Unsupervised Feature Orthogonalization for Learning Distortion-Invariant Representations

Sebastian Doerrich
sebastian.doerrich@uni-bamberg.de
Francesco Di Salvo
francesco.di-salvo@uni-bamberg.de
Christian Ledig
christian.ledig@uni-bamberg.de

xAILab Bamberg
University of Bamberg
Bamberg, Germany

Abstract

This study introduces unORANIC+, a novel method that integrates unsupervised feature orthogonalization with the ability of a Vision Transformer to capture both local and global relationships for improved robustness and generalizability. The streamlined architecture of unORANIC+ effectively separates anatomical and image-specific attributes, resulting in robust and unbiased latent representations that allow the model to demonstrate excellent performance across various medical image analysis tasks and diverse datasets. Extensive experimentation demonstrates unORANIC+'s reconstruction proficiency, corruption resilience, as well as capability to revise existing image distortions. Additionally, the model exhibits notable aptitude in downstream tasks such as disease classification and corruption detection. We confirm its adaptability to diverse datasets of varying image sources and sample sizes which positions the method as a promising algorithm for advanced medical image analysis, particularly in resource-constrained environments lacking large, tailored datasets. The source code is available at github.com/sdoerrich97/unoranic-plus.

1 Introduction

In recent years, significant progress has been made in deep learning via the introduction of novel training schemes and sophisticated network architectures [5, 11, 24]. However, achieving generalizability across diverse domains remains a challenge, limiting the impact of those advancements [9, 25]. This challenge is particularly pronounced in the medical domain, where data scarcity, inhomogeneities, and underrepresented demographics hinder model effectiveness [17, 18]. Moreover, domain shifts caused by variations in scanner models and imaging parameters further impede generalizability [13, 14, 22]. An example of this can be seen in Figure 1(a), which demonstrates distinct contrast and brightness variations across machines from different manufacturers (i–iii), different models of the same manufacturer (iv–vi), and the same model at different sites (vii–ix), despite displaying the same corresponding slice of T1-weighted MRI scans of the same healthy male subject from [19].

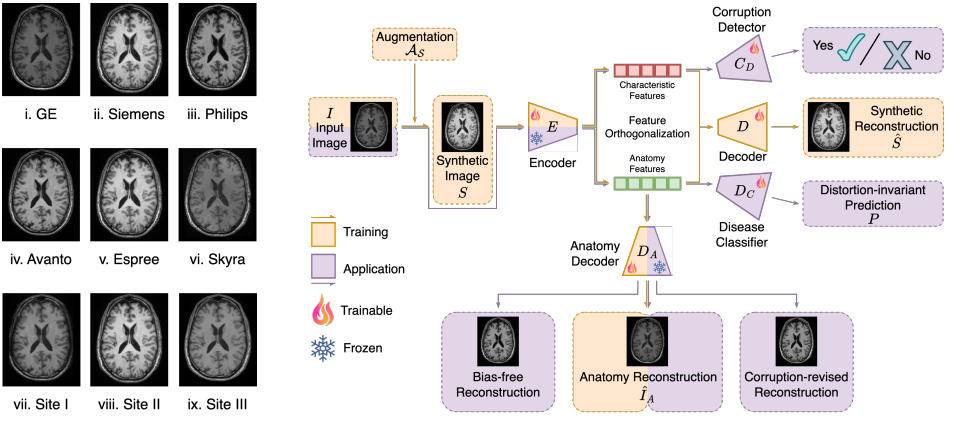

Figure 1: (a) Illustration of domain shifts in terms of different contrasts and brightness levels among manufacturers (i-iii), among models from the same producer (iv-vi), and among the same model at different sites (vii-ix) for the same slice of the same individual across multiple scans. (b) High-level overview of our proposed approach. During training, the encoder E is trained to orthogonalize anatomical and image-characteristic features in an input image (orange path). Once trained, the learned feature orthogonalization by the frozen encoder is used for various downstream tasks, including bias removal, corruption detection and revision in input images, as well as robust, distortion-invariant disease classification (purple path).

To tackle these issues, the work of unORANIC [7] has shown that unsupervised orthogonalization of anatomy and image-characteristic features can substantially improve robustness and generalizability without the need for domain knowledge, paired data, or labels. Building on this, we introduce unORANIC+, a simpler, more robust, and overall higher performant improvement. Figure 1(b) provides a high-level overview of our proposed approach. A single encoder is trained with a reconstruction objective to orthogonalize anatomical and image-characteristic information within an input image without requiring labels, image pairs, or additional information about the image domain (orange path). Once trained, the encoder is frozen, allowing the learned feature orthogonalization to be utilized for various downstream tasks, such as unbiased anatomical image reconstruction, detection and revision of corruptions in input images, as well as robust disease classification of biased or distorted images (purple path). Consequently, this disentanglement of image information enables unORANIC+ to generate robust latent representations even in the presence of data inhomogeneities and domain shifts. In summary, our main contributions include:

1. **Enhanced feature orthogonalization:** unORANIC+ synergizes unsupervised feature orthogonalization with a Vision Transformer's ability to capture global-local relationships for improved robustness and generalizability.

2. **Streamlined architecture:** with a single encoder, unORANIC+ effectively disentangles anatomical and image attributes, yielding robust latent representations to allow superior performance in a wide range of tasks.

3. **Versatility across datasets:** extensive quantitative experimentation across various datasets and medical conditions displays unORANIC+'s performance and versatility.

2 Related work

2.1 Orthogonalization of anatomy and image characteristics

Orthogonalization involves separating anatomical information from image-specific attributes in an image [7]. Anatomy features encompass the underlying anatomical structures like organs, tissues, or disease characteristics, while image-specific features include attributes such as contrast and brightness. This disentanglement allows for a focused assessment of anatomical elements, independent of image-specific biases or corruptions [4, 6, 29]. The concept of an entirely unsupervised separation of these feature classes was first introduced with unORANIC [7]. unORANIC+ builds upon this foundation, employing a Vision Transformer autoencoder to further advance that process and extend its applicability.

2.2 Vision Transformer autoencoder

Vision Transformers (ViTs) and autoencoding represent two key methodologies in the realm of representation learning and image processing [8, 28]. ViTs excel at capturing intricate local and global relationships [16], offering a promising avenue for tasks like image classification, segmentation, and object detection [8, 20]. Meanwhile, autoencoding is a classical technique known for its prowess in representation learning. It involves an encoder mapping inputs to a latent representation and a decoder reconstructing the original input. This process is typically conducted in a self- or unsupervised manner and holds significance in various applications including image denoising, and pre-training tasks [10, 21, 28].

Integrating both methods, particularly by employing ViTs for the encoder and decoder of the autoencoder, harnesses the strength of ViTs in discerning complex relationships to establish a representative latent space. Furthermore, this fusion enables the training of these traditionally data-hungry ViTs in a self- or even unsupervised fashion. This approach supports learning high-capacity models with strong generalization [10] and enables model training for datasets with limited labels. In contrast to the asymmetric design in [10], unORANIC+ adopts a symmetric configuration to enable high-resolution reconstructions. Moreover, a second decoder and an adapted loss function are introduced to allow the reconstruction of the original input and an anatomical, bias-free version of it.

3 Method

3.1 Baseline

The objective is to disentangle anatomy from image-specific features in an input image to learn robust, unbiased features in the latent space. For this, we harness the concept of unORANIC which originally employs a two-branch autoencoder [7]. During inference, the anatomical branch extracts core anatomical features to reconstruct a bias-free version of the input image, while the characteristic branch captures distinctive image details omitted by the first branch to allow the reconstruction of the original input image. For this, the anatomy encoder E_A is shared across a set of distorted variants (S, V_1, V_2) of the same input image I during training. E_A is updated via a combination of the L2-consistency loss $\mathcal{L}_{\mathrm{C}}$, forcing the feature embeddings of S, V_1, and V_2 to be the same, as well as the L2-reconstruction losses $\mathcal{L}_{\mathrm{R}_S}$ and $\mathcal{L}_{\mathrm{R}_I}$ to learn bias-robust anatomical representations, while the characteristic encoder E_C retains image-specific details. This process is displayed in Figure 2.

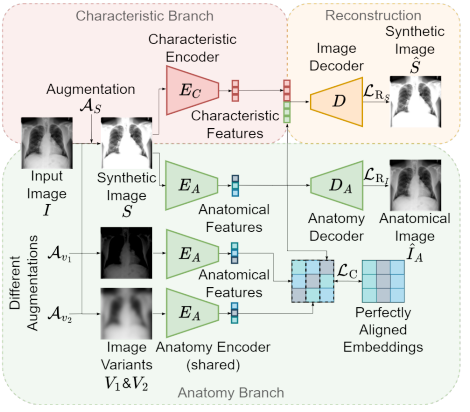

Figure 2: Schematic representation of the training pipeline for unORANIC (adapted from [7]). The input image I is assumed to be bias-free and uncorrupted. Random augmentations $\mathcal{A}_S$, $\mathcal{A}_{v_1}$, and $\mathcal{A}_{v_2}$ distort I to generate synthetic corrupted versions S, V_1, and V_2 with identical anatomical information but different distortions. These distorted images are processed by the shared anatomy encoder E_A, which uses the consistency loss $\mathcal{L}_C$ to learn anatomical, distortion-invariant features. Concurrently, S is processed by the characteristic encoder E_C to capture image-specific details such as contrast and brightness. Reconstruction losses $\mathcal{L}_{R_S}$ and $\mathcal{L}_{R_I}$ are applied to the reconstructed images $\hat{S}$ and $\hat{I}_A$ by decoder D and D_A, respectively, to ensure that E_A and E_C learn comprehensive, reliable features.

3.2 unORANIC+

In contrast to the former architecturally enforced orthogonalization, unORANIC+ employs only one encoder E which maps the input to a single, higher-dimensional latent space. The network comprises two consecutive decoders for which the first decoder D reconstructs the original input image, while the other D_A focuses on generating the bias-free anatomical reconstruction, akin to unORANIC. Inspired by the Masked Autoencoder (MAE) design from [11], we employ Vision Transformers (ViTs) for the encoder and both decoders. Specifically, we opt for a symmetric design in which the encoder and decoders are equal in size and depth. In line with standard ViT practices [8], we divide the image into regular non-overlapping patches. The encoder uses a linear projection with positional embeddings to embed the patches before processing them through a sequence of Transformer blocks. The two decoders will process the encoded patches through their own Transformer blocks to reconstruct the image.

3.3 Training and application

Following Figure 3, during training, we augment each input image I with a random set of distortions $\mathcal{A}_S$, taken from [3], to create a synthetic corrupted version S. S is afterward split into regular non-overlapping patches, which are flattened into 1D vectors before passing them through the encoder E. Positional embeddings are added to aid the learning of spatial dependencies. The resulting embeddings are passed through the encoder's Transformer blocks to create the latent representation, encompassing both anatomy and image-characteristic features. This enables the synthetic decoder D to reconstruct $\hat{S}$ with all added characteristics by

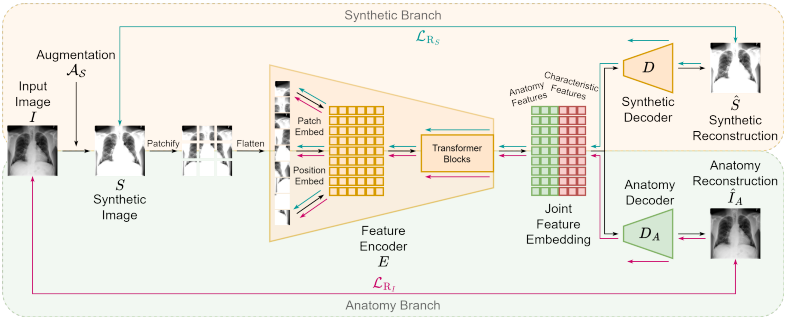

Figure 3: Schematic representation of the training pipeline for the refined unORANIC+ method on the example of chest X-ray images. The polar arrows illustrate the forward propagation and gradient flow, respectively. During training, an input image I is augmented with a random set of distortions, $\mathcal{A}_S$, to generate the synthetic, distorted image S. S is subsequently divided into non-overlapping patches before it is fed through the single Vision Transformer (ViT) encoder E to map the input image to a higher-dimensional latent space. Two ViT decoders, D and D_A, are used to reconstruct the original synthetic image $\hat{S}$ as well as a bias-free anatomical reconstruction $\hat{I}_A$, respectively. The two reconstruction losses $\mathcal{L}_{R_S}$ and $\mathcal{L}_{R_I}$ guide the separation of anatomical and image-characteristic features in the latent space and ensure a high quality of the reconstructions.

$\mathcal{A}_S$, while the anatomy decoder D_A can disregard these characteristics and reconstruct a bias-free anatomical image $\hat{I}_A$. To do so, the training exploits the two reconstruction losses $\mathcal{L}_{R_S}$ and $\mathcal{L}_{R_I}$ equally, which measure the mean squared error (MSE) in the pixel space between $\hat{S}$ and S as well as $\hat{I}_A$ and I, respectively. The latter enforces decoder D_A to focus solely on anatomical features, further promoting the feature orthogonalization within the latent space.

During inference, the model can be applied to any potentially (un-)biased or (un-)corrupted test image I to orthogonalize its anatomical and image-specific features, enabling tasks like bias-free reconstruction, corruption detection and revision, or robust downstream applications such as disease classification.

4 Experiments and results

We comprehensively evaluate unORANIC+ in terms of reconstruction quality, capability to revise existing corruptions, corruption robustness, and its effectiveness in downstream tasks such as disease classification and corruption detection. To allow a fair comparison with unORANIC, we utilize the same diverse selection of 28×28 biomedical 2D datasets from the MedMNIST v2 benchmark [27] the original method was evaluated on, including breastMNIST (546 training samples), retinaMNIST (1,080), pneumoniaMNIST (4,078), dermaMNIST (7,007), and bloodMNIST (11,959). Additionally, we assess all models on the larger chestMNIST dataset (78,468 training samples) as well. Examples for each dataset can be seen in Figure 4. Finally, in addressing a major limitation of unORANIC, which was exclusively evaluated on 28×28 images, we investigate unORANIC+'s potential to handle higher dimensional data as well. This is achieved by adopting higher resolution versions of the MedMNIST datasets, comprising images of 224×224 pixels, by using the original data samples [1, 2, 12, 15, 23, 26] in combination with the MedMNIST data splits.

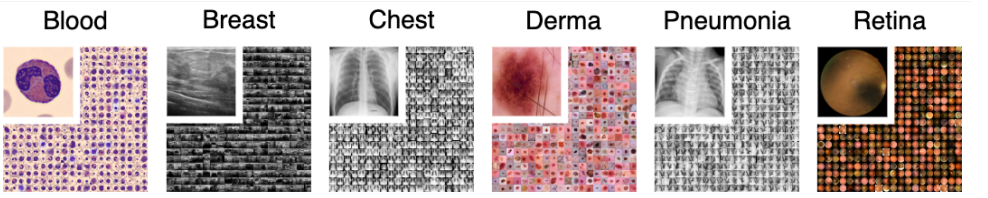

Figure 4: Examples from the datasets of the MedMNIST v2 benchmark [27] used for evaluating our approach (left to right: blood, breast, chest, derma, pneumonia, and retina)

4.1 Reconstruction and corruption revision

We first evaluate the reconstruction abilities of unORANIC+ and the original unORANIC model. For this, we train unORANIC+ including its encoder and both decoders, each composed of 12 layers and 16 attention heads, for 150 epochs with a batch size of 64, using the Adam optimizer with weight decay and learning rate warmup. The patch size is set to 4×4 due to the input size of 28×28, with a latent dimension of 128 per patch. The results displayed in Table 1 show the average Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Metric (SSIM) values on each dataset's test set for both models' anatomical reconstructions ($\hat{I}_A$) and reconstructions of the original input ($\hat{I}$) given an uncorrupted input image I. These findings clearly indicate that unORANIC+ substantially outperforms the original model for both reconstruction objectives.

Additionally, we assess both models based on their capability to revise corruptions in an input image via their anatomy branches. For this, we deliberately apply the same range of corruptions as used during training to all test images to generate synthetic distorted versions, similar to the experiment presented in [7]. Despite the distortions, both models successfully reconstruct the uncorrupted input images, as depicted in Figure 5 for the dermaMNIST dataset. Moreover, the figure illustrates unORANIC+'s enhanced corruption revision capability across all distortions in 5(a) and emphasizes its ability to preserve fine-grained details in 5(b) compared to its predecessor for distortions such as the displayed Gaussian noise.

Dataset	PSNR				SSIM			
	unORANIC		unORANIC+		unORANIC		unORANIC+	
	$\hat{I}_A$	$\hat{I}$	$\hat{I}_A$	$\hat{I}$	$\hat{I}_A$	$\hat{I}$	$\hat{I}_A$	$\hat{I}$
Blood	27.06	31.70	**35.88**	**49.15**	0.877	0.943	**0.987**	**0.999**
Breast	19.48	29.39	**26.21**	**33.55**	0.526	0.816	**0.889**	**0.957**
Chest	27.93	33.73	**35.30**	**56.02**	0.956	0.983	**0.995**	**0.999**
Derma	23.73	38.57	**30.07**	**45.09**	0.864	0.970	**0.971**	**0.995**
Pneumonia	24.00	36.04	**28.96**	**44.80**	0.901	0.977	**0.977**	**0.997**
Retina	27.50	36.31	**30.39**	**37.71**	0.888	0.954	**0.936**	**0.978**

Table 1: Comparison of average Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Metric (SSIM) values for the anatomical reconstructions ($\hat{I}_A$) and the reconstructions of the original input ($\hat{I}$) given an uncorrupted input image (I) between unORANIC and unORANIC+. The best performance per reconstruction task is indicated in **bold**.

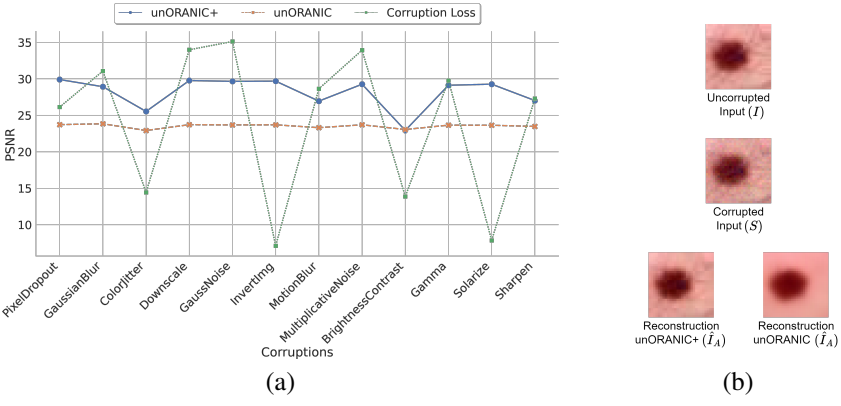

Figure 5: Corruption revision capabilities of unORANIC and unORANIC+. In (a), their reconstruction consistency is depicted despite the corruption-related image quality loss (PSNR between the original image I and the distorted variant S - "green dotted line"). (b) highlights the distortion correction capabilities of both methods using Gaussian noise as an example.

4.2 Disease classification and corruption detection

We evaluate the encoded feature embeddings of unORANIC+ to assess the degree of orthogonalization between anatomy and image-characteristic features for the tasks of disease classification and corruption detection. Depending on the used dataset, the disease classification comprises a binary classification (breast, chest, pneumonia) or a multi-class classification (blood, derma, retina) task. To this end, we freeze the encoder, replace the decoders with an architecturally alike ViT classifier for the particular task and dataset, and train only the classifier. The model's performance, measured in terms of Accuracy (ACC) and Area Under the ROC Curve (AUC) per dataset, is compared with the unORANIC model and supervised, end-to-end trained ResNet-18 and ViT baselines for each respective task. It is important to note that while these baselines provide a useful context for performance assessment, their results should not be compared directly to those of unORANIC and unORANIC+. Both baselines are trained fully supervised, end-to-end, while unORANIC and unORANIC+ both employ frozen encoder(s) and only train the classifier head for the specific task. The results in Table 2 for bloodMNIST demonstrate unORANIC+'s notable improvement over unORANIC, along with its comparable performance to the end-to-end trained, supervised models.

Methods	Disease Classification		Corruption Detection	
	ACC	AUC	ACC	AUC
ResNet-18†	0.958	0.998	0.973	0.998
ViT†	0.930	0.992	0.914	0.926
unORANIC	0.800	0.952	0.962	0.976
unORANIC+	**0.935**	**0.994**	**0.970**	**0.980**

Table 2: Comparison of the classification and corruption detection results on the bloodMNIST dataset. Fully supervised models, trained end-to-end, are indicated with †. A superior performance of unORANIC+ compared to unORANIC is indicated in **bold**.

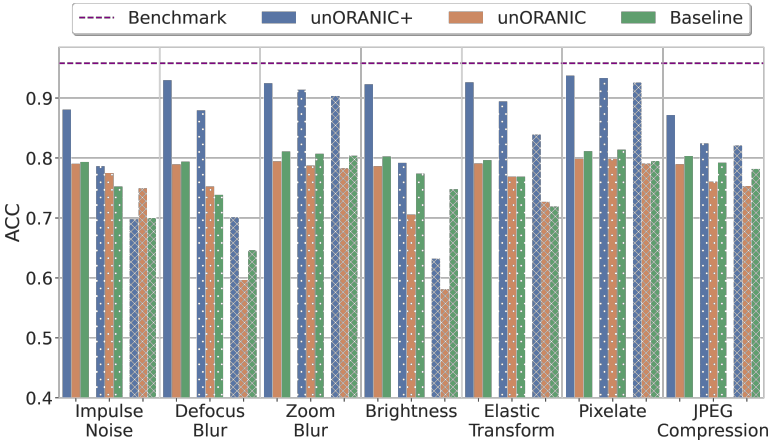

Figure 6: Visualization of unORANIC+'s resilience to unseen corruptions compared to the reference models (unORANIC and the supervised ResNet-18 "Baseline"), demonstrated by its robust disease classification performance on the bloodMNIST dataset as corruption severity increases. Different textures (, $\cdot$, $\times$) indicate distinct levels of severity for each presented corruption. For reference, the plot also includes the supervised classification "Benchmark" (i.e., ResNet-18) for uncorrupted images.

Furthermore, we test unORANIC+'s robustness to unseen corruptions that were not used during training, by applying varying severities of these corruptions to all test images before passing them through the trained models for the original disease classification task. We observe superior resilience of unORANIC+ compared to unORANIC and ResNet-18 ("Baseline") across all datasets, as illustrated in Figure 6 for bloodMNIST.

4.3 Evaluation on higher dimensional datasets

For the evaluation of unORANIC+ on the higher dimensional versions of the MedMNIST datasets, we adjusted the ViT configuration (using 16×16 patches instead of 4×4 and a latent dimension of 768 instead of 128). We retrained the model for each dataset individually and conducted the same experiments as previously described. Table 3 presents the evaluation for the reconstruction of the original input $\hat{I}$ and the anatomical, bias-free version $\hat{I}_A$ in terms of PSNR. Additionally, Table 4 presents the evaluation for disease classification and corruption detection (both in terms of AUC) for each individual dataset. To facilitate side-by-side comparisons across all six datasets, we omit the SSIM results for reconstruction and the ACC results for classification and detection, respectively. We remind that the ResNet-18† and ViT† reference models are trained end-to-end in a supervised manner solely for the classification and detection tasks. Therefore, we do not evaluate these models for reconstruction and only compare unORANIC+ with its predecessor in this case.

unORANIC+ maintains its accurate performance, showcasing robust anatomical reconstruction capabilities across all datasets and reliable corruption revision (consistently high PSNR between $\hat{I}_A$ and $\hat{I}$). Moreover, unORANIC+ excels beyond its already robust disease classification potential for the undistorted blood and chest datasets, even exceeding the end-to-end trained baselines. This performance boost can be attributed to the larger image sizes, enabling more effective learning of both global and local representations.

	Input Reconstruction ($\hat{I}$) [PSNR]						Anatomical Reconstruction ($\hat{I}_A$) [PSNR]					
	Blood	Breast	Chest	Derma	Pneumonia	Retina	Blood	Breast	Chest	Derma	Pneumonia	Retina
unORANIC	26.66	20.28	28.33	27.55	27.21	31.25	24.57	17.62	26.02	23.10	21.61	25.64
unORANIC+	**44.63**	**27.80**	**42.23**	**33.22**	**34.87**	**35.37**	**38.23**	**24.94**	**32.35**	**26.91**	**26.86**	**31.41**

Table 3: Comparison of average Peak Signal-to-Noise Ratio (PSNR) for the reconstructions of the original input ($\hat{I}$) and the clean anatomical reconstructions ($\hat{I}_A$) given an input image (I) between unORANIC and unORANIC+ on the test sets of all six higher dimensional datasets. The best performance per reconstruction task is indicated in **bold**.

	Disease Classification [AUC]						Corruption Detection [AUC]					
	Blood	Breast	Chest	Derma	Pneumonia	Retina	Blood	Breast	Chest	Derma	Pneumonia	Retina
ResNet-18†	0.840	0.794	0.518	0.608	0.977	0.684	0.941	0.833	0.876	0.842	0.867	0.869
ViT†	0.891	0.580	0.535	0.501	0.840	0.647	0.791	0.663	0.820	0.660	0.639	0.700
unORANIC	0.930	0.774	0.528	0.717	0.953	0.681	0.724	0.586	0.635	0.625	0.639	0.644
unORANIC+	**0.997**	0.757	**0.563**	0.563	**0.955**	0.675	**0.954**	**0.616**	**0.892**	**0.690**	**0.694**	**0.783**

Table 4: Comparison of the disease classification and corruption detection results across the higher dimensional datasets. Fully supervised models, trained end-to-end, are indicated with †. **Bold** highlights superior performance of unORANIC+ compared to unORANIC, while underlining indicates cases where unORANIC+ even surpasses the best supervised baseline.

While unORANIC+ preserves a high corruption detection performance across all datasets, the disease classification results reveal limitations of ViT-based architectures for comparatively small but higher dimensional datasets, such as breast, derma, and retina. Despite unORANIC+ surpassing the end-to-end trained ViT classifier, both lag behind the convolutional methods for the clean, undistorted versions of these datasets. Thus, it appears that the robust representations learned by unORANIC+ are particularly beneficial for medium-sized and potentially biased or distorted datasets of small or higher image dimension.

5 Discussion and Conclusion

Through the integration of Vision Transformers for unsupervised feature orthogonalization, unORANIC+ effectively disentangles anatomical and image-characteristic features, yielding robust latent representations. Our experiments across distinct datasets from various modalities, all showing different medical conditions, demonstrate superior reconstruction and corruption revision capabilities compared to the original unORANIC model, displaying its stable resilience against distortions. Additionally, we demonstrate its prowess in disease classification and corruption detection tasks as well as its adaptability to higher dimensional datasets, underscoring its potential for robust medical image analysis.

Compliance with ethical standards

This research study was conducted retrospectively using human subject data made available in open access by [19] and [27]. Ethical approval was not required as confirmed by the license attached with the open-access data.

Acknowledgments

This study was funded through the Hightech Agenda Bayern (HTA) of the Free State of Bavaria, Germany.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, 30:105474, 2020. ISSN 2352-3409.

[2] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. ISSN 2352-3409.

[3] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information 2020, Vol. 11, Page 125*, 11:125, 2020. ISSN 2078-2489.

[4] Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David E Newby, Rohan Dharmakumar, and Sotirios A Tsaftaris. Disentangled representation learning in cardiac image analysis. *Medical Image Analysis*, 58, 2019. ISSN 1361-8415.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186. Association for Computational Linguistics, 2019.

[6] Blake E. Dewey, Lianrui Zuo, Aaron Carass, Yufan He, Yihao Liu, Ellen M. Mowry, Scott D. Newsome, Jiwon Oh, Peter A. Calabresi, and Jerry L Prince. A disentangled latent space for cross-site mri harmonization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020.

[7] Sebastian Doerrich, Francesco Di Salvo, and Christian Ledig. unoranic: Unsupervised orthogonalization of anatomy and image-characteristic features. In *Machine Learning in Medical Imaging*, pages 62–71. Springer Nature Switzerland, 2023. ISBN 978-3-031-45673-2.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[9] Thomas Eche, Lawrence H. Schwartz, Fatima Zohra Mokrane, and Laurent Dercle. Toward generalizability in the deployment of artificial intelligence in radiology: Role

of computation stress testing to overcome underspecification. *Radiology: Artificial Intelligence*, 3, 2021. ISSN 26386100.

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June:15979–15988, 2022. ISSN 10636919.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[12] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalena Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A.N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172:1122–1131.e9, 2018. ISSN 0092-8674.

[13] Amjad Khan, Andrew Janowczyk, Felix Müller, Annika Blank, Huu Giao Nguyen, Christian Abbet, Linda Studer, Alessandro Lugli, Heather Dawson, Jean Philippe Thiran, and Inti Zlobec. Impact of scanner variability on lymph node segmentation in computational pathology. *Journal of Pathology Informatics*, 13:100127, 2022. ISSN 2153-3539.

[14] Maxime W. Lafarge, Josien P.W. Pluim, Koen A.J. Eppenhof, Pim Moeskops, and Mitko Veta. Domain-adversarial neural networks to address the appearance variability of histopathology images. *Lecture Notes in Computer Science*, 10553 LNCS:83–91, 2017. ISSN 16113349.

[15] Ruhan Liu, Xiangning Wang, Qiang Wu, Ling Dai, Xi Fang, Tao Yan, Jaemin Son, Shiqi Tang, Jiang Li, Zijian Gao, Adrian Galdran, J. M. Poorneshwaran, Hao Liu, Jie Wang, Yerui Chen, Prasanna Porwal, Gavin Siew Wei Tan, Xiaokang Yang, Chao Dai, Haitao Song, Mingang Chen, Huating Li, Weiping Jia, Dinggang Shen, Bin Sheng, and Ping Zhang. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, 3:100512, 2022. ISSN 2666-3899.

[16] José Maurício, Inês Domingues, and Jorge Bernardino. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, 13(9), 2023. ISSN 2076-3417.

[17] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature 2023 616:7956*, 616:259–265, 2023. ISSN 1476-4687.

[18] Natalia Norori, Qiyang Hu, Florence Marcelle Aellen, Francesca Dalia Faraci, and Athina Tzovara. Addressing bias in big data and ai for health care: A call for open science. *Patterns*, 2:100347, 2021. ISSN 26663899.

[19] Roland Opfer, Julia Krüger, Lothar Spies, Ann Christin Ostwaldt, Hagen H. Kitzler, Sven Schippling, and Ralph Buchert. Automatic segmentation of the thalamus using a massively trained 3d convolutional neural network: higher sensitivity for the detection of reduced thalamus volume by improved inter-scanner stability. *European radiology*, 2022. ISSN 1432-1084.

[20] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[21] Francisco Silva, Tania Pereira, Julieta Frade, José Mendes, Claudia Freitas, Venceslau Hespanhol, José Luis Costa, António Cunha, and Hélder P. Oliveira. Pre-training autoencoder for lung nodule malignancy assessment using ct images. *Applied Sciences*, 10(21), 2020. ISSN 2076-3417.

[22] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundstrom. Measuring domain shift for deep learning in histopathology. *IEEE Journal of Biomedical and Health Informatics*, 25:325–336, 2021. ISSN 21682208.

[23] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data 2018 5:1*, 5:1–9, 2018. ISSN 2052-4463.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. volume 2017-December, pages 5999–6009. Neural information processing systems foundation, 2017.

[25] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35 (8):8052–8072, 2023.

[26] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017.

[27] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data 2023 10:1*, 10:1–10, 2023. ISSN 2052-4463.

[28] Arezoo Zakeri, Yan Xia, Nishant Ravikumar, and Alejandro F. Frangi. Chapter 17 - deep learning for vision and representation learning. In Alejandro F. Frangi, Jerry L. Prince, and Milan Sonka, editors, *Medical Image Analysis*, The MICCAI Society book Series, pages 451–474. Academic Press, 2024. ISBN 978-0-12-813657-7.

[29] Lianrui Zuo, Blake E. Dewey, Yihao Liu, Yufan He, Scott D. Newsome, Ellen M. Mowry, Susan M. Resnick, Jerry L. Prince, and Aaron Carass. Unsupervised mr harmonization by learning disentangled representations using information bottleneck theory. *NeuroImage*, 243:118569, 2021. ISSN 1053-8119.