

Image reconstruction by localized inpainting for anomaly detection in ultrasound images

Sonia Petrini

October 27, 2022

1 Introduction

This work is framed in the context of the project 'automatic labeling of ultrasound images' developed by the EveryWere Lab of University of Milan. The goal is to detect the presence of effusion in ultrasound images of knees in individuals affected by hemophilia, a genetic disorder related to impairment of blood clotting mechanisms. Patients cannot always receive medical attention when suffering a contusion, so that medication is prescribed based on how the patient feels, possibly resulting in patients not receiving medication when needed, or in undue administration. Thus, automating the way in which a pathological knee is identified would enhance the process of treatment prescription, both providing better healthcare to patients suffering from this disorder, and allowing for a more efficient allocation of resources in the sanitary system.

Here, we address the issue of anomaly detection by adopting an approach based on image reconstruction by inpainting. An autoencoder-based convolutional network is trained to reconstruct anomaly-free images, where some pixels have been masked. Then, the trained network – which has learnt how to properly reconstruct healthy knees – is fed with ultrasound images of both normal and anomalous knees, and the output image is compared with the original one. The idea is that the network will use the context information from pixels around the inpainting mask to recreate an image without anomalies, so that the reconstruction error will be particularly high for anomalous instances, allowing for their identification. The main methodology and the model architecture build on the work presented in [3]. However, the masks used in the present work are localized on the area where the anomaly is expected. In fact, anomalies are related to the effusion of blood or synovial liquid in the knee bursa, whose estimated location on the image is provided by a model for localization built in the main project. We also implement their proposed weighted loss function based on gradient similarity, which proves to outperform Mean Squared Error in the task of image reconstruction. This loss function is obtained by a linear combination of the Structural Similarity Index (SSIM) [4] and the multi-scale Gradient Magnitude Similarity (MSGMS), plus MSE for regularization. In addition to the weights setting suggested in the article, we explore other possible combinations in a slight variant of the formula, namely one in which weights sum up to 1. This allows us to understand more about the relative importance of the structural features considered in the loss function, leading us to choose as best model for our purposes one in which only MSGMS is taken into account. This model obtains 82.3% $f\text{-}\beta$ score, 71.9% precision and 93.5% recall when the threshold for classification is chosen by optimizing $f\text{-}\beta$ with $\beta = 1.1$, thus giving more weight to recall.

2 Methods

Anomaly detection methods are based on the idea of an underlying normal distribution in the data, and the purpose is to identify instances that deviate from this distribution. If the model learns well how to recreate a 'normal' knee, it will struggle when applied to the ultrasound image of a knee with effusion, so that the distance between the original and the recreated image will be larger than in an anomaly-free knee. Thus, we first train a model for image reconstruction in a self-supervised fashion. Then, we create anomaly maps representing the distance between original and predicted images, and from the map we estimate an anomaly score for each image. Finally, based on the estimated anomaly scores we classify the images as presenting an anomalous bursa or not. In the following we report on the methodology adopted in each step.

2.1 Data

The dataset contains a total of 483 ultrasound images, with information on the presence of effusion ('yes', 'no'), the confidence of the annotation ('low', 'medium', 'high'), and a separate file containing the masks coordinates in YOLO format. For 4 images the coordinates annotations were missing and they have been dropped. Given that data is already scarce and the missing coordinates are few, an alternative approach would be to manually assign coordinates to these images. Other two files could not be read and they have been dropped as well. Finally, we had 354 negative and 123 positive examples, which have been split in training, validation, and test sets in such a way that the first two only contain anomaly-free instances, while the last contains approximately an equal number of pathological and anomaly-free images. Importantly, the dataset can contain multiple images for a single patient. Thus, the division has been made by making sure that each patient only appears in one among the sets, to avoid information leakage. This resulted in a train set of size 218, a validation set of size 61, and a test set of 198. The information about confidence is related to the noisiness of the labels. In fact, even for a human annotator it is often difficult to make a diagnosis and assign a label to an ultrasound image, given the high variability among individuals. As shown in figure 1 only few instances have low confidence, the majority of positives examples have medium confidence, and the majority of negative examples have high confidence. We do not consider this aspect in the present analysis, but future work could assign weights to images based on this attribute, so that more reliable labels have a larger weight. Before being fed to the model, all images are resized to 256x256 and normalized between 0 and 1.

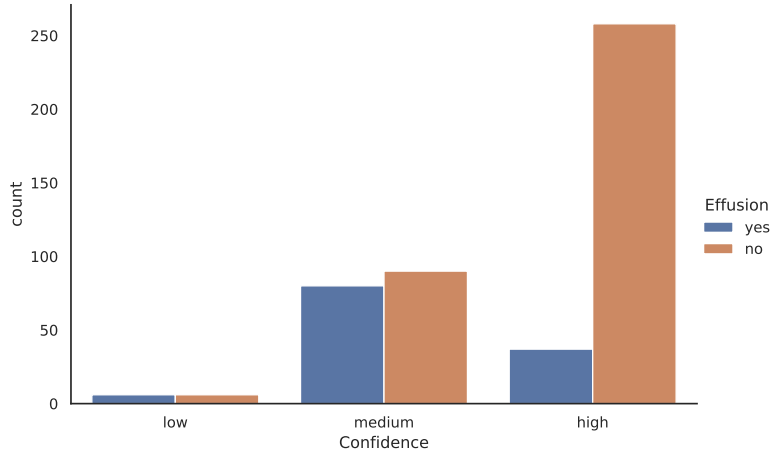


Figure 1: Data distribution by presence of effusion and confidence of the annotation.

2.2 Model architecture

The chosen architecture is a U-Net, the state-of-the-art model for image reconstruction [2]. U-Net is an autoencoder-based convolutional neural network, which owes its U-shape to the skip connections between encoding and decoding layers. The encoding layers progressively down-sample the image through convolutions down to the chosen latent dimension. The input is then decoded and reconstructed by up-sampling it again to its original dimension. Following [3] we set the latent dimension to 512. At each step of both the contraction and expansion paths we apply twice a convolution, batch normalization, and ReLU activation. As mentioned above, one of the main features of U-Net is the presence of skip connections, which facilitate image reconstruction by passing high level pattern information to the deeper layers. This is especially useful in our use case, as we want the network to be able to faithfully reconstruct the image around the mask. Once decoded, outputs are passed through a final convolutional layer with a sigmoid activation function. In the original paper a combination of tanh and normalization between -1 and 1 is used. However, on our data this resulted in some areas of the image not being properly reconstructed, and the sigmoid was chosen as it turned out to be better suited. Learning rate is fixed to 0.0001 and decreased by a factor of 10 after 250 epochs, and batch size is set to 8, as in the original study. After each training we select the model with the best weights with respect to validation loss, rather than the last model estimated.

2.3 Loss function

The traditional Mean Squared Error (MSE) has proven to perform poorly in visual quality assessment tasks [1]. In fact, the human visual system perceives anomalies by considering structural components of the image which are not reflected by MSE, as this is based on an assumption of independence between pixels. Thus, following [3] we consider a linear combination of two other loss functions, the structured similarity index (SSIM) [4] loss and the multi-scale gradient magnitude similarity (MSGMS) [5] loss, both able to detect structural differences between images. In particular, SSIM takes into account luminance similarity, contrast similarity, and structural similarity, while MSGMS detects similarity in images' gradients at multiple scales. Given that they are measures of similarity, the corresponding loss functions are defined by taking their av-

erage distance from 1 across pixels, so as to obtain $SSIM_{loss}$ and $MSGMS_{loss}$. The final loss function, which we refer to as 'Gradient Similarity loss' (\mathcal{L}_{gs}) is a weighted average of the two, plus L_2 loss for regularization, and has the following form:

$$\mathcal{L}_{gs} = \lambda_s * SSIM_{loss} + \lambda_m * MSGMS_{loss} + L_2 \quad (1)$$

where λ_s and λ_m are the weights of $SSIM_{loss}$ and $MSGMS_{loss}$ respectively. In the original formulation these weights could possible take on every value, but in order to tune them we introduce a modification for two of the models we develop. Namely, we define the weights in such a way that:

$$\lambda_s + \lambda_m = 1 \quad (2)$$

In this way, we can tune λ_s and let λ_m adjust consequently. Our goal in doing this is that of assessing which out of SSIM and MSGMS is more informative given the visual properties of the data, rather than finding the best combination of the two. In fact, the latter goal can also be more easily achieved once there is a better understanding of the former.

To implement $MSGMS_{loss}$ we had to modify the Tensorflow source code for Prewitt filtering. In fact, the first step to obtain $MSGMS$ is the computation of the gradient magnitude maps for both the original and the reconstructed images, which involve applying a square root. During training the gradients computed for this loss vanished, resulting in nan values when the square root was applied. To solve this problem, we simply added a very small constant to the operation under square root, following this Github issue. We also followed the unofficial code implementation of [3], available here, which is however based on the PyTorch framework.

2.4 Anomaly scores

The discrepancies between original and reconstructed images are localized and quantified through anomaly maps and scores. A Gradient Magnitude Similarity (GMS) [5] distance map is computed between the original and the predicted image at four different scales, and their pixel-wise average is returned. Thus, these MSGMS anomaly maps highlight the areas in which a structural difference is observed between the images. In order to obtain a single anomaly score for each image we aggregate the map values; ideally, this value should be low for healthy knees and high where effusion is present. Following [3] we aggregate by taking the maximum value. However, one could define other metrics that take into account the whole distribution of anomaly scores in the map. In the remainder of the work we often refer to this score as A_{score} .

2.5 Model selection procedure

The problem of anomaly detection is tackled in two stages. While the final goal is to detect the presence of effusion in knees, this is achieved by first training a network on an image reconstruction pretext task. Thus, the functioning of the anomaly detection algorithm is based on the assumption that the predictor which is better able to reconstruct negative instances will be the one making more mistakes on positive ones, thus allowing for the separation of the two classes. For this reason, we use the performance of the models on this self-supervised pretext task to select the best architecture to be tested on unseen data. Namely, for every model we compute the A_{score} between each image in the validation set and its reconstruction, and we select the model with the lowest scores. The reason is that the validation set only includes negative instances, which we want to be properly reconstructed (and thus to have a low A_{score}). Given

that for each model we have a distribution of values, we look at both the mean and the median scores, but we favour the latter given its greater robustness to outliers.

2.6 Classification

Once we have chosen the best model based on the pretext task we use it to reconstruct the images in the test set. By computing the anomaly scores between the reconstructions and the original images we can finally produce a classification of the examples. This is done by setting a threshold on the level of anomaly allowed in a 'normal' image, so that images with a score higher than this threshold are labeled as anomalous. In the framework of medical data it is common to put greater attention in avoiding false negative – that is not recognising pathological instances – as this usually has more adversary effect than treating someone which does not need it. However, in the scope of this project we aim not only at providing medication to patients that need it, but also at inducing a more efficient use of resources, so that the value of precision should not be underestimated. Hence, we consider both the traditional f1 score and its weighted version, the f- β score, in which we slightly increase the weight of recall in the harmonic mean to $\beta = 1.1$.

3 Discussion

Our analysis unfolds in two steps. First, we replicate the scenario proposed in [3], assessing the advantage of considering the Gradient Similarity based loss function in 1 as opposed to MSE. Thus, we build Model0 and Model1, respectively trained with MSE and with \mathcal{L}_{gs} . Then, having verified the advantage of including measures of structural difference in the computation of the loss, we modify the weights of the two terms composing \mathcal{L}_{gs} to understand their relative importance with respect to our task. Model2 and Model3 are thus built by modifying the value assigned to λ_s in 1. Every model is trained by keeping number of epochs, batch size, and learning rate fixed to the values used in the original paper.

3.1 Relevance of Gradient Similarity loss

Let us compare Model0 and Model1, in which the architecture is fixed but the loss functions are MSE and \mathcal{L}_{gs} (with equal weight for $SSIM_{loss}$ and $MSGMS_{loss}$) respectively. In figure 2 (a-b) we display the history of training and validation loss for these two models. Given that both Models contain a term for MSE, they clearly follow the same shape at different scales. However, we can also observe how validation loss in Model0 shows more stability compared to Model1, suggesting that the presence of $SSIM_{loss}$ and $MSGMS_{loss}$ introduces fluctuations in the model's performance. Table 1 displays the validation performance of these first two models, reporting mean and median A_{score} on unseen negative instances. As mentioned in the previous section, we focus on the median score and report the mean for robustness. Model1 achieves a better reconstruction accuracy, assigning an anomaly score lower or equal to approximately 0.61 to half of the images, while for Model0 this median value raises to roughly 0.64. We can visualize this in 3, displaying the density plots for the anomaly scores of each model. The distributions for the first two models largely overlap, but they are skewed towards opposite directions, with Model1 favoring lower values. Even given the longer training time – which increases by about 6 times when using \mathcal{L}_{gs} instead of MSE – the additional complexity of the Gradient Similarity loss seems to enhance the model's ability to reconstruct anomaly-free images.

3.2 Relative importance of structural similarity measures

We thus extend the analysis by considering the relative weight of the two considered measures of structural similarity included in \mathcal{L}_{gs} , SSIM and MSGMS. In fact, depending on the features of the analysed images one could be more suited than the other, and should therefore have a larger weight in the penalization. As described in section 2, in order to tune the weights parameters we introduced a simplification with respect to [3], which arises from the goal of evaluating the importance of one structural similarity measure with respect to the other. Given the long training times (around 6 hours for 300 epochs) and the reduced search space, instead of running a grid search in one shot we proceed in a progressive exploratory way. The loss of Model1 approximately corresponds to a configuration in which $\lambda_s = \lambda_m = 0.5$, so that we build Model2 by setting $\lambda_s = 0.25$, meaning that SSIM has a lower impact on the overall loss. As shown in table 1 and figure 3, this results in a reduction of A_{score} on validation, which – despite not being great in terms of median value – suggests that reducing the weight of SSIM might be the way to go. Hence, we continue moving in the same direction and build Model3 by setting $\lambda_s = 0$, namely only considering $MSGMS_{loss}$ (in addition to MSE for regularization) in \mathcal{L}_{gs} . This last experiment confirms what has been observed moving from Model1 to Model2 with respect to the greater power of $MSGMS_{loss}$ in reconstructing impainted images. In fact, Model3 clearly outperforms all the previous models in the reconstruction task on the validation set, with half of the negative images showing an A_{score} of 0.55 or lower, which is a great improvement from the value of 0.64 obtained with the simple MSE. Indeed, figure 3 shows how the distribution of anomaly scores from Model3 is substantially shifted towards lower values with respect to the others. We thus select Model3 as the best one in the scope of this work, but we cannot exclude that assigning different combinations of values to λ_s and λ_m (not constraint to sum to 1) could yield better results. Additional information about SSIM and MSGMS can be extrapolated from the history of their training. As previously mentioned, updating the models weights with respect to \mathcal{L}_{gs} introduced variability in validation loss values. We see in figure 2 how this variability progressively decreases with λ_s , suggesting it being mainly related to $SSIM_{loss}$.

Table 1: Models’ validation performance. For each model, we display the median and mean anomaly scores A_{score} on negative examples from the validation set. Loss functions for Model2 and Model3 are of the form: $\mathcal{L}_{gs}(\lambda_s)$.

model	loss	median A_{score}	mean A_{score}
model0	MSE	0.6390	0.6299
model1	\mathcal{L}_{gs}	0.6132	0.6234
model2	$\mathcal{L}_{gs}(0.25)$	0.6081	0.6096
model3	$\mathcal{L}_{gs}(0)$	0.5489	0.5496

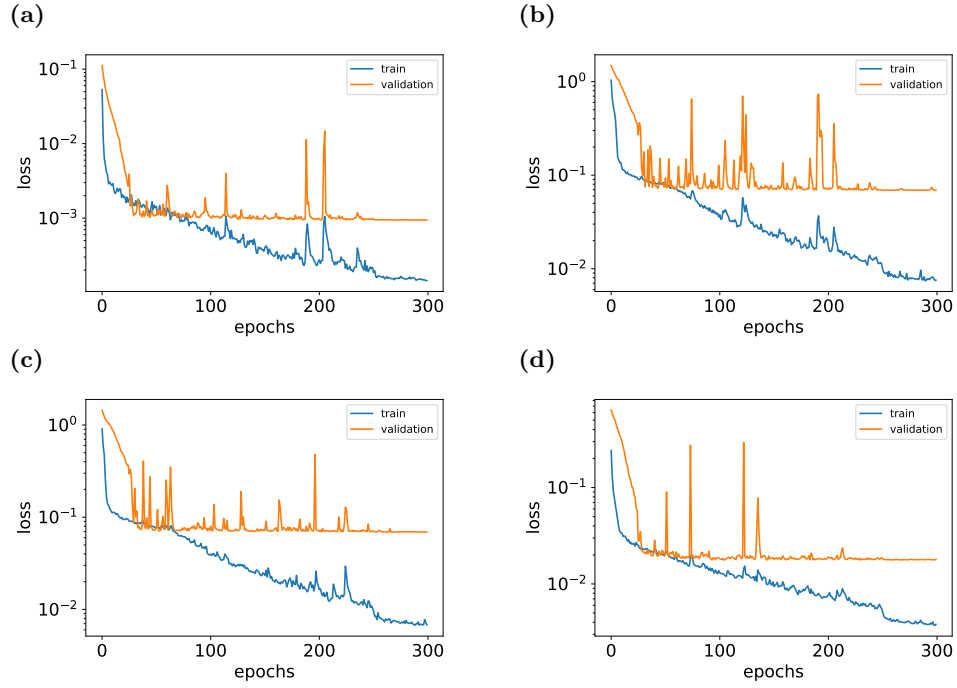


Figure 2: Evolution of training and validation loss over number o epochs. Model0 (a). Model1 (b). Model2 (c). Model3 (d).

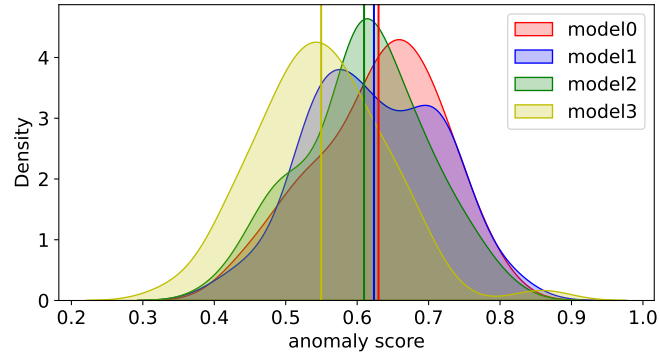


Figure 3: Anomaly scores density by loss function. Scores are computed on the validation set. The vertical lines mark the median values of the distribution.

3.3 Anomaly detection

Once the best model has been selected by means of its performance on the pretext-task, we can proceed to test it on the real task of anomaly detection. Thus, we reconstruct the impainted negative and positive images in the test set and compare them with the originals, computing the anomaly scores as in the training phase. In this scenario, given the presence of both normal and anomalous knees a good model would not be one that achieves the lowest or the largest scores, but one that is able to discriminate between positive and negative examples, assigning low scores to the latter and high scores to the former. Thus, to evaluate the performance of the model on this classification task we look at precision and recall, observing the model’s behaviour when either f_1 or f_β is optimized. We show the obtained metrics in 2. Optimizing f_1 leads as expected to a better balance between recall and precision, with both reaching a score larger than 80%, and an overall f_1 of 81.6. On the other hand, Optimizing f_β allows recall to reach 93.5%, while precision goes down, still remaining above 70%. Thus, depending on the way in which one wants to balance the costs of not prescribing due medication and wasting resources either of the two resulting thresholds can be chosen. Our personal opinion is that a precision value of 71.9% is still acceptable, especially considering the large gain in recall, and f_β should thus be preferred as a target metric. As previously mentioned, a model is able to correctly detect anomalies if it is able to well separate the distributions of positive and negative examples. In figure 5 we show the A_{score} density by presence of effusion for each of the models. While this information is not used to select the best model, it is interesting to observe that Model3 actually yields the smallest overlapping between the two distributions of positive and negative cases.

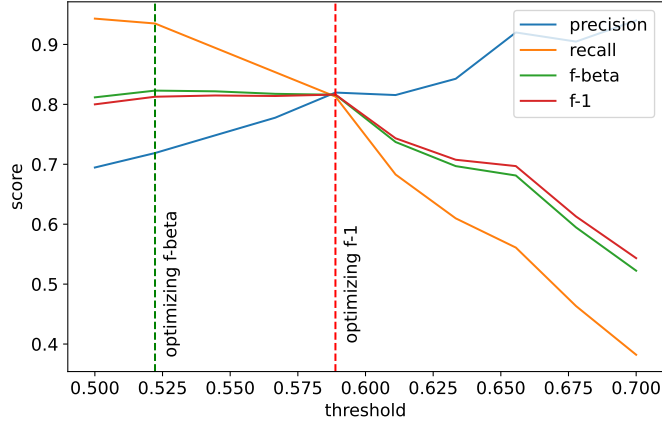


Figure 4: Classification performance for varying threshold. We include precision, recall, f_1 and f_β scores. The two dashed lines mark the thresholds chosen by optimizing each one of the f -scores.

Table 2: Model3 test performance. We display the f -score, precision, recall metrics and the best threshold obtained by optimizing either f_1 or f_β , where $\beta = 1.1$. Values are expressed in percentage.

target metric	f-score	precision	recall	threshold
f-1	81.6	82.0	81.3	0.59
f- β	82.3	71.9	93.5	0.52

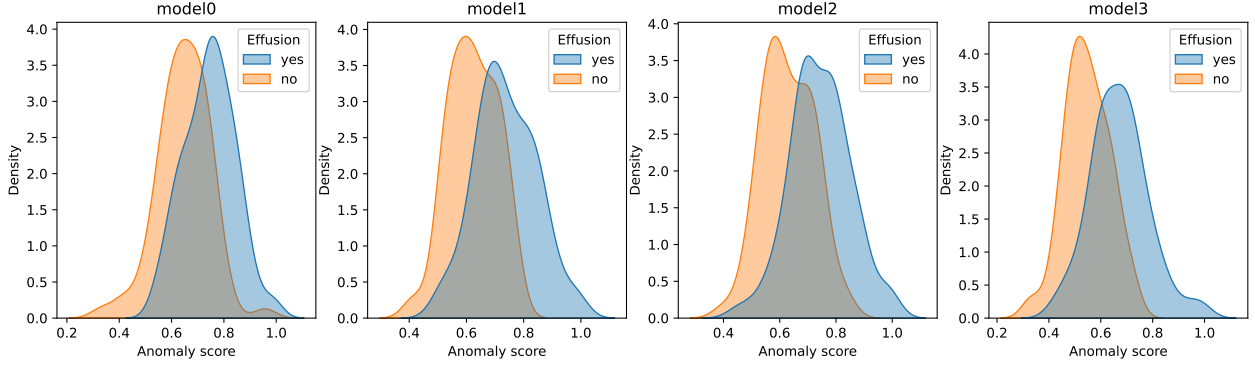


Figure 5: Anomaly scores density by model and class. Scores are computed on the test set.

3.4 Critical evaluation

The values reported in table 2 suggest that Model3 has quite a satisfactory performance in the task of anomaly detection, even if clearly inferior with respect to [3]. In fact, these results have to be contextualized with respect to our framework. First, the little numerosity and the nature of the data make this classification task complex. As it is often the case with medical data, the images that we used to train, validate, and test the network are only few hundreds, an order of magnitude less than those available in the original study. Moreover, the data used in [3] comes from industry, and thus contains many standardized images of the same kinds of objects, while this homogeneity is lacking in our dataset. Ultrasound images show a very high variability within each class: not only anomalies manifest in different forms and magnitudes, but physiological knees also present some natural variability. In addition to this there is no clear cut between a 'normal' and an 'anomalous' instance, so that the anomaly scores distribution of positive and negative cases are virtually inseparable. Extreme examples of the anomaly maps built with Model3 are displayed in figure 6: from left to right, we show the image with the lowest and the highest A_{score} in the negative class, and those with the lowest and the highest scores in the positive class. Larger A_{score} values are marked with yellow on the maps. Based on these instances we can make some considerations. As a first point – related to what just said about variability – even on negative examples the reconstruction ability of the model is not perfect. For instance, according to either of the chosen thresholds the second image, while being negative, is classified as positive. Second, anomalous images can have low anomaly scores too, especially if the bursa in which effusion occurs is not very dilated. As a final interesting point, the positive image with the highest anomaly score is not one in which the bursa is particularly expanded, but one in which the model tried to reconstruct the black borders, resulting in a wide and intense anomaly area. With this regard, images could be further processed by cropping out the borders where present. This would both increase data homogeneity and avoid this kind of mistake from happening. Another aspect to be taken into account concerns the masks. On one hand, the coordinates are obtained through an external localization model, which carries its potential flaws. While generally being very accurate in covering the bursa, some masks might leave some parts of it visible, thus reducing the ability of the network to reconstruct the impainted pixel generalizing from their neighbours. On the other hand, given that the reconstruction ability of the model is not perfect even on negative images, the larger the mask the higher the probability of obtaining a high anomaly score, as a wider area of the image has the possibility to be wrongly reconstructed. Thus, larger bursae have an a priori higher probability of being classified as positive. As a

final point, notice that both thresholds take values between 0.5 and 0.6; we suspect that this is related to the artificial balance created between positive and negative examples. While in the whole dataset positive cases account for about a fourth of the total, given the need of negative examples for training and validation the resulting test set has a similar proportion of examples, so that the threshold is located around 0.5. However, this is rarely the case in a real-world scenario.

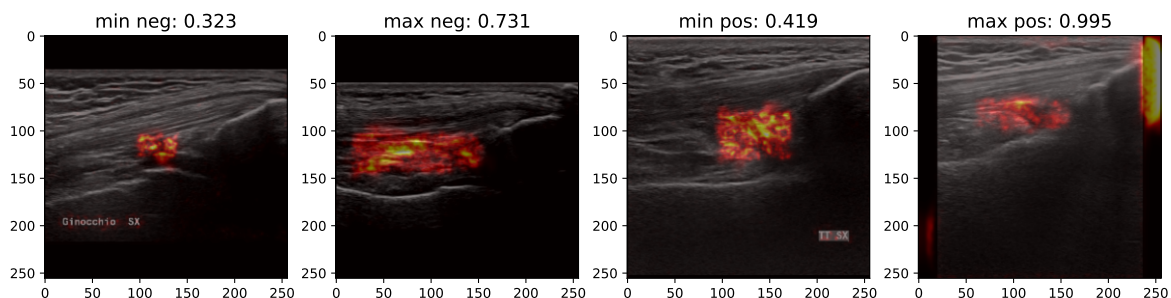


Figure 6: Anomaly maps over original images. Anomaly maps are computed between the original image and the one reconstructed by the chosen best model. Here we display the extreme cases of each class in terms of A_{score} : minimum of negative examples, maximum of negative examples, minimum of positive examples, maximum of positive examples.

3.5 Future work

Let us summarise the proposed improvements to the current work. Of course, a larger dataset would yield better results, but this factor cannot be controlled. Aspects that could be modified to improve performance are the number of epochs, batch size, and learning rate, out of which we expect the number of epochs to have the largest impact on reconstruction ability. Indeed, especially with a deep architecture like the U-net used in this work improvements may happen after a very large number of epochs. Then, future work will focus on better understanding the relation between SSIM and MSGMS, identifying the optimal balance between the two, or confirming that for the kind of data under examination SSIM is actually not well suited and can be excluded. Larger masks can be used in order to make sure that the whole area where the anomaly is present is properly covered. This should nevertheless take into account that a larger mask implies a greater chance of the image being classified as positive. Moreover, coordinates for the four images that are lacking them could be manually annotated in order to use all the available examples. Finally, the black borders could be cropped out of images, to improve homogeneity and avoid reconstruction errors on them. From a methodological point of view, information on the confidence of the annotations could be incorporated in the model, so as to create a more reliable predictor. Moreover, alternative aggregation techniques to obtain a single anomaly score from the maps could be investigated.

4 Conclusions

We here explored an approach to anomaly detection based on image reconstruction by localized inpainting. We managed to obtain a model with 82.3% f- β score, 93.5% recall and 71.9% precision on the anomaly detection task. Such model was obtained by progressively modifying the gradient similarity based loss function proposed in [3] in order to make it more suited to our specific problem. As a result, out of the two structural similarity measures proposed in the original formulation of the loss, using only multi-scale gradient magnitude similarity (MSGMS) turned out to yield better results in reconstructing ultrasound images. The test performance obtained by this model is promising, suggesting that this methodology might be a good direction to explore further.

5 Acknowledgements

We want to thank EveryWare Lab for giving us the opportunity to be a part of this important project, and in particular Marco Colussi for the support given in the development of this work.

References

- [1] Bernd Girod. *What's Wrong with Mean-Squared Error?*, page 207–220. MIT Press, Cambridge, MA, USA, 1993.
- [2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. cite arxiv:1505.04597Comment: conditionally accepted at MICCAI 2015.
- [3] Danijel Skočaj Vitjan Zavrtanik, Matej Kristan. Reconstruction by inpainting for visual anomaly detection. 2021.
- [4] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [5] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index, 2013.

“I/We declare that this material, which I/We now submit for assessment, is entirely my/our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my/our work. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.”