This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

UltraPIPS: Improving model perception in B-mode ultrasound with foundation models

Tal Grutman and Tali Ilovitsh

School of Biomedical Engineering, Tel Aviv University, Tel Aviv, Israel
ilovitsh@tauex.tau.ac.il

Abstract. In medical imaging, it is common to use learned perceptual image patch similarity (LPIPS) to compare images semantically in feature space. Although backbones pretrained on natural images are widely used for LPIPS computation, B-mode ultrasound images possess distinct speckle patterns and acoustic-specific image statistics that are fundamentally different from natural images and even from other images in radiology. Consequently, we propose that domain-specific models are needed to measure perceptual similarity in ultrasound data, a finding which is not necessarily the case for other imaging modalities. We compare deep feature distances across downstream tasks like classification, segmentation and reconstruction using natural image, medical generalist and ultrasound backbone models and show that backbone selection is a non-trivial design choice. In particular, the ultrasound backbone models were more correlated with downstream performance of supervised models than classical and natural image models, and optimization of a deep feature distance loss with an ultrasound backbone achieved the best reconstruction and texture statistics. Our code is available at https://github.com/talg2324/UltraPIPS and introduces the UltraPIPS library, a set of perceptual distance metrics based on the open-source foundation models analyzed in this paper.

Keywords: Ultrasound · Foundation models · Perceptual similarity · Computer vision

1 Introduction

Learned perceptual image patch similarity (LPIPS) [30] is an important perceptual metric that uses a pretrained image encoder to measure distance between image patches in the encoder's feature space. The resulting metric is grounded by the downstream task on which the encoder was trained, with the deep features extracted by the metric incorporating rich context with which to perceive the difference between patches. LPIPS is highly correlated to downstream classification and detection tasks, suggesting that it is a soft middle-ground between human perception and performance of deep models. One reason for this is that the LPIPS backbone models were pretrained in self-supervised or supervised learning tasks that require global context to be embedded in the feature space.

This provides a natural improvement from metrics like standard L_2 or structural similarity index measure (SSIM) [28]. Since it was first introduced, loss functions and evaluation criteria based on LPIPS have become well-established in image synthesis [22], implicit neural representations [4], and image reconstruction [12,24]. LPIPS is widely used in computer vision, and has been extended into medical imaging [5,2] often using the RadImageNet [15] backbone and dropping the learned calibration to serve as a deep feature distance.

There is no consensus on the importance of the backbone model to the effectiveness of the LPIPS metric, and the choice of backbone is not well studied in medical imaging. Backbone choice is conceptually related to transfer learning, where several studies found that in-domain training or pretraining added marginal performance or even performed worse than models pretrained on ImageNet [29,32]. More specifically in the context of deep feature distances, it has been shown that in-domain backbones did not improve MRI reconstruction [1] compared to ImageNet ones. These findings challenge the necessity of in-domain pretraining, but other notable works have found that in-domain medical image models learn vastly different features with much higher relevance in comparison to ImageNet-based models [15,21].

Some of the discrepancies in results can be attributed to the various domains within medical imaging, each possessing highly unique spatial resolutions, dynamic range, and signal-to-noise ratio. Ultrasound is well-known to be an especially difficult imaging domain, and foundation models trained exclusively on ultrasound images have been shown to outperform general-purpose radiology models [5,16] on a variety of tasks. Deep feature distance is particularly important in ultrasound models because deep backbones can accommodate the unique speckle characteristics common to B-mode ultrasound, which are quickly lost with pixel-wise losses like MSE [12] despite carrying important diagnostic meaning [27]. Thus, LPIPS and deep feature distance are widely used for training and evaluation of reconstruction-oriented ultrasound models [4,12,24]. We hypothesize that ultrasound reconstruction tasks like these can be improved by foundation models built for ultrasound images to power their distance metrics, as they are better accustomed to the unique properties of ultrasound.

In this work, we investigate the importance of ultrasound foundation models in perceptual similarity of B-mode ultrasound images compared to models trained on natural or general medical images. We utilize open-source, pretrained models and datasets freely available on the web to challenge the usage of non-ultrasound backbones in quantitative image comparison, using the relationship of deep backbone features and downstream model performance in classification, segmentation, and image reconstruction as a proxy for perceptual similarity. We test feasible candidates for LPIPS metrics in ultrasound and open-source the UltraPIPS library which contains multiple backbones and tests for perceptual ultrasound metrics.

2 Experiments

Several cross-domain encoder candidates exist for an ultrasound perceptual metric. Models like RadImageNet [15] and MedSAM [14] which were trained on large-scale medical image datasets (including but not exclusive to ultrasound) could potentially provide useful features based on their deep understanding of anatomy. This is especially true for MedSAM, which was trained specifically to segment diagnostically important information.

Alternatively, ultrasound foundation models can provide an ultrasound-first metric that can accommodate for the discrepancies between ultrasound and other modalities. To produce a useful model for perceptual metrics, such foundation models must be trained on a wide variety of organs with exposure to various imaging acquisition parameters. Several open-source candidates exist. Ultrasound Foundation Model (USFM) [8] is a ViT-b encoder that was trained in a masked auto-encoding framework. Texture ultrasound semantic analysis (TUSA) [5] used a learned dictionary of cross-anatomy texture kernels to reconstruct images, producing a SwinViT [13] backbone grounded in B-mode image texture characteristics. Ultrasound-CLIP [9] is an ultrasound foundation model trained to extract global semantic features from ultrasound images based on corresponding text reports. These models are attractive candidates for a deep feature backbone, and we hypothesize that with increased training grounded in the properties of ultrasound, they will outperform natural image or radiology models as perceptual metrics. We are not aware of any modern CNN-based ultrasound foundation model, limiting our analysis to ViT and SwinViT variants. We analyzed ImageNet variants of ViT and SwinViT to control for the impact of training data on these architectures. Similarly, we added the original CLIP [20] and the medical generalist BiomedCLIP [31] models to our analysis to isolate the effect of the training data compared to Ultrasound-CLIP and training strategy compared to self-supervised and classification models. L_2 and SSIM serve as classical metric baselines; candidate backbones are grouped by pre-training domain in table 1.

Table 1. Candidate backbone models grouped by pre-training domain.

Natural Image	Radiology	Ultrasound
AlexNet	RadImageNet	USFM
VGG-16	MedSAM	TUSA
ViT	BiomedCLIP	Ultrasound-CLIP
SwinViT		
CLIP		

The MONAI [3] library was used to calculate the RadImageNet LPIPS, and the original LPIPS [30] implementation was used for AlexNet and VGG-16. For ViT/SwinViT variants, features were extracted at the final layer to produce a

feature vector, in line with the LPIPS and MONAI implementations. The deep feature distance is the L_2 loss between channel-normalized feature vectors of two images.

Experiments were run on a Kubernetes cluster with the Run:ai (Tel-Aviv, Israel) resource manager on an NVIDIA RTX A5000 GPU (NVIDIA Corporation, Santa Clara, CA, USA) using PyTorch [18].

2.1 Supervised downstream tasks

We used the EchoGains [25] library to induce ultrasound-based augmentations on echocardiogram B-mode images. This unique and recent augmentation strategy presents a realistic scenario for image degradation in clinical context. EchoGains achieves this by performing geometric warping of ultrasound images, masking out pixels outside the original transducer field-of-view, and using a diffusion model to inpaint the missing pixels in the ultrasound sector. It offers augmentations of the imaging depth, sector width, probe translation and rotation. Since EchoGains is unique to cardiac data and requires left ventricle segmentation masks, we used CAMUS [11] and EchoNet [17] to analyze supervised tasks.

Cardiac View Classification The coarse view classifier published alongside EchoPrime [26] was evaluated on each echo dataset. The log-softmax output of the model at the correct view (either apical four chamber or apical two chamber) was used to isolate the model's confidence as a function of image perturbation.

Cardiac Chamber Segmentation UltraSam [16] was used to segment the left ventricle and left atrium with increasing deformation. Since UltraSam is a segment anything model [10], it relies on a bounding box prompt to segment images. This makes UltraSam far more robust to image degradations, and is more aligned with the strong human capacity for perception of images as they undergo degradation. Bounding boxes were extracted from geometrically warped true labels and provided to UltraSam.

2.2 Image Reconstruction

We trained implicit neural representations (INRs) for the USOVA 3D Follicles [19] and MicroSegNet Prostate [7] datasets. Since INR models are typically optimized with perceptual objectives in addition to pixel-wise losses, they are a natural setting for comparison of backbones. We train INRs with SIREN activations [23], the Adam optimizer, and an initial learning rate of 10^{-3} annealed to 10^{-6} over the course of 300 epochs according to a cosine schedule. Each INR had a layer width of 256 with six intermediate SIREN layers, and skip connections at every even intermediate layer. A final tanh activation mapped the output to [-1, 1]. INRs were trained for each of the 16 follicle volumes and 75 prostate volumes using L_2 loss with an additional equally weighted candidate LPIPS metric. We elected to exclude MedSAM from this analysis, as its input 1024^2 resolution is

impractical to use as a loss function for such tasks. After training, SSIM [28] was calculated on each frame of the volume to measure the reconstruction quality of the perceptual metric. To emphasize speckle accuracy, we computed the high-frequency error norm $\mathrm{HFEN}(x, \hat{x}) = \|\mathrm{LoG}(x) - \mathrm{LoG}(\hat{x})\|_2$ with $\sigma = 2.5$ [1], where $\mathrm{LoG}(x)$ is the Laplacian of Gaussian filter with standard deviation σ.

Although SSIM and HFEN are useful image reconstruction metrics for measuring the recovery of structure by a trained INR, they do not measure how realistic is a reconstructed image, which is especially notable in ultrasound reconstruction due to the unique distribution of speckles. To isolate these effects, we also computed the gray-level co-occurrence matrix (GLCM) [6] to quantify texture realism. Five standard features were extracted from a symmetric GLCM computed at distances of 1 and 3 and angles 0°–135° with 32 gray levels: contrast, dissimilarity, homogeneity, energy, and correlation. GLCM features were used to measure the realism of reconstructed images by calculating the L_1 distance between the reference GLCM features and the reconstructed ones.

2.3 Results

Supervised tasks

Classification Spearman correlation and standard deviation of the various distance metrics to downstream classifier confidence is shown in Table S1 for the various deformations in our study, calculated across the axis of degradation per-frame and averaged across datasets. The three CNN models in our dataset (AlexNet, VGG-16, RadImageNet) all had high correlation to downstream confidence, outperforming classical metrics and several of the ViT models. However, the highest correlations were achieved by ultrasound models, which were more correlated to downstream confidence than other ViT variants. Among CLIP models, BiomedCLIP and UltrasoundCLIP both outperformed CLIP, suggesting that exposure to ultrasound images may improve correlation. Surprisingly, MedSAM and USFM, both ViT-b variants, were among the least correlated models and were even outperformed by the ImageNet ViT.

Segmentation Spearman correlation and standard deviation of the various distance metrics to downstream dice score is shown in Table S2 for the various deformations in our study, calculated across the axis of degradation per-frame and averaged across datasets. Interestingly, all models saw marginal but clear improvement in correlation to rotation relative to classification, but failed entirely to correlate when tested on depth deformations. This is because both CAMUS and EchoNet are within UltraSam's training data, and small perturbation of depth creates downsized heart chambers that are still within the distribution of the dataset. Other augmentations like rotation and sector width immediately caused significant drops in UltraSam's dice score leading to higher correlation with perceptual metrics.

Major trends noted in classification remained the same in segmentation, with TUSA and Ultrasound-CLIP scoring the highest correlations followed by CNN

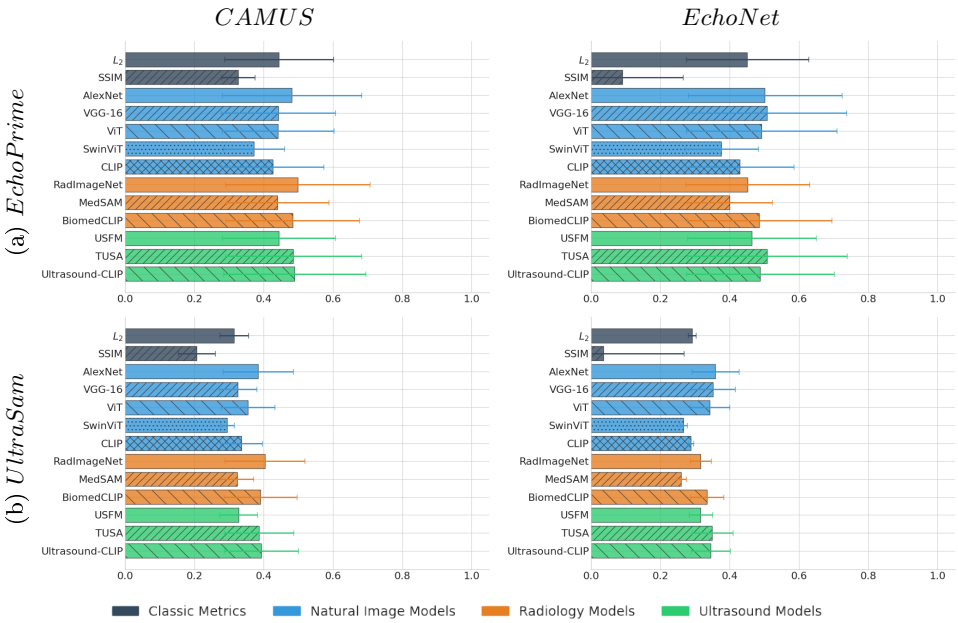

Fig. 1. Spearman correlations of supervised models on CAMUS and EchoNet datasets under ultrasound augmentation. In the view classification task (a) correlation is calculated between the log-softmax confidence score of EchoPrime's view classifier and each distance metric, averaged across augmentations and frames. For segmentation (b), the correlation is calculated between UltraSam's dice score and each of the distance metrics, averaged across augmentations and frames.

models, and BiomedCLIP. The weak correlation of MedSAM to segmentation is particularly notable because it is a segmentation model and was fine-tuned from the same source model as UltraSam. The failure of MedSAM to correlate with the performance of such a similar model may indicate that its training data is heavily biased towards other radiology applications, or that the image embedding alone does not contain any indication of uncertainty, which may be induced by cross-attention with prompts provided to SAM models.

Overall distance metrics and standard deviations averaged across the various augmentations are shown for each model on each supervised task and dataset in fig. 1. CNNs or models with high exposure to ultrasound data tended to outperform supervised alternatives, with the exception of MedSAM and USFM. Only a handful of models consistently beat L_2 loss, indicating that not all backbones are useful compared to classical metrics. SSIM was particularly noisy, as it sometimes correlated negatively with downstream metrics.

To provide a statistical analysis of the different backbone groups, we evaluate the null hypothesis that ultrasound foundation model metrics are no more correlated to downstream performance than each of the other metric groups using

Table 2. Wilcoxon test of deep feature distance computed with ultrasound models vs. comparison groups.

Null Hypothesis	Classification	Segmentation
$\rho_{\text{US}} \leq \rho_{\text{Classical}}$	7.8×10^{-3}	7.8×10^{-3}
$\rho_{\text{US}} \leq \rho_{\text{Natural Images}}$	0.016	0.039
$\rho_{\text{US}} \leq \rho_{\text{Radiology}}$	0.078	0.078

Table 3. Wilcoxon signed-rank test p-values of deep feature distance vs. comparison groups for INR reconstruction metrics.

			GLCM Texture Features				
Null Hypothesis	**SSIM**	**HFEN**	**Contrast**	**Dissimilarity**	**Homogeneity**	**Energy**	**Correlation**
$\rho_{\text{US}} \leq \rho_{\text{Classical}}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	2.1×10^{-7}	$< 10^{-12}$
$\rho_{\text{US}} \leq \rho_{\text{Natural Images}}$	$< 10^{-12}$	$< 10^{-12}$	8.6×10^{-4}	2.5×10^{-8}	7.4×10^{-3}	$< 10^{-12}$	1.7×10^{-9}
$\rho_{\text{US}} \leq \rho_{\text{Radiology}}$	$< 10^{-12}$	$< 10^{-12}$	2.2×10^{-7}	0.132	5.1×10^{-11}	0.260	2.1×10^{-6}

a Wilcoxon signed-rank test, and report the p-values in table 2. We reject the null hypothesis for classical metrics and LPIPS with natural image models, but not for radiology models. One explanation for this is that although ultrasound is outnumbered in radiology datasets, radiology models still benefit from some degree of exposure to ultrasound data. Nonetheless, it is clear that our perceptual metrics do not uniformly correlate with downstream model perception, suggesting that backbone choice may be an important implementation choice.

INR Reconstruction A sample image is shown from each distance metric (with the exclusion of MedSAM) in fig. 2. L_2 loss provides faithful image restoration but loses speckle quality, as models converge to the low-frequency solution. This effect is largely overcome by SSIM, which forces the INR to include high detail information. Distance metrics from natural and medical generalist backbones induce hallucination of fine details, bending the tissue between the two largest follicles, distorting acoustic shadow, or failing to reconstruct the image altogether. All of the ultrasound backbones faithfully reconstruct the input image with various levels of detail. USFM and Ultrasound-CLIP generate highly realistic speckle detail reminiscent of the SSIM reconstruction baseline.

The null hypothesis that ultrasound foundation model metrics outperform other metrics is evaluated individually for each reconstruction task in table 3. Ultrasound models convincingly outperform all other groups in SSIM, HFEN, and GLCM texture statistics, failing only to achieve significance on GLCM dissimilarity and energy compared to radiology models. Per-model results from the two 3D datasets are shown in Table S3, showing that this is caused by an outlier ultrasound model underperforming on these specific metrics. Medical backbones score lower in reconstruction but higher in texture statistics. This effect is visible in fig. 2 where models like BiomedCLIP and RadImageNet that fail to reconstruct the anatomy still create meaningful texture. Ultrasound backbones

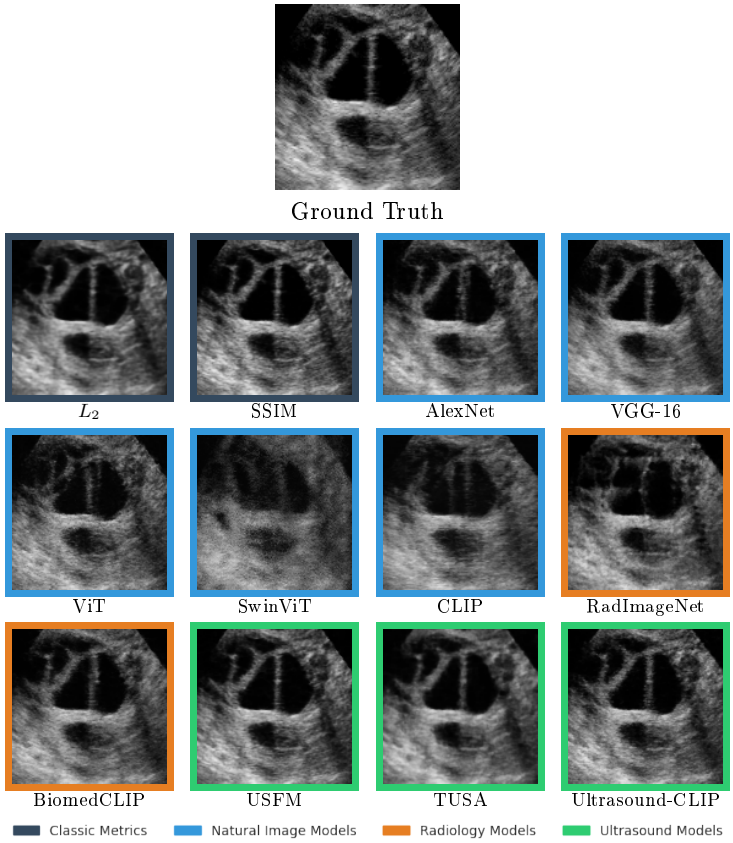

Fig. 2. Implicit Neural Representation (INR) visual comparison on the follicles dataset. Models are grouped and color-coded by their pre-training domain: Classical (dark blue), Natural Images (light blue), Radiology (orange), and Ultrasound (green).

provide a balance between reconstruction quality and texture, suggesting that they may be strong candidates for deep features in reconstruction tasks.

3 Conclusion

In this paper, we studied the importance of ultrasound models in perceptual metrics for B-mode ultrasound images. We found that in supervised tasks, CNN backbones were highly correlated with downstream performance, but the highest correlations were achieved by ViT variants with exposure to ultrasound images. This was especially notable in CLIP models that were all trained in the same framework with the same architecture, but had increasing improvement in both downstream correlation and reconstruction with additional exposure to ultrasound images. In image reconstruction, ultrasound models provided a balance

between reconstruction quality and realistic texture, and were the only backbones we tested that did not distort anatomic details upon qualitative inspection. This work is a first step in challenging the widespread usage of arbitrary LPIPS backbones, but it is important to note that correlation to model performance is a proxy for visual perception, and further work is needed to quantify the relationship between human perception of B-mode images and the various backbones studied here. Additional work can be done on adding more datasets and downstream models to be tested, but this is currently limited by the need for realistic augmentations, which do not meaningfully exist outside of echocardiography. Nonetheless, our findings suggest that exposure to ultrasound images may improve the suitability of models to serve as deep feature backbones, especially in image reconstruction.

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

References

1. Adamson, P.M., Desai, A.D., Dominic, J., Varma, M., Bluethgen, C., Wood, J.P., Syed, A.B., Boutin, R.D., Stevens, K.J., Vasanawala, S., Pauly, J.M., Gunel, B., Chaudhari, A.S.: Using deep feature distances for evaluating the perceptual quality of MR image reconstructions. Magnetic Resonance in Medicine **94**(1), 317–330 (2025)
2. Cahan, N., Klang, E., Aviram, G., Barash, Y., Konen, E., Giryes, R., Greenspan, H.: X-ray2ctpa: leveraging diffusion models to enhance pulmonary embolism classification. npj Digital Medicine **8**(439) (2025)
3. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Darestani, M.Z., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B.S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P.F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L.A.D., Roth, H.R., Xu, D., Bericat, D., Floca, R., Zhou, S.K., Shuaib, H., Farahani, K., Maier-Hein, K.H., Aylward, S., Dogra, P., Ourselin, S., Feng, A.: Monai: An open-source framework for deep learning in healthcare (2022), https://arxiv.org/abs/2211.02701
4. Grutman, T., Bismuth, M., Glickstein, B., Ilovitsh, T.: Implicit neural representation for scalable 3d reconstruction from sparse ultrasound images. npj Acoustics **1**(14), 1–11 (2025)
5. Grutman, T., Shinar, C., Ilovitsh, T.: A texture-based framework for foundational ultrasound models (2026), https://arxiv.org/abs/2602.01444
6. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics **SMC-3**(6), 610–621 (1973). https://doi.org/10.1109/TSMC.1973.4309314
7. Jiang, H., Imran, M., Muralidharan, P., Patel, A., Pensa, J., Liang, M., Benidir, T., Grajo, J.R., Joseph, J.P., Terry, R., et al.: MicroSegNet: A deep learning approach for prostate segmentation on micro-ultrasound images. Computerized Medical Imaging and Graphics **112**, 102326 (2024)

8. Jiao, J., Zhou, J., Li, X., Xia, M., Huang, Y., Huang, L., Wang, N., Zhang, X., Zhou, S., Wang, Y., Guo, Y.: USFM: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Medical Image Analysis **96**, 103202 (2024)
9. 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. In: CVPR (2026). https://doi.org/10.48550/arXiv.2604.01749, https://arxiv.org/abs/2604.01749
10. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023), https://arxiv.org/abs/2304.02643
11. Leclerc, S., Smistad, E., Pedrosa, J., Ostvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P., Grenier, T., Løvstakken, 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
12. Li, X., Navab, N., Jiang, Z.: Speckle2self: Self-supervised ultrasound speckle reduction without clean data. Medical Image Analysis p. 103755 (2025)
13. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
14. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
15. Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, T.M., Yang, Y., et al.: Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence **4**(5), e210315 (2022)
16. Meyer, A., Murali, A., Zarin, F., Mutter, D., Padoy, N.: UltraSAM: a foundation model for ultrasound using large open-access segmentation datasets. International Journal of Computer Assisted Radiology and Surgery **20**(1), 1–10 (2025)
17. Ouyang, D., He, B., Ghorbani, A., Lungren, M.P., Ashley, E.A., Liang, D.H., Zou, J.Y.: Video-based ai for beat-to-beat assessment of cardiac function. Nature **580**(7802), 252–256 (2020). https://doi.org/10.1038/s41586-020-2145-8
18. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library (2019), https://arxiv.org/abs/1912.01703
19. Potočnik, B., Munda, J., Reljič, M., Rakić, K., Knez, J., Vlaisavljević, V., Sedej, G., Cigale, B., Holobar, A., Zazula, D.: Public database for validation of follicle detection algorithms on 3d ultrasound images of ovaries. Computer Methods and Programs in Biomedicine **196**, 105621 (2020)
20. 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. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), https://proceedings.mlr.press/v139/radford21a.html
21. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. In: NeurIPS (2019)
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)

23. Sitzmann, V., Martel, J.N.P., Bergman, A.W., Lindell, D.B., Wetzstein, G.: Implicit neural representations with periodic activation functions (2020), https://arxiv.org/abs/2006.09661
24. Stevens, T.S.W., Nolan, O., Somphone, O., Robert, J.L., van Sloun, R.J.G.: High volume rate 3d ultrasound reconstruction with diffusion models. IEEE Transactions on Medical Imaging (2025). https://doi.org/10.1109/TMI.2025.3645849
25. Van De Vyver, G., Lenz, A.T., Smistad, E., Olaisen, S.H., Grenne, B., Holte, E., Dalen, H., Løvstakken, L.: Generative augmentations for improved cardiac ultrasound segmentation using diffusion models (Feb 2025), https://arxiv.org/abs/2502.20100, arXiv:2502.20100
26. Vukadinovic, M., Tang, X., Yuan, N., Cheng, P., Li, D., Cheng, S., He, B., Ouyang, D.: Echoprime: A multi-video view-informed vision-language model for comprehensive echocardiography interpretation (2024), https://arxiv.org/abs/2410.09704
27. Wagner, R.F., Smith, S.W., Sandke, J.M., Lopez, R.H.: Statistics of speckle in ultrasound B-scans. IEEE Transactions on Sonics and Ultrasonics **30**(3), 156–163 (1983)
28. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
29. Yuan, H., Zhu, M., Yang, R., Liu, H., Li, I., Hong, C.: Rethinking domain-specific pretraining by supervised or self-supervised learning for chest radiograph classification: A comparative study against ImageNet counterparts in cold-start active learning. Health Care Science **4**(2), 110–143 (2025)
30. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
31. 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 (2023). https://doi.org/10.48550/arXiv.2303.00915, https://arxiv.org/abs/2303.00915
32. Zu, W., Xie, S., Chen, H., Ma, L.: Pre-trained models succeed in medical imaging with representation similarity degradation. arXiv preprint arXiv:2503.07958 (2025)