ATTENTION ENHANCED U-NET MODELS FOR IMAGE SEGMENTATION

Nikhil Kumar Boddu, Li Zhang

Royal Holloway, University of London, Surrey, UK
E-MAIL: nikhil.boddu.2023@live.rhul.ac.uk, li.zhang@rhul.ac.uk

Abstract:

Deep learning advances computer vision applications significantly, improving challenging deployments such as autonomous driving and clinical image analysis. However, exact pixel-level segmentation remains hard because of its complex image structures and variations across datasets. To address these challenges, this research explores several U-Net-based architectures and proposes enhanced models integrating attention mechanisms like SE, CBAM, and ECA. The proposed models show crucial enhancements in segmentation performance when evaluated against the baseline U-Net architectures, evaluated using various metrics across the investigated datasets.

Index Terms: Image Segmentation, Deep Learning, CBAM, SE, ECA.

1. Introduction

Image segmentation is an essential assignment in computer vision which contains splitting an image into meaningful parts for accurate interpretation and analysis. It is broadly used in applications like autonomous driving, medical image analysis, and agricultural monitoring. Traditional segmentation techniques based on thresholding, edge detection, and clustering often struggle with noisy images and complex structures. Lately, deep learning techniques, specifically CNNs, have considerably enhanced segmentation results by directly learning hierarchical characteristics from the data of images.

Among deep learning models, U-Net has become one of the majorly used frameworks because its architecture has the shrinking and expanding paths, which will help to keep the layout of picture intact. Even though, the standard U-Net still struggles when dealing with complex textures, irregular object boundaries, and diverse datasets. To address these limitations, in this work, we investigated several U-Net-inspired architectures and improved attention-based models for image segmentation. The key findings of this study are as follows:

- We compared baseline U-Net with its modified versions(R-2-U-Net, Attention-U-Net, Attention-/r2-U-Net) to check which handles image splitting well.
- We proposed enhanced U-Net architectures by incorporating attention strategies including CBAM, SE blocks, and ECA to boost image representation and partitioning accuracy.
- We analysed our new architecture using ISIC Skin Lesion dataset and the Carvana Image Masking Challenge dataset using segmentation measures such as the Dice score, mIoU, and the accuracy percentage.

2. Related Work

Latest progress in neural networks has boosted the quality of image mapping, specially through CNN-based architectures. Among this **U-Net**, developed by Ronneberger et al. [1][2],which is the combination of contracting-expanding pipeline that safeguard essential image. Because of its effectiveness and capability to work under limited training data, U-Net became widely used in biomedical image segmentation and other computer vision applications. Several improved variants of U-Net have been introduced to improve feature representation and segmentation accuracy. **R2U-Net** [3][4] integrates recurrent convolutional layers and residual connections into the U-Net architecture, enabling better contextual feature extraction while maintaining efficient parameter usage. This architecture improves gradient flow during training and elevates the systems's capability to collect fine image features. To strengthen segmentation results, attention mechanisms have been incorporated into U-Net architectures. **Attention U-Net** [5][6] introduces attention gates that make the architecture to emphasize on chosen image segments by reducing irrelevant extra noise features. This mechanism has shown improved segmentation performance in several medical imaging tasks.

Building on these ideas, **Attention R2U-Net** [5][6][7] combines the benefits of recurrent residual connections with attention mechanisms, allowing the network to learn richer contextual features and improve segmentation precision across complex datasets. In addition to architectural modifications, attention modules have been integrated into U-Net to improve feature learning. The **CBAM** [8][9] introduces spatial and channel attention mechanisms to optimize activation maps and highlight important regions in an image. Similarly, **SE blocks** [10] enhance channel-wise feature relationships by recalibrating feature maps according to their importance. Another lightweight attention mechanism, **ECA** [11][12], improves channel interaction without introducing significant computational complexity.

These attention-based improvements have shown consistent improvements in image-splitting jobs by improving model's capability to emphasize on relevant image structures. There are also various other segmentation models with dynamic architecture generation and hyper-parameter optimization [13-20].

3. Proposed Approach

The baseline U-Net framework, a popular paradigm for the image segmentation. In this project baseline framework and its variants were implemented to enhance feature learning and localization. The following frameworks were proposed extending the standard model.

The baseline model: It has the standard contracting-expanding pathway architecture with direct layer links, as shown in Figure 1.

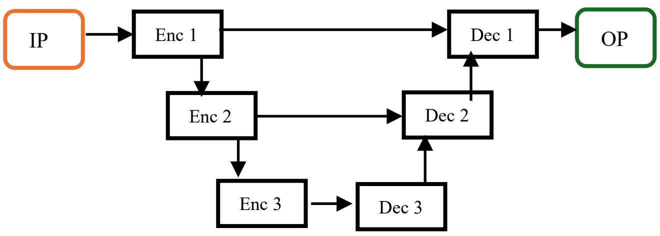

Figure 1: Architecture of baseline model.

CBAM enhanced with baseline model: Incorporates CBAM with standard model for **extracting** spatial and channel-wise dependency. Figure 2 shows the proposed CBAM-enhanced with baseline model.

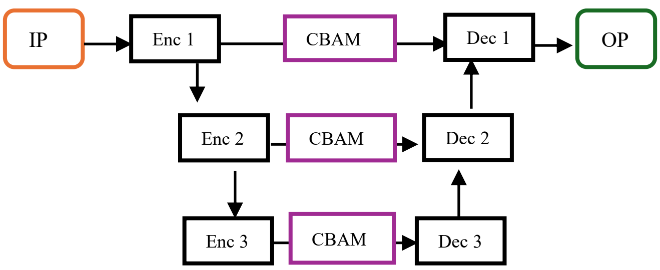

Figure 2: Architecture of CBAM-enhanced with baseline model.

CBAM [21] integrates inter-channel relationship modeling. The feature wise weighting block focuses on identifying the most informative feature channels by modeling channel-wise interactions, On the other hand spatial attention model points out essential regions within the feature maps. By combining spatial and channel attention, CBAM suppresses background noise and extracts vital visual cues. This is specifically beneficial in image segmentation tasks where accurate localization and boundary detection are required, leading to improved segmentation performance in structured datasets.

SE enhanced with baseline model: U-Net is integrated with Squeeze-and-Excitation block to fine-tune channel-wise importance. The SE block [22] is designed to enhance channel-wise feature representation by openly modeling the inter-channel dependencies. Figure 3 shows the proposed SE-enhanced with baseline model.

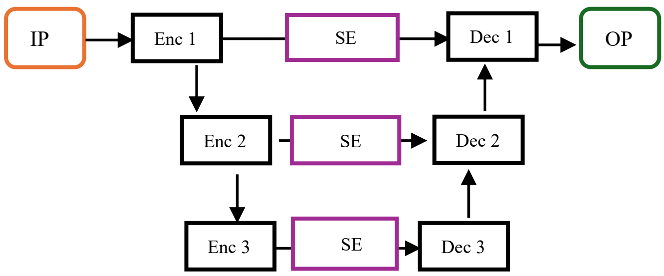

Figure 3: SE-enhanced with baseline model.

The SE attention [22] comprises a squeeze operation and an excitation step for attention weight generation. This approach allows the network to allocate greater significance to more meaninful features while reducing the influence of less relevant ones. Although SE improves feature discrimination, it does not incorporate spatial details, which may limit its effectiveness in tasks requiring precise localization of complex structures.

ECA enhanced with baseline model: U-Net is combined with Efficient Channel Attention to accomplish less weight channel-wise attention without reducing the dimensionality. Figure 4 illustrates the detailed architecture of ECA-based U-Net. ECA [23] is an effective channel attention method that enhances upon traditional channel attention methods by avoiding dimensionality reduction. Rather than using fully connected layers, ECA employs a rapid one dimensional operation to extract immediate inter-channel dependencies. This approach preserves feature information while maintaining low computational complexity. As a result, ECA provides an effective balance between performance and efficiency, enabling better generalization, especially in datasets with limited samples or high variability. Its lightweight design enables it to be well-suited for improving segmentation performance without significantly increasing model complexity.

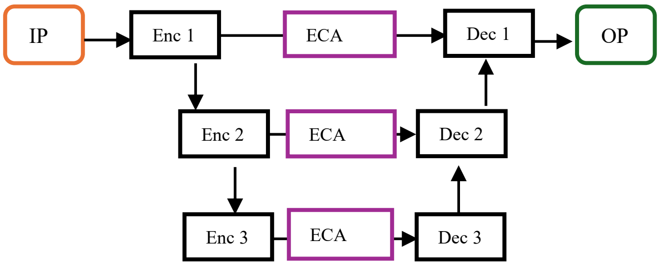

Figure 4: Architecture of ECA-enhanced with baseline model.

4. Experimental Studies

We evaluate the efficiency of the three U-Net variants developed against the baseline model for the pixel level parsing task in this section. In this study, U-Net and its variants were trained and validated on two different datasets, namely the ISIC 2017 Challenge dataset and the Carvana Image Masking Challenge dataset. To examine the efficiency of the presented models, we trained each architecture using different training configurations with varying numbers of epochs.

ISIC 2017 is a challenging skin lesion dataset which contains 2,000, 150 and 600 images of the training, validation and testing respectively. The Carvana Image Masking Challenge dataset is a vehicle segmentation dataset which contains 4,070, 1,018, and 100,064 images of train, validation and test respectively.

A hybrid loss is adopted which combines Dice Loss and Binary Cross-Entropy was used to effectively train the models:

Dice Loss:

$$L_{Dice} = 1 - Dice \tag{1}$$

Combined Loss:

$$\mathcal{L} = \mathcal{L}_{BCE} + \mathcal{L}_{Dice} \tag{2}$$

Segmentation performance was examined via below evaluation criteria.

Dice score:

$$Dice = \frac{2|P \cap G|}{|P| + |G|} \tag{3}$$

mIoU:

$$IoU = \frac{TP}{TP+FP+FN} \tag{4}$$

mIoU for multiple classes is given by

$$mIoU = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c+FP_c+FN_c} \tag{5}$$

Pixel Accuracy:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

where, TP, TN, FP, FN, defined as true positives, true negatives, false positives, false negatives respectively.

4.1. Evaluation Results for ISIC 2017

We present the comprehensive experimental results for ISIC 2017 in Table 1.

TABLE 1. output evaluation for ISIC 2017 dataset

Model	LR = 1e-2, Weight decay = 1e-5, epochs = 35, gamma = 0.3		
	Best Dice	*mIoU*	*Accuracy%*
U-Net	0.819	0.73	92.9
R2_U-Net	0.683	0.57	88.8
Attention U-Net	0.79	0.70	92.7
Attention R2_U-Net	0.655	0.54	88.4
U-Net + CBAM (Ours)	**0.823**	**0.74**	**93.4**
U-net + SE (Ours)	0.70	0.61	90.2
U-Net + ECA (Ours)	**0.827**	**0.75**	**93.7**

For the ISIC 2017 dataset, among the evaluated models, U-Net with ECA and CBAM performed comparatively well, outperforming all baseline models. The U-Net + ECA model achieves the highest Dice score was 0.827, with an mIoU of 0.75 and an accuracy rate of 93.7%, indicating the efficiency pf ECA for important channel extraction in tackling segmentation tasks with complex borders. U-NET + CBAM also obtains a competitive Dice score of 0.823, an mIoU of 0.74 and an accuracy rate of 93.4%, indicating the channel and spatial context learning in locating fuzzy and ambiguous boundaries of various segmentation tasks. Example segmentation results are shown in Table 2.

4.2. Evaluation Results for Carvana Dataset

We summarized the detailed evaluation findings of the Carvana dataset in Table 3. Each model performed very well with the Carvana dataset. U-Net + CBAM outperformed all other baseline and variant models with the highest Dice score was 0.995, with an mIoU of 0.99 and an accuracy rate of 99.8%, followed by U-Net + SE with the second-best dice score of

0.995, an mIoU of 0.99 and an accuracy rate of 99.7%. Example pixel-level parsing outcomes are given in Table 4.

TABLE 2. Output evaluation of proposed U-Net variants and baseline models for the ISIC dataset

	Image 1	Image 2	Image 3
Original Image			
GT			
U-Net + CBAM (Ours)			
U-net + SE (Ours)			
U-Net + ECA (Ours)			
R2_U-Net			
Attention U-Net			
Attention R2_U-Net			

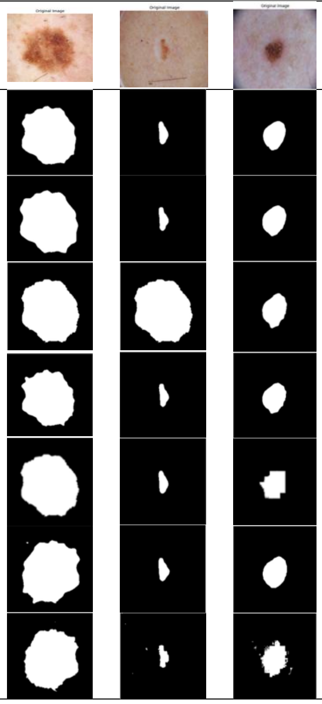

TABLE 3. Output evalulation for the Carvana dataset

Model	LR = 1e-3, Weight decay = 1e-4, epochs = 30, gamma = 0.3		
	Best Dice	*mIoU*	*Accuracy%*
U-Net	0.985	0.98	99.7
R2_U-Net	0.919	0.85	96.9
Attention U-Net	0.986	0.98	99.8
Attention R2_U-Net	0.937	0.88	97.2
U-Net + CBAM (Ours)	**0.995**	**0.99**	**99.8**
U-net + SE (Ours)	**0.995**	**0.99**	**99.7**
U-Net + ECA (Ours)	0.994	0.99	99.7

TABLE 4. Output evaluation of proposed U-Net variants and baseline models for the Carvana dataset

	Image 1	Image 2	Image 3
Original Image			
GT			
U-Net + CBAM (Ours)			
U-net + SE (Ours)			
U-Net + ECA (Ours)			
R2_U-Net			
Attention U-Net			
Attention R2_U-Net			

Previous studies have shown that baseline model combined with gating mechanisms generally performs better than the standard U-Net, particularly for complex assignments such as skin anomaly pixel wise classification. In particular, U-Net + ECA and U-Net + CBAM achieved higher Dice and mIoU scores on both ISIC 2017 and Carvana datasets. Table 5 shows comparison with other existing studies for both datasets.

TABLE 5. Comparison with existing studies

Dataset	Method	Dice
ISIC 2017	U-Net [1]	0.76
	Transformer based segmentation	0.82
	Our U-net + ECA	0.827
	Our U-Net + CBAM	0.823
Carvana Image Masking	Advanced U-Net variant [5]	0.994
	U-Net + CBAM	0.995
	U-Net + SE	0.995

5. Conclusion

Within this study, we present several attention reliant U-Net variants**,** including baseline model incorporated with CBAM, SE, and ECA, for complex pixel level parsing tasks in areas like autonomous driving and medical imaging. The models were evaluated on the ISIC 2017 and Carvana Image Masking datasets. Experimental results show that U-Net integrated with attention mechanisms and channel-wise feature enhancement achieved better performance in contrast to standard model**.** On the ISIC 2017 dataset**,** ECA incorporated with standard model obtained the strongest results with a Dice score of 0.827, while on the Carvana dataset**,** Baseline model with CBAM and SE outperformed other models with a Dice of 0.995**.** For future work, we will incorporate evolutionary algorithm-based architecture and hyper-parameter optimization [24-28] to further improve the adaptation capabilities of the proposed baseline model with gating mechanism for tackling diverse pixel-level parsing tasks (e.g. 3D clinical multi-organ segmentation) with various irregular boundaries and challenging factors. We will also test the proposed attention encoder architectures for other audio-visual classification tasks [29-36] to further test its efficiency.

References

[1] Ronneberger, O., Fischer, P. and Brox, T. (2015) 'U-Net: Convolutional Networks for Biomedical Image Segmentation', in Navab, N., Hornegger, J., Wells, W. and Frangi, A. (eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, vol. 9351, Springer, Cham, pp. 234–241.

[2] Ergun, H., 2021. Segmentation of Rays in Wood Microscopy Images Using the U-Net Model. *BioResources*, *16*(1).

[3] Alom, M.Z., Yakopcic, C., Taha, T.M. and Asari, V.K., 2018, July. Nuclei segmentation with recurrent residual convolutional neural networks based U-Net (R2U-Net). In *NAECON 2018-IEEE National Aerospace and Electronics Conference* (pp. 228-233). IEEE.

[4] Alom, M.Z., Yakopcic, C., Hasan, M., Taha, T.M. and Asari, V.K., 2019. Recurrent residual U-Net for medical image segmentation. *Journal of medical imaging*, *6*(1), pp.014006-014006.

[5] Gupta, A., Dixit, M., Mishra, V.K., Singh, A. and Dayal, A., 2022, December. Brain tumor segmentation from MRI images using deep learning techniques. In *International Advanced Computing Conference* (pp. 434-448). Cham: Springer Nature Switzerland.

[6] Chu, B., Zhao, J., Zheng, W. and Xu, Z., 2025. (DA-U) 2Net: double attention U2Net for retinal vessel segmentation. *BMC ophthalmology*, *25*(1), p.86.

[7] Zhang, M., Kong, Z., Zhu, W., Yi, Y., Zhou, Y. and Yan, F., 2022, December. Pelvic segmentation based MultiR2UNet. In *International Symposium on Artificial Intelligence and Robotics 2022* (Vol. 12508, pp. 118-124). SPIE.

[8] Shah, H.A. and Kang, J.M., 2023. An optimized multi-organ cancer cells segmentation for histopathological images based on CBAM-residual U-Net. *IEEE Access*, *11*, pp.111608-111621.

[9] Wang, J., Yu, Z., Luan, Z., Ren, J., Zhao, Y. and Yu, G., 2021, November. DA-ResUNet: a novel method for brain tumor segmentation based on U-Net with residual block and CBAM. In *2021 International Conference on Image, Video Processing, and Artificial Intelligence* (Vol. 12076, pp. 86-91). SPIE.

[10] Liu, L., Wu, K., Wang, K., Han, Z., Qiu, J., Zhan, Q., Wu, T., Xu, J. and Zeng, Z., 2024. SEU2-Net: multi-scale U2-Net with SE attention mechanism for liver occupying lesion CT image segmentation. *PeerJ Computer Science*, *10*, p.e1751.

[11] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W. and Hu, Q., 2020. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11534-11542).

[12] Wei, S., Hu, Z. and Tan, L., 2025. Res-ECA-UNet++: an automatic segmentation model for ovarian tumor ultrasound images based on residual networks and channel attention mechanism. *Frontiers in Medicine*, *12*, p.1589356.

[13] Lem, H. and Zhang, L., 2023. Mask R-CNN Transfer Learning Variants for Multi-Organ Medical Image Segmentation. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 1209-1216).

[14] Joy, C.P., Mistry, K., Pillai, G. and Zhang, L., 2022. A Hybrid-DE for Automatic Retinal Image-Based Blood Vessel Segmentation. In *Recent Advances in AI-enabled Automated Medical Diagnosis* (pp. 157-172). CRC Press.

[15] Zhang, L. and Lim, C.P., 2020. Intelligent optic disc segmentation using improved particle swarm optimization and evolving ensemble models. *Applied Soft Computing*, 92, p.106328.

[16] Tan, T.Y., Zhang, L., Lim, C.P., Fielding, B., Yu, Y. and Anderson, E., 2019. Evolving ensemble models for image segmentation using enhanced particle swarm optimization. *IEEE access*, *7*, pp.34004-34019.

[17] Slade, S., Zhang, L., Huang, H., Asadi, H., Lim, C.P., Yu, Y., Zhao, D., Lin, H. and Gao, R., 2023. Neural inference

search for multiloss segmentation models. *IEEE Transactions on Neural Networks and Learning Systems*.
[18] Zhang, L., Slade, S., Lim, C.P., Asadi, H., Nahavandi, S., Huang, H. and Ruan, H., 2023. Semantic segmentation using Firefly Algorithm-based evolving ensemble deep neural networks. *Knowledge-Based Systems*, *277*, p.110828.
[19] Chin Neoh, S., Srisukkham, W., Zhang, L., Todryk, S., Greystoke, B., Peng Lim, C., Alamgir Hossain, M. and Aslam, N., 2015. An intelligent decision support system for leukaemia diagnosis using microscopic blood images. *Scientific reports*, *5*(1), p.14938.
[20] Banugariya, P., Zhang, L., Yu, Y. and Sedov, V., 2025, June. Enhanced U-Net Models with Hybrid Attention Elements for Medical Image Segmentation. In *2025 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-9). IEEE.
[21] Woo, S., Park, J., Lee, J.Y. and Kweon, I.S., 2018. CBAM: Convolutional Block Attention Module. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.3-19.
[22] Hu, J., Shen, L. and Sun, G., 2018. Squeeze-and-Excitation Networks. Proceedings of *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.7132-714.
[23] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W. and Hu, Q., 2020. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11534-11542).
[24] Zhang, Y., Zhang, L. and Hossain, M.A., 2015. Adaptive 3D facial action intensity estimation and emotion recognition. *Expert systems with applications*, *42*(3), pp.1446-1464.
[25] Cunha, L., Zhang, L., Sowan, B., Lim, C.P. and Kong, Y., 2024. Video deepfake detection using Particle Swarm Optimization improved deep neural networks. *Neural Computing and Applications*, *36*(15), pp.8417-8453.
[26] Slade, S., Zhang, L., Asadi, H., Lim, C.P., Yu, Y., Zhao, D., Panesar, A., Wu, P.F. and Gao, R., 2025. Cluster search optimisation of deep neural networks for audio emotion classification. *Knowledge-Based Systems*, *314*, p.113223.
[27] Zhang, L., Zhao, D., Lim, C.P., Asadi, H., Huang, H., Yu, Y. and Gao, R., 2024. Video deepfake classification using particle swarm optimization-based evolving ensemble models. *Knowledge-Based Systems*, *289*, p.111461.
[28] Slade, S., Zhang, L., Yu, Y. and Lim, C.P., 2022. An evolving ensemble model of multi-stream convolutional neural networks for human action recognition in still images. *Neural computing and applications*, *34*(11), pp.9205-9231.
[29] Zhang, L., Lim, C.P. and Liu, C., 2023. Enhanced bare-bones particle swarm optimization based evolving deep neural networks. *Expert systems with applications*, 230, p.120642.
[30] Zhang, L., Lim, C.P., Yu, Y. and Jiang, M., 2022. Sound classification using evolving ensemble models and Particle Swarm Optimization. *Applied soft computing*, 116, p.108322.
[31] Tan, T.Y., Zhang, L. and Lim, C.P., 2020. Adaptive melanoma diagnosis using evolving clustering, ensemble and deep neural networks. *Knowledge-Based Systems*, *187*, p.104807.
[32] Kondrotas, K., Zhang, L., Lim, C.P., Asadi, H., and Yu. Y., 2025. Audio-Visual Emotion Classification Using Reinforcement Learning Enhanced Particle Swarm Optimisation. *IEEE Transactions on Affec tive Computing*.
[33] Zhang, L., Lim, C.P. and Yu, Y., 2021. Intelligent human action recognition using an ensemble model of evolving deep networks with swarm-based optimization. Knowledge-based systems, 220, p.106918.
[34] Choi, H., Zhang, L. and Watkins, C., 2025. Dual representations: A novel variant of Self-Supervised Audio Spectrogram Transformer with multi-layer feature fusion and pooling combinations for sound classification. *Neurocomputing*, 623, p.129415.
[35] Tan, T.Y., Zhang, L., Neoh, S.C. and Lim, C.P., 2018. Intelligent skin cancer detection using enhanced particle swarm optimization. Knowledge-based systems, 158, pp.118-135.
[36] Ye, Z., Zhao, D., Zhang, L., Yan, X., Bi, Q., Zhao, W. and Flynn, D., 2026. Enhancing Data Efficiency With a Trustworthy Counterfactual Generative Model. *IEEE Transactions on Industrial Informatics*.