CausCLIP: Causality-Adapting Visual Scoring of Visual Language Models for Few-Shot Learning in Portable Echocardiography Quality Assessment

Yiran Li[1], Xiaoxiao Cui[2], Yankun Cao[2], Yuezhong Zhang[3] Huihui Wang[3], Lizhen Cui[2] (✉), Zhi Liu[1] (✉), and Shuo Li[4]

[1] School of Information Science and Engineering, Shandong University, Qingdao, Shandong, China
[2] Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR), Shandong University, Jinan, Shandong, China
{clz,liuzhi}@sdu.edu.cn
[3] Shandong Provincial Hospital, Jinan, shandong, China
[4] Case Western Reserve University, Cleveland, OH, USA

Abstract. How do we transfer Vision Language Models (VLMs), pre-trained in the source domain of conventional echocardiography (Echo), to the target domain of few-shot portable Echo (fine-tuning)? Learning image causality is crucial for few-shot learning in portable echocardiography quality assessment (PEQA), due to the domain-invariant causal and topological consistency. However, the lack of significant domain shifts and well-labeled data in PEQA present challenges to get reliable measurements of image causality. We investigate the challenging problem of this task, i.e., learning a consistent representation of domain-invariant causal semantic features. We propose a novel VLMs based PEQA network, Causality-Adapting Visual Scoring CLIP (CausCLIP), embedding causal diposition to measure image causality for domain-invariant representation. Specifically, Causal-Aware Visual Adapter (CVA) identifies hidden asymmetric causal relationships and learns interpretable domain-invariant causal semantic consistency, thereby improving adaptability. Visual-Consistency Contrastive Learning (VCL) focuses on the most discriminative regions by registing visual-causal similarity, enhancing discriminability. Multi-granular Image-Text Adaptive Constraints (MAC) adaptively integrate task-specific semantic multi-granular information, enhancing robustness in multi-task learning. Experimental results show that CausCLIP outperforms state-of-the-art methods, achieving absolute improvements of 4.1%, 9.5%, and 8.5% in view category, quality score, and distortion metrics, respectively.

Keywords: Automated quality assessment · Causal learning · Pre-trained visual language models · Portable echocardiography · Transfer learning.

1 Introduction

How do we transfer Vision Language Models (VLMs), pre-trained in the source domain of conventional echocardiography (Echo), to the target domain of few-

shot portable Echo (fine-tuning)? Learning image causality is crucial for few-shot learning (FSL) in portable Echo quality assessment (PEQA). PEQA faces challenges due to significant domain shifts and limited well-labeled data. Transfer learning methods, represented by VLMs, becomes a natural choice. However, it faces a key question: Can PEQA benefit from specific knowledge during the transfer learning? Therefore, finding a causality-driven domain transfer method is crucial for PEQA [8, 7, 13]. Once successful, it will bring great adaptability of pre-trained networks, and robust transferring for downstream task [10].

Causal disposition [5] has significant potential in domain transfer, yet it still faces challenges in the learning of image causality in PEQA. 1) Domain mismatches hinder the capture of causal consistency and the adaptation of out-of-distribution data. In Echo imaging, especially with portable devices, images from different vendors exhibit significantly distinct features. And domain shift from natural to Echo images diminishes effectiveness of VLMs when applied directly. Therefore, it is crucial to find an efficient method that bridges domain mismatches and possesses superior domain transfer capabilities for understanding causal consistency in Echo. 2) Hidden visual-causal similarities, challenging to articulate textually, frequently overlooked by VLMs. The deformable appearance and poor spatial resolution of Echo make it difficult to identifying discriminative visual-causal features via common graphical attributes in VLMs. Especially in FSL, the overlooked visual information often proves critical. 3) Weak correspondence between causal visual-language distribution cannot be accurately

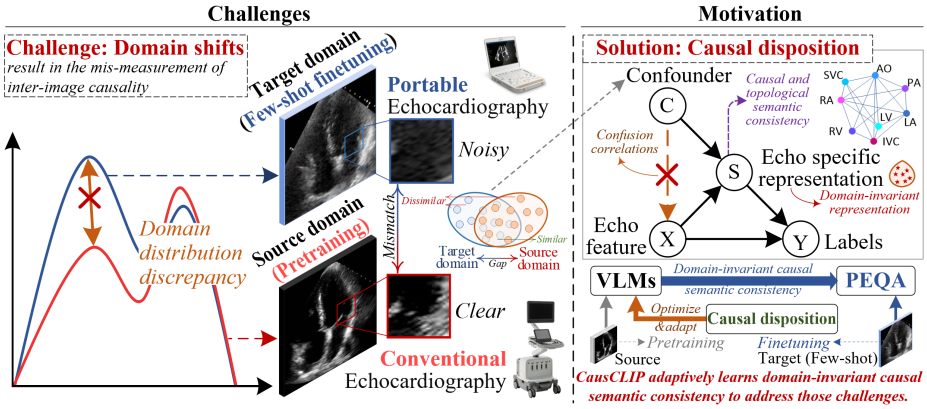

Fig. 1. How do we transfer Vision Language Models (VLMs), pre-trained in the conventional echocardiography (Echo) source domain, to the portable Echo target domain (fine-tuning)? Our CausCLIP achieves domain transfer by mastering domain invariance, focusing on domain-invariant causal semantic consistency in portable Echo quality assessment (PEQA). Confounder C dictates the relationships among objects in Echo-views but fails to identify key features, causing confusion in $X \rightarrow Y$. M represents the Echo specific representation of C. CausCLIP eliminates C, clarifying the causal path from $X \rightarrow Y$.

represented in multi-task learning (MTL), thereby hindering knowledge transfer in FSL. VLMs struggle to grasp and depict the complex semantics of Echo. Weak correspondences hinder generalization and sensitivity to task and feature variations when capturing semantic consistency from limited labeled data during domain transfer.

To sum up the above limitations, the mis-measurement for image causality is the key challenge for FSL in PEQA, interfering the discovery of domain invariance. Cardiac anatomical structure invariance ensures topological semantic consistency in Echo. Therefore, deriving and amplifying causality from the interaction between statistical features and domain-invariant topological semantic consistency enables the achievement of domain-invariant causal semantic consistency, which is crucial for understanding image causality.

In this paper, we propose a novel Causality-Adapting Visual Scoring (Caus) of contrastive language-image pre-training (CLIP) [9], aimed at enhancing PEQA. It addresses the above limitations via three key innovations: 1) Causal-Aware Visual Adapter (CVA) learns asymmetric causal relationships in weak causal signals to adapt domain mismatches. By building on the extensive pre-trained knowledge from CLIP, it learns interpretable and adaptable domain-invariant causal consistency. 2) Visual-consistency contrastive learning (VCL) learns the visual semantic registration for hidden visual-causal similarity. By extracting detailed visual information relevant to PEQA, VCL concentrates on the most discriminative regions. 3) Multi-granular Image-Text Adaptive Constraints (MAC) adaptively learn the weak correspondence between causal semantic information and text distribution, facilitating effective visual grouping in MTL. By providing multi-level textual information with multi-granular anchors, MAC enhances the understanding of complex semantic information and improves the balance in MTL, effectively utilizing auxiliary data.

Our contributions are summarized as follows: 1) For the first time, our CausCLIP advances the learning of image causality in PEQA and promotes domain-invariant causal semantic consistency, enhancing domain-invariant representability in VLMs. 2) For the first time, our novel CVA integrates causal learning into adaptive learning of VLMs, effectively addressing domain mismatches and thus improving adaptability. 3) Our novel VCL effectively reduces the registration error of visual-causal similarity, improving discriminability. 4) Our powerful MAC achieves collaborative optimization of multi-granular visual grouping, enhancing the robustness of knowledge transfer.

2 Causality-Adapting Visual Scoring CLIP (CausCLIP)

Preliminaries Given a set of portable Echo $\mathcal{E} \in \mathbb{R}^H$, image $x \sim \mathcal{E}$ are randomly sampled from $\mathcal{Q}$. The goal of CausClip $\nu : \mathbb{R}^H \to \mathbb{R}^Q$ is to predict the quality of x, aiming for it to approximate the quality assessment result $\mathcal{Y} \in \mathbb{R}^Q$. We use conventional Echo to construct source domain dataset $\mathcal{S} \in \mathbb{R}^S$ and fine-tune the CLIP model. We develop prompts for PEQA based on three key dimensions: classification of Echo views, assessment of quality scores, and identification of

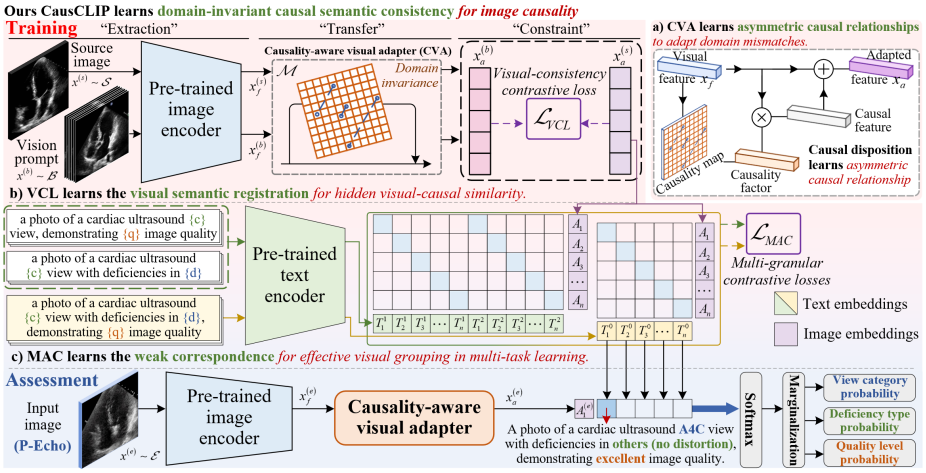

Fig. 2. The framework of our CausCLIP. Our CausCLIP learns the domain-invariant causal semantic consistency in PEQA, thus driving the learning of image causality via the gradient in backpropagation.

ultrasound distortion. We consider echocardiographic 7-views: $c \in \mathcal{C} = \{$'PLAX', 'A2C', 'A3C', 'A4C', 'A5C', PSAX-PM', 'PSAX-AV'$\}$. We selected 7 views from conventional Echo to compose the benchmark set $\mathcal{B} \in \mathbb{R}^B$. We also work with five quality levels: $q \in \mathcal{Q} = \{$'excellent', 'good', 'fair', 'poor', 'bad'$\}$. An Echo may exhibit multiple types of quality distortions, we focus on on identifying the predominant ones: $d \in \mathcal{D} = \{$'depth-gain', 'chamber clarity', 'zoom', 'offset', 'chamber integrity','other'$\}$. The 'others' category includes images with no distortions. The ultimate prompt integrates the aforementioned three specified metrics: a photo of a(n) cardiac ultrasound $\{c\}$ view with deficiencies in $\{d\}$, demonstrating $\{q\}$ image quality. For example: a photo of a cardiac ultrasound **PLAX** view with deficiencies in **offset**, demonstrating **bad** image quality.

Causal disposition formulation: By counting the number $C(A, B)$ of images in which the causal dispositions of artifacts A and B is such that B disappears if one removes A, one can assume that the artifact A causes the presence of artifact B when $C(A, B)$ is greater than the converse $C(B, A)$ [1, 12]. We can infer that any causal disposition leads to asymmetric causal relationships among features, representing weak causality signals about images. The interaction between causality and domain-invariant topological consistency enables generalization to new distributions.

Causal-aware visual adapter (CVA) The CVA adaptively reveals hidden asymmetric causal relationships to learn interpretable causal features from visual (x_f) and textual (x_t) features. Subsequently, CVA utilizes the causal-aware module $\mathcal{M}$ to concentrate on weak causal signals in x_f, identifying asymmetric causal relationships and adaptively weighting features. The ReLU operations

guarantees the feature maps F contain only non-negative numbers and normalize these numbers to the interval $[0, 1]$ by dividing of each them to the maximal possible value $MAX(F)$. Subsequently, F is fed into a process that computes pairwise conditional probabilities, we get k features $F^1, F^2, \ldots, F^k$ represented by $n \times n$ feature maps and generates $k \times k$ causality map, thereby representing the probability of a feature appearing at a specific location. Specifically, for a pair of features maps F^i and F^j, we connect the conditional probability with the joint probability $P\left(F^i \mid F^j\right) = \frac{P\left(F^i, F^j\right)}{P(F^j)}$. And, we apply the generalized average function (Lehmer means) to estimate the conditional probabilities between features pairs:

$$P\left(F^i \mid F^j\right)_\alpha = \frac{LM_\alpha\left(F^i \times F^j\right)}{LM_\alpha\left(F^j\right)} \tag{1}$$

where $F^i \times F^j = \left\{F_{11}^i \cdot F_{11}^j, F_{11}^i \cdot F_{12}^j, \ldots, F_{11}^i \cdot F_{nn}^j, \ldots, F_{nn}^i \cdot F_{nn}^j\right\}$ is the vector of pairwise multiplications between two $n \times n$ feature maps. LM_α is the Lehmer means with trainable parameter α, which can produce values spanning from the minimum to the maximum, across a simple average among the operands of a vector $LM_\alpha(x) = \frac{\sum_{k=1}^{n} x_k^{\alpha+1}}{\sum_{k=1}^{n} x_k^{\alpha}}$. Formular 1 could be used to estimate asymmetric causal relationships between F^i and F^j, where typically, $P\left(F^i \mid F^j\right) \neq P\left(F^j \mid F^i\right)$. By estimating pairwise causal relationships between quantities for every pair i and j of the k feature maps, we obtain the $k \times k$ causality map. Causality maps are flattened and concatenated with flattened feature maps, enabling the adapter to learn their impact on PEQA. To prevent information loss, the adapter employs the adaptive residual connection to integrate causal features with the original visual features x_v.

$$x_a = \lambda \mathcal{M}\left(x_v\right) + (1 - \lambda) x_v \tag{2}$$

where x_a is the adapted features and λ is the weighting parameter.

Summary advantage: Our CVA proposes a novel causal representation framework with high adaptability and causal preservation ability. It models the weak causal signals and utilizes asymmetric causal relationship for cross-domain alignment. Therefore, it has effectively improved the domain-invariant causal consistency with the limitation of domain mismatches.

Visual-consistency contrastive learning (VCL) To uncover hidden visual-causal similarities, we utilize VCL to register the causal-adapted feature x_a, ensuring it retains more discriminative information from visual-semantic features. Constructing vision prompts enables the model to precisely comprehend complex visual-semantic relationships that are challenging to articulate textually. To maximize the cosine similarity between x_a and the positive vision prompt and minimize it between x_a and the negative vision prompt, we formulate the visual-consistency contrastive loss:

$$\mathcal{L}_{VCL} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(sim\left(x_a, x_v^{+}\right)/\tau\right)}{\exp\left(sim\left(x_a, x_v^{(j)}\right)/\tau\right)} \tag{3}$$

where N is the number of samples in the batch. x_v^{+} is the embedding of the positive vision prompt and $x_v^{(j)}$ represents the negative ones. $sim(\cdot)$ denotes the cosine similarity between two embeddings. τ is a temperature scaling parameter.

Summary advantage: Our VCL proposes a novel registration framework with high discriminability. It utilized the the visual semantic registration and densely samples their visual correlation. Therefore, it has effective improve the visual-causal similarity.

Multi-granular image-text adaptive constraints (MAC) To enhance the correspondence between causal semantic information and text distribution in MTL, MAC sets multi-granular anchors from the original prompt based on three key metrics - category, quality, and distortion - and creates additional text prompts while dynamically adjusting weights in the multi-granular loss function. The choice of multi-granularity aims to enable as precise an image-text correspondence as possible for the specified task. In addition to the original prompt, we apply multi-granular contrastive losses among the additional prompts, such as $\mathcal{L}_{ap}^{(cd)}$, $\mathcal{L}_{ap}^{(cq)}$, and $\mathcal{L}_{ap}^{(og)}$, and employ uncertainty weighting [4] to adaptively adjust the weights for each prompt, ensuring alignment with causality:

$$\mathcal{L}_{MAC} = \sum \frac{1}{\boldsymbol{\sigma}^2} \odot \boldsymbol{L} + \log \Pi\boldsymbol{\sigma} \tag{4}$$

where $\boldsymbol{L} = \left[\mathcal{L}_{ap}^{(cd)}, \mathcal{L}_{ap}^{(cq)}, \mathcal{L}_{ap}^{(og)}\right]$ is loss vectors, $\boldsymbol{\sigma} = [\sigma_{cd}, \sigma_{cs}, \sigma_{cd}]$ is uncertainty weight vector, $\odot$ represents the element-wise multiplication of vectors, Π represents the product of vector elements.

Summary advantage: Our MAC proposes a novel registration-based text generation program with higher diversity. It utilizes the multi-granular anchor to precisely align task features with specific multi-level text. Therefore, it enhances the robustness of PEQA.

3 Experiments

Data Collection We collect a dataset (P-Echo) comprising 5146 portable echocardiographic images across 7-views for few-shot learning: parasternal long-axis (PLAX), apical two-chamber (A2C), three-chamber (A3C), four-chamber (A4C), five-chamber (A5C), papillary muscle level (PSAX-PM), and parasternal short-axis at the aortic valve level (PSAX-AV). P-Echo comprises completely anonymized data from three community health check-ups, with the dataset split into 350 images for few-shot training and 4796 images for testing.And we utilized a dataset comprising 16,572 conventional echocardiography images (C-Echo) of

Table 1. The quantitative evaluation demonstrates the superiority of our CausCLIP. Our CausCLIP achieves the highest performance in PEQA.

Model	PLCC	SRCC	KRCC	ACC_{view}	$ACC_{quality}$	ACC_{dists}
KonCept	$0.727_{\pm 0.051}$	$0.719_{\pm 0.029}$	$0.701_{\pm 0.027}$	–	$0.623_{\pm 0.063}$	–
HyperIQA	$0.835_{\pm 0.007}$	$0.829_{\pm 0.003}$	$0.820_{\pm 0.024}$	–	$0.716_{\pm 0.012}$	–
TRIQ	$0.822_{\pm 0.017}$	$0.821_{\pm 0.033}$	$0.785_{\pm 0.016}$	$0.866_{\pm 0.016}$	$0.707_{\pm 0.021}$	$0.607_{\pm 0.038}$
IQT	$0.831_{\pm 0.040}$	$0.827_{\pm 0.039}$	$0.816_{\pm 0.025}$	$0.879_{\pm 0.027}$	$0.739_{\pm 0.069}$	$0.632_{\pm 0.059}$
MEON	$0.867_{\pm 0.024}$	$0.869_{\pm 0.013}$	$0.833_{\pm 0.010}$	$0.887_{\pm 0.009}$	$0.763_{\pm 0.037}$	$0.675_{\pm 0.036}$
LIQE	$0.907_{\pm 0.003}$	$0.905_{\pm 0.002}$	$0.877_{\pm 0.003}$	$0.916_{\pm 0.015}$	$0.791_{\pm 0.029}$	$0.752_{\pm 0.022}$
CausCLIP	$0.921_{\pm 0.020}$	$0.922_{\pm 0.007}$	$0.893_{\pm 0.015}$	$0.957_{\pm 0.011}$	$0.886_{\pm 0.033}$	$0.837_{\pm 0.027}$

varying quality during the CLIP fine-tuning stage, gathered from routine ultrasound examinations at two hospitals. The above datasets show significant style differences and have been annotated by two expert sonographers.

Experimental Settings We employ the pre-trained CLIP (ViT-B/32) for image and text encoding, implemented in Pytorch on an NVIDIA GeForce RTX 3080. The model is pre-trained on the C-Echo dataset for 100 epochs with a mini-batch size of 4. For the training stage, we employ the Adam optimizer with a base learning rate of 5×10^{-6}, set the weight decay to 0.005, and use five-fold cross-validation. In the assessment, we utilize the Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-Order Correlation Coefficient (SRCC), Kendall Rank-Order Correlation Coefficient (KRCC), and Accuracy (ACC).

Comparison with state-of-the art We compare CausCLIP with KonCept [3], HyperIQA [11], TRIQ [14], IQT [2], MEON [6], and LIQE [15], using publicly available implementations. Competing models are retrained on our datasets using the training codes provided by their respective authors. In the TRIQ and IQT, we added [CLASS] token and [DIST] token to their transformer architectures for MTL analysis. Compared to other quality assessment models, CausCLIP shows superior performance in three key aspects (Table 1): 1) Efficient transfer of image domain invariance. CausCLIP achieved superior performance in all experiments, surpassing the next best method, LIQE, with average accuracy improvements of 4.1%, 9.5%, and 8.5% in view categories, quality, and distortion metrics, respectively. This success is due to its capability to learn image causality, enhancing domain-invariant causal semantic consistency. 2) Effective transfer of causal relationships. Qualitatively, we generated Grad-CAM heatmaps to highlight regions significantly influencing the results. The heatmaps reveal distinct patterns between the base CLIP and CausCLIP (Figure 3). The former primarily focuses on smaller central areas of the chamber structures, whereas the CausCLIP encompasses a broader A4C structure, covering both interior and edge tissues of the chambers. This demonstrates that CausCLIP excels in locating key areas, offering more interpretable features for decision-making. 3) Superiority in auxiliary knowledge of VLMs. In all experiments, VLM-based

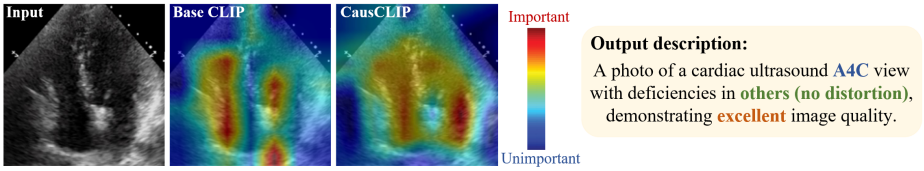

Fig. 3. Impact of causality on the echocardiography. From left to right, the columns represent the input test images, the Grad-CAM activations for the base CLIP (not causality-adaption), the Grad-CAM activations for CausCLIP. The yellow part is the description of the input test images.

Table 2. Ablation results on CausCLIP variants demonstrate the great contributions of our innovation. We evaluate the zero-shot and few-show performance simultaneously.

	CVA	VCL	MAC	ACC_{view}	$ACC_{quality}$	ACC_{dists}
Zero-shot	√	√	√	$0.899_{\pm 0.023}$	$0.734_{\pm 0.016}$	$0.693_{\pm 0.039}$
Few-shot				$0.927_{\pm 0.036}$	$0.816_{\pm 0.005}$	$0.764_{\pm 0.017}$
	√			$0.946_{\pm 0.012}$	$0.863_{\pm 0.007}$	$0.787_{\pm 0.009}$
		√	√	$0.952_{\pm 0.010}$	$0.859_{\pm 0.023}$	$0.792_{\pm 0.013}$
	√	√		$0.950_{\pm 0.006}$	$0.869_{\pm 0.011}$	$0.812_{\pm 0.021}$

models showed significantly superior performance, with CausCLIP achieving an average precision increase of at least 12.3% compared to other quality assessment methods. This superior performance is attributed to VLM-based methods enhancing the network's generalization ability through pre-trained auxiliary knowledge. The improvements in PLCC, SRCC, and KRCC metrics further confirm VLMs' capability to learn superior auxiliary knowledge from visual-language correspondences.

Ablation study The ablation studies on CausCLIP variants demonstrate the great improvement of our innovations (Tabel 2). We explore the impact of the number of training samples on CausCLIP's generalization performance. Without portable Echo samples in the training process (zero-shot), we observe a decline in all three accuracy metrics. Additionally, we examined the impact of individual components within CausCLIP. The results indicate that removing any component leads to decreased performance and reduced generalization ability. Significantly, the causal module demonstrates heightened sensitivity to quality accuracy, attributable to its ability to focus on more discriminative features.

4 Conclusion

In this paper, we advance image causality learning for FSL in PEQA and introduce Causality-Adapting Visual Scoring CLIP (CausCLIP). This approach significantly enhances the representation of image causality, achieving powerful domain-invariant representation for domain-invariant causal semantic con-

sistency. The proposed CVA learns asymmetric causal relationships, thereby enhancing domain-invariant causal consistency with adaptability. Our VCL enhances visual-causal similarity, improving the visual semantic discriminability. Our MAC constrains multi-granular causal visual-text information, thereby enhancing robustness. Extensive experiments achieving state-of-the-art results on the MTL task showcase the powerful performance of our CausCLIP in few-shot quality assessment. We believe CausCLIP will advance the field of causal learning in medical image analysis.

Acknowledgments. This work was supported in part by Joint fund for smart computing of Shandong Natural Science Foundation under Grant ZR2020LZH013; the Major Scientific and Technological Innovation Project in Shandong Province under Grant 2021CXGC010506 and 2022CXGC010504; "New Universities 20 items" Funding Project of Jinan under Grant 2021GXRC108; Shandong Provincial Natural Science Foundation under Grant ZR2022LZH007; Qingdao key technology research and industrialization-Future Industry Cultivation Special Project 22-3-4-xxgg5-nsh.

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

References

1. Carloni, G., Pachetti, E., Colantonio, S.: Causality-driven one-shot learning for prostate cancer grading from mri. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 2616–2624 (October 2023)
2. Cheon, M., Yoon, S.J., Kang, B., Lee, J.: Perceptual image quality assessment with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 433–442 (June 2021)
3. Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Transactions on Image Processing **29**, 4041–4056 (2020). https://doi.org/10.1109/TIP.2020.2967829
4. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
5. Lopez-Paz, D., Nishihara, R., Chintala, S., Scholkopf, B., Bottou, L.: Discovering causal signals in images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6979–6987 (2017)
6. Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., Zuo, W.: End-to-end blind image quality assessment using deep neural networks. IEEE Transactions on Image Processing **27**(3), 1202–1213 (2018). https://doi.org/10.1109/TIP.2017.2774045
7. Poudel, K., Dhakal, M., Bhandari, P., Adhikari, R., Thapaliya, S., Khanal, B.: Exploring Transfer Learning in Medical Image Segmentation using Vision-Language Models. arXiv e-prints arXiv:2308.07706 (Aug 2023). https://doi.org/10.48550/arXiv.2308.07706
8. Qin, Z., Yi, H.H., Lao, Q., Li, K.: MEDICAL IMAGE UNDERSTANDING WITH PRETRAINED VISION LANGUAGE MODELS: A COMPREHENSIVE STUDY. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=txlWziuCE5W

9. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
10. Shrestha, P., Amgain, S., Khanal, B., Linte, C.A., Bhattarai, B.: Medical Vision Language Pretraining: A survey. arXiv e-prints arXiv:2312.06224 (Dec 2023). https://doi.org/10.48550/arXiv.2312.06224
11. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
12. Terziyan, V., Vitko, O.: Causality-aware convolutional neural networks for advanced image classification and generation. Procedia Computer Science **217**, 495–506 (2023). https://doi.org/https://doi.org/10.1016/j.procs.2022.12.245, 4th International Conference on Industry 4.0 and Smart Manufacturing
13. Wang, J., Liu, Z., Zhao, L., Wu, Z., Ma, C., Yu, S., Dai, H., Yang, Q., Liu, Y., Zhang, S., Shi, E., Pan, Y., Zhang, T., Zhu, D., Li, X., Jiang, X., Ge, B., Yuan, Y., Shen, D., Liu, T., Zhang, S.: Review of large vision models and visual prompt engineering. Meta-Radiology **1**(3), 100047 (2023). https://doi.org/https://doi.org/10.1016/j.metrad.2023.100047
14. You, J., Korhonen, J.: Transformer for image quality assessment. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 1389–1393 (2021). https://doi.org/10.1109/ICIP42928.2021.9506075
15. Zhang, W., Zhai, G., Wei, Y., Yang, X., Ma, K.: Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14071–14081 (June 2023)