TargetIQA: Anatomy-Aware Ultrasound Image Quality Assessment

George Lansdown[1,2(✉)], Sotirios A. Tsaftaris[2], Alison Q. Smithard[1,2], and Antanas Kascenas[1]

[1] Canon Medical Research Europe, Edinburgh, UK
[2] The University of Edinburgh, Edinburgh, UK
george.lansdown@mre.medical.canon

Abstract. Poor medical image quality hinders accurate clinical decision making. Image quality in ultrasound is modulated by many variables, including positioning and acquisition parameter choices. Optimising these variables is slow because human judgement of image quality is required. Existing automatic image quality assessment (IQA) methods are overly general, not adaptable, or are designed for static tasks. We argue that the clinical value of an image depends on the quality of specific anatomical regions rather than the image as a whole. Based on this, we propose an IQA method that conditions the quality score on the operator's anatomical focus at scan time. By exploiting a contrastively pre-trained biomedical vision-language model (BiomedCLIP), TargetIQA requires no dedicated training and outperforms trained organ-specific IQA methods, a pre-trained VLM with a 138× higher parameter count, and general IQA methods across a variety of tasks. Our ablations show that TargetIQA is sensitive to the anatomical context, and this context meaningfully changes its quality scores. These findings show anatomical conditioning is a lightweight but effective basis for clinically meaningful IQA.

Keywords: Ultrasound Imaging · Image Quality Assessment · Vision-Language Models · Zero-Shot Learning

1 Introduction

In standard ultrasound imaging, a probe is manually placed near the anatomy of interest to collect 2D planar images with a high refresh rate. Small probe movements shift the imaged plane and change the image on the scanner. Suitable acquisition parameters must then be selected based on the depth of the object of interest (OOI), the tissue composition of the patient, etc. Capturing ultrasound images of high clinical value requires a well-trained clinician to position the transducer, and adjust acquisition parameters – and dynamically update these in response to challenges like probe or patient motion [21]. For procedures like ultrasound-assisted prostate biopsy where a biopsy tool must also be positioned, two clinicians are typically required. Thus, this manual optimisation cycle can slow the procedure and increase clinical staffing requirements.

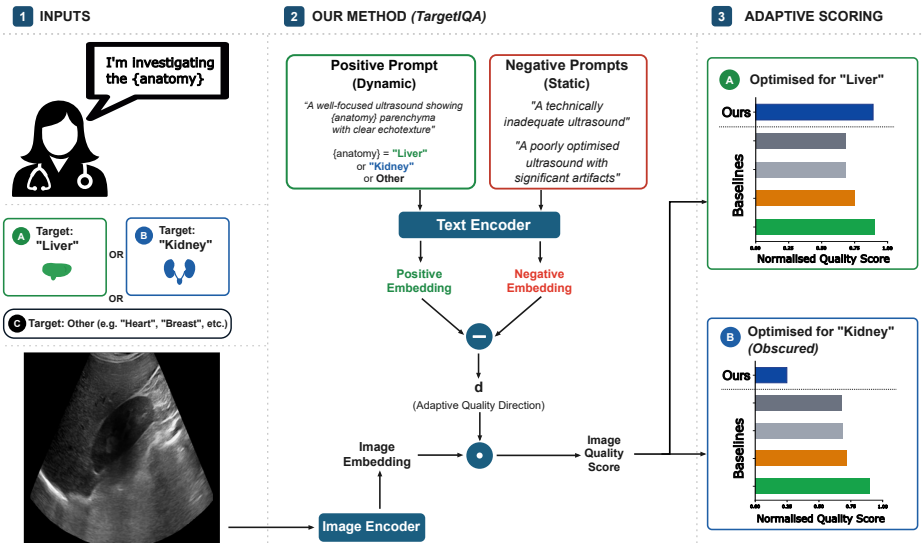

Fig. 1. Anatomically insensitive IQA methods produce inappropriate scores in scans with multiple organs. We demonstrate TargetIQA using an example ultrasound image with two potential anatomical intents. The kidney is obscured by shadowing whereas the liver's echotexture is clear and high quality. Only TargetIQA adapts to these differences. The image and scores are drawn from our experiments.

Assistive tools for producing high-quality images could have substantial clinical impact by reducing the need for clinicians' manual judgement of image quality. Closing this gap requires an automatic image quality assessment (IQA) metric. Crucially, clinical image quality is specific to the target anatomy, pathology, or surgical device, and ultrasound workflows are often dynamic – requiring the clinician to focus on different OOIs over time or in response to image findings. A clinically useful IQA metric should therefore be sensitive to the current OOI and adapt to changes in task during the scan.

Two widely used metrics are BRISQUE [13] and NIQE [14], which model natural scene statistics; NIQE in particular has correlated with clinical quality perception in ultrasound [11]. CLIP-IQA [19] is an alternative embedding-based metric that scores the alignment between an image and antonym captions such as "Good photo" and "Bad photo" using contrastively pre-trained vision-language models (CLIP [16]). Both of these methods are anatomy-agnostic by design and do not adapt to the operator's current focus.

In contrast, several methods are trained on examples for a specific organ or clinical task [2, 3, 6, 7]. TinyUSFM-NRQ [7], for instance, introduces a family of models trained to rate image quality across eleven target organs. A significant limitation of such methods is that deploying and switching across multiple bespoke IQA models – one per organ – is impractical for dynamic workflows requiring coverage of multiple organs. A different line of work adapts frame-by-

frame by automatically defining a region of interest (ROI) targeting the OOI and computing a ROI-local quality metric [4, 20]. These methods are OOI-specific but may be challenging to deploy on a low-resource ultrasound scanner due to the additional computational cost of real-time ROI localisation, particularly when multiple OOIs are investigated during a procedure.

We present TargetIQA, a simple and training-free IQA method (Figure 1) that scores image quality relative to the operator's anatomical target, specified in natural language at scan time. The method measures the alignment between the image embedding and a dynamically generated text embedding containing information describing the current clinical task. Our contributions are:

- We propose anatomy-conditioned IQA, a formulation in which the quality score adapts to the operator's stated anatomical target — reflecting the clinical reality that image quality is task-specific and workflows are dynamic.
- We propose TargetIQA, a training-free realisation of this formulation that adapts to clinical intent through natural-language prompting at scan time.
- We provide an evaluation across three tasks, showing TargetIQA outperforms trained organ-specific, general-purpose, and VLM baselines, with ablations isolating anatomical conditioning as the most impactful factor.

2 Method

The main idea behind our method is that it computes image quality by estimating how distinctly an image depicts the target anatomy. Our intuition is that very high-quality ultrasound images look clearly like the anatomy being scanned, whereas very low-quality ultrasound images are indistinct and resemble one another regardless of target. The method captures this in BiomedCLIP's [23] shared image-text embedding space by constructing a direction vector at scan time that points from the area of generic low-quality scans toward distinct, high-quality scans of the named anatomy. It scores an image by its cosine similarity with this direction. The direction is adapted from a clinician-supplied anatomical prompt and a fixed set of negative prompts describing undesirable characteristics.

TargetIQA: BiomedCLIP [23] provides an image encoder $f_I : \mathcal{X} \to \mathbb{R}^d$ and a text encoder $f_T : \mathcal{T} \to \mathbb{R}^d$ that map into a shared, ℓ_2-normalised embedding space, such that $\|f_I(x)\| = \|f_T(t)\| = 1$. We adopt BiomedCLIP [23] rather than the original CLIP [16] for its broader coverage of ultrasound images (we demonstrate results comparing the two models in the supplementary material).

We define a positive prompt t_a^+ describing the anatomy a desired by the clinician, and a negative prompt t^- describing undesirable image characteristics. Subtracting their embeddings gives an anatomy-specific direction vector $\mathbf{d}_a$ in the joint embedding space, as $\mathbf{d}_a = f_T(t_a^+) - f_T(t^-)$.

This vector approximates movement from the region of indistinct, low-quality images toward distinct, high-quality images which match the desired anatomy clearly. TargetIQA estimates the image quality Q via cosine similarity of the image embedding with the ℓ_2-normalised direction $Q(x; a) = f_I(x)^\top \frac{\mathbf{d}_a}{\|\mathbf{d}_a\|}$.

Changing a at scan time alters $\mathbf{d}_a$ whilst the embedding model remains unchanged, enabling zero-shot adaptation to the changes in anatomical intent common in dynamic ultrasound workflows. To develop a general method, we search for a single prompt template to evaluate across all our experiments. The optimal phrasing and terms are selected via combinatorial search across our validation datasets (70%/30% validation/test partitions, split patient-wise where possible) using candidate descriptive templates, and ultrasound quality descriptors gathered from ultrasound image quality literature [1, 21].

We also investigate averaging across multiple prompts in an embedding sweep, with the sweep determining the optimal number and combination of negative prompts. We find that the best performing configuration uses two negatives, appearing to approximate the region of low-quality images better than any individual prompt, $\mathbf{d}_a = f_T(t_a^+) - \frac{1}{2}\left(f_T(t_1^-) + f_T(t_2^-)\right)$.

The top-performing prompts are:

$t_a^+ =$ ‘‘`A well-focused ultrasound showing {anatomy} parenchyma with clear echotexture`’’
$t_1^- =$ ‘‘`A technically inadequate ultrasound`’’
$t_2^- =$ ‘‘`A poorly optimised ultrasound with significant artifacts`’’

Our prompts are deliberately asymmetric. The positive prompt t_a^+ entangles anatomical and quality semantics, whereas the negatives only describe quality failures. We expect that a high-quality image will always depict the OOI clearly, whereas in low-quality images the OOI can be poorly depicted. This departs from the antonym-pair construction of CLIP-IQA [19], in which t^+ and t^- describe opposing ends of a single quality axis (e.g. “Good photo” / “Bad photo”).

3 Experiments

Baseline Methods: To evaluate general IQA methods with no anatomical sensitivity, we adopt NIQE [14], BRISQUE [13], and CLIP-IQA [19]. CLIP-IQA is a similar embedding-based baseline which relies on natural language prompting [19]. To strengthen this baseline, we perform prompt tuning and evaluate both CLIP and BiomedCLIP as potential backbones, selecting the strongest antonym pair and model. We find that “Well-visualized Image” and “Poorly-visualized Image” are the best performing antonym pair when combined with BiomedCLIP (details in supplementary material). To evaluate the performance of organ-specific IQA methods, we adopt the TinyUSFM-NRQ [7] family of models. They train a series of IQA models based on the TinyUSFM foundation model [12]. We apply two versions of the method: the matched variant applies the correct model for the task (e.g. the liver variant for the liver images), whilst the mixture variant applies eleven bespoke models and averages their scores. To evaluate the performance of open-source VLMs, we adopt Qwen3.6-27B [15]. We chose a 27B variant because it is at the upper limit of what is practically deployable on a single dedicated GPU in an ultrasound machine to satisfy latency

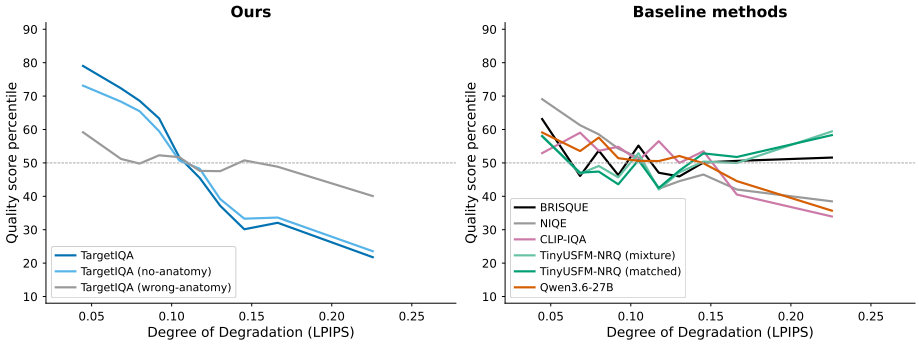

Fig. 2. TargetIQA shows the strongest and most monotonic sensitivity to image degradation, while anatomically insensitive baselines are weaker. Scores were computed on original and synthetically degraded ultrasound images (shadow, haze, speckle, attenuation, excessive gain) across three organs. For each image pair, the change in score was ranked within organ to remove inter-organ differences, then macro-averaged across organs and binned into degradation severity deciles.

and privacy requirements. At inference time, Qwen3.6-27B [15] receives a simple prompt describing the task with the same anatomical information given to TargetIQA (details provided in supplementary material).

Ablations: We also evaluate two ablated versions of TargetIQA: "no-anatomy", which removes the anatomical context from our prompts; and "wrong-anatomy", which replaces the correct anatomy with a different one at random (of the set explored in our experiments). For instance, if the standard prompt is "A well-focused ultrasound showing liver parenchyma with clear echotexture", then the prompt for the "no-anatomy" ablation would be: "A well-focused ultrasound with clear echotexture", and the prompt for the "wrong-anatomy" ablation might be: "A well-focused ultrasound showing kidney parenchyma with clear echotexture".

Experiment 1 – CAMUS Clinical Quality Classification: To evaluate the clinical alignment of each method, we classify the clinical utility of echocardiography images annotated by cardiologists. The CAMUS echocardiography dataset [10] contains images labelled as "Good" (n=1010) and "Medium" (n=753) which were clinically useable, and "Poor" (n=232) which were sub-clinical quality.

Experiment 2 – Synthetic Degradation Correlation: To evaluate the sensitivity of each method, and estimate their utility for gradient-based optimisation tasks, we precisely control the quality of images by applying synthetic degradation to high-quality images. To quantify the difference between the original and degraded images, we use LPIPS [22] – a measure of perceptual similarity. We assume that the change in an IQA metric before and after synthetic degradation

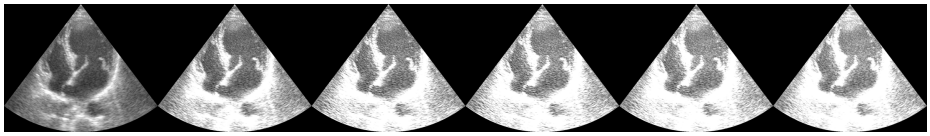

Fig. 3. Examples of progressively increasing synthetic degradation. The strength of our "excessive gain" and "speckle" augmentations are varied from zero (original image) up to their maximum strength. Original image sampled from the CAMUS dataset [10].

should be strongly correlated with the degree of perceptual difference between the original, high-quality image and the degraded version. To estimate the utility of each method for gradient-based optimisation, we are interested in the monotonicity and proportionality of the method's scoring. We use Spearman's Rank Correlation Coefficient (SRCC) to measure monotonicity, and Pearson's Linear Correlation Coefficient (PLCC) to measure proportionality. To synthetically degrade the images we use USAugment [17], a Python library which implements several ultrasound-specific image augmentation techniques. Each augmentation method simulates some undesirable change to the image. We use the standard depth attenuation, gaussian shadow, and haze artifact augmentations, and we adapt some of their methods to simulate increased intensity of speckle noise and excessive gain settings (see supplementary material for details and examples). We select 75 images labelled as high-quality, split equally across liver, kidney, and heart datasets. Fifty synthetic versions of each high-quality image are generated with randomly varied levels of degradation giving 3750 images total. Each augmentation is applied with severity randomly sampled from USAugment's default ranges. Figure 3 shows a range of synthetic degradation with further examples.

Experiment 3 – Paired Quality Assessment: To evaluate subtle differences in image quality, each method ranks pairs of patient-matched images captured on expensive hospital-grade ultrasound machines and inexpensive point-of-care ultrasound machines (POCUS). POCUS devices have inherent limitations due to their low cost and size which generally reduce the image quality [5, 9]. We use the 2023 USEnhance challenge dataset [18] which involved generating the high-quality ultrasound images from the low-quality ones. The dataset contains 1050 paired images across liver, kidney, breast, carotid artery, and thyroid anatomies. Each IQA method is evaluated by its accuracy in ranking the quality of the images in each pair – the score for the high-quality image should be highest.

4 Results and Discussion

Experiment 1: This experiment tests whether each method's quality scores agree with clinical utility labels on the CAMUS [10] dataset. Table 1 reports AUROC on the CAMUS Good+Medium vs Poor classification task. TargetIQA achieves 0.839, a margin of 0.137 AUROC over the strongest baseline (Qwen3.6-27B [15] at 0.702) and 0.191 over the matched TinyUSFM-NRQ variant (0.648).

Table 1. Exp1: CAMUS cardiac quality classification, AUROC (Good+Medium vs. Poor). **Exp2**: synthetic degradation sensitivity, Spearman's Rank Correlation Coefficient (SRCC) and Pearson's Linear Correlation Coefficient (PLCC) between score delta and LPIPS, macro-averaged over organs. **Exp3**: paired quality ranking, macro-averaged paired accuracy over 5 organs. The +/− values denote 95% bootstrap CI (clustered on patients for Exp1 and on original images for Exp2). **Bold**: best; underline: next best.

	Experiment 1	**Experiment 2**		**Experiment 3**
Method	**AUROC** (↑)	**SRCC** (↑)	**PLCC** (↑)	**Macro PA %** (↑)
General (no anatomy)				
BRISQUE [13]	0.583 $^{+0.080}_{-0.080}$	0.041 $^{+0.092}_{-0.114}$	0.111 $^{+0.100}_{-0.103}$	69.6 $^{+5.1}_{-5.5}$
NIQE [14]	0.578 $^{+0.095}_{-0.099}$	0.287 $^{+0.118}_{-0.060}$	0.333 $^{+0.099}_{-0.068}$	<u>83.4</u> $^{+3.5}_{-3.8}$
CLIP-IQA [19]	0.606 $^{+0.076}_{-0.085}$	0.214 $^{+0.139}_{-0.104}$	0.252 $^{+0.148}_{-0.099}$	69.0 $^{+4.7}_{-4.8}$
TinyUSFM-NRQ (mixture) [7]	0.619 $^{+0.081}_{-0.085}$	-0.038 $^{+0.069}_{-0.071}$	0.050 $^{+0.083}_{-0.090}$	44.7 $^{+5.2}_{-5.6}$
Organ-specific				
TinyUSFM-NRQ (matched) [7]	0.648 $^{+0.075}_{-0.082}$	-0.054 $^{+0.058}_{-0.066}$	0.073 $^{+0.076}_{-0.096}$	45.0 $^{+5.8}_{-5.7}$
Qwen3.6-27B [15]	0.702 $^{+0.047}_{-0.050}$	0.203 $^{+0.094}_{-0.077}$	0.204 $^{+0.105}_{-0.071}$	38.9 $^{+5.4}_{-5.1}$
Ours				
TargetIQA	**0.839** $^{+0.045}_{-0.049}$	**0.645** $^{+0.067}_{-0.061}$	**0.633** $^{+0.067}_{-0.055}$	**84.4** $^{+3.6}_{-3.8}$
TargetIQA (no-anatomy)	<u>0.824</u> $^{+0.049}_{-0.059}$	<u>0.559</u> $^{+0.105}_{-0.090}$	<u>0.530</u> $^{+0.118}_{-0.073}$	72.8 $^{+4.9}_{-4.5}$
TargetIQA (wrong-anatomy)	0.584 $^{+0.075}_{-0.080}$	0.118 $^{+0.031}_{-0.036}$	0.120 $^{+0.033}_{-0.039}$	70.5 $^{+5.2}_{-5.3}$

BRISQUE and NIQE sit at 0.583 and 0.578 respectively, just above chance. The ablation ordering across TargetIQA (0.839), no-anatomy (0.824) and wrong-anatomy (0.584) shows that the anatomical signal does not drastically improve the method for this task but the wrong anatomical signal significantly degrades performance. Echocardiography images are captured with standard views where the heart anatomy is distinct and clear. This may explain the overlapping confidence intervals of TargetIQA and the "no-anatomy" ablation: the standard views used in echocardiography fill most of the frame with relevant cardiac tissue so the anatomical prompt has little effect. Its contribution should be larger in regions such as the abdomen, where the anatomy of interest may be surrounded by more background tissue or other irrelevant organs, and the anatomical prompt specifies which regions of the image are relevant.

Experiment 2: This experiment tests whether each method's score responds monotonically and proportionally to controlled synthetic degradation across three organs. Table 1 reports SRCC and PLCC between degradation severity (LPIPS) and the change in quality score, macro-averaged across anatomies. TargetIQA achieves the strongest correlations at 0.645 SRCC and 0.633 PLCC, followed by our no-anatomy ablation. Figure 2 shows that TargetIQA's score is more sensitive to the degree of degradation than other methods, many of which begin to invert at higher levels of degradation. The two general-purpose baselines (BRISQUE, NIQE) and the trained TinyUSFM-NRQ variants are all weak on both metrics, indicating poor response to varied degradation. The same ablation ordering as the first experiment holds: TargetIQA (0.645/0.633) beats no-anatomy (0.559/0.530) which beats wrong-anatomy (0.118/0.120), with the

wrong-anatomy variant approaching no correlation with degradation at all. The poor performance of TinyUSFM-NRQ [7] may be the result of a mismatch between its clean training data and our significantly degraded evaluation data.

Experiment 3: This experiment tests whether each method can correctly rank patient-matched ultrasound pairs by device quality across five organs. Table 1 reports paired ranking accuracy (PA) on the USEnhance dataset. TargetIQA achieves the highest macro PA at 84.4%, followed closely by NIQE at 83.4%, with overlapping 95% confidence intervals. The ablation ordering in this experiment remains the same but the differences are smaller than in the other experiments. TargetIQA (84.4%) leads, but no-anatomy (72.8%) and wrong-anatomy (70.5%) have very similar performance with overlapping 95% confidence intervals. NIQE's strong performance on this experiment compared to its weak performance on the other experiments likely shows better sensitivity to image variations caused by machine characteristics. TargetIQA's strong performance here continues to show no clear weakness across any of the experiments.

Limitations: TargetIQA establishes anatomical conditioning as an effective basis for ultrasound IQA, and several natural extensions remain open: we only explore anatomical descriptions of task, but other types of desirable image characteristics are a natural extension; a recent CLIP-style model [8] trained to be sensitive to a specific ontology of other task characteristics may enable formulations such as "vascular echogenic lesions on the liver" rather than just "liver". We only evaluate static image quality due to the availability of annotated datasets, but metric stability during continuous scans should be evaluated too – particularly in situations with patient or probe motion.

5 Conclusion

This work introduces TargetIQA, a zero-shot ultrasound IQA method that conditions image quality on the operator's anatomical intent at scan time. Across three tasks – clinical quality classification, sensitivity to synthetic degradation, and paired ranking – TargetIQA performs strongly and consistently (AUROC 0.839, SRCC 0.645, and 84.4% paired accuracy respectively) while every other method suffers a clear weakness on one of the tasks. TargetIQA requires no training, yet outperforms an organ-specific baseline (TinyUSFM-NRQ [7]) and a VLM with 138× more parameters (Qwen3.6-27B [15]). If embedded in an ultrasound scanner, an anatomy-conditioned IQA method like TargetIQA could shorten scans and reduce operator variability.

Acknowledgments. This work was supported by the UKRI EPSRC Centre for Doctoral Training in Applied Photonics [EP/Y035437/1]

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

Bibliography

[1] Adams, R.B.: Ultrasound scanning techniques. Surgery Open Science **10**, 182–207 (2022). https://doi.org/10.1016/j.sopen.2022.09.002

[2] Annangi, P., Ravishankar, H., Patil, R., Tore, B., Aase, S.A., Steen, E.: AI assisted feedback system for transmit parameter optimization in Cardiac Ultrasound. In: 2020 IEEE International Ultrasonics Symposium (IUS). pp. 1–4 (2020). https://doi.org/10.1109/IUS46767.2020.9251501

[3] Baum, Z.M.C., Bonmati, E., Cristoni, L., Walden, A., Prados, F., Kanber, B., Barratt, D.C., Hawkes, D.J., Parker, G.J.M., Wheeler-Kingshott, C.A.M.G., Hu, Y.: Image quality assessment for closed-loop computer-assisted lung ultrasound. In: Linte, C.A., Siewerdsen, J.H. (eds.) Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling. vol. 11598, p. 115980R. SPIE (2021). https://doi.org/10.1117/12.2581865

[4] Chen, H., Wu, L., Dou, Q., Qin, J., Li, S., Cheng, J.Z., Ni, D., Heng, P.A.: Ultrasound standard plane detection using a composite neural network framework. IEEE Transactions on Cybernetics **47**(6), 1576–1586 (2017). https://doi.org/10.1109/TCYB.2017.2685080

[5] Haider, S.J.A., diFlorio Alexander, R., Lam, D.H., Cho, J.Y., Sohn, J.H., Harris, R.: Prospective Comparison of Diagnostic Accuracy Between Point-of-Care and Conventional Ultrasound in a General Diagnostic Department: Implications for Resource-Limited Settings. Journal of Ultrasound in Medicine **36**(7), 1453–1460 (2017). https://doi.org/10.7863/ultra.16.06084

[6] He, D., Wang, H., Yaqub, M.: Advancing Fetal Ultrasound Image Quality Assessment in Low-Resource Settings (2025), arXiv:2507.22802 [cs]

[7] Huang, Z., Li, B., Ma, C., Liu, T., Zhai, Y., Xu, H., Guo, Y., Li, Z., Wang, Y.: Defining Robust Ultrasound Quality Metrics via an Ultrasound Foundation Model (2026), arXiv:2604.19512 [eess.IV]

[8] Jin, J., Chai, H., Huang, X., Guo, X., Zheng, Z., Zhou, Z., Wang, J., Wang, X., Liu, J., Zhou, B.: Ultrasound-CLIP: Semantic-Aware Contrastive Pre-training for Ultrasound Image-Text Understanding (2026), arXiv:2604.01749 [cs.CV]

[9] Karnenbeek, L.M.v., Pol, H.G.A.v.d., Wijkhuizen, M., Poelman, E., Drukker, C.A., Ruers, T., Geldof, F., Dashtbozorg, B.: A Paired Point-of-Care Ultrasound Dataset for Image Quality Enhancement and Benchmarking via a cGAN Baseline (2026), arXiv:2605.08282 [eess.IV]

[10] Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., Lartizien, C., D'hooge, J., Lovstakken, L., Bernard, O.: Deep Learning for Segmentation using an Open Large-Scale Dataset in 2D Echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (2019). https://doi.org/10.1109/TMI.2019.2900516

[11] Loizou, C.P., Murray, V., Pattichis, M.S., Pantziaris, M., Nicolaides, A.N., Pattichis, C.S.: Despeckle Filtering for Multiscale Amplitude-Modulation Frequency-Modulation (AM-FM) Texture Analysis of Ultrasound Images of the Intima-Media Complex. International Journal of Biomedical Imaging **2014**, 518414 (2014). https://doi.org/10.1155/2014/518414

[12] Ma, C., Jiao, J., Liang, S., Fu, J., Wang, Q., Li, Z., Wang, Y., Guo, Y.: TinyUSFM: Towards Compact and Efficient Ultrasound Foundation Models (2025), arXiv:2510.19239 [eess]

[13] Mittal, A., Moorthy, A.K., Bovik, A.C.: No-Reference Image Quality Assessment in the Spatial Domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012). https://doi.org/10.1109/TIP.2012.2214050

[14] Mittal, A., Soundararajan, R., Bovik, A.: Making a "Completely Blind" Image Quality Analyzer. Signal Processing Letters, IEEE **20**, 209–212 (2013). https://doi.org/10.1109/LSP.2012.2227726

[15] Qwen Team: Qwen3.6-27B: Flagship-level coding in a 27B dense model. https://qwen.ai/blog?id=qwen3.6-27b (Apr 2026), accessed: 2026-07-02

[16] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision (2021), arXiv:2103.00020 [cs]

[17] Tupper, A., Gagné, C.: Revisiting Data Augmentation for Ultrasound Images (2025), arXiv:2501.13193 [eess]

[18] Ultrasound Image Enhancement Challenge Organizers: Ultrasound image enhancement challenge 2023 (USenhance). https://ultrasoundenhance2023.grand-challenge.org/ (2023), accessed: 2026-07-02

[19] Wang, J., Chan, K.C.K., Loy, C.C.: Exploring CLIP for Assessing the Look and Feel of Images (2022), arXiv:2207.12396 [cs.CV]

[20] Wu, L., Cheng, J.Z., Li, S., Lei, B., Wang, T., Ni, D.: FUIQA: Fetal Ultrasound Image Quality Assessment With Deep Convolutional Networks. IEEE transactions on cybernetics **47**(5), 1336–1349 (2017). https://doi.org/10.1109/TCYB.2017.2671898

[21] Zander, D., Hüske, S., Hoffmann, B., Cui, X.W., Dong, Y., Lim, A., Jenssen, C., Löwe, A., Koch, J.B., Dietrich, C.F.: Ultrasound Image Optimization ("Knobology"): B-Mode. Ultrasound International Open **6**(1), E14–E24 (2020). https://doi.org/10.1055/a-1223-1134

[22] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric (2018), arXiv:1801.03924 [cs.CV]

[23] Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M.P., Naumann, T., Wang, S., Poon, H.: BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs (2025), arXiv:2303.00915 [cs]