Focus on Texture: Rethinking Pre-training in Masked Autoencoders for Medical Image Classification

Chetan Madan[1*], Aarjav Satia[1*], Soumen Basu[1†][0000−0002−3915−7545], Pankaj Gupta[2], Usha Dutta[2], and Chetan Arora[1]

[1] Indian Institute of Technology, Delhi
[2] PGIMER Chandigarh

Abstract. Masked Autoencoders (`MAEs`) have emerged as a dominant strategy for self-supervised representation learning in natural images, where models are pre-trained to reconstruct masked patches with a pixel-wise mean squared error (`MSE`) between original and reconstructed `RGB` values as the loss. We observe that `MSE` encourages blurred image reconstruction, but still works for natural images as it preserves dominant edges. However, in medical imaging, when the texture cues are more important for classification of a visual abnormality, the strategy fails. Taking inspiration from `Gray Level Co-occurrence Matrix` (`GLCM`) feature in Radiomics studies, we propose a novel `MAE` based pre-training framework, `GLCM-MAE`, using reconstruction loss based on matching `GLCM`. `GLCM` captures intensity and spatial relationships in an image, hence proposed loss helps preserve morphological features. Further, we propose a novel formulation to convert matching `GLCM` matrices into a differentiable loss function. We demonstrate that unsupervised pre-training on medical images with the proposed `GLCM` loss improves representations for downstream tasks. `GLCM-MAE` outperforms the current state-of-the-art across four tasks – gallbladder cancer detection from ultrasound images by 2.1%, breast cancer detection from ultrasound by 3.1%, pneumonia detection from x-rays by 0.5%, and COVID detection from CT by 0.6%. Source code and pre-trained models are available at: https://github.com/ChetanMadan/GLCM-MAE.

Keywords: Medical Image classification · Masked Autoencoders · Image reconstruction

1 Introduction

Masked Autoencoders (`MAEs`) have emerged as a powerful approach for self-supervised representation learning in natural images [14,34]. In `MAE` based pre-training, one masks parts of input image, and then trains a neural network model to reconstruct the masked regions. Importantly, the whole training process

*Joint first authors
†Soumen contributed while he was at IIT Delhi

does not require any supervised labels, which allows one to use large amount of training images available on the internet, thus helping a model to learn robust image representation from highly varied data. Several researchers have shown that MAE based pre-training followed by fine-tuning on downstream task helps achieve state-of-the-art performance on variety of tasks [4,25].

Whereas, several variations [4,35,37] of MAE pre-training have been proposed, all these use pixel-wise mean squared error (MSE) between the original and reconstructed patch to compute the loss, and then back-propagate the loss to train the model. We observe that such pre-training is not effective for medical images. Our analysis reveals that the problem lies in the pixel-wise MSE based reconstruction loss. The particular loss encourages model to generate smooth images which works for contour or dominant edge based classification in natural images.

In medicine, radiomics is a technique which extracts an array of texture features from medical images. These features, known as radiomic features, are then used to train simple machine learning algorithms like SVMs and decision trees. GLCM is one such radiomic feature which analyzes how often gray levels (intensities) occur together within an image, considering only the neighboring pixels. By counting co-occurrences in a specific pixel offset (often 1 or 2), GLCM effectively captures the spatial arrangement of textures. In this work we propose a novel pre-training framework for medical images, which uses MAE with GLCM guided reconstruction loss. This enables a model to focus on subtle gray-level variations within each patch, and better preserve morphological features.

Contributions. (1) We investigate the reasons behind relative ineffectiveness of the MAE based pre-training on medical imaging tasks. We observe that pixel-wise MSE based reconstruction over-smooth subtle, but semantically important, texture features in the image representation, and degrade performance on medical image classification. **(2)** We propose a novel MAE technique tailored for medical imaging, incorporating a GLCM-guided loss. GLCM is non-differentiable. However, we do a clever work-around based on a differentiable joint histogram of an image, shifting the image by one pixel to generate a GLCM-like representation. This design ensures that the learned representations preserve critical morphological information, which is essential for downstream tasks such as classification and segmentation in medical imaging. **(3)** We demonstrate the efficacy of GLCM-MAE with the proposed GLCM loss, outperforming existing SOTA on four medical image analysis tasks: gallbladder cancer detection from ultrasound by 2.1%, breast cancer detection from ultrasound by 3.1%, pneumonia detection from x-rays by 0.5%, and COVID detection from CT by 0.6%. **(4)** We are releasing source-code and pre-trained models for the community.

2 Proposed Method

MSE Loss. Let $\mathbf{x}$ denote the original image, and $\mathbf{x}_m$ the masked patches. MSE loss, $\mathcal{L}$, for RGB-based reconstruction is computed as: $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_m^{(i)} - \hat{\mathbf{x}}_m^{(i)}\|_2^2$. Here, N is the number of masked patches, $\mathbf{x}_m^{(i)}$ is the ground truth RGB value of the i-th masked patch, and $\hat{\mathbf{x}}_m^{(i)}$ is the reconstructed RGB value.

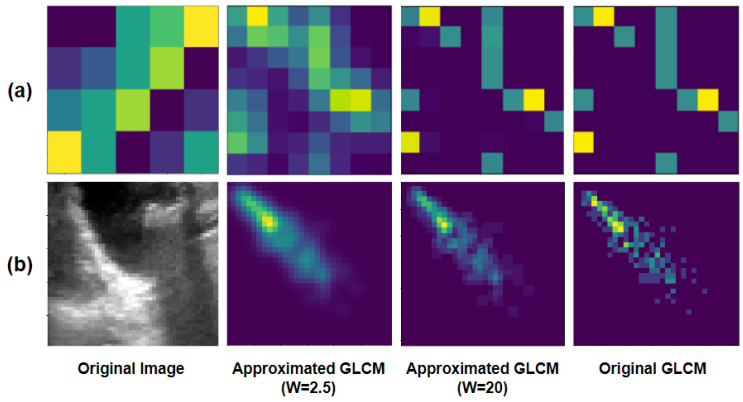

Fig. 1. Differentiable approximate GLCM construction converges to original GLCM with sufficiently large kernel bandwidth (W). Row (a) contains a sample image, followed by the `GLCM` approximations for different values of W, and the original `GLCM`. Row (b) shows an image patch from the GBCU dataset along with the `GLCM` approximations and original `GLCM`.

Revisiting Gray-Level Co-Occurrence Matrix (`GLCM`). Let i and j represent pixel intensity values having G gray levels (e.g. for 8-bit grayscale image, $G = 256$), d is the distance between the pixel pairs, and θ is the angle (e.g., 0°, 45°, 90°, or 135°). `GLCM` is a $G \times G$ matrix, denoted by $\mathcal{G}(v, w \mid d, \theta)$, where each element (v, w) represents the number of times a pixel with intensity v is paired with a pixel of intensity w at the given distance and direction. Given a grayscale patch i with pixel intensities $I_i(x, y)$ where x, y represent pixel locations, the `GLCM` matrix records the co-occurrence frequency of intensity values across pixel pairs within a defined spatial offset d. We denote the co-occurrence count for a pair of intensity values (v, w) as:

$$\mathcal{G}^{(i)}(v, w) = \sum_{(x,y),(x',y')} \delta(I_i(x, y) = v) \cdot \delta(I_i(x', y') = w), \tag{1}$$

where δ is the Kronecker delta function that is 1 if its argument is true, and 0 otherwise. Additionally, (x', y') represent the neighbor pixel of (x, y).

Differentiable Approximation of `GLCM`. Note that, `GLCM` is non-differentiable. Hence, we compute a differentiable approximation of `GLCM` matrix for each masked patch independently. We use Kernel Density Estimation (`KDE`) technique proposed by [3] to compute differentiable histograms. Let $I_i \in [0, 255]$ denote a gray level value of an image pixel i. `KDE` for estimating the intensity density f_I of an image I with N pixels is given by: $\hat{f}_I(x) = \frac{1}{NW} \sum_{i=1}^{N} \mathcal{K}\left(\frac{I_i - x}{W}\right)$, where $x \in [0, 255]$, $\mathcal{K}(\cdot)$ denotes the kernel, and W is the bandwidth. We select the kernel as the derivative of sigmoid function: $\mathcal{K}(z) = \sigma'(z) = \sigma(z)\sigma(-z)$ where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function. For a given patch I, we calculate two offset images – I_1 = dropping the last column of I, and I_2 = dropping the first

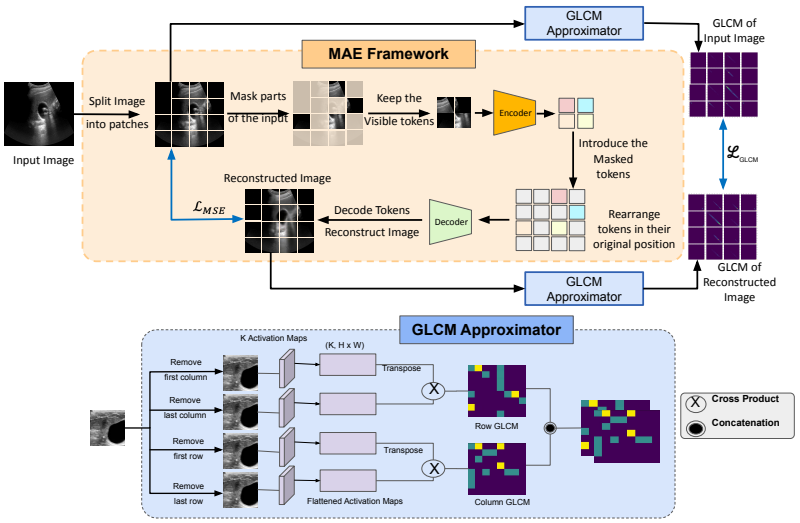

Fig. 2. Overview of the proposed GLCM-MAE pipeline. We propose novel GLCM loss guided reconstruction for improved representation learning in medical imaging tasks. The *black lines* represent forward pass and the *blue lines* represent loss computation.

column of I. Following [3], we calculate a differentiable joint histogram between I_1 and I_2 as: $H(I_1, I_2) = \frac{1}{N} P_1 P_2^T$, where, the KDE-based probabilities P_1 and P_2 for each intensity bin k are defined as:

$$P_1 = \sigma\left(\frac{I_1 - \mu_k + L/2}{W}\right) - \sigma\left(\frac{I_1 - \mu_k - L/2}{W}\right), \quad P_2 = \sigma\left(\frac{I_2 - \mu_k + L/2}{W}\right) - \sigma\left(\frac{I_2 - \mu_k - L/2}{W}\right). \quad (2)$$

Here, μ_k is the center of the k^{th} intensity bin, W is the bandwidth which controls the smoothness, $L = \frac{2}{K}$ is the length of bins, where K is the total number of bins. The sigmoid function, creates a soft, continuous approximation of bin boundaries. For sufficiently large bandwidth W, $H(I_1, I_2)$ converges to a smooth approximation of the GLCM. This smooth joint histogram, illustrated in Fig. 1, retains key structural properties of the GLCM while being differentiable. To extend this approach vertically, we repeat the above formulation to compute $V(I_1, I_2)$, representing the GLCM calculated from vertically adjacent pixel pairs. This allows for directional, and differential GLCM computation compatible with gradient-based learning.

GLCM Loss Function. Once the normalized GLCM is obtained, we calculate the GLCM loss by comparing the predicted and ground-truth GLCM matrices for each patch. Let $\hat{\mathcal{G}}_p^{(i)}$ and $\hat{\mathcal{G}}_g^{(i)}$ denote the differentiable GLCM computed for predicted and ground truth patch i. The GLCM loss is computed as the mean squared error between these matrices:

$$\mathcal{L}_{\text{GLCM}} = \frac{1}{|M|} \sum_{i \in M} \sum_{v,w} \left(\hat{\mathcal{G}}_p^{(i)}(v,w) - \hat{\mathcal{G}}_g^{(i)}(v,w)\right)^2 \quad (3)$$

where M refers to the set of masked patches.

Adding MSE and SSIM Loss for Stable Pre-training. While the GLCM loss helps capture subtle intensity relationships, it lacks global consistency due to its focus on local intensity relationships. Hence we further add pixel-wise MSE loss, and Structural Similarity Index (SSIM) loss to preserve visual structures in an image. We define the final loss function as: $\mathcal{L} = \alpha\mathcal{L}_{\text{MSE}} + \beta\mathcal{L}_{\text{GLCM}} + \gamma\mathcal{L}_{\text{SSIM}}$, where α, β, and γ are scaling factors. During the later phase of training, we scale down α to 0.1 to reduce the dominance of MSE loss, and preserve morphological features due to GLCM and SSIM losses. Fig. 2 shows the proposed pipeline.

2.1 GLCM-MAE Architecture

Fig. 2 shows a brief depiction of the proposed GLCM-MAE framework. Below we discuss the different parts of the proposed architecture.

Masked Token Generation. Similar to vanilla MAE, we divide the image into patches of 16×16 pixels and randomly mask 75% of the patches using uniform sampling. We choose a high masking ratio of 75% to prevent any data leakage while reconstructing the masked patches.

Encoder. We use a standard ViT architecture with embedding dimension of 768, encoder depth of 12 layers and 12 heads. Similar to the vanilla MAE strategy, only the visible image patches are passed to the encoder and the masked patches are dropped at the encoder to allow the encoder to train effectively on a fraction of the full image.

Decoder. The encoded visible patches, along with the masked patches are fed to the decoder, which is a shallow ViT and has a depth of 8, number of heads as 16 and embedding dimensions as 512. The decoder is only used during the GLCM-MAE pre-training and not during the finetuning of the downstream task.

Warm-up Phase. Training is divided into two phases. In the initial 200 epochs, we use only MSE reconstruction loss to provide stable gradients and improve structural learning ($\beta=0, \gamma=0$). This phase helps the model establish a coherent base representation, reducing instability in later training.

MAE Pre-training Phase. In the next 200 epochs, we activate the GLCM and SSIM losses, with the MSE loss scaled down by a factor of 0.1 ($\alpha=0.1, \beta=1, \gamma=1$). This phase promotes learning of both subtle intensity patterns (via GLCM) and structural integrity (via SSIM).

3 Experiments and Results

Datasets Used. We use four problems/datasets to demonstrate the efficacy of our approach, viz, GU: gallbladder cancer detection from ultrasound [5,8], BU: breast cancer detection from ultrasound [2], PX: pneumonia detection from chest X ray [17], and CC: COVID-19 detection from lung CT-scan dataset [27]. Due to restriction on paper length, we skip the problem/dataset details and request the reader to refer to the respective original papers.

Table 1. (Left) The 5-fold cross-validation (Mean) accuracy, specificity, and sensitivity of baselines and `GLCM-MAE` in detecting GBC from the US. `GLCM-MAE` achieves the best accuracy and sensitivity, which is much desired for GBC detection. (Right) The AUROC, Accuracy, and F1 of baselines and `GLCM-MAE` for identifying Pneumonia from chest X-Ray images. We report the metrics on the test set consisting of 390 pneumonia and 234 normal images.

Method	Acc.	Spec.	Sens.
ResNet50 [15]	0.867	0.926	0.672
VGG16 [26]	0.693	0.960	0.495
InceptionV3 [29]	0.844	0.953	0.807
RetinaNet [21]	0.749	0.867	0.791
EfficientDet [30]	0.739	0.881	0.858
ViT [11]	0.803	0.901	0.860
DEIT [32]	0.829	0.900	0.875
PVTv2 [33]	0.824	0.887	0.894
GBCNet [5]	0.921	**0.967**	0.919
US-USCL [9]	0.921	0.926	0.900
RadFormer [6]	0.921	0.961	0.923
MAE [14]	0.928	0.937	0.921
OmniMAE [12]	0.855	0.927	0.621
`GLCM-MAE` (Ours)	**0.949**	0.957	**0.934**

Method	AUROC	Acc.	F1.
VAE [19]	0.618	0.640	0.774
GANomaly [1]	0.780	0.700	0.790
f-MemAE [13]	0.778	0.565	0.826
MNAD [23]	0.773	0.736	0.793
IGD [10]	0.734	0.740	0.809
SQUID [36]	0.876	0.803	0.847
CheXzero [31]	0.927	0.830	0.875
Xplainer [24]	0.899	0.782	0.850
CoOp [38]	0.946	0.846	0.886
MaPLe [18]	0.951	0.894	0.895
PPAD [28]	**0.967**	0.894	0.918
MAE [14]	0.941	0.949	0.922
OmniMAE [12]	0.953	0.950	**0.968**
`GLCM-MAE` (Ours)	0.964	**0.955**	0.964

Table 2. The 5-fold cross validation (Mean±SD) accuracy, specificity, and sensitivity of baselines and `GLCM-MAE` in detecting breast cancer from ultrasound data.

	ResNet50[15]	DenseNet121[16]	RadFormer[6]	MAE[14]	OmniMAE[12]	**Ours**
Accuracy	-	-	0.842±0.063	0.892±0.013	0.832±0.050	**0.923±0.014**
Specificity	**0.975±0.024**	0.968±0.018	0.881±0.069	0.945±0.019	0.947±0.034	0.936±0.029
Sensitivity	0.829±0.088	0.824±0.027	0.738±0.105	0.748±0.059	0.614±0.124	**0.852±0.028**

Implementation and Evaluation. (1) Pretraining with `GLCM-MAE`: We implemented our experiments using PyTorch. The pretraining is done in two stages: we first train an MAE with ViT-B backbone using pixel-wise `MSE` reconstruction loss for 200 epochs. For the second stage, we resume training the model with a combination of the proposed loss in Eq. (3) for 200 epochs. For both stages, we mask 75% patches and train using a learning rate of 10^{-6} with weight decay of 0.05. We use a batch size of 8 and AdamW optimizer for pretraining. **(2) Finetuning for Downstream Tasks:** We finetune our pre-trained ViT encoder for 100 epochs for downstream classification task. The learning rate starts from 10^{-3} and decays to a minimum value of 10^{-6} with a weight decay of 0.05 with AdamW optimizer and a batch size of 128. We select the model that gives the highest accuracy on the validation set. **(3) Evaluation Metrics:** We used accuracy, specificity (true negative rate), and sensitivity (true positive rate/ recall) for GU, and BU, AUROC for CC, and AUROC and F1-score for PX evaluation.

Efficacy of `GLCM-MAE` over `SOTA` Baselines. To validate the effectiveness of our `GLCM-MAE`, we apply it on four disease detection tasks: `GU`, `BU`, `PX`, and `CC`. Our experiments reveal that `GLCM-MAE` beats the task specific `SOTA` on all four datasets.

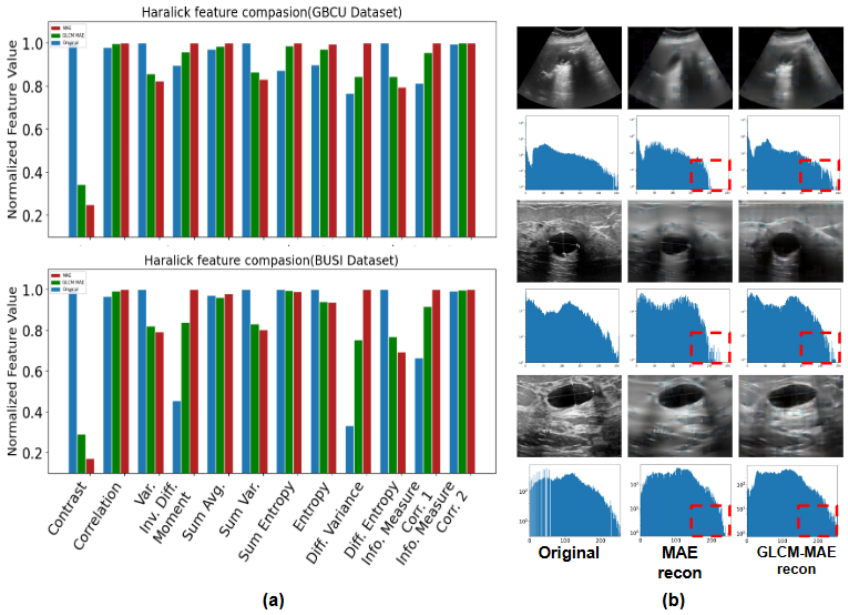

Fig. 3. (a) Shows the Haralick feature values derived from Original image (in blue), `GLCM-MAE` reconstructions (in green) and `MAE` reconstructions (in red) averaged over the test entire set of GBCU and BUSI dataset. The plot suggests better texture preservation of `GLCM-MAE` reconstruction relative to `MAE` over 12 texture based Haralick features. **(b)** Shows histogram comparisons between reconstructions of `MAE` and `GLCM-MAE` for images from GBCU and BUSI dataset, clearly depicting the loss of subtle gray level intensity values in `MAE` reconstructions, which are better conserved by `GLCM-MAE`.

We have used `SOTA` mask-based representation learning baselines namely, OmniMAE (CVPR2023) [12] along with the original MAE [14] for comparison with our proposed `GLCM-MAE`. Note that the MAE baseline methods use `MSE` loss based reconstruction in contrast to our `GLCM`-guided loss. **Quantitative Results:** We show the quantitative evaluation results in Tabs. 1 and 2. Additionally, we also report the results on COVID classification using CT images, where we achieve an AUROC of 99.19, exceeding the current SOTA, CropMixPaste[20] by 7.8% and exceed other MAE variants' [14] performance by 0.6%. **Statistical Significance:** The p-values of `GLCM-MAE` vs. MAE (`GU` = 0.036; `BU` = 0.005; `PX` = 0.034; `CC` = 0.048) show that `GLCM-MAE` gains are significant.

Masked Patch Reconstruction Analysis. We extract the Haralick features from the original, and reconstructed images using `MAE` and `GLCM` visually compare the outputs (Fig. 3). We see that reconstructions from `GLCM` have Haralick features similar to the original, providing conclusive evidence of `GLCM`'s superior performance. The `GLCM-MAE` reconstructions are visually closer to the original image, particularly in regions with subtle texture and intensity variations.

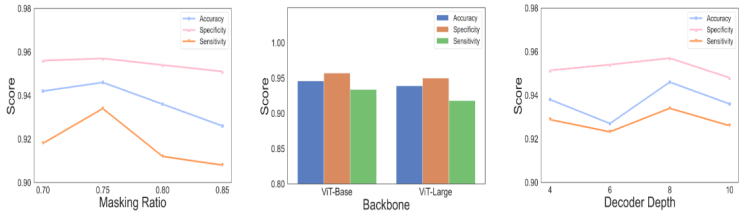

Fig. 4. Ablation study. We report the mean scores over 5-fold cross-validation for GBC detection. (a) Effect of varying the masking ratio. (b) Using different encoder backbones. (c) Effect of varying the decoder depth.

Table 3. Ablation Study Results. (Left) Effect of the kernel bandwidth W value on GBCU data. (Right) Effect of different reconstruction loss functions on `GLCM-MAE` on the GBCU Dataset.

Bandwidth (W)	Acc.	Spec.	Sens.
$W = 5$	0.939	0.941	0.924
$W = 15$	0.940	0.947	0.924
$W = 30$	**0.949**	**0.957**	**0.934**

Loss Function	Acc.	Spec.	Sens.
MSE	0.928	0.937	0.919
GLCM	0.930	0.947	0.919
GLCM + MSE	0.939	0.950	0.919
GLCM + SSIM	0.914	0.944	0.852
GLCM + MSE + SSIM	**0.949**	**0.957**	**0.934**

Ablation Study on the GBCU dataset. (1) Selection of Reconstruction Loss Function: Tab. 3(right) shows the effect of choosing the loss functions: Mean Squared Error (`MSE`), `GLCM` loss, Structural Similarity Index (`SSIM`), and their combinations, on the quality of the representation learned for the downstream GBC detection task. **(2) Bandwidth Selection:** Tab. 3(left) shows the effect of different kernel bandwidths (W) for GLCM computation. Kernel size of $W = 30$ offers a minor improvement over lower kernel sizes. **(3) Masking Ratio:** Fig. 4 (a) compares different masking ratios for `GLCM-MAE`. We found that a masking ratio of 0.75 achieved the best balance between learning generalizable features and preserving sufficient image context. **(4) Encoder:** Fig. 4 (b) shows the effect of choosing different Vision Transformer (ViT) variants as encoders for the token encoding task. **(5) Decoder Depth:** We conduct experiments by adjusting the decoder blocks, with results shown in Fig. 4(c). Performance improves when the decoder depth increases from 4 to 8. However, performance declines when the depth is extended to 10, aligning with findings in [7,22].

4 Conclusion

In this work, we introduced `GLCM-MAE`, a novel texture-aware self-supervised pretraining framework tailored for medical image classification. Unlike traditional `MAEs` relying on pixel-wise `MSE`-based reconstruction, we integrate a `GLCM`-guided loss to capture essential textures and spatial relationships in medical images. With `GLCM` guidance, our model effectively learns subtle intensity variations

that are often missed in `MSE` based reconstructions, enhancing feature extraction suited for medical imaging needs. In summary, `GLCM-MAE` sets a benchmark for texture-sensitive self-supervised learning, opening avenues for texture-based losses in medical imaging applications.

Acknowledgments. We thank the funding support from the Department of Biotechnology (grant number BT/PR51120/AI/133/179/2024), and Ministry of Education funding (grant number IITJMU/CPMU-AI/2024/0002) for the Centre of Excellence for AI in healthcare. Chetan Arora's travel is supported by Yardi School of AI Travel Grant.

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

References

1. Akcay, S., Atapour-Abarghouei, A., Breckon, T.P.: Ganomaly: Semi-supervised anomaly detection via adversarial training. In: Asian conference on computer vision. pp. 622–637. Springer (2018) 6
2. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020) 5
3. Avi-Aharon, M., Arbelle, A., Raviv, T.R.: Differentiable histogram loss functions for intensity-based image-to-image translation. IEEE TPAMI **45**(10), 11642–11653 (2023) 3, 4
4. Bandara, W.G.C., Patel, N., Gholami, A., Nikkhah, M., Agrawal, M., Patel, V.M.: Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In: CVPR. pp. 14507–14517 (2023) 2
5. Basu, S., Gupta, M., Rana, P., Gupta, P., Arora, C.: Surpassing the human accuracy: Detecting gallbladder cancer from usg images with curriculum learning. In: CVPR. pp. 20886–20896 (2022) 5, 6
6. Basu, S., Gupta, M., Rana, P., Gupta, P., Arora, C.: Radformer: Transformers with global–local attention for interpretable and accurate gallbladder cancer detection. Medical Image Analysis **83**, 102676 (2023) 6
7. Basu, S., Gupta, M., Madan, C., Gupta, P., Arora, C.: Focusmae: Gallbladder cancer detection from ultrasound videos with focused masked autoencoders. In: CVPR. pp. 11715–11725 (2024) 8
8. Basu, S., Papanai, A., Gupta, M., Gupta, P., Arora, C.: Gall bladder cancer detection from us images with only image level labels. In: International Conference on MICCAI. pp. 206–215. Springer (2023) 5
9. Basu, S., Singla, S., Gupta, M., Rana, P., Gupta, P., Arora, C.: Unsupervised contrastive learning of image representations from ultrasound videos with hard negative mining. In: MICCAI. pp. 423–433. Springer (2022) 6
10. Chen, Y., Tian, Y., Pang, G., Carneiro, G.: Deep one-class classification via interpolated gaussian descriptor. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 383–392 (2022) 6
11. Dosovitskiy, A., Beyer, L., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 6
12. Girdhar, R., El-Nouby, A., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Omnimae: Single model masked pretraining on images and videos. In: CVPR. pp. 10406–10417 (2023) 6, 7

13. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: ICCV. pp. 1705–1714 (2019) 6
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 16000–16009 (2022) 1, 6, 7
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 6
16. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR. pp. 4700–4708 (2017) 6
17. Kermany, D., Zhang, K., Goldbaum, M.: Labeled optical coherence tomography (oct) and chest x-ray images for classification (2018). https://doi.org/10.17632/rscbjbr9sj.2 5
18. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: CVPR. pp. 19113–19122 (2023) 6
19. Kingma, D.P.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 6
20. Li, W., Liu, G.H., Fan, H., Li, Z., Zhang, D.: Self-supervised multi-scale cropping and simple masked attentive predicting for lung ct-scan anomaly detection. IEEE TMI **43**(1), 594–607 (2024). https://doi.org/10.1109/TMI.2023.3313778 7
21. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proc. IEEE ICCV. pp. 2980–2988 (2017) 6
22. Madan, C., Gupta, M., Basu, S., Gupta, P., Arora, C.: Lq-adapter: Vit-adapter with learnable queries for gallbladder cancer detection from ultrasound images. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 557–567. IEEE (2025) 8
23. Park, H., Noh, J., Ham, B.: Learning memory-guided normality for anomaly detection. In: CVPR (June 2020) 6
24. Pellegrini, C., Keicher, M., Özsoy, E., Jiraskova, P., Braren, R., Navab, N.: Xplainer: From x-ray observations to explainable zero-shot diagnosis. In: International Conference on MICCAI. pp. 420–429. Springer (2023) 6
25. Ristea, N.C., Croitoru, F.A., Ionescu, R.T., Popescu, M., Khan, F.S., Shah, M., et al.: Self-distilled masked auto-encoders are efficient video anomaly detectors. In: Proceedings of the IEEE/CVF CVPR. pp. 15984–15995 (2024) 2
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 6
27. Soares, E., Angelov, P., Biaso, S., Froes, M.H., Abe, D.K.: Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. medRxiv (2020). https://doi.org/10.1101/2020.04.24.20078584, https://www.medrxiv.org/content/early/2020/05/14/2020.04.24.20078584 5
28. Sun, Z., Gu, Y., Liu, Y., Zhang, Z., Zhao, Z., Xu, Y.: Position-Guided Prompt Learning for Anomaly Detection in Chest X-Rays . In: proceedings of MICCAI 2024. vol. LNCS 15001. Springer Nature Switzerland (October 2024) 6
29. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR. pp. 2818–2826 (2016) 6
30. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proc. IEEE CVPR. pp. 10781–10790 (2020) 6
31. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering **6**(12), 1399–1406 (2022) 6

32. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICML. pp. 10347–10357. PMLR (2021) 6
33. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvtv2: Improved baselines with pyramid vision transformer (2021) 6
34. Wei, C., Mangalam, K., Huang, P.Y., Li, Y., Fan, H., Xu, H., Wang, H., Xie, C., Yuille, A., Feichtenhofer, C.: Diffusion models as masked autoencoders. In: Proceedings of the IEEE/CVF ICCV. pp. 16284–16294 (2023) 1
35. Wei, Y., Gupta, A., Morgado, P.: Towards latent masked image modeling for self-supervised visual representation learning. In: ECCV. pp. 1–17 (2025) 2
36. Xiang, T., Zhang, Y., Lu, Y., Yuille, A.L., Zhang, C., Cai, W., Zhou, Z.: Squid: Deep feature in-painting for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF CVPR. pp. 23890–23901 (June 2023) 6
37. Yao, X., Zhang, C., Li, R., Sun, J., Liu, Z.: One-for-all: Proposal masked cross-class anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 4792–4800 (2023) 2
38. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022) 6