FlowCLAS: Enhancing Normalizing Flow-Based Anomaly Segmentation Via Contrastive Learning

Supplementary Material

6. Methodology

6.1. Normalizing flow architecture details

This section describes the structure of the affine coupling operation Affine [13] used in FlowCLAS which incorporates 2D subnets following [45]. As depicted in Fig. 4, the operation bisects the input $\mathbf{x}$ along the channel dimension. The first half serves as input to a subnetwork that generates affine transformation parameters which are applied to transform the second half. This process can be formalized as follows:

$$\begin{aligned} \mathbf{x}_1, \mathbf{x}_2 &= \text{split}\left(\mathbf{x}\right) \\ \mathbf{s}, \mathbf{t} &= \text{split}\left(\text{subnet}\left(\mathbf{x}_1\right)\right) \\ \mathbf{x}_2' &= \mathbf{x}_2 \cdot \exp\left(\mathbf{s}\right) + \mathbf{t} \\ \mathbf{x}' &= \text{concat}\left(\mathbf{x}_1, \mathbf{x}_2'\right) \end{aligned} \tag{9}$$

where subnet is a residual network described in Fig. 5

6.2. Mask-based Anomaly Score Smoothing

A detailed pseudo-code of the mask-based anomaly score smoothing pipeline can be found in Algorithm 1.

Algorithm 1 Mask-based Anomaly Score Smoothing

$masks \leftarrow$ `SAM 2`(I) ▷ Segment regions using SAM 2
for mask M in $masks$ **do**
  $(h, bins) \leftarrow$ `Histogram`$(scores_M, N)$
  $bin^* \leftarrow \arg\max_b\ h(b)$
  $score^* \leftarrow$ `Mean`$(\{s \in bin^*\})$
  for pixel p in M **do**
    if score(p) not in bin^* **then**
      $S(p) \leftarrow score^*$
    end if
  end for
end for
return $scores$

7. Experiments & results

7.1. Threshold bins for binary classification metrics

Conventional anomaly segmentation test sets, due to their relatively small size, allow each pixel-level anomaly score to function as an individual threshold bin, enabling tractable precise metric calculation. However, to overcome memory limitations and enable online processing with the extensive ALLO test set [29], we employ predefined threshold bins instead of the traditional method of post-processing all the scores simultaneously. Furthermore, to ensure compatibility with these bins, we constrain the anomaly scores to the $[0, 1]$ interval by applying the $\text{sigmoid}(\cdot)$ function prior to metric computation.

L	ALLO		FS-L&F		Road Anomaly		# of params
	AUPRC ↑	FPR_{95} ↓	AUPRC ↑	FPR_{95} ↓	AUPRC ↑	FPR_{95} ↓	
8	63.0	63.1	79.2	1.7	88.5	5.4	25.0 M
12	62.8	63.7	77.1	2.1	88.0	5.6	37.4 M
16	63.6	64.2	80.7	1.4	90.6	5.0	49.8 M

Table 8. Summary of results with varying flow steps L and the total count of trainable parameters. DINOv2-B was used for ALLO and DINOv2-B-Rein was used for FS-L&F and Road Anomaly.

7.2. COCO images for ALLO experiments

For the ALLO experiments, the backgrounds of COCO images [30] were excluded during training and not used to construct the set $\mathcal{B}$ described in Sec. 3. This decision was made to avoid confusion between the predominantly black space background and the colorful real-world backgrounds present in COCO images.

7.3. Fine-tuning with Rein

We used Rein [44] to fine-tune DINOv2-B,L for inlier semantic segmentation on ALLO and Cityscapes. This process follows the semantic segmentation pre-training procedure performed by the supervised baselines. The fine-tuning was conducted with a batch size of 8, a learning rate of 6×10^{-5} using polynomial decay, the AdamW optimizer, and 40,000 training iterations.

7.4. Extra visualizations

Figure 8 compares anomaly heatmaps predicted by UNO and FlowCLAS on Fishyscapes Lost&Found (FS L&F) [4] and Road Anomaly [31]. In both FS L&F examples (top), FlowCLAS more accurately detects the baskets than UNO, the previous supervised state-of-the-art. Similarly, FlowCLAS successfully identifies the cow in the Road Anomaly example (bottom-left) and reduces false-positive detections in the road region surrounding the peccary (bottom-right).

7.5. Additional ablation studies

Our analysis of the flow step count L in FlowCLAS, detailed in Table 8, reveals a non-monotonic relationship between model depth and performance, highlighting a crucial

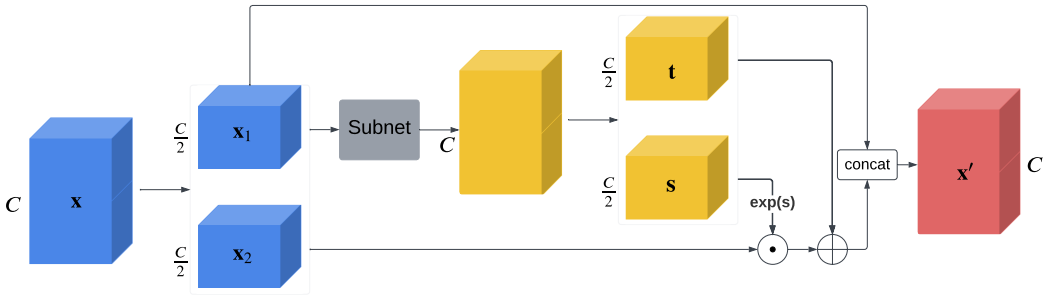


Figure 4. The feature or latent maps $\mathbf{x}$ are bisected along the channel dimension. The first half $\mathbf{x}_1$ is processed by a learnable module, subnet which generates affine transformation parameters $\mathbf{s}$ and $\mathbf{t}$. These parameters are then applied to the second half $\mathbf{x}_2$, yielding a transformed $\mathbf{x}_2'$. The final output is obtained by concatenating $\mathbf{x}_1$ and $\mathbf{x}_2'$ along the channel axis.

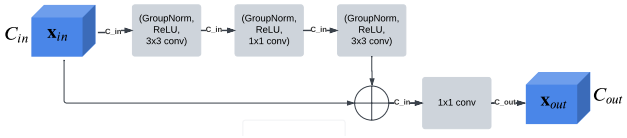


Figure 5. The residual network used in the affine coupling layer, denoted as 'subnet' in Fig. 4.

τ	FS-L&F		Road Anomaly	
	AUPRC ↑	FPR_{95} ↓	AUPRC ↑	FPR_{95} ↓
0.07	80.0	1.7	88.9	5.5
0.10	81.1	1.4	89.4	5.2
0.13	80.7	1.4	90.6	5.0

Table 9. Summary of results with varying temperature scale parameters τ.

trade-off between expressivity and optimization. Theoretically, increasing L enhances the normalizing flow's expressive power. However, our results show that increasing L from 8 to 12 leads to a slight degradation in performance across the road anomaly benchmarks (*e.g.* AUPRC drops from 79.2 to 77.1 on FS-L&F). This counter-intuitive result suggests that while model capacity increases (from 25.0M to 37.4M parameters), the model may become more difficult to optimize or prone to overfitting. The more complex transformation might model the training inliers too tightly, potentially degrading its ability to generalize for outlier separation without a corresponding increase in expressivity that provides a tangible benefit. However, increasing the step count further to $L = 16$ consistently yields the best performance across all datasets, achieving the highest AUPRC and lowest FPR_{95} on FS-L&F and Road Anomaly. This suggests that with a sufficient increase in capacity, the model can finally leverage its enhanced expressivity to learn a superior representation of the complex normal data distribution, overcoming the optimization hurdles observed at $L = 12$. Based on these findings, we selected $L = 16$ for our main experiments to maximize performance, accepting the trade-off in computational cost.

Our analysis of the temperature scale hyperparameter τ, reveals that FlowCLAS is relatively insensitive to the choice of τ within the tested range of 0.07 to 0.13. Across both the FS-L&F and Road Anomaly benchmarks, the AUPRC varies by less than 2 points, indicating that the framework's performance is not overly sensitive to this hyperparameter. Furthermore, the optimal temperature appears to be dataset-dependent. On the FS-L&F dataset, the performance is non-monotonic, peaking at $\tau = 0.10$, while on the Road Anomaly dataset, both AUPRC and FPR_{95} show a clear monotonic improvement as τ increases. This suggests that a higher temperature, which "softens" the contrastive loss, is more beneficial for the Road Anomaly dataset's distribution. Given that $\tau = 0.13$ yields the best overall performance profile—achieving the top AUPRC on Road Anomaly and the best FPR_{95} across both datasets—it was selected as the default value for our main experiments.

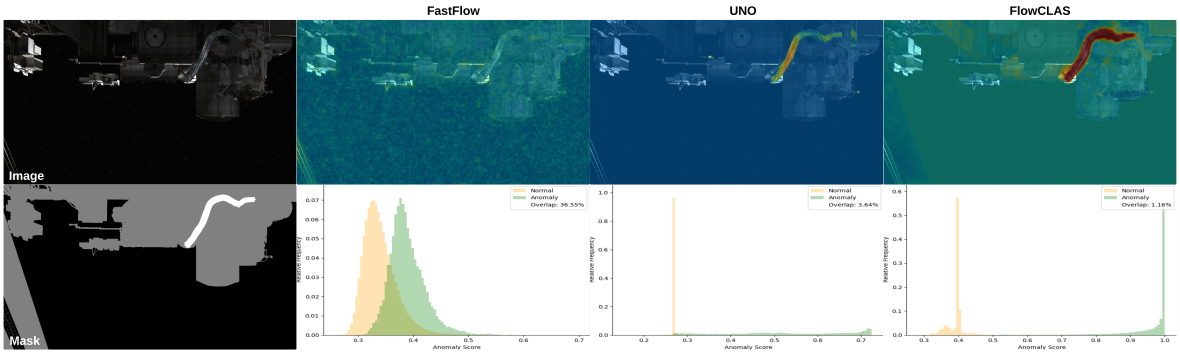


Figure 6. Predicted heatmaps (top) and anomaly score histograms (bottom) for an ALLO test image. While the leading unsupervised method, FastFlow [45], fails to detect the anomalous cable, and the supervised SOTA, UNO [10], detects a part of the main body, FlowCLAS provides a more complete segmentation that captures the entire object structure. The histograms below visually illustrate this, showing the superior anomaly score separation between the two classes achieved by FlowCLAS compared to both methods.

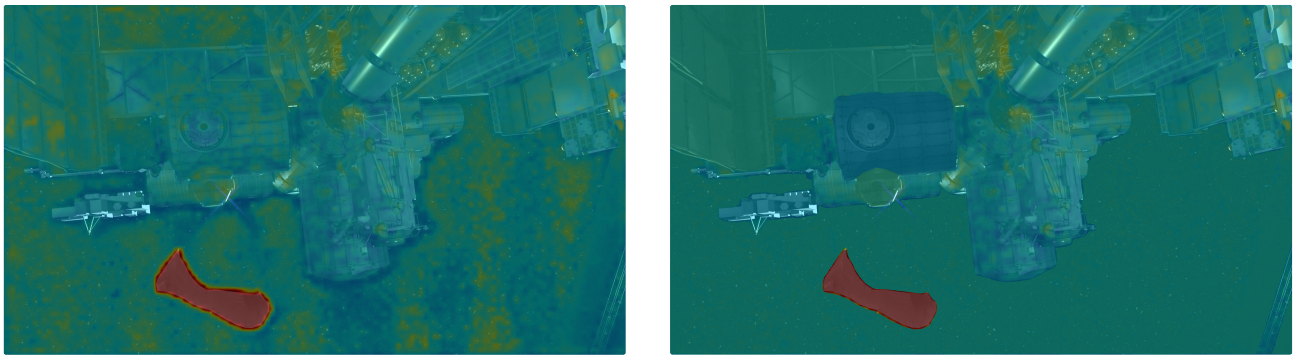
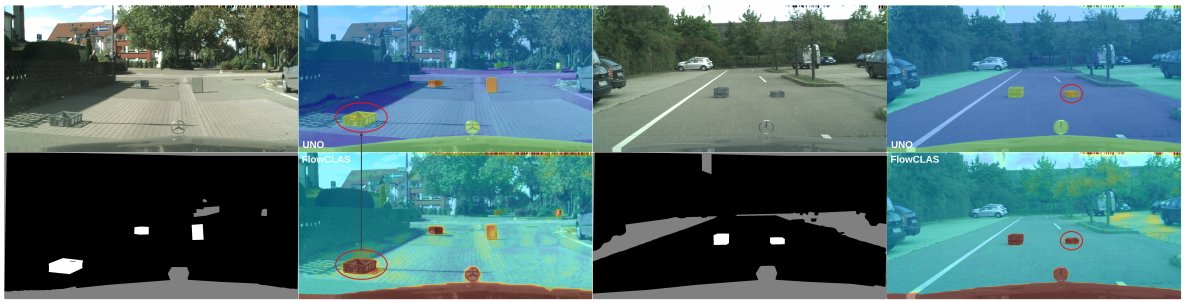
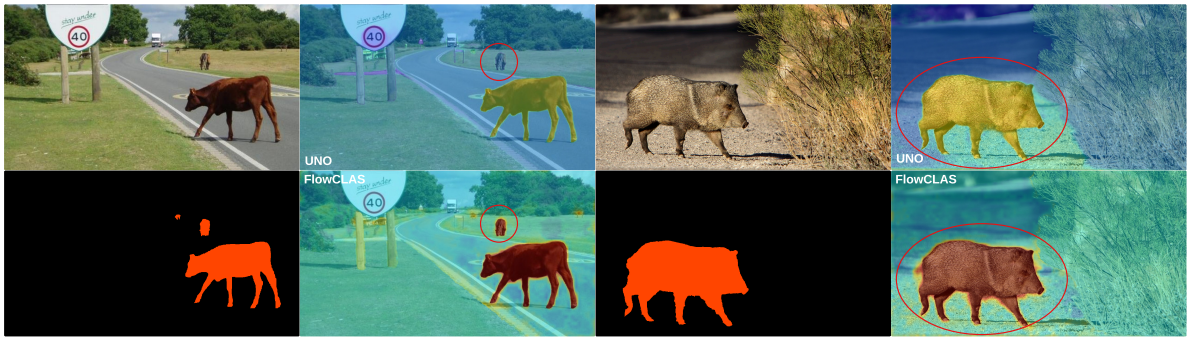


Figure 7. Visual comparison of heatmaps from FlowCLAS without (left) and with (right) mask-based score refinement.



(a) Fishyscapes Lost&Found



(b) Road Anomaly

Figure 8. Predicted heatmaps from UNO [10] and FlowCLAS for examples from the Fishyscapes Lost&Found [4] (top) and Road Anomaly [31] (bottom) validation sets.