International Journal of Innovative Information Systems & Technology Research 19(4):76-83, Oct-Dec, 2026

© SEAHI PUBLICATIONS, 2026 www.seahipublications.org ISSN: 2467-8562

DOI:10.5281/zenodo.23357546

Supervised U-Net Segmentation for Industrial Surface Anomaly Localisation in Cleanliness Assessment

[1]E. G. Wonodi *[2]C. G. Igiri *[3]V. I. E. Anireh

[1,2&3]Department of Computer Science, Rivers State University, Port-Harcourt, Nigeria
Email: [1]grace.wonodi@rsu.edu.ng, [2]igiri.chima@ust.edu.ng, [3]anireh.ike@ust.edu.ng

ABSTRACT
Industrial cleanliness inspection often depends on visual assessment, whose consistency varies with illumination, surface reflectivity, residue visibility, and operator judgement. A supervised U-Net localises anomalous regions in industrial surface images for cleanliness assessment. The ImageNet-pretrained ResNet-34 encoder processes 256×256 RGB images and produces a pixel-wise anomaly probability map. Normal images receive zero masks, while anomalous images retain binary ground-truth masks. Image preprocessing converts BGR images to RGB, resizes inputs with area interpolation, scales values to $[0,1]$, and resizes masks by nearest-neighbour interpolation before binarisation. The 5,354 MVTec AD images are assigned to deterministic 70:15:15 training, validation, and held-out test partitions. The model is trained for 25 epochs with binary cross-entropy with logits and Adam optimisation. A sigmoid activation and a fixed 0.50 threshold produce binary masks. The test evaluator aggregates true-positive, false-positive, true-negative, and false-negative pixel counts after thresholding. On 803 held-out test images, the model achieved 89.86% precision, 68.70% recall, 77.87% Dice/F1, 63.76% intersection-over-union, and 99.57% pixel accuracy. The MPS implementation recorded mean inference time of 19.79 ms per image.
Keywords: industrial cleanliness assessment, anomaly localisation, semantic segmentation, convolutional neural network, MVTec AD

1. INTRODUCTION

Industrial cleanliness inspection supports product quality, equipment reliability, and workplace safety. Surface residues, stains, particulate deposits, oil films, and other visible contaminants can indicate inadequate cleaning or developing process problems. In many facilities, inspection still depends on visual checks by operators. The outcome can vary with inspector experience, fatigue, viewing angle, illumination, surface reflectivity, and the subtlety of the contamination. These sources of variation make consistent assessment difficult, particularly when large numbers of surfaces must be examined repeatedly [1].

Conventional machine-vision approaches can improve repeatability, but rule-based processing and handcrafted features are sensitive to changes in surface texture, lighting, camera position, and contaminant appearance. Convolutional neural networks (CNNs) provide an alternative by learning visual representations from images rather than relying only on fixed thresholds or manually specified features [2], [3]. Residual CNN architectures further support the training of deeper image models [4]. Their use is especially relevant to industrial imagery, where anomalous regions can be small, heterogeneous, and visually close to normal surface patterns.

The MVTec Anomaly Detection (MVTec AD) dataset provides a widely used benchmark for this setting. It contains 5,354 high-resolution colour images across 15 object and texture categories, with defect-free training images, normal and anomalous test images, and pixel-precise annotations for anomalous regions [5]. Its defect types include scratches, dents, contaminations, and structural changes, making it suitable for evaluating image-based industrial surface assessment under a controlled benchmark setting.

The method treats cleanliness assessment as supervised anomaly localisation. RGB preprocessing, an ImageNet-pretrained ResNet-34 encoder, and binary-mask supervision produce pixel-level anomaly masks on a deterministic MVTec AD split. Evaluation uses aggregate confusion counts, precision, recall, Dice/F1, and IoU.

2. RELATED WORKS

Industrial visual inspection includes anomaly detection, defect detection, and defect localisation. The first task determines whether an image contains an irregularity, whereas localisation identifies the affected region. MVTec AD is central to this area because it contains high-resolution images from object and texture categories, more than 70 defect types, and pixel-precise annotations for anomalous test samples [5]. Its conventional protocol trains on defect-free images and evaluates image-level anomaly detection and pixel-level localisation on a separate test set. The masks supplied for anomalous test samples also support supervised segmentation, but that setting differs from the original normal-only benchmark protocol.

Normal-only methods address industrial settings in which defective examples are rare, costly to collect, or unavailable. Deep one-class classification constructs a compact description of normal observations and assigns larger anomaly scores to samples outside that description [6]. CutPaste creates local synthetic perturbations and trains a representation to distinguish altered patches from their normal context [7]. Student-teacher methods model normal feature representations with a teacher network and use disagreement between teacher and student outputs as anomaly evidence [8]. These approaches remove the need for manually annotated defect masks, but their output maps are inferred from deviations in appearance rather than learned directly from observed defect boundaries.

Reconstruction and distillation approaches also learn normal appearance without using annotated defect masks. DRÆM introduces synthetic anomalies during training, reconstructs a normal-looking image, and combines reconstruction residuals with discriminative segmentation to identify anomalous pixels [9]. Reverse distillation trains a student decoder to reproduce the teacher's normal feature embeddings; feature-reconstruction errors form the anomaly map [10]. Both methods generate dense localisation output, but the anomaly signal comes from reconstruction failure or synthetic corruption. Their learning objective therefore differs from supervised segmentation, where each positive pixel is paired with an observed binary annotation.

Feature-distribution and memory-bank methods represent normality at the patch level. PaDiM estimates a multivariate Gaussian distribution for normal patch embeddings and scores patches with low probability under that distribution [11]. SPADE retrieves related normal features across a deep feature pyramid and uses correspondence differences for sub-image anomaly detection [12]. PatchCore stores a coreset of nominal patch features and measures the distance from a test patch to its closest retained normal feature [13]. These methods can deliver both image-level and local anomaly scores, but their effectiveness depends on the quality of pretrained embeddings, the coverage of normal reference features, and the selected scoring rule.

Normalising-flow models estimate feature likelihoods instead of using a memory bank. CFLOW-AD learns conditional flows over multi-scale encoder features and treats low-likelihood feature locations as anomalous [14]. Fully convolutional cross-scale flows extend this idea by modelling dependencies between feature scales for image-based defect detection [15]. Flow-based methods provide probabilistic anomaly scores and dense maps, although they require a learned feature distribution and remain within the normal-only anomaly-detection setting.

Supervised segmentation provides a different route when pixel annotations are available. U-Net uses an encoder-decoder structure and skip connections to recover spatial detail during upsampling [16]. Residual networks enable deeper feature extraction through residual connections and offer multi-level representations suitable for an encoder [4]. A ResNet34-U-Net can therefore map an RGB surface

image directly to a binary anomaly mask. Its performance is measured against the annotation through pixel precision, recall, Dice/F1, and IoU, with Dice/F1 and IoU reflecting foreground overlap rather than the dominant background class.

Deployment work introduces speed, model-size, and explanation considerations. EfficientAD uses a compact student-teacher design to target millisecond-scale anomaly detection [17]. EfficientNet and vision transformers offer alternative CNN and transformer backbones for visual representation learning [18], [19]. Grad-CAM produces a post-hoc visualisation of CNN decision regions from gradients at a selected convolutional layer [20]. The ResNet34-U-Net configuration records per-image inference time as an execution measure.

Taken together, the reviewed methods show two distinct routes to industrial anomaly localisation. Normal-only approaches derive anomaly scores from deviations in appearance, features, or likelihood distributions. Supervised segmentation uses binary masks to optimise the spatial correspondence between predicted regions and annotated anomalies. The ResNet34-U-Net configuration follows the supervised route, combining residual image features with U-Net decoding to produce dense anomaly masks for pixel-level assessment.

3. METHODOLOGY

A ResNet34-U-Net performs supervised binary segmentation on normal and anomalous MVTec AD images. Normal images receive all-zero masks, while anomalous images retain their binary ground-truth masks. The positive class contains annotated MVTec anomaly and defect pixels, including contamination, scratches, dents, and structural defects. The resulting protocol uses binary-mask supervision rather than the dataset's conventional normal-only training arrangement.

Figure 1 presents the model-centred architecture. A 256×256 RGB surface image enters the ResNet-34 encoder, whose multi-scale features pass through the feature bridge to the U-Net decoder. Skip connections transfer encoder features directly to the decoder to retain spatial localisation detail. The decoder produces a pixel-wise anomaly probability map through a sigmoid activation, and thresholding converts the map to a binary anomaly mask.

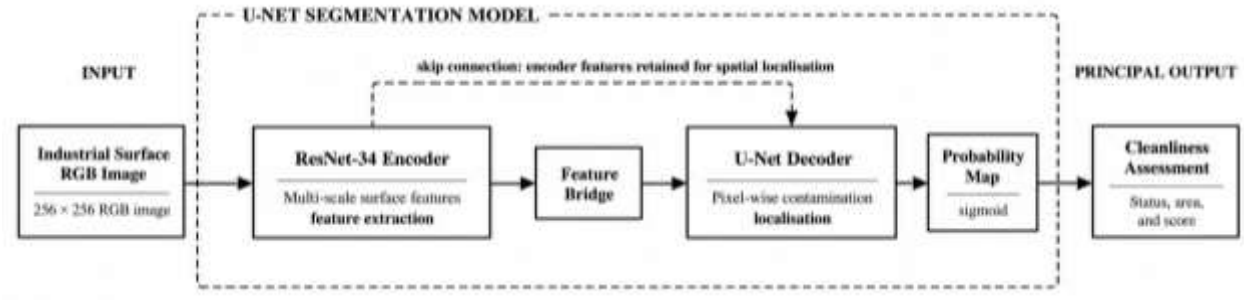

Figure 1: U-Net architecture for automated industrial cleanliness assessment.

Each image is read with OpenCV, converted from BGR to RGB, resized to 256×256, and scaled to $[0,1]$. Corresponding masks are resized with nearest-neighbour interpolation and binarised. A fixed seed of 42 controls the NumPy, Python, and PyTorch random generators. The complete 5,354-image dataset is shuffled and divided into permanent 70%, 15%, and 15% training, validation, and test partitions, containing 3,748, 803, and 803 images, respectively.

The model is a U-Net with an ImageNet-pretrained ResNet-34 encoder, three input channels, and one output channel [4], [16]. It predicts one logit per image pixel. A sigmoid function converts logits to probability maps, and a fixed threshold of 0.50 produces binary anomaly masks. Training uses binary cross-entropy with logits and Adam optimisation for 25 epochs. The checkpoint with the lowest validation loss is retained for held-out test evaluation.

Table 1: Experimental training and evaluation configuration.

Component	Configuration
Framework	PyTorch
Execution backend	Apple Metal Performance Shaders (MPS)
Segmentation model	U-Net with ImageNet-pretrained ResNet-34 encoder
Input and output	256×256 RGB input; one binary mask output channel
Dataset and split protocol	5,354 MVTec AD images; seed 42; permanent 70:15:15 split
Partition sizes	Training: 3,748; validation: 803; held-out test: 803
Loss and optimiser	Binary cross-entropy with logits; Adam
Training duration	25 epochs
Model selection	Minimum validation loss
Pixel threshold	0.50
Mask post-processing	None

For each pixel logit z, the predicted anomaly probability is $p = \frac{1}{1+e^{-z}}$.

MVTec AD contains 5,354 high-resolution RGB images across 15 industrial object and texture categories, including 3,629 defect-free training images and 1,725 test images [5]. In this supervised experiment, normal and anomalous images are used with binary masks to form the permanent partitions. Predicted masks and ground-truth masks yield pixel true-positive, false-positive, true-negative, and false-negative counts. Pixel precision, recall, Dice/F1, intersection-over-union, and accuracy are calculated from the aggregate held-out pixel counts.

4. Results and Discussion

The validation-selected U-Net is evaluated on 803 held-out images without augmentation. Each of the $803 \times 256 \times 256 = 52{,}625{,}408$ pixels is assigned an anomaly probability. Applying the fixed 0.50 threshold without post-processing produces the aggregate pixel confusion counts in Table 2.

Table 2: Held-out pixel-level confusion counts.

Outcome	Pixel count
True positive	401,132
False positive	45,252
True negative	51,996,244
False negative	182,780
Total	52,625,408

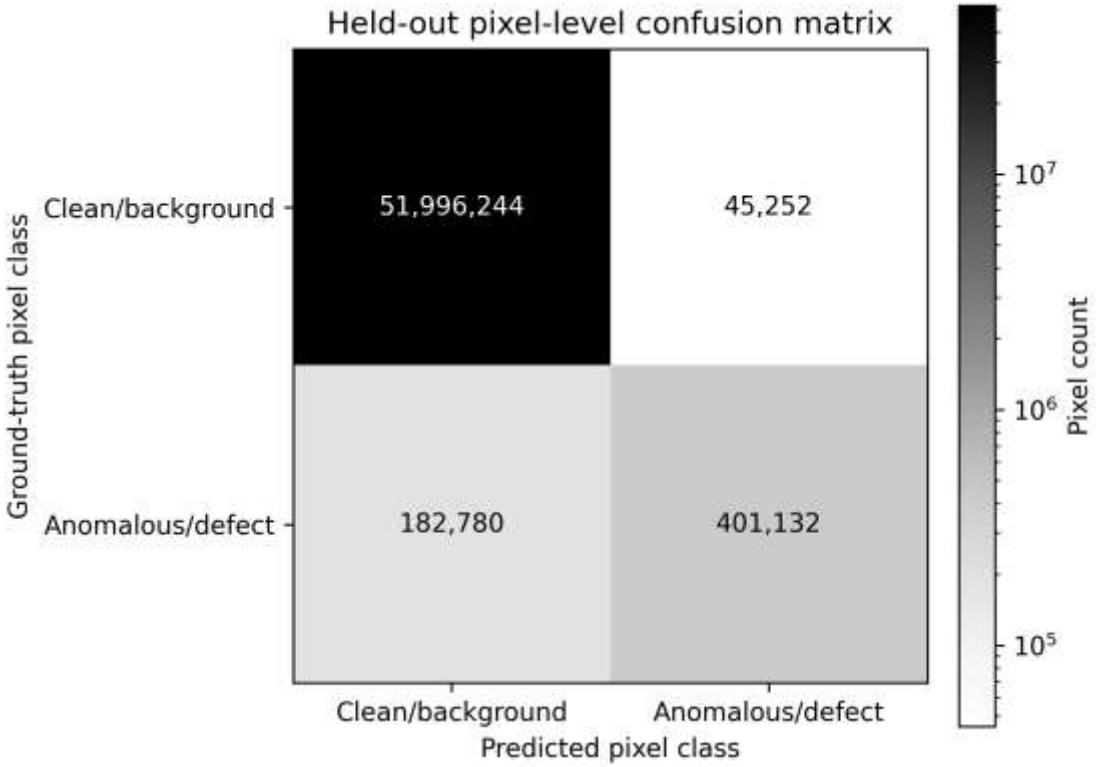

Figure 2: Held-out pixel-level confusion matrix.

4.1 Pixel-Level Segmentation Performance

Table 3 reports micro-averaged segmentation metrics from the aggregate held-out pixel counts. The model reached 89.86% precision, 68.70% recall, 77.87% Dice/F1, and 63.76% IoU at the fixed threshold.

Table 3: Held-out pixel-level segmentation performance.

Metric	Result
Pixel accuracy	99.57%
Precision	89.86%
Recall	68.70%
Dice/F1	77.87%
Intersection-over-union	63.76%
Mean inference time	19.79 ms per image

The held-out partition contained 52,041,496 clean/background pixels and 583,912 annotated anomalous pixels. Mean inference time was 19.79 ms per image on the MPS execution backend.

At the 0.50 threshold, the model produced 45,252 false-positive pixels and 182,780 false-negative pixels. Precision therefore exceeded recall by 21.16 percentage points. The results describe supervised training on a random split of MVTec AD images, not normal-only anomaly detection under the conventional MVTec AD protocol.

4.2 Training Dynamics

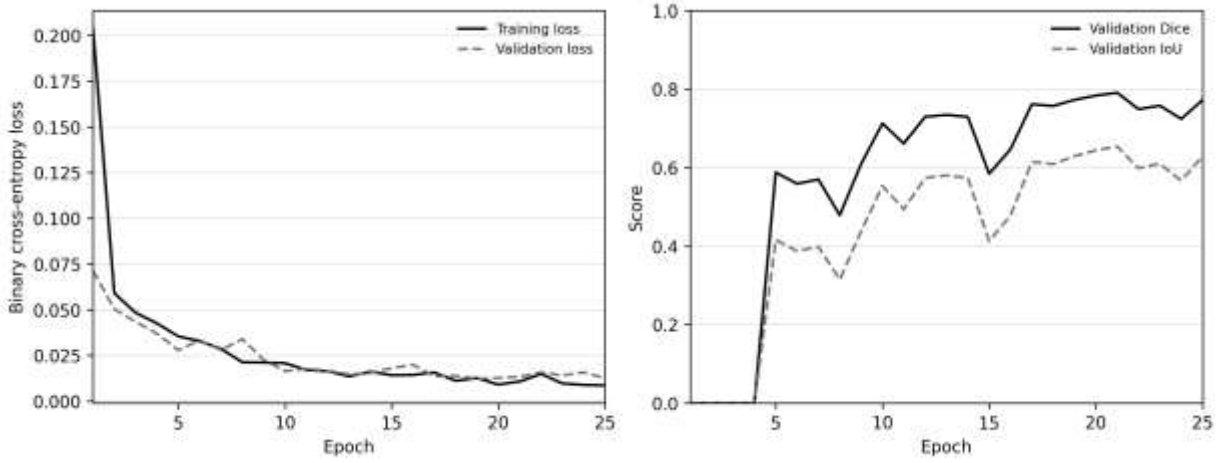

Figure 3: Training and validation metrics across 25 epochs.

Figure 3 shows the training and validation trajectories over 25 epochs. Validation loss declined from 0.0714 in the first epoch to its minimum value of 0.0122 at epoch 19, which produced the retained checkpoint. Validation Dice became non-zero at epoch 5 and reached 0.7905 at epoch 21; validation IoU reached 0.6536 in the same epoch. Across epochs 22–25, validation Dice ranged from 0.7237 to 0.7717 and validation IoU ranged from 0.5669 to 0.6283.

4.3 Discussion

The held-out results show a clear precision–recall difference. Of the 446,384 pixels predicted as anomalous, 401,132 were true positives, producing 89.86% precision. The 182,780 false-negative pixels reduced recall to 68.70%, indicating that the fixed 0.50 threshold omitted part of the annotated anomaly area. The corresponding Dice/F1 score of 77.87% and IoU of 63.76% quantify the overlap between the predicted and annotated foreground regions.

The test partition was strongly dominated by clean/background pixels: 52,041,496 of 52,625,408 evaluated pixels belonged to that class. This distribution accounts for the 99.57% pixel accuracy and makes the foreground measures more informative for the anomaly-localisation outcome. The confusion matrix shows that false positives were limited relative to the clean/background region, while the false-negative count remained substantial within the much smaller annotated anomaly region.

The training trajectory supports the checkpoint-selection result. Validation loss reached its minimum at epoch 19, while validation Dice and IoU reached their highest values at epoch 21. The difference reflects the use of binary cross-entropy loss for checkpoint selection and overlap measures for segmentation assessment. The held-out Dice/F1 of 77.87% falls within the late-epoch validation range shown in Figure 3, providing a consistent test-set result after model selection.

The 19.79 ms mean inference time was recorded on the MPS backend for 256×256 inputs. The measurement covers model inference only; image acquisition, preprocessing, display, and deployment integration were not timed. A complete pixel-level mask was produced during each measured inference interval.

PaDiM, PatchCore, DRÆM, and EfficientAD are commonly evaluated with MVTec AD's normal-only training protocol, whereas the ResNet34-U-Net uses annotated binary masks and a random image partition. The two settings use different training data, supervision, and split construction. The reported pixel metrics therefore characterise supervised binary-mask segmentation rather than normal-only anomaly detection.

5. Conclusion

A supervised ResNet34-U-Net was evaluated for pixel-level localisation of annotated industrial surface anomalies relevant to cleanliness assessment. The configuration used 256×256 RGB

inputs, binary-mask supervision, binary cross-entropy with logits, Adam optimisation, and a fixed 0.50 prediction threshold. A permanent 70:15:15 split of 5,354 MVTec AD images provided 3,748 training images, 803 validation images, and 803 held-out test images.

On the held-out partition, the model achieved 89.86% pixel precision, 68.70% pixel recall, 77.87% Dice/F1, 63.76% IoU, and 99.57% pixel accuracy. The aggregate confusion counts contained 401,132 true-positive pixels, 45,252 false-positive pixels, and 182,780 false-negative pixels. These results show that the model aligned most predicted anomaly pixels with annotated regions, while some annotated anomaly area remained undetected at the selected threshold. The 19.79 ms mean inference time on the MPS backend demonstrates the recorded per-image computation time for the evaluated configuration.

The measured values characterise supervised segmentation on the stated random split, binary-mask labels, 256×256 image resolution, and MPS execution environment. Category-specific reporting, field-collected cleanliness images, cross-dataset evaluation, threshold analysis, alternative losses, and common-protocol model comparisons would extend the evidence beyond this benchmark configuration. Such evaluations would establish how well binary-mask segmentation transfers from annotated benchmark anomalies to operational cleanliness inspection.

References

[1] Szeliski, R. (2022). *Computer vision: Algorithms and applications* (2nd ed.). Springer.

[2] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

[3] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 25, 1097–1105.

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

[5] Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., & Steger, C. (2021). The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4), 1038–1059.

[6] Ruff, L., Vandermeulen, R. A., Görnitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., & Kloft, M. (2018). Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning*, 4393–4402.

[7] Li, C.-L., Sohn, K., Yoon, J., & Pfister, T. (2021). CutPaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9664–9674.

[8] Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2020). Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4183–4192.

[9] Zavrtanik, V., Kristan, M., & Skočaj, D. (2021). DRÆM: A discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8330–8339.

[10] Deng, H., & Li, X. (2022). Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9737–9746.

[11] Defard, T., Setkov, A., Loesch, A., & Audigier, R. (2021). PaDiM: A patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges*, 12664, 475–489.

[12] Cohen, N., & Hoshen, Y. (2020). Sub-image anomaly detection with deep pyramid correspondences. *arXiv*, 2005.02357.

[13] Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., & Gehler, P. (2022). Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328.

[14] Gudovskiy, D., Ishizaka, S., & Kozuka, K. (2022). CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 98–107.

[15] Rudolph, M., Wehrbein, T., Rosenhahn, B., & Wandt, B. (2022). Fully convolutional cross-scale-flows for image-based defect detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1088–1097.

[16] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 9351, 234–241.

[17] Batzner, K., Heckler, L., & König, R. (2024). EfficientAD: Accurate visual anomaly detection at millisecond-level latencies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 128–138.

[18] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, 6105–6114.

[19] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

[20] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.