SUPPLEMENTARY MATERIAL

Shaurya Gupta[1] *Neil Gautam*[1*] *Anurag Malyala*[1]

[1] HyperVerge AI

1. IMPLEMENTATION DETAILS

This paper showcases various models developed with Python 3.8 [1] and PyTorch 1.12 [2]. To train and test these models, we employed the Torchvision ImageNet-pretrained models as backbones, with our default setting utilising the ResNet50-backbone [3] in conjunction with ATAC-Net. We also conducted comparisons using other techniques under the same conditions.

1.1. ATAC Net Specifics

The method's Attention-based cropping and corresponding training procedure are explained in detail in algorithms Algorithm 1 and Algorithm 2, as referenced in the paper's equations. To better understand how ATAC-Net performs anomaly generalisation, we compare it to other weak-supervised methods on various anomaly forms, presented in Table 2. We observe varying results by training and comparing only one class of anomaly on a known object/surface for detecting other anomaly classes on the same object/surface. The model is highly accurate, highlighting its required robustness to determine a new anomaly in a closed setting. Results demonstrate that ATAC-Net outperforms Dev-Net and DRA in detecting known anomalies on surface-level tests for the "Tile" and object-level tests for the "Metal Nut" class.

Algorithm 1 Attention Based cropping

Input: $C_{attn}, \omega, x, iH, iW$
Output: Attention Cropped inputs x_c
Parameters: $\theta \in \{ \theta_f, \theta_a, \theta_s \}$

$\omega_c = \omega * max(C_{attn})$
$C_{attn} = \frac{C_{attn}}{max(C_{attn}}$
$C_{mask} = upsample(C_{attn}, iH, iW) >= \omega_c$
$C_{mask} = non - zero(C_{mask})$
$(i_{min}, j_{min}) = \arg\min_{x,y}(C_{mask})$
$(i_{max}, j_{max}) = \arg\max_{i,j}(C_{mask})$
$x_c = x[((i_{min}, j_{min})), (i_{max}, j_{max})]$

Algorithm 2 Training ATAC-Net

Input: $\{x\}_{i=0}^{N+M} \in \mathbb{R}^{H \times W \times C}; where\ N >> M$
Output: Anomaly scorer network $\gamma(x; \theta)$
Parameters: $\theta \in \{ \theta_f, \theta_a, \theta_s \}$

Initialize γ with He Initialization.
$X_N = \{x\}_{i=0}^{N}$; $N \in$ Normal samples
$X_T = \{x\}_{i=0}^{M} + cutmix(X_N)$; $M \in$ Anomaly samples
for $i = 1$ to $epochs = n$ **do**
  for $j = 1$ to $steps = k$ **do**
    $m \leftarrow \{X_N i\}_{i=1}^{\frac{|m|}{2}} \cup \{X_T j\}_{j=1}^{\frac{|m|}{2}}$
    $a_{mp1} \leftarrow \tau(\cdot; \theta_s) \circ \sigma(\cdot; \theta_a) \circ \phi(x_m; \theta_f)$
    $x_{cm} \leftarrow$ attention cropping of x_m.
    $a_{mp2} \leftarrow \tau(\cdot; \theta_s) \circ \sigma(\cdot; \theta_a) \circ \phi(x_{cm}; \theta_f)$
    $a_{mp} \leftarrow \frac{(a_{mp1} + a_{mp2})}{2}$
    $dev(a_{mp}) \leftarrow \frac{\gamma(x;\theta) - \mu_R}{\sigma_R}$
    $\mathcal{L}_{dev} = (1 - y_i)|dev(a_{mp})| + y_i(k - |dev(a_{mp})|)$
    Update Gradients: $\theta' := \theta - \alpha\nabla(\mathcal{L}_{dev})$
  end for
end for

2. DATA PRE-PROCESSING AND EVALUATION

Before making any predictions using the proposed architecture, we apply a set of pre-processing steps to the image. Firstly, we use Contrastive Equalisation to improve the contrast of the image. Next, we normalize the image between the range of 0 and 1, and then standardize it channel-wise using the pre-computed Image Net mean and standard deviation $\mu = [0.48145466, 0.4578275, 0.40821073]^T; \sigma = [0.26862954, 0.26130258, 0.27577711]^T$. Finally, to ensure consistency, each image is resized to a spatial resolution of 224x224 using Bi-Cubic interpolation through the Pillow Library [4]. Additionally, we use the Cut-Mix operation [5] for each batch to add more anomalous samples to the dataset. To test the model performance, we use AUC-ROC scores on image-level testing, while keeping the same image resolution of 224x224 across all the compared methods, in line with other state-of-the-art models.

*Work done during intership

3. ABLATION STUDY

Starting with a basic setting of 10 reference anomaly samples for training, we conduct experiments using the original genuine training set.

1. We adopt from [6] the ResNet-50 [3] architecture with the last anomaly mapping layer connected to a classification head and start with testing a baseline with just the CUT-MIX augmentation [5] and a few anomalies samples over the binary cross-entropy loss as the optimization target.

2. After the previous step, we updated to use weighted cross-entropy to evaluate the effectiveness of a penalized classifier for the same task. Specifically, we assigned anomaly and genuine class weights of 20 and 0.5, respectively.

3. After conducting the last two tests, we removed the classification layer and used the deviation loss [6]. With the top-k score of the anomaly mapped layer over the last steps, we observed how reframing of the loss term handles the issue caused by imbalance. Consequently, we achieved better results as shown in Table 1.

4. We have conducted an experiment to test the effectiveness of retraining a network over a zoomed view of the activated regions using Grad-CAM [7] on a set of images. However, we found that Grad-CAM [7] is not a reliable method to test this, as the activated regions are randomly selected. As a result, zooming in on these regions and retraining the network can lead to confusion and unreliable results, as demonstrated in the samples shown in Fig. 1.

5. Rather than treating zoomed augmentations as a separate process using conditional back-propagation, we incorporate zooming into the learning process by introducing a self-attention convolution block [8] before the anomaly mapping layer. We use the channel wise mean of this block and interpolate it to create the zoomed augmentation, this added pipeline we refer to as ATAC-Net. We calculate our final anomaly score by taking the mean of the outputs from both the original and zoomed samples. Since the zoomed sample has the same ground truth as the original sample, it helps the attention layer differentiate between zoomed regions during training. This allows the model to learn the activations where the anomaly lies without requiring explicit information about the region of interest in the sample.

A few samples are compared in Fig. 1, to show the anomaly and genuine sample activations given by ATAC-Net and Grad-CAM [7] over step IV. The training steps and settings for all these experiments are the same as mentioned in the Training details section of the paper (Sec-4.2).

Further observing Fig. 1, the high activation maps for genuine samples show that the full image is cropped so no zooming and thus anomaly score remains unaffected even with guided zooming re-iteration. The overlaid activation maps for GRAD-CAM [7] can be seen to be unreliable for the given cases, thus rejecting the use of it for the given purpose.

The Attention based cropping and the corresponding training procedure of the method is described through Algorithm 1 and Algorithm 2, synchronous to the equations in mentioned in paper. Further for better understanding of the anomaly generalisation we compare ATAC-Net with other weak-supervised methods on different forms of anomaly, the same is presented in Table 2, we observe how the results vary if we train and compare only one class of the anomaly of a known object/surface for detection of other anomaly classes of the same object/surface, the model are pretty accurate highlighting the required robustness to determine whether the proposed technique can determine a new anomaly in a closed setting. The observed results show that ATAC-Net is more robust to known anomalies than Dev-Net and DRA, for surface level test on "Tile" and object level on "Metal Nut" class.

4. REFERENCES

[1] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. 1

[2] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019. 1

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. 1, 2

[4] A. Clark, "Pillow (pil fork) documentation," 2015. 1

[5] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "Cutmix: Regularization strategy to train strong classifiers with localizable features," pp. 6022–6031, 10 2019. 1, 2

[6] G. Pang, C. Ding, C. Shen, and A. van den Hengel, "Explainable deep few-shot anomaly detection with deviation networks," *CoRR*, vol. abs/2108.00462, 2021. 2

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. 2

[8] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10073–10082, 2020. 2

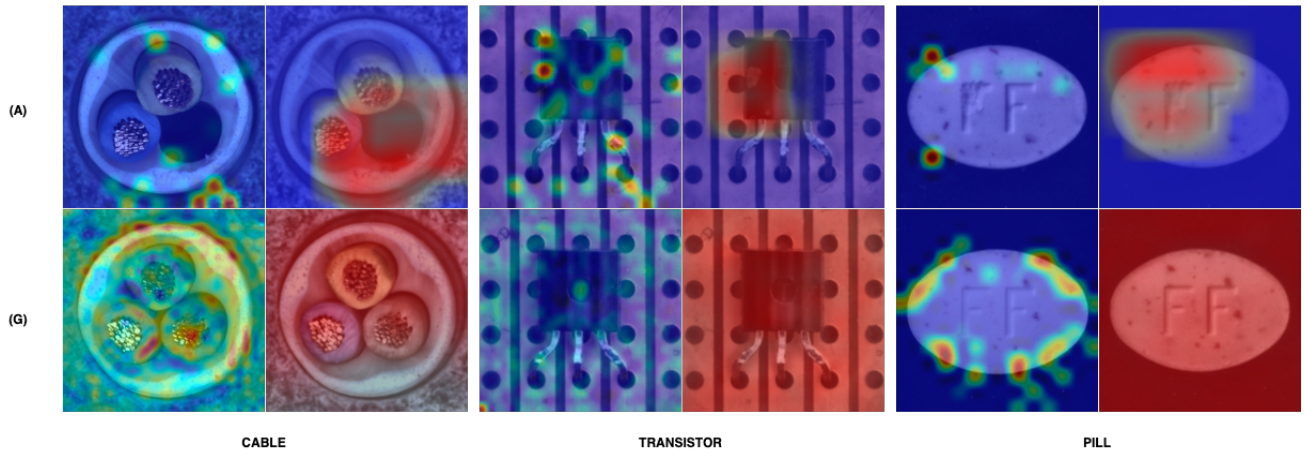

Fig. 1. The following are some examples comparing the activation maps produced for anomalous and genuine samples of Cable (missing one wire), Transistor (damaged case), and Pill (faulty imprint). The first row (A) represents anomalous samples, while the second row (G) represents genuine samples. For each example, the left column represents Gad-CAM, and the right column represents ATAC-Net outputs. The activated regions for ATAC-Net would zoom onto those areas and reiterate for better anomaly score detection.

Dataset	No. of Anomalies	CUT-MIX	Weighted CE	Deviation Loss	**ATAC-Net**
Carpet	5	0.764	0.803	0.852	0.924
Grid	5	0.729	0.931	0.947	0.988
Leather	5	0.941	0.984	0.993	1.000
Tile	5	0.902	0.935	0.987	1.000
Wood	5	0.944	0.988	0.985	0.996
Bottle	3	0.985	0.991	0.994	1.000
Capsule	5	0.777	0.914	0.911	0.934
Pill	7	0.824	0.852	0.866	0.921
Transistor	4	0.815	0.903	0.923	0.969
Zipper	7	0.927	0.989	0.990	1.000
Cable	8	0.870	0.864	0.892	0.983
Hazelnut	4	0.984	1.000	1.000	1.000
Metal nut	4	0.763	0.952	0.989	1.000
Screw	5	0.855	0.966	0.970	0.997
Toothbrush	1	0.863	0.901	0.861	0.879
MVTec-AD	15	0.862	0.932	0.944	0.973

Table 1. Ablative Comparison of ATAC-net over different experiments conducted over the ResNet-50 backbone on MVTEC dataset, showing the effectiveness of guided zoom over baseline approaches

Class	Anomaly Type	One Anomaly			Ten Anomaly		
		Dev-Net	DRA	ATAC-Net	Dev-Net	DRA	ATAC-Net
Tile	Crack	0.926	**0.975**	0.969	0.947	**0.986**	0.973
	Glue Strip	0.763	0.872	**0.904**	0.879	0.942	**0.953**
	Gray Stroke	0.621	0.905	**0.913**	0.884	**0.947**	0.943
	Oil	0.794	0.891	**0.909**	0.863	0.933	**0.941**
	Rough	0.752	**0.970**	0.968	0.932	0.959	**0.984**
	Mean	**0.771**	**0.923**	**0.933**	**0.901**	**0.935**	**0.959**
Metal Nut	Bent	0.797	0.952	**0.954**	0.904	0.990	**0.993**
	Color	0.909	0.946	**0.961**	0.978	0.967	**0.982**
	Flip	0.764	0.921	**0.945**	0.987	0.913	**0.991**
	Scratch	0.952	0.909	**0.964**	0.991	0.911	**0.982**
	Mean	**0.855**	**0.932**	**0.956**	**0.965**	**0.945**	**0.987**

Table 2. Comparison of ATAC-net with DRA and Deviation net on one class variations training to check the generalizability over different types of anomalies present within same class.