interesting approaches rely on pretrained models especially on imageNet, using the feature extraction of pretrained network to extract useful information about a given sample. The idea is to extract features with a pretrained model and then train a normalizing flow model on good samples, so that the model is ready to find out if a given sample is an anomaly by looking at the reconstruction error. An advantage of normalizing flow is the reversible aspect which is useful to locate the anomaly pixel-wise. Many techniques based on this concept have been proposed such as differNet (Rudolph et al., 2021a) and CS-FLOW (Rudolph et al., 2021b) which consider multi-scale normalizing flow and FastFlow (Yu et al., 2021) based on a 2D normalizing flow.

Recently, the concept of knowledge distillation has also been used for unsupervised anomaly detection. The student-teacher method consists of training a student teacher based on the output of a larger teacher model which is pretrained on ImageNet. The student network will learn to imitate the teacher on good samples only. Then, when an abnormal sample is tested, the teacher will be able to generalize and the student won't be, the difference between the output of the teacher and the output of the student will allow the detection of the anomaly. On the MVTEC dataset, four methods have been implemented, STPM (Wang et al., 2021) which trained the student on the 3 first layers of ResNet-18, RSTPM (Yamada and Hotta, 2022) which is basically the same method but with an attention layer, reverse distillation (Deng and Li, 2022) and CFA (Lee et al., 2022).

3 LAYER SELECTION AND REDUCED STUDENT

In this section, after a comparative study of layer selection methods for optimal texture anomaly detection, we present a new student architecture that both increases performance and reduces the inference time.

3.1 Layer Selection

In deep neural networks, a common observation is that deep layer features contain context information

and shallow layer features contain color, texture and contour information. In a case of detection of defects on the fabric or on a generic texture, the context information is less important than the texture information, therefore, we will turn to shallow layer features. As reported in table 1, different combinations of shallow layers have been tried in order to select the optimal architecture with respect to detection performance evaluated by the AUC.

Table 1: Layers selection results.

Measures	Layer 1 and 2 AUC	Layer 2 and 3 AUC
Mean objects	0.876	0.910
Mean textures	0.990	0.971

3.2 Reduced Student

ResNet-18 architecture has been retained for the teacher network. As texture specific anomaly detection is the main objective of this work, we propose to add the ResNet-18 first layer after the first convolution to extract even more textural information. The second objective was to alleviate the student architecture to decrease inference time and possibly performance. As ResNet-18 presents several residual blocks with two identical convolutional layers, we first decided to take only one layer for each block in our student architecture. The classifier bias is another known problem while dealing with pretrained classifier and we tackled this problem by reducing features size with an adaptive average pooling layer in each Resnet residual block's output as presented in figure 1.

Given a training dataset of images without anomaly $D = [I_1, I_2, ..., I_n]$, our goal is to extract the information of the L bottom layers. For an image $I_k \in R^{w*h*c}$ where w is the width, h the height and c the number of channels, the teacher outputs features $F_t^l(I_k) \in R^{w_l*h_l*c_l}$ and $F_s^l(I_k) \in R^{w_l/2*h_l/2*c_l/2}$ with $l > 1$ and $F_s^l(I_k) \in R^{w_l*h_l*c_l}$ if $l = 1$. For the loss function, we took the $l2$ distance of normalized feature vectors like in the STPM original paper (Wang et al., 2021) while using an adaptive average pooling on teacher features where $l > 1$ and just sum all feature maps of all layers to obtain our loss with the same ratio for all layers (Eq.1).

$$F_t^{l>1}(I_k) = AAP(F_{Resnet18}^{l>1}(I_k)) \tag{1}$$

where AAP refers to the Adaptive Average Pooling. Pixel loss is defined in the following Eq.2:

$$loss^l(I_k)_{ij} = \frac{1}{2}\|norm(F_t^l(I_k)_{ij}) - norm(F_s^l(I_k)_{ij})\| \tag{2}$$

and for the layer l, the loss is defined as:

$$loss^l(I_k) = \frac{1}{w_l h_l}\sum_{i=1}^{w_l}\sum_{j=1}^{h_l} loss_{resNet}^l(I_k)_{ij} \tag{3}$$

and finally for the total loss is written as:

$$loss(I_k) = \sum^{l} loss^l(I_k) \tag{4}$$

Performance and inference speed are later reported in section 5 with comparison with SOTA networks on anomaly detection.

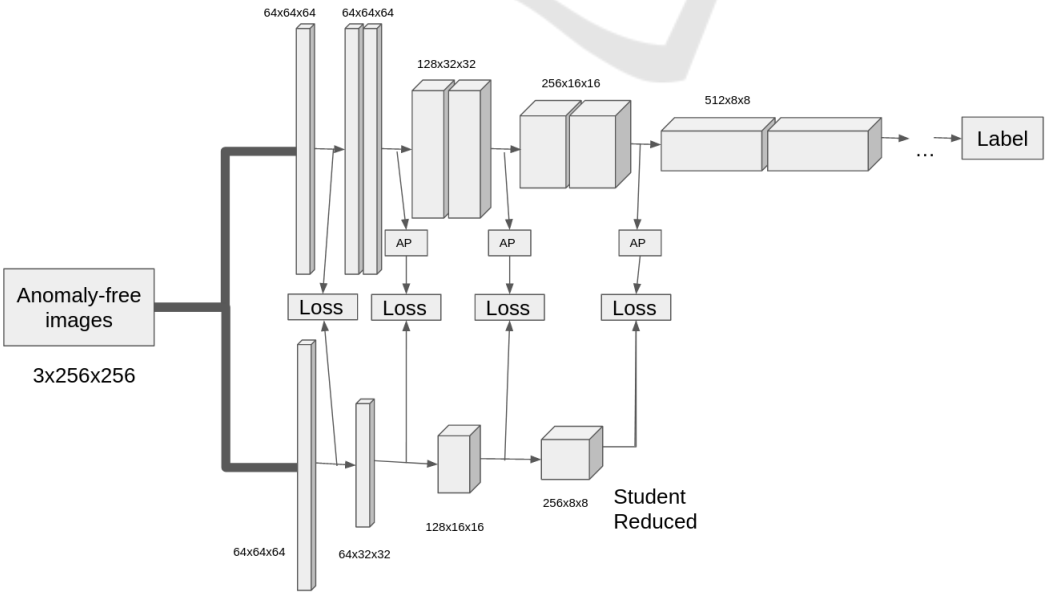

Figure 1: Reduced student architecture with AP for adaptive average pooling.

4 MIXED TEACHER

In this section, we introduce our new student teacher network structure that combines two teachers with the purpose of reducing the classifier bias, taking benefits from the two networks and exploiting the different layers in an optimal way.

4.1 Observation and Main Ideas

While testing our new student reduced architecture on the MVTEC AD textures, we obtained good results, but some noise still degrade results in terms of default localisation on specific images or texture-specific normal variation. Different teacher network architectures have been tested to conclude that ResNet-18 remains the best in terms of average precision and speed. However, interesting behaviors have been observed on the noise localisation for each architecture. In fact, every classifier had the capacity to locate the anomaly, but with output noise and anomaly detection mistakes.

The combination of two pretrained classifier networks has therefore been proposed with the purpose of interpolating their defect localisation to cancel noise and false detection/segmentation.

EfficientNet-b0 has been proposed as the second teacher when considering its performance in terms of precision and speed. For this network, it has been observed that for the bottom layers, one has good localisation but with a huge noise and with top layers, a coarse defect localisation but with minimal noise has been obtained, as illustrated in figure 2.

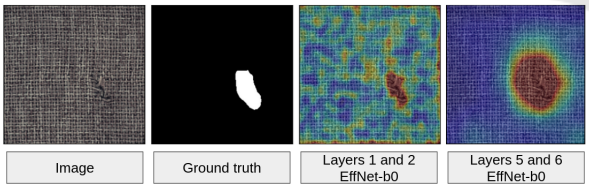

Figure 2: Difference between top layers and bottom layers for EfficientNet-b0 architecture.

4.2 Method Description

The learning architecture is composed of two teachers: the ResNet-18 as main teacher and EfficientNet-b0 as a localisation confirmation teacher. For the ResNet-18 part, the reduced student proposed in section 3 is used in order to ensure a good inference speed and precision on texture samples. For EfficientNet-b0 student, we used one convolution for each efficientnet block without pooling because we used deepest layers. In the student architecture, there is no communication between the networks except for the loss function as shown in figure 3.

For the training loss function, we used basically the same loss function as the one for the reduced teacher and we add an α factor to smooth the layer activation difference from the two teacher networks. As feature difference in efficientNet was about 10 times bigger than in ResNet-18, α has been set to 0.1.

$$loss^{l=5,6}_{effNet}(I_k)_{ij} = \frac{1}{2}\|norm(F^l_t(I_k)_{ij}) - norm(F^l_s(I_k)_{ij})\| \tag{5}$$

and

$$loss^{l=5,6}_{effNet}(I_k) = \frac{1}{w_l h_l}\sum_{i=1}^{w_l}\sum_{j=1}^{h_l} loss^l_{effNet}(I_k)_{ij} \tag{6}$$

and for Resnet-18 part :

$$loss^{l=1,2,3}_{resNet}(I_k) = \frac{1}{w_l h_l}\sum_{i=1}^{w_l}\sum_{j=1}^{h_l} loss^l_{resNet}(I_k)_{ij} \tag{7}$$

with $loss^l_{resNet}(I_k)_{ij}$ defined as in section 3. For the total loss with the α factor :

$$loss_{tot}(I_k) = \sum_{l=1}^{3} loss^l_{resNet}(I_k) + \alpha\sum_{l=5}^{6} loss^l_{effNet}(I_k) \tag{8}$$

As in every knowledge distillation method, the loss only impacts the student.

4.3 Anomaly Score and Localisation

In the test phase (inference), we want an anomaly map M of the original image size where every pixel at position (i, j) has an anomaly score M_{ij}. With a test image I and $F^l_{tResnet}$, $F^l_{tEffNet}$ the two teachers features of lth layer and $F^l_{sResnet}$, $F^l_{sEffNet}$ their corresponding lth layer student features, we perform an upsample on the difference between the corresponding layers. The coarse localisation output of the efficientNet layers is obtained by summing each layer's anomaly map.

The anomaly map is obtained in the same way for the resnet part. Respectively :

$$A_{mapEffnet} = \sum_{l=5}^{6} Upsample(F^l_{tEffNet} - F^l_{sEffNet}) \tag{9}$$

and :

$$A_{mapResnet} = \sum_{l=1}^{3} Upsample(F^l_{tResnet} - F^l_{sResnet}) \tag{10}$$

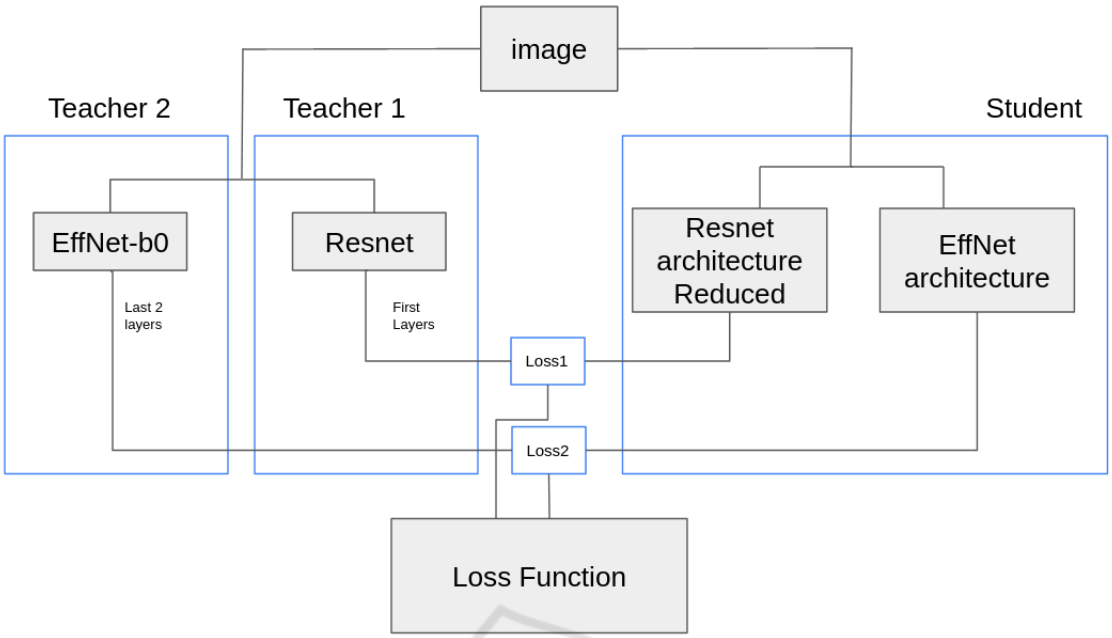

Figure 3: MixedTeacher architecture.

We then multiply the resnet anomaly map by the normalization of the effnet anomaly map multiplied by its mathematical extent. With $A_{mapEffnet}$, the anomaly map of efficientNet layers and $A_{mapResnet}$ the anomaly map of resnet layers, the final anomaly map is then defined as :

$$M = A_{mapResnet} * (max(A_{mapEffnet}) - min(A_{mapEffnet}))A_{mapEffnet} \quad (11)$$

The anomaly score is defined as :

$$score = \sum_{i=1}^{w} \sum_{j=1}^{h} M_{i,j} \quad (12)$$

with w and h are respectively the width and height of the anomaly map.

5 EXPERIMENTS

5.1 Datasets

We experiment our methods on the textures from the **MVTEC AD** (Bergmann et al., 2019) dataset which consists of 15 categories : 5 textures and 10 objects with a total of more than 5000 high resolution images. This dataset is used for unsupervised anomaly detection therefore it contains only anomaly free images for the training. For the test part, it shows a good variety of defects with ground truth masks for anomaly localisation. We also used the texture of the **BTAD** (Mishra et al., 2021) dataset which is an unsupervised anomaly dataset with three different categories including one texture figure 4.

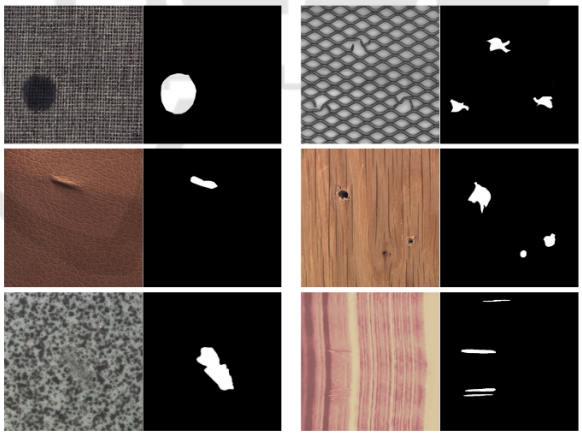

Figure 4: Overview of textures from MVTEC AD and BTAD dataset, samples with anomaly and ground truth. These images are only used for testing and unseen during the training.

The performance is evaluated with AUROC metric image-level and pixel-level to compare our results with other methods.

5.2 Implementation and Training Metrics

Training and inference were done on an rtx 2080ti.

To test the student reduced, we used the features of the first three blocks and the layer before the first block of ResNet-18. The Resnet network was pretrained on imageNet. We used stochastic gradient descent with a learning rate of 0.4 for 100 epochs with a batch size of 16. To test the MixedTeacher, we used the output features of the first two blocks and the layer before the first block of ResNet-18 and the output features of block 5 and 6 of EfficientNet-b0. We used stochastic gradient descent with a learning rate of 0.4 for 200 epochs with a batch size of 16. Both networks are pretrained on imageNet. We resized all the images to a size of 256x256 keeping 80% for training and 20% for validation. We kept the checkpoint with the lowest validation loss.

5.3 Reduced Student

5.3.1 Performance Results

In this paragraph, we will compare reduced student AUROC results to SOTA methods. In 2, we present AUROC performance results of CFA (Lee et al., 2022), PatchCore (Roth et al., 2021), FastFlow (Yu et al., 2021), STPM (Wang et al., 2021), CutPaste (Li et al., 2021) and our reduced student on MVTEC AD textures.

Table 2: Image-AUROC comparison on MVTEC AD : Reduced Student.

Category	CutPaste	CFA	PatchCore	STPM	FastFlow	Ours
carpet	100	97.3	98.7	95.4	99.4	100
tile	100	99.4	98.7	94.9	100	98.7
grid	96.2	99.2	98.2	98.2	100	99.7
wood	99.1	99.7	99.2	96.1	99.2	99.6
leather	95.4	100	100	98.9	99.9	99.7
Mean	98.1	99.1	99.0	96.7	99.7	99.5

For FastFlow, we choosed to take the results from Anomalib as we were not able to reproduce their paper results (99.9 AUROC in paper). As seen in table 2, our reduced student is better than CFA for texture anomaly detection, which is the best actual knowledge distillation unsupervised anomaly detection method and is close to the SOTA results. We manage to gain 2.8 points against classic STPM with a network reduction and a wise layer selection aiming for texture specific anomaly detection.

5.3.2 Inference Time Results

In table 3, we compare the reduced student inference time to other SOTA methods. The main purpose of reduced student was to propose a high processing speed to manage real time for several high resolution images. To get inference time results, we employ Anomalib. All the additional results come from this library to make sure the tests were carried out under the same conditions.

Table 3: Inference time results.

Category	PatchCore	FastFlow	STPM	Ours
FPS	5.8	21.8	83.2	**108.1**
Latency (ms)	172	45.9	12	**9.2**

The presented results are based on Anomalib inference time. In a self made code, we were able to obtain a 10x better inference time for STPM and reduced student. The most important thing to consider is that the STPM is by far the fastest anomaly detector and reduced student reduced its inference time by 30%.

5.4 MixedTeacher

5.4.1 Performance Results

Unlike the reduced student, the MixedTeacher main purpose is performance and not inference time. In table 4 we compared AUROC of several SOTA methods in texture anomaly detection.

Table 4: Image-AUROC comparison on MVTEC AD : MixedTeacher.

Category	CutPaste	CFA	PatchCore	FastFlow	ReducedStudent	Ours
carpet	100	97.3	98.7	99.4	**100**	99.8
tile	100	99.4	98.7	100	98.7	**100**
grid	96.2	99.2	98.2	**100**	99.7	99.7
wood	99.1	**99.7**	99.2	99.2	99.6	99.6
leather	95.4	100	100	99.9	99.7	**100**
Mean	98.1	99.1	99.0	99.7	99.5	**99.8**

Our method is the new state of the art texture anomaly detection on the MVTEC AD dataset.

5.4.2 Anomaly Localisation

Even though anomaly localisation was not our main purpose, our approach uses EfficientNet-b0 with the objective of making the location more precise. To this end, we present in table 5 and table 6, our anomaly location results on textures from MVTEC AD dataset and BTAD respectively and we compare these results to the SOTA methods.

Table 5: Pixel-AUROC comparison on MVTEC AD : MixedTeacher.

Category	CutPaste	PatchCore	FastFlow	Ours
carpet	98.3	98.9	99.1	99.0
tile	90.5	95.6	96.6	95.9
grid	97.5	98.7	99.2	97.5
wood	95.5	95	94.1	94.9
leather	99.5	99.3	99.6	99.4
Mean	96.2	97.5	97.7	97.3

Table 6: Image-AUROC comparison on BTAD: MixedTeacher.

Category	FastFlow	Ours
1 (wood from btad)	96.0	**97.0**

5.4.3 Inference Time Results

In terms of inference speed, our MixedTeacher is 3x slower than the reduced student since it used two teacher networks and a more complex student architecture.

6 CONCLUSION

In this paper, we proposed two methods for efficient unsupervised anomaly detection using the principle of knowledge distillation applied to unsupervised anomaly training. Both methods offer different benefits. The reduced student proposes a high speed texture anomaly detector with an AUROC performance close to the state of the art, this method is to be used in situations where inference time is the most important priority (mobile device, low computational power, cost efficiency). The MixedTeacher propose the highest actual performance on anomaly detection with a localisation close to the state of the art on the MVTEC AD textures with still a fast inference. This method is to be used in situations where performance is the priority and the computational power is big enough (monitoring server etc ...)

REFERENCES

Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. MVTec AD — a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9584–9592. IEEE.

Deng, H. and Li, X. Anomaly detection via reverse distillation from one-class embedding.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition.

Huang, J., Li, C., Lin, Y., Lian, S., and Innovation, A. Unsupervised industrial anomaly detection via pattern generative and contrastive networks.

Iglesias, F. and Zseby, T. Analysis of network traffic features for anomaly detection. 101(1):59–84.

Kähler, F., Schmedemann, O., and Schüppstuhl, T. Anomaly detection for industrial surface inspection: application in maintenance of aircraft components. 107:246–251.

Lee, S., Lee, S., and Song, B. C. CFA: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization.

Li, C.-L., Sohn, K., Yoon, J., and Pfister, T. CutPaste: Self-supervised learning for anomaly detection and localization.

Liang, Y., Zhang, J., Zhao, S., Wu, R., Liu, Y., and Pan, S. Omni-frequency channel-selection representations for unsupervised anomaly detection.

Mei, S., Wang, Y., and Wen, G. Automatic fabric defect detection with a multi-scale convolutional denoising autoencoder network model. 18(4):1064.

Minhas, M. S. and Zelek, J. AnoNet: Weakly supervised anomaly detection in textured surfaces.

Mishra, P., Verk, R., Fornasier, D., Piciarelli, C., and Foresti, G. L. VT-ADL: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06.

Nguyen, Q. P., Lim, K. W., Divakaran, D. M., Low, K. H., and Chan, M. C. GEE: A gradient-based explainable variational autoencoder for network anomaly detection.

Pourreza, M., Mohammadi, B., Khaki, M., Bouindour, S., Snoussi, H., and Sabokrou, M. G2d: Generate to detect anomaly. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2002–2011. IEEE. event-place: Waikoloa, HI, USA.

Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., and Gehler, P. Towards total recall in industrial anomaly detection.

Rudolph, M., Wandt, B., and Rosenhahn, B. Same same but DifferNet: Semi-supervised defect detection with normalizing flows. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1906–1915. IEEE. event-place: Waikoloa, HI, USA.

Rudolph, M., Wehrbein, T., Rosenhahn, B., and Wandt, B. Fully convolutional cross-scale-flows for image-based defect detection.

Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G., and Schmidt-Erfurth, U. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. 54:30–44.

Tan, M. and Le, Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks.

Wang, G., Han, S., Ding, E., and Huang, D. Student-teacher feature pyramid matching for anomaly detection.

Yamada, S. and Hotta, K. Reconstruction student with attention for student-teacher pyramid matching.

Yu, J., Zheng, Y., Wang, X., Li, W., Wu, Y., Zhao, R., and Wu, L. FastFlow: Unsupervised anomaly detection and localization via 2d normalizing flows.