# Regressing Simulation to Real: Unsupervised Domain Adaptation for Automated Quality Assessment in Transoesophageal Echocardiography

Jialang Xu[1](✉), Yueming Jin[4], Bruce Martin[2], Andrew Smith[2], Susan Wright[3], Danail Stoyanov[1], and Evangelos B. Mazomenos[1](✉)

[1] UCL Wellcome/EPSRC Centre for Interventional and Surgical Sciences,
Department of Medical Physics and Biomedical Engineering,
University College London, London, UK
{jialang.xu.22,e.mazomenos}@ucl.ac.uk
[2] St Bartholomew's Hospital, NHS Foundation Trust, London, UK
[3] St George's University Hospitals, NHS Foundation Trust, London, UK
[4] Department of Biomedical Engineering and Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore

**Abstract.** Automated quality assessment (AQA) in transoesophageal echocardiography (TEE) contributes to accurate diagnosis and echocardiographers' training, providing direct feedback for the development of dexterous skills. However, prior works only perform AQA on simulated TEE data due to the scarcity of real data, which lacks applicability in the real world. Considering the cost and limitations of collecting TEE data from real cases, exploiting the readily available simulated data for AQA in real-world TEE is desired. In this paper, we construct the first simulation-to-real TEE dataset, and propose a novel Simulation-to-Real network (SR-AQA) with unsupervised domain adaptation for this problem. It is based on uncertainty-aware feature stylization (UFS), incorporating style consistency learning (SCL) and task-specific learning (TL), to achieve high generalizability. Concretely, UFS estimates the uncertainty of feature statistics in the real domain and diversifies simulated images with style variants extracted from the real images, alleviating the domain gap. We enforce SCL and TL across different real-stylized variants to learn domain-invariant and task-specific representations. Experimental results demonstrate that our SR-AQA outperforms state-of-the-art methods with 3.02% and 4.37% performance gain in two AQA regression tasks, by using only 10% unlabelled real data. Our code and dataset are available at https://doi.org/10.5522/04/23699736.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43996-4_15.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 154–164, 2023.
https://doi.org/10.1007/978-3-031-43996-4_15

**Keywords:** Automated quality assessment · Unsupervised domain adaptation regression · Transoesophageal echocardiography · Uncertainty · Consistency learning · Style transfer

## 1 Introduction

Transoesophageal echocardiography (TEE) is a valuable diagnostic and monitoring imaging modality with widespread use in cardiovascular surgery for anaesthesia management and outcome assessment, as well as in emergency and intensive care medicine. The quality of TEE views is important for diagnosis and professional organisations publish guidelines for performing TEE exams [5,22]. These guidelines standardise TEE view acquisition and set benchmarks for the education of new echocardiographers. Computational methods for automated quality assessment (AQA) will have great impact, guaranteeing quality of examinations and facilitating training of new TEE operators.

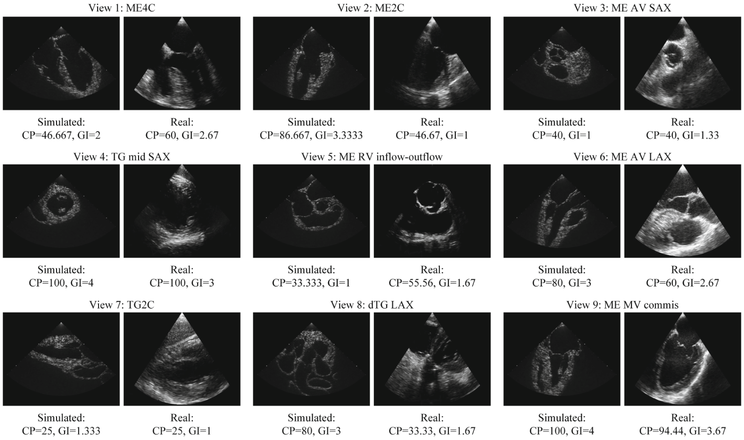

**Fig. 1.** Overview of our proposed dataset. The 9 TEE views in our dataset: 1: Mid-Esophageal 4-Chamber, 2: Mid-Esophageal 2-Chamber, 3: Mid-Esophageal Aortic Valve Short-Axis, 4: Transgastric Mid-Short-Axis, 5: Mid-Esophageal Right Ventricle inflow-outflow, 6: Mid-Esophageal Aortic Valve Long-Axis, 7: Transgastric 2-Chamber, 8: Deep Transgastric Long-Axis, 9: Mid-Esophageal Mitral Commissural

Deep models for AQA have shown promise in transthoracic echocardiography (TTE) and other ultrasound (US) applications [1,9,13,14]. Investigation of such methods in the real TEE domain remains underexplored and has been restricted to simulated datasets from Virtual Reality (VR) systems [15]. Although VR technology is useful for developing and retaining US scanning skills [7,10,16,17,20], AQA methods developed on simulation settings cannot meet real-world usability without addressing the significant domain gap. As shown in Fig. 1, there are significant content differences between simulated and real TOE images. Simulated images are free of speckle noise and contain only the heart muscle, ignoring tissue in the periphery of the heart. In this work, we take the first step in exploring AQA in the real TEE domain. We propose to leverage readily accessible simulated data, and transfer knowledge learned in the simulated domain, to boost performance in real TEE space.

To alleviate the domain mismatch, a feasible solution is unsupervised domain adaptation (UDA). UDA aims to increase the performance on the target domain by using labelled source data with unlabelled target data to reduce the domain shift. For example, Mixstyle [24], performs style regularization by mixing instance-level feature statistics of training samples. The most relevant UDA work for our AQA regression task is representation subspace distance (RSD) [4], which aligns features from simulated and real domains via representation sub-spaces. Despite its effectiveness in several tasks, performing UDA on the simulation-to-real AQA task of TEE has two key challenges that need to be addressed. From Fig. 1, it is evident that: 1) there are many unknown *intra-domain* shifts in real TEE images due to different scanning views and complex heart anatomy, which requires uncertainty estimation; 2) the *inter-domain* gap (heart appearance, style, and resolution) between simulated and real data is considerable, necessitating robust, domain-invariant features.

In this paper, we propose a novel UDA regression network named SR-AQA that performs style alignment between TEE simulated and real domains while retaining domain-invariant and task-specific information to achieve promising AQA performance. To estimate the uncertainty of *intra-domain* style offsets in real data, we employ uncertainty-aware feature stylization (UFS) utilizing multivariate Gaussians to regenerate feature statistics (i.e. mean and standard deviation) of real data. To reduce the *inter-domain gap*, UFS augments simulated features to resemble diverse real styles and obtain real-stylized variants. We then design a style consistency learning (SCL) strategy to learn domain-invariant representations by minimizing the negative cosine similarity between simulated features and real-stylized variants in an extra feature space. Enforcing task-specific learning (TL) in real-stylized variants allows SR-AQA to keep task-specific information useful for AQA. Our work represents the original effort to address the TEE domain shift in AQA tasks. For method evaluation, we present the *first* simulation-to-real TEE dataset with two AQA tasks (see Fig. 1), and benchmark the proposed SR-AQA model against four state-of-the-art UDA methods. Our proposed SR-AQA outperforms other UDA methods, achieving 2.13%–5.08% and 4.37%–16.28% mean squared error (MSE) reduction for two AQA tasks, respectively.

## 2 Methodology

### 2.1 Dataset Collection

We collected a dataset of 16,192 simulated and 4,427 real TEE images from 9 standard views. From Fig. 1, it is clear that significant style differences (e.g.

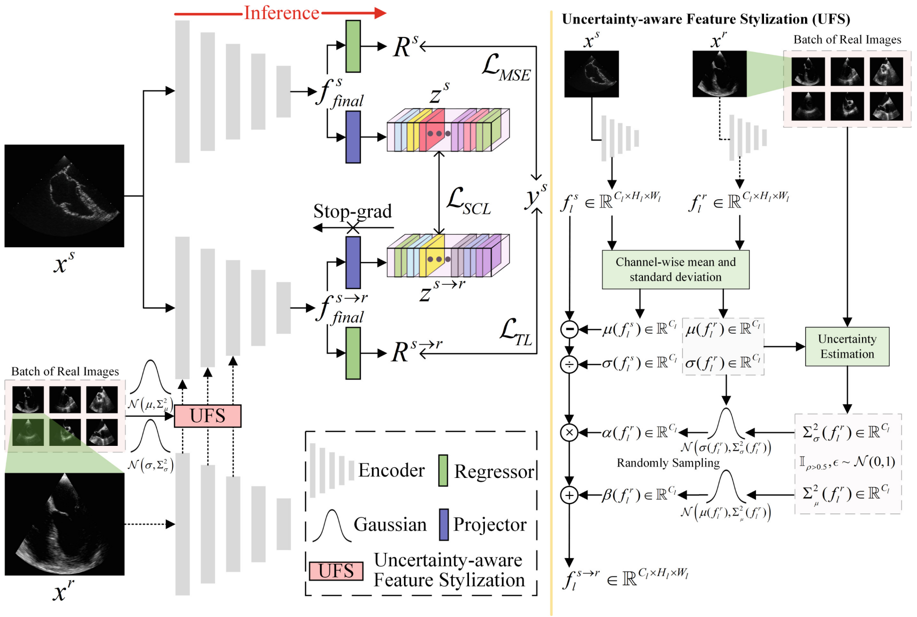
Inference
$R^s$
$\mathcal{L}_{MSE}$
$f^s_{final}$
$z^s$
$\mathcal{L}_{SCL}$
Stop-grad
$y^s$
$x^s$
$f^{s\to r}_{final}$
$z^{s\to r}$
$\mathcal{L}_{TL}$
$R^{s\to r}$
Batch of Real Images
$\mathcal{N}(\mu,\Sigma^2_\mu)$
UFS
$\mathcal{N}(\sigma,\Sigma^2_\sigma)$
$x^r$
Encoder
Regressor
Gaussian
Projector
UFS
Uncertainty-aware Feature Stylization
Uncertainty-aware Feature Stylization (UFS)
$x^s$
$x^r$
Batch of Real Images
$f^s_l \in \mathbb{R}^{C_l\times H_l\times W_l}$
$f^r_l \in \mathbb{R}^{C_l\times H_l\times W_l}$
Channel-wise mean and standard deviation
$\mu(f^s_l) \in \mathbb{R}^{C_l}$
$\mu(f^r_l) \in \mathbb{R}^{C_l}$
$\sigma(f^s_l) \in \mathbb{R}^{C_l}$
$\sigma(f^r_l) \in \mathbb{R}^{C_l}$
Uncertainty Estimation
$\alpha(f^r_l) \in \mathbb{R}^{C_l}$
$\mathcal{N}(\sigma(f^r_l),\Sigma^2_\sigma(f^r_l))$
$\Sigma^2_\sigma(f^r_l) \in \mathbb{R}^{C_l}$
Randomly Sampling
$\mathbb{I}_{\rho>0.5}, \epsilon \sim \mathcal{N}(0,1)$
$\beta(f^r_l) \in \mathbb{R}^{C_l}$
$\mathcal{N}(\mu(f^r_l),\Sigma^2_\mu(f^r_l))$
$\Sigma^2_\mu(f^r_l) \in \mathbb{R}^{C_l}$
$f^{s\to r}_l \in \mathbb{R}^{C_l\times H_l\times W_l}$

**Fig. 2.** Overall architecture of our proposed SR-AQA. The right part shows uncertainty-aware feature stylization (UFS) for one simulated image $x^s$ with real images at $l^{th}$ layer. No gradient calculation is performed on the dotted line. The red line indicates the path of model inference.

brightness, contrast, acoustic shadowing, and refraction artifact) exist between simulated and real data, posing a considerable challenge to UDA. Simulated images were collected with the HeartWorks TEE simulation platform from 38 participants of varied experience asked to image the 9 views. Fully anonymized real TEE data were collected from 10 cardiovascular procedures in 2 hospitals, with ethics for research use and collection approved by the respective Research Ethics Committees. Each image is annotated by 3 expert anaesthetists with two independent scores w.r.t. two AQA tasks for TEE. The criteria percentage (CP) score ranging from "0–100", measuring the number of essential criteria, from the checklists (provided in supplementary material) of the ASE/SCA/BSE imaging guidelines [5], met during image acquisition and a general impression (GI) score ranging from "0–4", representing overall US image quality.

As the number of criteria thus the maximum score varies for different views, we normalise CP as a percentage to provide a consistent measure across all views. Scores from the 3 raters were averaged to obtain the final score for each view. The Pearson product-moment correlation coefficients between CP and GI are 0.81 for simulated data and 0.70 for real data, indicating that these two metrics are correlated but focus on different clinical quality aspects. Inter-rater variability is assessed using the two-way mixed-effects interclass correlation coefficient (ICC) with the definition of absolute agreement. Both CP and GI, show very good agreement between the 3 annotators with ICCs of 0.959, 0.939 and 0.813, 0.758 for simulated and real data respectively.

### 2.2 Simulation-to-Real AQA Network (SR-AQA)

**Overview of SR-AQA.** Illustrated in Fig. 2, the proposed SR-AQA is composed of ResNet-50 [6] encoders, regressors and projectors, with shared weights. Given the simulated $x^s$ and real TEE image $x^r$ as input, SR-AQA first estimates the uncertainty in the real styles of $x^s$, from the batch of real images and transfers real styles to simulated features by normalizing their feature statistics (i.e. mean and standard deviation) via UFS. Then, we perform style consistency learning with $\mathcal{L}_{SCL}$ and task-specific learning with $\mathcal{L}_{TL}$ for the final real-stylized features $f^{s\rightarrow r}_{final}$ and the final simulated features $f^s_{final}$ to learn domain-invariant and task-specific information. Ultimately, the total loss function of SR-AQA is $\mathcal{L}_{total} = \mathcal{L}_{MSE} + \lambda_1\mathcal{L}_{SCL} + \lambda_2\mathcal{L}_{TL}$, where $\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(R^s_i - y^s_i\right)^2$ is the MSE loss calculated from the simulated data result $R^s$ and its label $y^s$, while $\lambda_1$ and $\lambda_2$ are parameters empirically set to "10" and "1" to get a uniform order of magnitude at the early training stage. The input is fed into one encoder and regressor to predict the score during inference.

**Uncertainty-Aware Feature Stylization (UFS).** The UFS pipeline is shown in the right part of Fig. 2. Different domains generally have inconsistent feature statistics [12,21]. Since style is related to the features' means and standard deviations [2,8,11,24], simulated features can be augmented to resemble real styles by adjusting their mean and standard deviation with the help of unlabelled real data. Let $f^s_l$ and $f^r_l$ be the simulated features and real features extracted from the $l^{th}$ layer of the encoder, respectively. We thus can transfer the style of $f^r_l$ to $f^s_l$ to obtain real-stylized features $f^{s\rightarrow r}_l$ as:

$$f^{s\rightarrow r}_l = \sigma(f^r_l)\frac{f^s_l - \mu\left(f^s_l\right)}{\sigma\left(f^s_l\right)} + \mu(f^r_l), \tag{1}$$

where: $\mu(f)$ and $\sigma(f)$ denote channel-wise mean and standard deviation of feature $f$, respectively.

However, due to the complexity of real-world TEE, there are significant intra-domain differences, leading to uncertainties in the feature statistics of real data. To explore the potential space of unknown intra-domain shifts, instead of using fixed feature statistics, we generate multivariate Gaussian distributions to represent the uncertainty of the mean and standard deviation in the real data. Considering this, the new feature statistics of real features $f^r_l$, i.e. mean $\beta(f^r_l)$ and standard deviation $\alpha(f^r_l)$, are sampled from $\mathcal{N}\left(\mu(f^r_l), \Sigma^2_\mu(f^r_l)\right)$ and $\mathcal{N}\left(\sigma(f^r_l), \Sigma^2_\sigma(f^r_l)\right)$, respectively and computed as:

$$\beta(f^r_l) = \mu\left(f^r_l\right) + \left(\epsilon\Sigma_\mu\left(f^r_l\right)\right)\cdot\mathbb{I}_{\rho>0.5}, \alpha(f^r_l) = \sigma\left(f^r_l\right) + \left(\epsilon\Sigma_\sigma\left(f^r_l\right)\right)\cdot\mathbb{I}_{\rho>0.5}, \tag{2}$$

where: $\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{1})$[1], variances $\Sigma^2_\mu\left(f^r_l\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\mu(f^r_l) - \mathbb{E}_b[\mu(f^r_l)]\right)^2$ and $\Sigma^2_\sigma\left(f^r_l\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\sigma(f^r_l) - \mathbb{E}_b[\sigma(f^r_l)]\right)^2$ are estimated from the mean and standard deviation of the batch $B$ of real images, $\mathbb{I}_{\rho>0.5}$ is an indicator function and

[1] Re-parameterization trick is applied here to make the sampling operation differentiable.

$\rho \sim \mathcal{U}(\mathbf{0}, \mathbf{1})$. Finally, our UFS for $l^{th}$ layer is defined as:

$$f_l^{s\to r} = \alpha(f_l^r)\frac{f_l^s - \mu\left(f_l^s\right)}{\sigma\left(f_l^s\right)} + \beta(f_l^r). \tag{3}$$

To this end, the proposed UFS approach can close the reality gap by generating real-stylized features with sufficient variations, so that the network interprets real data as just another variation.

**Style Consistency Learning (SCL).** Through the proposed UFS, we obtain the final real-stylized features $f_{final}^{s\to r}$ that contain a diverse range of real styles. The $f_{final}^{s\to r}$ can be seen as style perturbations of the final simulated features $f_{final}^{s}$. We thus incorporate a SCL step, that maximizes the similarity between $f_{final}^{s}$ and $f_{final}^{s\to r}$ to enforce their consistency in the feature level, allowing the encoder to learn robust representations. Specifically, the SCL adds a projector independently of the regressor to transform the $f_{final}^{s}$ ($f_{final}^{s\to r}$) in an extra feature embedding, and then matches it to the other one. To prevent the Siamese encoder and Siamese projector (i.e. the top two encoders and projectors in Fig. 2) from collapsing to a constant solution, similar to [3], we adopt the stop-gradient (stop-grad) operation for the projected features $z^s$ and $z^{s\to r}$. The SCL process is summarized as:

$$\mathcal{L}_{SCL} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}\mathcal{D}\left(f_{final}^{s}, \text{stopgrad}\left(z^{s\to r}\right)\right) + \frac{1}{2}\mathcal{D}\left(f_{final}^{s\to r}, \text{stopgrad}\left(z^{s}\right)\right)\right), \tag{4}$$

where: $\mathcal{D}\left(f^1, f^2\right) = -\frac{f^1}{\|f^1\|_2} \cdot \frac{f^2}{\|f^2\|_2}$ is the negative cosine similarity between the input features $f^1$ and $f^2$, and $\|\cdot\|_2$ is $L2$-normalization.

The SCL guides the network to learn domain-invariant features, via various style perturbations, so that it can generalize well to the different visual appearances of the real domain.

**Task-Specific Learning (TL).** While alleviating the style differences between the simulated and real domain, UFS filters out some task-specific information (e.g. semantic content) encoded in the simulated features, as content and style are not orthogonal [11,19], resulting in performance deterioration. Therefore, we embed TL to retain useful representations for AQA. Specifically, $f_{final}^{s\to r}$ should retain task-specific information to allow the regressor to predict results $R^{s\to r}$ that correspond to the quality scores (CP, GI) in the simulated data. In TL, simulated labels $y^s$ are used as the supervising signal:

$$\mathcal{L}_{TL} = \frac{1}{N}\sum_{i=1}^{N}\left(R_i^{s\to r} - y_i^s\right)^2. \tag{5}$$

The TL performs AQA tasks for style variants to complement the loss of information due to feature stylization.

# 3 Experiments

## 3.1 Experimental Settings

Experiments are implemented in PyTorch on a Tesla V100 GPU. The maximum training iteration is 40,000 with a batch size of 32. We adopted the SGD optimizer with a weight decay of 5e−4, a momentum of 0.9 and a learning rate of 1e-4. Input images are resized to 224 × 224, the CP and GI scores are normalized to [0, 1]. The MSE is adopted as the evaluation metric for both the CP and GI regression tasks. Following the standard approach for UDA [4,18,23], we use all 16,192 labelled simulated data and all 4,427 unlabelled real data for domain adaptation, and test on all 4,427 real data. To further explore the data efficiency of UDA methods on our simulation-to-real AQA tasks, we also conduct experiments with fractions (10%, 30%, and 50%) of unlabeled real data for domain adaptation, randomly selected from the 10 TEE real cases.

## 3.2 Comparison with State-of-the-Arts

We compare the proposed SR-AQA with MixStyle [24], MDD [23], RSD [4], and SDAT [18]. All methods are implemented based on their released code and original literature, and fine-tuned to fit our tasks to provide a basis for a fair comparison.

As shown in Table 1, all UDA methods show better AQA overall performance (lower MSE) compared to the model trained only with simulated data ("Simulated Only"), demonstrating the effectiveness of UDA methods in TEE simulation-to-real AQA. The proposed SR-AQA achieves superior performance with the lowest MSE in all CP experiments and all but one GI experiment (50%), in which is very close (0.7648 to 0.76) to the best-performing SDAT. We calculate the MSE reduction percentage between our proposed SR-AQA and the second-best method, to obtain the degree of performance improvement. Specifically, on the CP task among the five real data ratio settings, the MSE of our method dropped by 2.13%–5.08% against the suboptimal method SDAT. It is evident that even with a small amount (10%) of unlabelled real data used for UDA, our SR-AQA still achieves a significant MSE reduction, of at least 3.02% and 4.37% compared to other UDA methods, on the CP and GI tasks respectively, showcasing high data efficiency. We also conduct paired t-tests on MSE results, from multiple runs with different random seeds. The obtained p-values ($p < 0.05$ in all but one case, see supplementary material Table S3), validate that the improvements yielded by the proposed SR-AQA are statistically significant. In Table 2, we report the performance over different (low, medium, and high) score ranges with SR-AQA, obtaining promising results among all ranges[2].

## 3.3 Ablation Study

We first explore the impact of the amount of UFS on generalization performance. As shown in the left part of Table 3, the performance continues to improve as UFS

[2] Example GI and CP results are provided in the supplementary material

**Table 1.** MSE results for CP and GI scores, in different unlabelled real data ratios. Lower MSE means better model performance. The 'Reduction' row lists the MSE reduction percentage of our proposed SR-AQA compared to the second-best method. Top two results are in **bold** and underlined.

| Methods | CP Task (MSE) | | | | GI Task (MSE) | | | |
|---|---|---|---|---|---|---|---|---|
| | Unlabelled Real Data Ratio | | | | Unlabelled Real Data Ratio | | | |
| | 100% | 50% | 30% | 10% | 100% | 50% | 30% | 10% |
| SR-AQA (Ours) | **411** | **411** | **413** | **417** | **0.6992** | 0.7648 | **0.7440** | **0.7696** |
| Reduction | −5.08% | −2.38% | −2.13% | −3.02% | −16.28% | +0.63% | −7.55% | −4.37% |
| SDAT [18] | 433 | 421 | 422 | 430 | 0.9792 | **0.7600** | 0.8048 | 1.1024 |
| RSD [4] | 540 | 507 | 501 | 513 | 0.8768 | 0.8768 | 0.9392 | 0.8048 |
| Mixstyle [24] | 466 | 474 | 465 | 496 | 0.8352 | 1.0048 | 0.8736 | 1.0400 |
| MDD [23] | 766 | 787 | 755 | 742 | 1.1696 | 1.1360 | 1.1328 | 1.1936 |
| Simulated Only | 913 | | | | 1.2848 | | | |

**Table 2.** MSE results on AQA tasks for different score ranges. The full real dataset (without labels) is used for unsupervised domain adaptation.

| Methods | CP Task (MSE) | | | GI Task (MSE) | | |
|---|---|---|---|---|---|---|
| | Low | Medium | High | Low | Medium | High |
| | (0,30] | (30, 60] | (60, 100] | (0,1.2] | (1.2, 2.4] | (2.4, 4] |
| SR-AQA (Ours) | **1668** | **403** | **286** | **2.3200** | **0.4080** | **0.4832** |
| Reduction | −2.68% | −2.66% | −1.38% | −13.38% | −19.56% | −1.30% |
| SDAT [18] | 1886 | 441 | 290 | 3.2512 | 0.6400 | 0.4896 |
| RSD [4] | 1881 | 489 | 427 | 2.6784 | 0.5168 | 0.6912 |
| Mixstyle [24] | 1752 | 414 | 359 | 2.7152 | 0.5072 | 0.5632 |
| MDD [23] | 1714 | 550 | 771 | 3.1376 | 0.8208 | 0.8480 |
| Simulated Only | 2147 | 850 | 831 | 3.4592 | 0.8896 | 1.0336 |

**Table 3.** Ablation results on two AQA tasks via different UFS settings (left part), and for the proposed uncertainty-aware, $\mathcal{L}_{SCL}$, and $\mathcal{L}_{TL}$ (right part). Headings 1, 1-2, 1-3, 1-4, 1-5 refer to replacing the $1^{st}$, 1-$2^{nd}$, 1-$3^{rd}$, 1-$4^{th}$, 1-$5^{th}$ batch normalization layer(s) of ResNet-50 with the UFS. The full real dataset (without labels) is used for unsupervised domain adaptation.

| UFS | 1 | 1-2 | 1-3 | 1-4 | 1-5 |
|---|---|---|---|---|---|
| CP Task (MSE) | 421 | 412 | **411** | 422 | 430 |
| GI Task (MSE) | 0.7808 | 0.7488 | **0.6992** | 0.7312 | 0.7840 |

| Feature Stylization | $\mathcal{L}_{SCL}$ | $\mathcal{L}_{TL}$ | CP Task (MSE) | GI Task (MSE) |
|---|---|---|---|---|
| w/o Uncertainty | ✓ | | 426 | 0.8080 |
| | ✓ | ✓ | 419 | 0.7408 |
| w/ Uncertainty | ✓ | | 422 | 0.7840 |
| | ✓ | ✓ | **411** | **0.6992** |

is applied to more shallow layers, but decreases when UFS is added to deeper layers. This is because semantic information is more important than style in the deeper layers. Using a moderate number of UFS to enrich simulated features with real-world styles, without corrupting semantic information, improves model generalization. Secondly, we study the effect of uncertainty-aware, SCL, and TL, as shown in the right part of Table 3, removing each component leads to performance degradation.

## 4 Conclusion

This paper presents the first annotated TEE dataset for simulation-to-real AQA with 16,192 simulated images and 4,427 real images. Based on this, we propose a novel UDA network named SR-AQA for boosting the generalization of AQA performance. The network transfers diverse real styles to the simulated domain based on uncertainty-award feature stylization. Style consistency learning enables the encoder to learn style-independent representations while task-specific learning allows our model to naturally adapt to real styles by preserving task-specific information. Experimental results on two AQA tasks for CP and GI scores show that the proposed method outperforms state-of-the-art methods with at least 5.08% and 16.28% MSE reduction, respectively, resulting in superior TEE AQA performance. We believe that our work provides an opportunity to leverage large amounts of simulated data to improve the generalisation performance of AQA for real TEE. Future work will focus on reducing negative transfer to extend UDA methods towards simulated-to-real TEE quality assessment.

**Acknowledgements.** This work is supported by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145Z/16/Z and NS/A000050/1]; EPSRC-funded UCL Centre for Doctoral Training in Intelligent, Integrated Imaging in Healthcare (i4health) [EP/S021930/1]; Horizon 2020 FET [863146]; a UCL Graduate Research Scholarship; and Singapore MoE Tier 1 Start up grant (WBS: A-8001267-00-00). Danail Stoyanov is supported by a RAE Chair in Emerging Technologies [CiET1819/2/36] and an EPSRC Early Career Research Fellowship [EP/P012841/1].

## References

1. Abdi, A.H., et al.: Automatic quality assessment of echocardiograms using convolutional neural networks: feasibility on the apical four-chamber view. IEEE Trans. Med. Imaging **36**(6), 1221–1230 (2017)
2. Chen, C., Li, Z., Ouyang, C., Sinclair, M., Bai, W., Rueckert, D.: MaxStyle: adversarial style composition for robust medical image segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13435, pp. 151–161. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_15
3. Chen, X., He, K.: Exploring simple Siamese representation learning. In: CVPR 2021, pp. 15745–15753 (2021)
4. Chen, X., Wang, S., Wang, J., Long, M.: Representation subspace distance for domain adaptation regression. In: ICML 2021, pp. 1749–1759 (2021)

5. Hahn, R.T., et al.: Guidelines for performing a comprehensive transesophageal echocardiographic examination: recommendations from the American society of echocardiography and the society of cardiovascular anesthesiologists. J. Am. Soc. Echocardiogr. **26**(9), 921–964 (2013)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR 2016, pp. 770–778 (2016)
7. Hempel, C., et al.: Impact of simulator-based training on acquisition of transthoracic echocardiography skills in medical students. Ann. Card. Anaesth. **23**(3), 293 (2020)
8. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV 2017, pp. 1501–1510 (2017)
9. Labs, R.B., Vrettos, A., Loo, J., Zolgharni, M.: Automated assessment of transthoracic echocardiogram image quality using deep neural networks. Intell. Med. (2022)
10. Le, C.K., Lewis, J., Steinmetz, P., Dyachenko, A., Oleskevich, S.: The use of ultrasound simulators to strengthen scanning skills in medical students: a randomized controlled trial. J. Ultrasound Med. **38**(5), 1249–1257 (2019)
11. Lee, S., Seong, H., Lee, S., Kim, E.: WildNet: learning domain generalized semantic segmentation from the wild. In: CVPR 2022, pp. 9926–9936 (2022)
12. Li, X., Dai, Y., Ge, Y., Liu, J., Shan, Y., Duan, L.: Uncertainty modeling for out-of-distribution generalization. In: ICLR 2022 (2022)
13. Liao, Z., et al.: On modelling label uncertainty in deep neural networks: automatic estimation of intra-observer variability in 2D echocardiography quality assessment. IEEE Trans. Med. Imaging **39**(6), 1868–1883 (2019)
14. Lin, Z., et al.: Multi-task learning for quality assessment of fetal head ultrasound images. Med. Image Anal. **58**, 101548 (2019)
15. Mazomenos, E.B., Bansal, K., Martin, B., Smith, A., Wright, S., Stoyanov, D.: Automated performance assessment in transoesophageal echocardiography with convolutional neural networks. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 256–264. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_30
16. Mazomenos, E.B., et al.: Motion-based technical skills assessment in transoesophageal echocardiography. In: Zheng, G., Liao, H., Jannin, P., Cattin, P., Lee, S.-L. (eds.) MIAR 2016. LNCS, vol. 9805, pp. 96–103. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43775-0_9
17. Montealegre-Gallegos, M., et al.: Imaging skills for transthoracic echocardiography in cardiology fellows: the value of motion metrics. Ann. Card. Anaesth. **19**(2), 245 (2016)
18. Rangwani, H., Aithal, S.K., Mishra, M., Jain, A., Radhakrishnan, V.B.: A closer look at smoothness in domain adversarial training. In: ICML 2022, pp. 18378–18399 (2022)
19. Shen, Y., Lu, Y., Jia, X., Bai, F., Meng, M.Q.H.: Task-relevant feature replenishment for cross-centre polyp segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13434, pp. 599–608. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16440-8_57
20. Song, H., Peng, Y.G., Liu, J.: Innovative transesophageal echocardiography training and competency assessment for Chinese anesthesiologists: role of transesophageal echocardiography simulation training. Curr. Opin. Anaesthesiol. **25**(6), 686–691 (2012)
21. Wang, X., Long, M., Wang, J., Jordan, M.: Transferable calibration with lower bias and variance in domain adaptation. In: NeurIPS 2020, pp. 19212–19223 (2020)

22. Wheeler, R., et al.: A minimum dataset for a standard transoesphageal echocardiogram: a guideline protocol from the British society of echocardiography. Echo Res. Pract. **2**(4), G29 (2015)
23. Zhang, Y., Liu, T., Long, M., Jordan, M.: Bridging theory and algorithm for domain adaptation. In: ICML 2019, pp. 7404–7413 (2019)
24. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. In: ICLR 2021 (2021)