machines

MDPI

Article

Automated Weld Defect Detection in Radiographic Images Using Normalizing Flows

Morteza Mahvelatishamsabadi and Sudong Lee *

Department of Industrial Engineering, University of Ulsan, Ulsan 44610, Republic of Korea; morteza1987@ulsan.ac.kr
* Correspondence: sudonglee@ulsan.ac.kr

Abstract

Anomaly detection is a pressing issue, particularly in industrial images. Detecting weld defects in radiographic images is a challenge due to the small signal-to-noise ratio (SNR) and the limited availability of data. In this paper, we propose an automated weld defect detection method using Normalizing Flows (NFs). We employed various state-of-the-art NF architectures with different feature extractors to detect defects in radiographic images of welds, comprehensively comparing the results with radiographic images of welded steel pipes collected from industrial sites. The results show that the combination of CFlow-AD with a wide residual network-50-2 (WRN-50-2) outperformed the other methods, indicating its effectiveness in anomaly detection.

Keywords: anomaly detection; Normalizing Flows; weld defect detection

check for updates

Academic Editors: Feng Zhang, Feifei Zhao and Fan Yang

Received: 4 July 2025
Revised: 22 August 2025
Accepted: 3 September 2025
Published: 9 September 2025

Citation: Mahvelatishamsabadi, M.; Lee, S. Automated Weld Defect Detection in Radiographic Images Using Normalizing Flows. *Machines* **2025**, *13*, 836. https://doi.org/10.3390/machines13090836

Correction Statement: This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Detecting weld defects is essential for maintaining the integrity and safety of welded constructions in manufacturing. Defects in welds compromise the structural integrity of welded joints, potentially leading to failures in critical applications such as pipelines, pressure vessels, and structural elements across both the construction and manufacturing sectors [1]. Identifying and addressing defects early on is crucial to prevent costly repairs and severe breakdowns. Radiographic Testing (RT) is a commonly used nondestructive testing (NDT) technique that utilizes penetrating radiation, such as X-rays or gamma rays, to examine the internal structure of materials and components without damaging the inspected object [2]. RT is especially effective in detecting defects in welded components, as it identifies a variety of internal defects, including cracks, porosity, lack of fusion, and other irregularities. RT is highly accurate, suitable for various component shapes and sizes, and can detect subsurface discontinuities in welds [3]. Generally, trained human inspectors evaluate radiographic images of weld defects. However, as manufacturing processes advance, there is a growing need for automated systems capable of detecting weld defects using artificial intelligence. These automated systems can enhance the accuracy and efficiency of NDT by reducing human error.

Studies show that deep learning is a practical approach for identifying defects in images and has attracted significant interest within the realm of RT images. Convolutional neural networks (CNNs), autoencoders (AE), and generative adversarial networks (GANs) have been proven effective in detecting defects in images, with CNNs enhancing their detection and measurement by analyzing images using convolutions. However, such supervised learning requires an enormous training dataset consisting of both normal and

defective data. In contrast, an AE is an unsupervised learning method, as it learns simplified representations of normal data and uses the reconstruction error to detect defects. For instance, a deep spatial autoencoder model can effectively capture normal anatomical variations in complete 2D brain magnetic resonance imaging (MRI) slices [4]. However, AE suffers from an overgeneralization problem in that defective patterns are accurately reconstructed by the AE trained on normal data [5]. GAN is another unsupervised anomaly detection approach that uses both generative and discriminative loss functions for defect detection. Although the discriminator often improves anomaly detection accuracy, its learning is unstable and requires significantly more data than AEs do [6].

A significant challenge in applying deep learning to automated weld defect detection in radiographic images stems from the inherent properties of the data itself [7], SNR, and the limited availability of defective samples. Typically, weld defect patterns are tiny, and the radiographic images are characterized by low contrast and high noise levels, which makes visual defect detection by human inspectors challenging and necessitates advanced automated methods. The extremely low SNR requires a highly sophisticated feature extractor that can effectively learn and represent the subtle, often faint, defect patterns. The second major obstacle is data scarcity; collecting a comprehensive dataset of defective welds with accurate labels is often impractical in industrial settings [8]. Therefore, we address these challenges by proposing an unsupervised anomaly detection framework that learns a powerful defect detection model using solely normal (non-defective) data, thereby eliminating the need for labeled defective samples during training. In this paper, we propose deep learning-based weld defect detection models using NFs because of the following advantages:

1. The accurate estimation of the likelihood density distribution of complex data in a tractable and interpretable function.
2. The representation of normal images to differentiate defective instances from different distributions.
3. Compatibility with various feature extractors for images, such as convolutions and transformers.

Herein, we employ different combinations of state-of-the-art NF architectures with various feature extractors to detect defects in radiographic images of welds. The remainder of this paper is organized as follows. Section 2 reviews related work on deep learning approaches for industrial image anomaly detection, Section 3 discusses different NF methods, and Section 4 presents our dataset and experimental results. Finally, Section 5 contains a discussion and conclusion.

2. Related Work

Deep learning approaches for defect detection are typically categorized as either supervised or unsupervised. Supervised methods train on a fully labeled dataset to learn a classification boundary between normal and defective samples. However, this approach is often impractical for weld defect detection due to the diverse nature and scarcity of real-world defect data. Consequently, this study focuses on unsupervised anomaly detection, which does not require labeled defect samples. These methods are broadly divided into two categories: reconstruction-based and feature embedding-based. We propose a combined method as follows.

2.1. Reconstruction-Based Methods

Reconstruction-based methods operate on the principle of training a model to accurately reconstruct normal data. Defects are then identified by pinpointing inputs that the model fails to reconstruct accurately, resulting in a high reconstruction error. Autoencoders

serve as a fundamental model in this category, utilizing an encoder–decoder architecture to first compress data into a lower-dimensional representation and then reconstruct it [9]. More advanced methods have been built upon this foundation, as follows.

- DRAEM [10] enhances anomaly detection by incorporating a discriminative subnetwork alongside the reconstruction model. It trains the discriminative component on artificially generated anomalous samples from normal training data.
- FAVAE [11] aims to improve reconstruction fidelity not only in the pixel space but also across multiple feature spaces by leveraging a pre-trained CNN. This approach ensures the model learns more semantically meaningful representations by penalizing reconstruction errors in both the image and its perceptual feature maps.
- AE-GAN [12] is a combination of AE and GAN. In its setting, it utilizes parameter sharing between the GAN's generator and the AE's decoder to preserve their distribution as closely together as possible, which helps stabilize the generator training.

While generative models, such as GANs and variational autoencoders (VAEs), have shown impressive results in learning image distributions for reconstruction-based anomaly detection, their practical application can be limited. A key issue is that they do not allow for the exact evaluation of the probability density of new data points. Furthermore, their training processes are often unstable, facing challenges such as posterior collapse, mode collapse, vanishing gradients, and training instability [13].

2.2. Feature Embedding-Based Methods

Feature embedding-based methods aim to detect defects by transforming high-dimensional input data, such as images, into a compact, low-dimensional feature space where normal and defective data are more easily separable. In this learned representation, the model establishes a distribution of normal patterns; samples that deviate significantly from this distribution are then classified as defects. The effectiveness of these methods is critically dependent on the quality of the feature extractor, which significantly impacts the final accuracy of anomaly detection [14,15]. CNNs are widely adopted for this role due to their proven ability to extract hierarchical features from images efficiently and recognize complex patterns with high accuracy [16].

Several unsupervised approaches leverage this embedding concept. For instance, CutPaste [17] trains a model on self-supervised tasks, using cut-and-paste augmentations to learn an image-level representation that distinguishes normal patterns from artificially created defective samples. Similarly, NFs are powerful deep generative models that learn explicit probability density for the complex distribution of normal data. By applying a series of invertible transformations, NFs map the intricate distribution of feature embeddings from normal images to a simple, tractable base distribution, such as a standard Gaussian, as illustrated in Figure 1. Unlike other generative models, NFs allow for the exact calculation of likelihood, providing a precise anomaly score based on the probability of a given sample belonging to the learned normal distribution [9]. For an input image, its anomaly score is determined by the distance between its embedding and the learned distribution of normal embeddings in this transformed space.

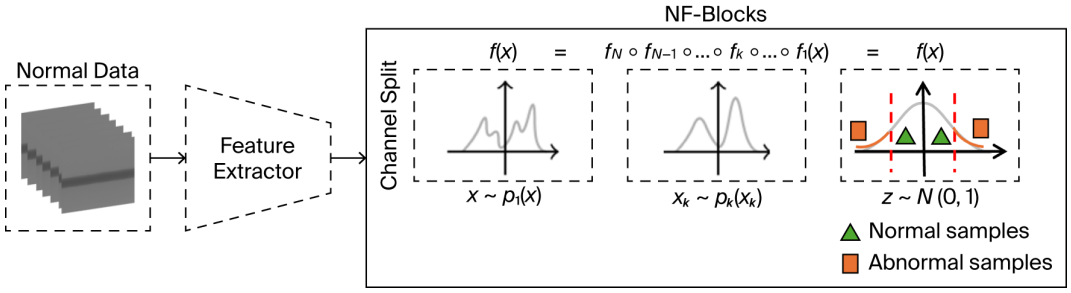


Figure 1. Overview of an NF model with a feature extractor and NF blocks. The feature extractor first extracts informative features, which are then fed into the NF blocks. A transformation sequence converts the complex distribution into a standard distribution, such as a Gaussian distribution.

3. Method

In this section, we first describe the fundamental principles of NFs. We then detail the various components of our proposed method, including the feature extractors, coupling block architecture, and the anomaly scoring mechanism. Finally, we introduce the specific NF models evaluated in this study: CFlow-AD, FastFlow, and CSFlow.

3.1. Normalizing Flows

Normalizing Flows (NFs) are deep generative models that transform a simple base distribution, such as a standard Gaussian, into a complex target distribution through a sequence of invertible mappings (Figure 1). Let $x \in \mathbb{R}^d$ be an observed sample and $z = f_\theta(x)$ its latent representation after applying K transformations $f = f_K \circ \cdots \circ f_1$. By the change-of-variables rule, the density of x [18] is

$$p_X(x) = p_Z(z) \prod_{k=1}^{K} \left| det\left(\frac{\partial f_k}{\partial h_{k-1}}\right)\right|, \quad h_0 = x, \quad h_K = z \tag{1}$$

and the corresponding log-likelihood becomes

$$logp(x) = -\frac{1}{2}\|z\|^2 - \frac{d}{2}log(2\pi) + \sum_{k=1}^{K} log\left|det\ \frac{\partial f_k}{\partial h_{k-1}}\right| \tag{2}$$

where $z \sim \mathcal{N}(0, I)$. For affine coupling layers, the Jacobian matrix is triangular, so its log-determinant simplifies to a sum over scale outputs:

$$\log\left|det\left(\frac{\partial f_k}{\partial h_{k-1}}\right)\right| = \sum_i s_k(h_{k-1})_i \tag{3}$$

In simple terms, the Gaussian term measures how well the transformed data fit the base space, while the Jacobian terms adjust for how each transformation stretches or compresses the space. This tractable formulation enables exact and efficient likelihood computation, which is crucial for anomaly detection in image data.

Finally, the NF is trained by maximizing the total log-likelihood of the training set:

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} logp(x_n) \tag{4}$$

3.2. Feature Extractor

The choice of feature extractor is critical to the performance of NF models, as it provides the informative and discriminative representations required for effective anomaly detection. To ensure fair and meaningful comparisons, we employed consistent and powerful backbone architectures, primarily from the CNN and Vision Transformer families. A key challenge was accommodating models like CSFlow [19], which require multiscale features,

with standard transformer architectures (e.g., Data-efficient image Transformer (DeiT) [20] and Class-attention image Transformer (CaiT) [21]) that produce single-scale representations. We addressed this by utilizing the Pyramid Vision Transformer (PVT) [22], whose inherent hierarchical structure naturally generates multiscale feature maps, ensuring compatibility.

Furthermore, we explored a range of robust backbones, including Vision Transformer (ViT) [23] for its powerful global attention, EfficientNet-B5 [24] for its balance of efficiency and performance, and ResNeXt [25] for its enhanced representational capacity. Recognizing that an ideal feature extractor would combine the large receptive fields of CNNs with the dynamic attention of transformers, we also incorporated the Convolutional Vision Transformer (CVT) [26] into our approach. By merging convolutional inductive biases with self-attention, the CvT architecture enhances semantic representation and improves the model's ability to focus on defective regions. This curated selection of feature extractors strengthens the robustness and representational power of our NF-based anomaly detection framework.

3.3. Coupling Blocks

To construct complex invertible transformations while maintaining a tractable Jacobian determinant, our NF models are built using affine coupling blocks. In each block, the input vector x is split into two disjoint parts, x_1 and x_2. The transformation is then defined as

$$\begin{cases} y_1 = x_1 \\ y_2 = x_2 \odot exp(s(x_1)) + t(x_1) \end{cases} \tag{5}$$

where y_1 and y_2 are the corresponding outputs, and $y = concat(y_1, y_2)$ as an output of a coupling block and the input of the next block, and $\odot$ denotes element-wise multiplication. The scale function $s(\cdot)$ and shift function $t(\cdot)$ are arbitrary complex functions, typically implemented as neural networks, that take x_1 as an input. This transformation is easily invertible, resulting in a triangular Jacobian matrix. This makes the log-determinant computation highly efficient, as it simplifies to the sum of the outputs of the scale function

$$log\left|det\left(\frac{\partial y}{\partial x}\right)\right| = \sum s(x_1) \tag{6}$$

This computational efficiency is crucial for training deep NF models.

3.4. Anomaly Score

To detect defective instances, we utilize the log-likelihood values provided by the trained NF model. Since the model is trained exclusively on normal data, it will assign a high log-likelihood to normal samples and a low log-likelihood to defective samples that deviate from the learned distribution. The process is as follows:

1. For a given input image, the NF model outputs a feature map where each element represents the log-likelihood of the corresponding image region.
2. We define the anomaly score as the negative log-likelihood. This way, a higher score corresponds to a higher degree of anomaly.
3. This score map is then up-sampled to the original image resolution using bilinear interpolation to create a pixel-wise score heatmap.
4. The final image-level anomaly score is determined by the maximum value in this heatmap. If this score exceeds a predefined threshold, the image is classified as anomalous.

3.5. NF Models

We evaluated three state-of-the-art NF models, each with a distinct architectural approach to anomaly detection.

- CFlow-AD [27] is a conditional NF that enhances distribution mapping by incorporating positional information. It uses a CNN feature extractor with multiscale pyramid pooling and injects 2D sinusoidal positional encoding vectors into its coupling layers, making it sensitive to the spatial location of features. Additionally, it utilizes RealNVP [28] with linear subnets.
- FastFlow [29] is designed for speed and efficiency. It employs a 2D NF structure with fully convolutional networks. This enables it to model both local and global distributions effectively and generate anomaly maps directly from 2D feature maps, resulting in faster inference times. It is compatible with various deep feature extractors, such as ResNet [30] and DeiT.
- CSFlow, which stands for Cross-Scale Flow, is designed to handle features at multiple scales explicitly. It jointly estimates the likelihood across a multiscale feature pyramid extracted by a pre-trained network, enabling it to localize defects of varying sizes and complexities better.

4. Experiments

This section details our experimental setup and presents the results of our evaluation. We first describe the dataset used for training and testing, followed by the implementation details and performance metrics. We then provide a comprehensive quantitative and qualitative comparison of the evaluated anomaly detection models.

4.1. Data Description

The dataset used in this study consists of RT images of welded steel pipes collected from industrial sites. As detailed in Table 1, we split the dataset into training, validation, and inference sets. By the principles of unsupervised anomaly detection, the training set consists exclusively of 10,000 normal (non-defective) images. The validation and inference sets contain a mixture of normal and defective samples, which are used for hyperparameter tuning and final evaluation, respectively. The dataset comprises 11 distinct types of weld defects, external welding defects, including crack (CR), porosity (PO), Undercut (UC), and internal welding defects like Incomplete Penetration (IP), Slag Inclusion (SI), and lack of fusion (LF), as illustrated in Figure 2. Table 2 provides a detailed breakdown of the number of defective instances available for training and testing, along with their descriptive properties across these categories. The highly imbalanced nature of the dataset, with a vast number of normal samples and a very small number of defective samples, reflects a realistic industrial scenario.

Table 1. Description of the dataset. The table displays the number of trainings, validations, and inference data used in this study, along with the proportion of normal and defective data at each stage.

Dataset Split	Normal Images	Defective Images	Total Images
Training Set	10,000	0	10,000
Validation Set	2500	134	2634
Test set	14,934	66	15,000
Total	27,434	200	27,634

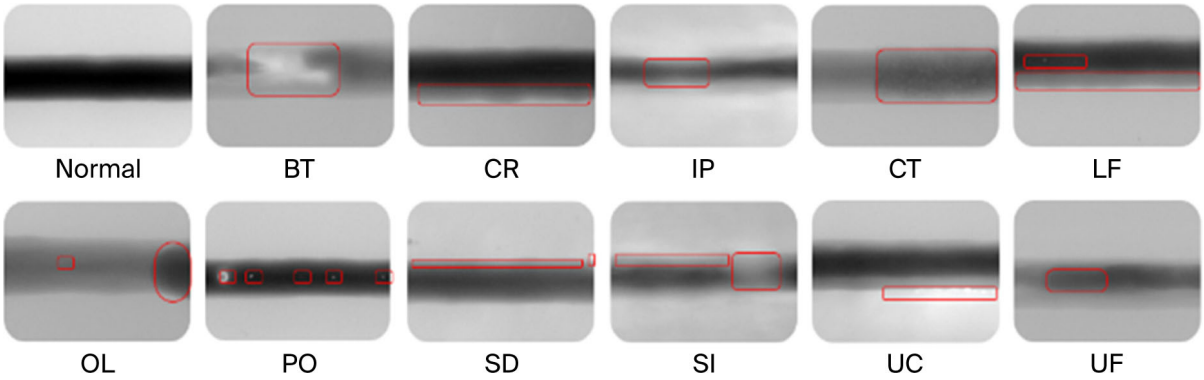


Figure 2. Example of normal and defective images. The image in the top left corner is normal, while the rest are defective. In the first row, from left to right, the categories of weld defects are BT, CR, IP, CT, and LF. In the second row, the categories are OL, PO, SD, SI, UC, and UF.

Table 2. Description of the defect types in our datasets.

Defect Types	Number of Training Samples	Number of Test Samples	Defect Description
Burn Through (BT)	2	1	Localized melting and collapsing of the base metal causes a hole in the weld.
Crack (CR)	3	1	Linear fractures in the weld or heat-affected zone, either during solidification (hot cracks) or after cooling (cold cracks).
Contamination (CT)	43	10	Foreign substances such as oil, grease, oxide, or dirt on or inside the weld metal can weaken the bond.
Incomplete Penetration (IP)	20	5	Weld metal not extending through the joint thickness, leaving unbonded root areas.
Lack of Fusion (LF)	2	0	Incomplete bonding between the weld metal and base material is caused by inadequate heat or poor technique.
Overlap (OL)	26	13	Weld metal flows over the base surface without fusing, forming a lip at the weld toe
Porosity (PO)	75	22	Gas-filled cavities or bubbles trapped in the weld metal due to contamination or poor shielding.
Surface Defect (SD)	15	1	Visible irregularities on the surface of the weld bead include issues such as pits, roughness, and irregular profile.
Slag Inclusion (SI)	5	1	Solid non-metallic material, such as flux residue, is trapped in the weld due to insufficient cleaning or low heat input.
Undercut (UC)	64	11	A groove in the base metal along the weld remains unfilled, which reduces joint strength.
Underfill (UF)	20	1	The weld bead surface is below the level of the adjacent base metal due to insufficient filler material.

4.2. Experimental Setting

To ensure a comprehensive and fair evaluation, we carefully configured each model–feature extractor combination. We determined the input image resolutions that were compatible with the backbone architecture, adjusting the number of coupling layers to balance model complexity with detection performance. Key hyperparameters, such as

learning rate and batch size, were optimized using the validation set. This rigorous setup allowed us to robustly assess the performance of each architecture and identify the optimal configurations for detecting weld defects in challenging radiographic images Figure 3.

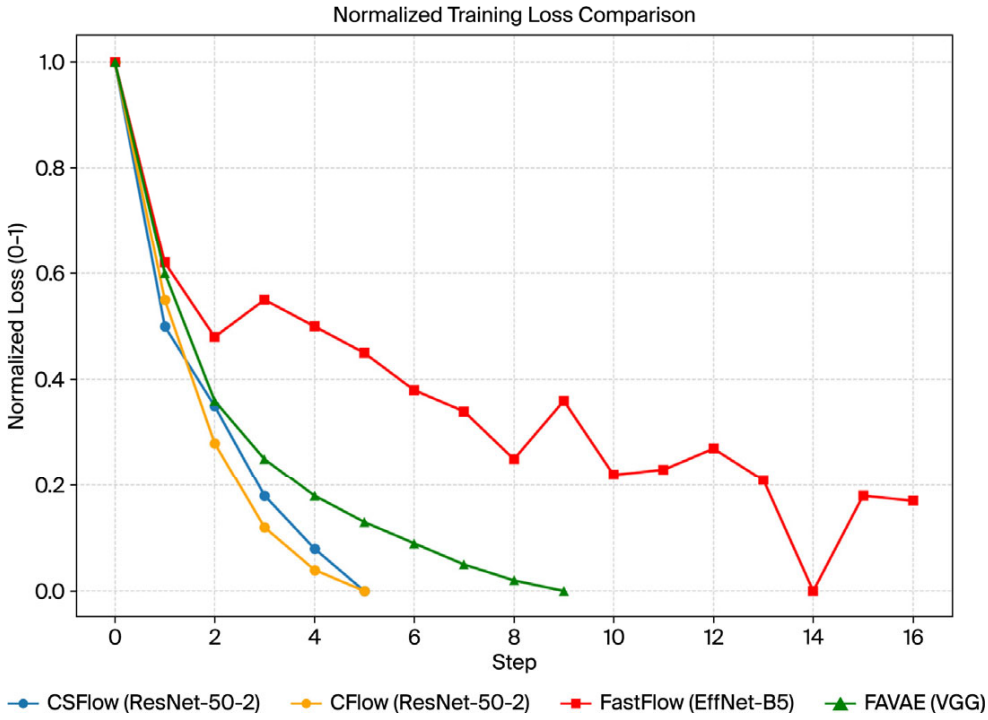


Figure 3. CFlow achieves the fastest and most stable convergence, consistently reducing loss to nearly zero in fewer steps. CSFlow follows, with a steady but slower decline, while FAVAE and FastFlow converge less efficiently and exhibit more fluctuations. These results highlight CFlow's superior training stability and reliability, making it the most effective framework for robust anomaly detection.

4.3. Performance Measures

For model comparison, we selected decision thresholds by maximizing the F1-score, which offers a balanced measure of precision and recall in class-imbalanced conditions. This approach allowed for a fair and consistent benchmark across different models. Additionally, we report the training likelihood loss for four different models: CFlow-AD, FastFlow, CSFlow, and FAVAE, based on their AUPRC performance, as shown in Figure 3. We also present confusion matrices for four models that achieved the highest recall in their respective categories, as indicated in Figure 4. Besides, AUROC as a discriminative metric is calculated in Table 3 and AUPRC curves in Figure 5, which are threshold-independent and illustrate the recall–precision trade-off across a wide range of operating points. We note that in real-world safety-critical scenarios, threshold selection would reasonably be shifted toward prioritizing recall, even if it reduces precision. For a more detailed analysis at a specific decision point, we assess four confusion matrices in Figure 4 and calculate accuracy, precision, recall, and the F1-score.

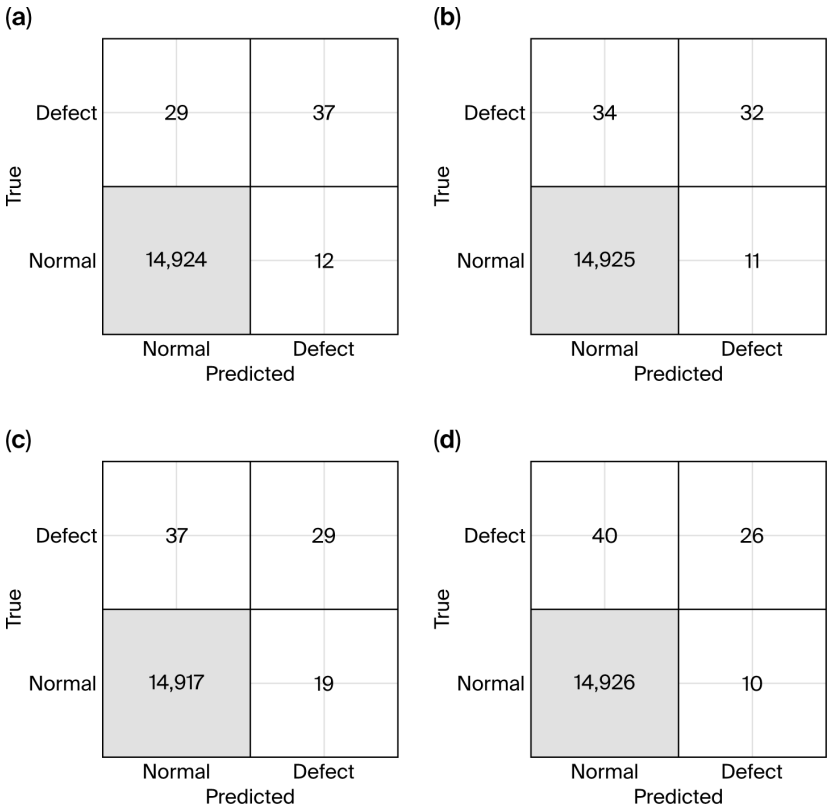


Figure 4. Confusion matrices and recall–precision results for four models. (**a**) CFlow (WRN-50-2) achieved the best balance, with high recall and precision, capturing more defects while minimizing false alarms. (**b**) FastFlow (CaiT-m48-448) showed similar precision but lower recall. (**c**) CSFlow (CvT) exhibited lower precision and recall. (**d**) FAVAE (VGG-16) had higher precision but missed many defects. These results highlight each approach's strengths and the trade-off between sensitivity and precision in anomaly detection.

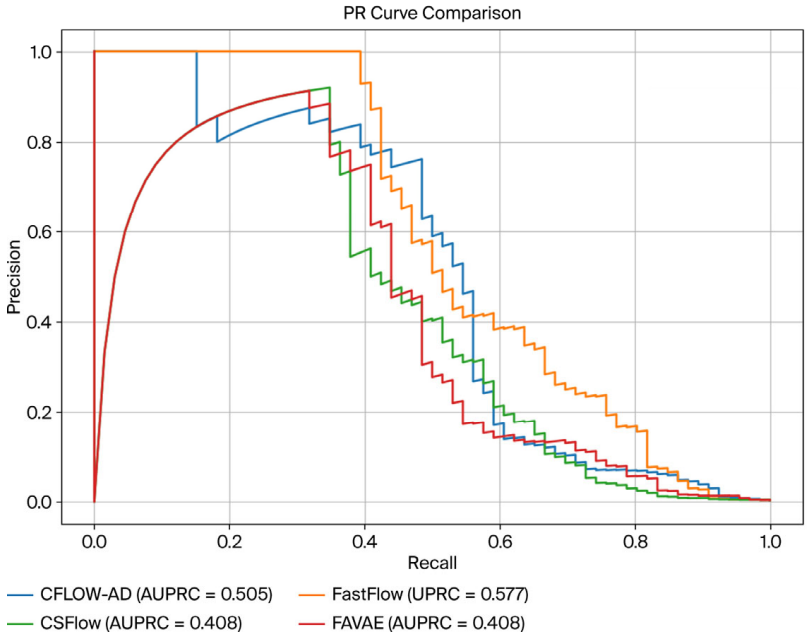


Figure 5. Plotting the highest AUPRC of the four top models. The blue, orange, green, and red lines represent the performance of CFlow, FastFlow, CSFlow, and FAVAE, respectively, using their specific extractors: WRN-50-2, EfficientNet-b5, WRN-50-2, and VGG-16.

Table 3. Experimental results of the defect detection models.

Learning Method	Model	Feature Extractor	Metrics					
			AUROC	AUPRC	Accuracy	Precision	Recall	F1-Score
Reconstruction-based	DRAEM	CNN	0.810	0.140	0.995	0.300	0.250	0.270
	FAVAE	VGG-16	0.927	0.410	0.997	0.722	0.394	0.510
	AE-GAN	CNN	0.520	0.010	0.990	0.005	0.040	0.040
Embedding-based	CutPaste	CNN	0.530	0.010	0.994	0.011	0.045	0.065
	CSFlow	WRN-50-2	0.903	0.410	0.996	0.710	0.410	0.520
		ResNeXt50-32×4d	0.930	0.380	0.996	0.630	0.340	0.441
		EfficientNet-B5	0.912	0.291	0.995	0.406	0.391	0.398
		PVT	0.911	0.390	0.996	0.600	0.440	0.507
	CFlow-AD	WRN-50-2	**0.958**	0.505	**0.998**	0.755	**0.561**	0.643
		ResNeXt50-32×4d	0.940	0.460	0.966	0.625	0.530	0.573
		EfficientNet-B5	0.941	0.536	0.997	0.710	0.520	0.600
		CvT	0.923	0.466	0.966	0.627	0.484	0.546
		ViT-384	0.954	0.501	0.997	0.833	0.530	**0.648**
	FastFlow	WRN-50-2	0.910	0.405	0.997	0.727	0.484	0.581
		EfficientNet-B5	0.924	**0.580**	0.997	**0.870**	0.410	0.557
		CVT	0.912	0.468	0.996	0.660	0.470	0.549
		DeiT-patch16-384	0.925	0.441	0.997	0.770	0.455	0.572
		CaiT-m48-448	0.943	0.540	0.997	0.750	0.485	0.590

The bold values indicate the highest value in each metric.

4.4. Quantitative Comparison

The experimental results are summarized in Table 3, providing a comprehensive comparison across all evaluated models.

4.4.1. Baseline Reconstruction Models

The performance of reconstruction-based models varied significantly. While methods like AE-GAN and CutPaste performed poorly on this dataset, FAVAE, which leverages a pre-trained VGG-16 [23] backbone to guide reconstruction, achieved a respectable AUROC of 0.927. Nevertheless, its performance provided a strong baseline, which the top-performing NF models ultimately surpassed.

4.4.2. Normalizing Flow Models

We thoroughly evaluated three NF models—CFlow-AD, FastFlow, and CSFlow—with various feature extractors.

- CFlow-AD consistently demonstrated superior performance across most metrics. The combination with WRN-50-2 yielded the highest AUROC of 0.958 and a strong F1-score of 0.643, representing an optimal trade-off between recall and precision. When paired with ViT-384, it achieved the highest precision of 0.833 and the best F1-score of 0.648, indicating its effectiveness in accurately identifying actual defects with a low false positive rate.
- FastFlow also showed impressive results, particularly when using transformer-based backbones. With CaiT-m48-448, it reached an AUROC of 0.943 and a high precision of 0.750, but its lower recall of 0.485 resulted in a more moderate F1-score of 0.590. This suggests that while FastFlow is precise, it may miss some defective instances compared to CFlow-AD.
- CSFlow displayed the most modest performance among the NF models. Its best configuration, WRN-50-2, achieved an AUROC of 0.903 and an F1-score of 0.520. While still effective, its comparatively lower scores suggest a limited ability to handle the subtle defect patterns present in our dataset.

4.5. Qualitative Comparison

To assess the localization ability of the models visually, we examined the pixel-by-pixel anomaly score maps (heatmaps) shown in Figure 6. These heatmaps offer qualitative evidence of how effectively the models highlight defective regions. To move from visualization to quantitative assessment, we applied a threshold to the anomaly scores, classifying pixels above the threshold as part of a defect region and those below as background. This binarization allowed us to measure the length (L) of the detected defect region along its longest axis, the width (W) across its shortest axis, and the aspect ratio ($\frac{L}{W}$). The results for the CFlow-AD model with Wide ResNet-50-2 are summarized in Table 4. By connecting the qualitative heatmaps with these quantitative metrics, we demonstrate that the model not only localizes defects but also captures their geometric dimensions, which are critical for industrial defect evaluation.

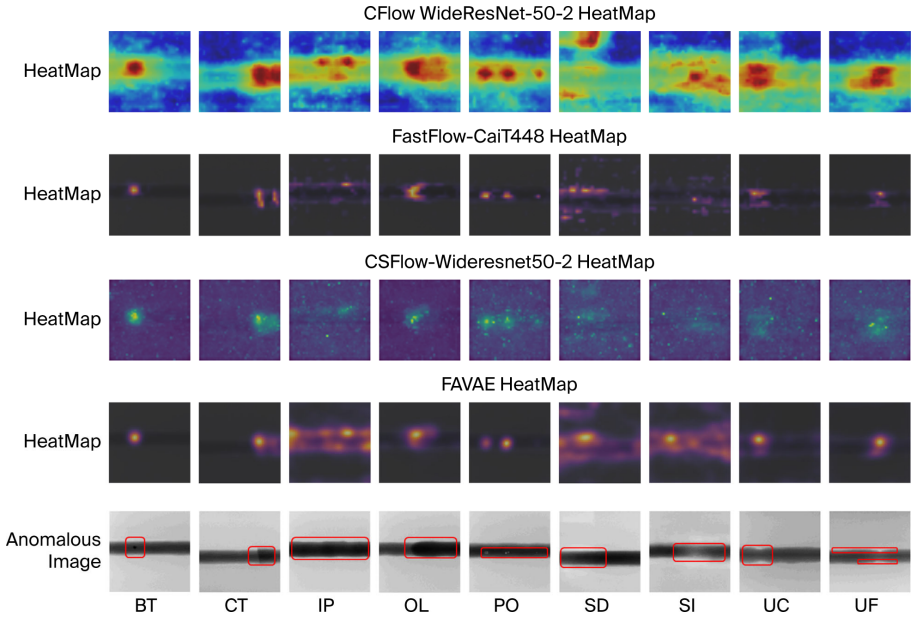


Figure 6. Pixel-wise anomaly heatmaps are overlaid on weld radiographs (**top**), with the original and contrast-enhanced images below for reference. Warmer colors indicate higher anomaly scores. For visual consistency, the color scale is normalized within this 9-image panel. Dimensional readouts—length (longest extent), width (shortest extent), and area—are derived from the thresholded heatmap masks.

Table 4. Quantitative evaluation of weld defect localization was performed using thresholded heatmaps from the CFlow-AD model with WRN-50-2. After applying a cut-off threshold to the anomaly scores, connected defect regions were identified, and their length (L), width (W), and aspect ratio (L/W) were calculated. These measurements complement the qualitative visualizations in Figure 6, confirming that the proposed approach not only highlights defective regions but also captures their geometric size and shape, which are essential for safety-critical inspection tasks.

Defect Types	Threshold	Length	Width	Ratio ($\frac{L}{W}$)	Unit	Area
BT	0.822	76.322	55.886	3277	pixel	0.822
CT	0.949	86.989	51.369	3277	pixel	0.949
IP	0.818	123.721	35.119	2687	pixel	0.818
OL	0.953	85.405	47.736	3116	pixel	0.953
PO	0.868	144.750	34.472	3241	pixel	0.868
SD	0.856	78.545	54.918	3277	pixel	0.856
SI	0.855	55.549	38.170	1618	pixel	0.855
UC	0.951	66.511	65.112	3277	pixel	0.951
UF	0.962	81.184	57.251	3277	pixel	0.961

- CFlow-AD-WRN-50-2 achieves the best coverage–leakage trade-off and results in the lowest leakage across classes, indicating sharper, more reliable localization. It produces strong, well-localized anomaly signals across most defect types (e.g., OL, PO, BT, UF), demonstrating its effectiveness in precisely identifying anomalous regions.
- FastFlow-CaiT-m48-448 tends to underestimate the length of elongated defects (e.g., UC/BT), producing finer-grained but less intense heatmap activations. This trait can be helpful for detecting subtle defects, but it may sometimes fail to generate a strong signal for more dispersed defects, such as BT and UC.
- CSFlow-WRN-50-2 exhibits more diffuse, low-contrast activations, and heatmaps appear weaker and less focused compared to the other NF models. The dispersed, low-intensity signals indicate a lower confidence in anomaly localization.
- FAVAE-VGG16 slightly overestimates W for diffuse classes (e.g., PO/SI), showing intermediate performance, successfully identifying compact defects like PO and SI, but with less precise boundaries than CFlow-AD, which reflects its moderate quantitative scores.

5. Discussion

The results of this study demonstrate the significant potential of NF models, particularly CFlow-AD architecture, for the challenging task of detecting weld defects in radiographic images. Our detailed experimental evaluation revealed that CFlow-AD, particularly when combined with powerful feature extractors such as WRN-50-2 and ViT-384, consistently outperformed competing NF models and reconstruction-based baselines. This section discusses the interpretation of these findings, as well as the broader implications of our work and its limitations.

The superior performance of CFlow-AD can be attributed to several factors. Its conditional architecture, which injects positional encoding into the flow, appears to be highly effective for this task, where the spatial location of a defect is critical. This enables the model to learn a more precise distribution of normal features conditioned on their position, thereby increasing its sensitivity to localized defective instances. In contrast, while FastFlow showed high precision, its lower recall suggests it may struggle with more diverse or subtly expressed defect types. The moderate performance of CSFlow indicates that its cross-scale feature integration may not have provided a significant advantage for the specific types of defects present in our dataset, which are often small and localized. This highlights a crucial insight: the optimal NF architecture is strongly dependent on the nature of the target anomalies.

From an industrial perspective, the successful application of an unsupervised method, such as CFlow-AD, is highly significant. It addresses the critical bottleneck of data scarcity in manufacturing environments and provides a pathway for developing robust, automated quality control systems for nondestructive testing. Such systems can reduce inspection costs, minimize human error, and ultimately enhance the structural safety and integrity of welded components. Academically, our work reinforces the importance of the interaction between feature extraction backbones and generative models. The strong results obtained by combining WRN-50-2 and ViT-384 with NFs pave the way for future research into more sophisticated hybrid architectures.

Despite the promising results, this study has limitations. As observed in the qualitative analysis, the model's localization performance was less distinct for certain defect categories, such as UC. This suggests that more complex or subtle defect patterns still pose a challenge. Future research should focus on enhancing the model's sensitivity to these hard-to-detect defects, perhaps by incorporating attention mechanisms directly within the NF coupling blocks or by exploring more advanced data augmentation techniques. Furthermore, extending this methodology to learn from multimodal distributions could allow a single model

to represent an even wider variety of defect types and normal manufacturing variations, further improving its real-world applicability

6. Conclusions

This study investigates the application of NF models for unsupervised weld defect detection in radiographic images, a critical task made challenging by SNR and data scarcity. Through a comprehensive set of experiments, we demonstrated that an NF-based approach, specifically CFlow-AD paired with a WRN-50-2 feature extractor, achieves state-of-the-art performance, outperforming other advanced unsupervised methods. The model exhibited an AUROC of 0.958 and an F1-score of 0.643, effectively detecting and localizing various types of weld defects using only normal data for training.

Our findings demonstrate that the combination of powerful feature extractors and carefully selected NF architecture yields a robust and highly effective solution for industrial anomaly detection. Future work will focus on enhancing performance on more subtle defect patterns and expanding the model's ability to handle diverse data distributions, further paving the way toward automated and reliable quality control in manufacturing.

Author Contributions: Conceptualization, M.M. and S.L.; methodology, M.M. and S.L.; software, M.M.; validation, M.M. and S.L.; formal analysis, M.M. and S.L.; investigation, M.M. and S.L.; resources, S.L.; data curation, S.L.; writing—original draft preparation, M.M.; writing—review and editing, M.M. and S.L.; visualization, M.M.; supervision, S.L.; project administration, S.L.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work and the APC were funded by the research fund of the University of Ulsan (Title: Image anomaly detection for manufacturing processes using Normalizing Flows/Grant Number: 2024-0327).

Data Availability Statement: The data used in the paper is the property of a private company and cannot be made public.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shaloo, M.; Schnall, M.; Klein, T.; Huber, N.; Reitinger, B. A review of non-destructive testing (NDT) techniques for defect detection: Application to fusion welding and future wire arc additive manufacturing processes. *Materials* **2022**, *15*, 3697. [CrossRef] [PubMed]
2. Gunasekaran, S. Evaluation of Automatic Ultrasonic Testing (AUT) in Lieu of Radiography for Weld Inspection. Indian National Seminar & Exhibition on Non-Destructive Evaluation NDE 2016, Dec, Thiruvananthapuram. *e-J. Nondestruct. Test.* **2017**, *22*.
3. Abas, A.A.; Shamsudin, M.K.S.; Azaman, N. Application of Phased Array Ultrasonic Testing (PAUT) on single V-butt weld integrity determination. In Proceedings of the Nuclear Technical Convention 2015 (NTC 2015), Bangi, Malaysia, 3–5 November 2015.
4. Baur, C.; Wiestler, B.; Albarqouni, S.; Navab, N. Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. In *Revised Selected Papers, Part I, Proceedings of the Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, Granada, Spain, 16 September 2018*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 161–169.
5. Spigler, G. Denoising autoencoders for overgeneralization in neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 998–1004. [CrossRef] [PubMed]
6. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*; Neural Information Processing Systems Foundation, Inc. (NeurIPS): La Jolla, CA, USA, 2016.
7. Moore, C.; Wood, T.; Saunderson, J.; Beavis, A. Correlation between the signal-to-noise ratio improvement factor (KSNR) and clinical image quality for chest imaging with a computed radiography system. *Phys. Med. Biol.* **2015**, *60*, 9047. [CrossRef] [PubMed]
8. Harley, J.B.; Sparkman, D. Machine learning and NDE: Past, present, and future. *AIP Conf. Proc.* **2019**, *2102*, 90001. [CrossRef]
9. Rezende, D.; Mohamed, S. Variational inference with normalizing flows. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 1530–1538.

10. Zavrtanik, V.; Kristan, M.; Skočaj, D. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 8330–8339.
11. Tukra, S.; Hoffman, F.; Chatfield, K. Improving visual representation learning through perceptual understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 14486–14495.
12. Sevyeri, L.R.; Fevens, T. On the effectiveness of generative adversarial network on anomaly detection. *arXiv* **2021**, arXiv:2112.15541. [CrossRef]
13. Kobyzev, I.; Prince, S.J.; Brubaker, M.A. Normalizing flows: An introduction and review of current methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3964–3979. [CrossRef] [PubMed]
14. Arsov, N.; Mirceva, G. Network embedding: An overview. *arXiv* **2019**, arXiv:1911.11726. [CrossRef]
15. Huang, X.; Song, Q.; Yang, F.; Hu, X. Large-scale heterogeneous feature embedding. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 3878–3885.
16. O'shea, K.; Nash, R. An introduction to convolutional neural networks. *arXiv* **2015**, arXiv:1511.08458. [CrossRef]
17. Li, C.-L.; Sohn, K.; Yoon, J.; Pfister, T. Cutpaste: Self-supervised learning for anomaly detection and localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 9664–9674.
18. Rudolph, M.; Wandt, B.; Rosenhahn, B. Same same but differnet: Semi-supervised defect detection with normalizing flows. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Online, 5–9 January 2021; pp. 1907–1916.
19. Rudolph, M.; Wehrbein, T.; Rosenhahn, B.; Wandt, B. Fully convolutional cross-scale-flows for image-based defect detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 1088–1097.
20. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; pp. 10347–10357.
21. Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; Jégou, H. Going deeper with image transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 32–42.
22. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 568–578.
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
24. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
25. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
26. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 22–31.
27. Gudovskiy, D.; Ishizaka, S.; Kozuka, K. 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, Waikoloa, HI, USA, 3–8 January 2022; pp. 98–107.
28. Dinh, L.; Sohl-Dickstein, J.; Bengio, S. Density estimation using real nvp. *arXiv* **2016**, arXiv:1605.08803.
29. Yu, J.; Zheng, Y.; Wang, X.; Li, W.; Wu, Y.; Zhao, R.; Wu, L. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv* **2021**, arXiv:2111.07677. [CrossRef]
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.