RTN: Reinforced Transformer Network for Coronary CT Angiography Vessel-level Image Quality Assessment

Yiting Lu[1*], Jun Fu[1*], Xin Li[1], Wei Zhou[1], Sen Liu[1], Xinxin Zhang[2], Congfu Jia[2], Ying Liu[2], and Zhibo Chen[1†]

[1] University of Science and Technology of China, Hefei, Anhui, China
{luyt31415, fujun, lixin666, weichou}@mail.ustc.edu.cn
elsen@iat.ustc.edu.cn
chenzhibo@ustc.edu.cn
[2] The First Affiliated Hospital of Dalian Medical University, Dalian, Liaoning, China

Abstract. Coronary CT Angiography (CCTA) is susceptible to various distortions (e.g., artifacts and noise), which severely compromise the exact diagnosis of cardiovascular diseases. The appropriate CCTA Vessel-level Image Quality Assessment (CCTA VIQA) algorithm can be used to reduce the risk of error diagnosis. The primary challenges of CCTA VIQA are that the local part of coronary that determines final quality is hard to locate. To tackle the challenge, we formulate CCTA VIQA as a multiple-instance learning (MIL) problem, and exploit **T**ransformer-based **MIL** module (termed as T-MIL) to aggregate the multiple instances along the coronary centerline into the final quality. However, not all instances are informative for final quality. There are some quality-irrelevant/negative instances intervening the exact quality assessment(*e.g.*, instances covering only background or the coronary in instances is not identifiable). Therefore, we propose a **P**rogressive **R**einforcement learning based **I**nstance **D**iscarding module (termed as PRID) to progressively remove quality-irrelevant/negative instances for CCTA VIQA. Based on the above two modules, we propose a **R**einforced **T**ransformer **N**etwork (RTN) for automatic CCTA VIQA based on end-to-end optimization. The experimental results demonstrate that our proposed method achieves the state-of-the-art performance on the real-world CCTA dataset, exceeding previous MIL methods by a large margin.

Keywords: Image Quality Assessment · CCTA · Reinforced Learning · Transformer.

1 Introduction

Coronary Computed Tomography Angiography (CCTA) plays an important role in the diagnosis of cardiovascular diseases for providing vital visual clues. However, the CCTA images are easily degraded by various factors (*i.e.*, breathing

*The first two authors contribute equally to this work.
†Corresponding Author

motion artifacts and insufficient contrast agent dose) and contain hybrid distortions [11], which inevitably affects the subsequent analysis of doctors [5]. When artifacts appear in the coronary artery stenosis, it is difficult for doctors to diagnose stenosis [8]. To ensure accurate diagnosis, it is necessary to provide doctors with high-quality CCTA images. Therefore, there is an urgent need to develop CCTA Vessel-level Image Quality Assessment (CCTA VIQA) algorithms.

With the rapid development of machine learning, the seminal work [18] maps hand-crafted global and local features (*i.e.*, noise, contrast, misregistration scores, and un-interpretability index) of coronary arteries onto quality scores through machine learning algorithms. However, its input features are not rich since they only include four types of image characteristics, which always causes the sub-optimal performance and lacks of enough flexibility. Also, quality metric [13,14] designed for natural image are not suitable for medical image. During the dataset annotation process, the professional doctors only provide the vessel-level label when browsing the complete CT. So no position labels are provided for quality relevant regions and the key local parts that determine the vessel-level quality are hard to locate, which shows CCTA VIQA is an obvious weakly-supervised problem [24]. So the quality relationship between various local parts of coronary arteries in CCTA image can be excavated by modeling CCTA VIQA as a multiple-instance learning (MIL) problem. Therefore, we propose Transformer-based MIL backbone (T-MIL) in CCTA VIQA. Specifically, since the quality of CCTA images is only associated with the coronary arteries, we utilize the centerline tracking algorithm [21] to detect the regions of coronary arteries. Then we define 3D cubes cropped along the vessel centerline as instances. Finally, the discriminative features from multiple instances extracted by 3D convolutional neural networks are aggregated into the quality space through the latest network architecture, *i.e.*, transformer. Recently, there are various *instance* aggregators in MIL methods, like attention [6,15,10], RNN [2], sparse convolution [9], and graph [23]. Specially, transformer-based MIL frameworks [7,17,19,22] have achieved remarkable success in a broad of medical tasks, such as whole slide image classification.

Although the instances (*i.e.*, cubes) have covered all possible quality-associated contents, the quality-irrelevant contents also infiltrate the instances severely, which is detrimental for the estimation of overall quality. For instance, the quality-related cubes only take a small proportion of all cubes. According to our observation, there are three typical cases of quality-irrelevant instances *i.e.*, the instance that does not match the vessel-level label, the coronary in instances is not identifiable, and the instance contains only background. To remove these negative instances while mining the most informative instances, we propose a **P**rogressive **R**einforcement learning based **I**nstance **D**iscarding module (termed as PRID) to preserve informative instances as the inputs of the transformer. The reinforcement learning (RL) agent from PRID accepts the output feature embedding of transformer as states, and selects one instance to discard. Then we input the new instance set into T-MIL to obtain the states (both in training and testing) for the next iteration and the reward (just for training) to refine cur-

rent action. We call the T-MIL together with PRID as **R**einforced **T**ransformer **N**etwork, which is denoted as RTN. We summarize our contributions as follows.

- To our knowledge, we propose the first fully automatic CCTA VIQA algorithm RTN based on end to end optimization. We formulate the CCTA VIQA as the typical MIL problem, and introduce transformer to aggregate multiple instances and map them to final quality.
- To elide the intervention from quality-irrelevant/negative isntances, we propose a progressive reinforced learning based instance discarding strategy (*i.e.*, PRID) to mine the most informative instances for transformer network.
- Extensive experimental results reveal that our proposed RTN achieves the state-of-the-art (SOTA) performance on hospital-built CCTA dataset, exceeding previous MIL methods by a large margin.

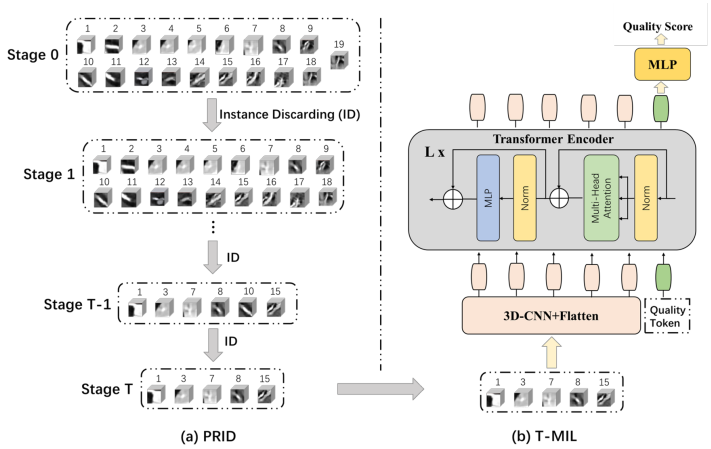

Fig. 1. Our RTN includes two modules: (a) Progressive Reinforcement Learning based Instance Discarding (PRID), (b) Transformer-based MIL Backbone (T-MIL).

2 Methods

Fig. 1 depicts the overall framework of RTN for the CCTA VIQA task, which is composed of two basic modules *i.e.*, Progressive Reinforcement Learning based Instance Discarding (PRID) and Transformer-based MIL Backbone (T-MIL). Given one CCTA image, we first collect the cubes cropped along the coronary centerline as instances. Then PRID module employs a reinforcement learning (RL) agent to determine which instance should be discarded progressively. After obtaining the most informative instances, T-MIL is devoted to classifying the final vessel-level quality grade. In the following sections, we will clarify the T-MIL and PRID of our RTN from both implementation and principal perspectives.

2.1 Transformer-based MIL

Multiple-instance learning (MIL) is a strong tool to solve weakly-supervised problem. In the definition of MIL, a set of multiple instances can be regarded as a bag and only bag-level label is provided. In our method, we define the i^{th} 3D cube sampled on the coronary artery centerline as an instance x_i, and the whole coronary artery region is taken as a bag $\mathcal{B} = \{\boldsymbol{x_i}|1 \leq i \leq n\}$. Then the perceptual quality y of whole coronary $\mathcal{B}$ is:

$$y(\mathcal{B}) = h(f(\boldsymbol{x_1}), f(\boldsymbol{x_2}), ..., f(\boldsymbol{x_i}), ..., f(\boldsymbol{x_n})),\ 1 \leq i \leq n \tag{1}$$

Where, $x_i \in \mathbb{R}^{C_1 \times D \times H \times W}$ is the i^{th} instance in the bag $\mathcal{B}$. T-MIL contains $f(.)$ and $h(.)$, which are separately as instance feature extractor and transformer-based aggregator. In this paper, the instance feature extractor f is composed of several 3D convolution based residual blocks [3] and flatten operation.

Transformer-based Aggregator. To capture the long-range dependency between different instances, we employ the transformer architecture in ViT [4] as the aggregator of MIL. As shown in Fig. 1 each transformer encoder layer is consist of multi-head self-attention (MHSA) layer and feed-forward (FF) layer. We follow the ViT [4] and add the quality token $\boldsymbol{c_0}(\mathcal{B})$ to the instance token groups. The input token embeddings can be written as:

$$\boldsymbol{z_0} = [\boldsymbol{c_0}(\mathcal{B}), f(\boldsymbol{x_1}), f(\boldsymbol{x_2}), ..., f(\boldsymbol{x_i}), ..., f(\boldsymbol{x_n})],\ 1 \leq i \leq n. \tag{2}$$

In MHSA, we firstly transform instance embedding to key K, query Q and value V, and then calculate the similarity of key and query as attention weight matrix. The matrix's each item means dependencies between any pair of instances. The output of MHSA contains aggregation information, especially quality token embedding that aggregates the contribution of each instance to final vessel-level quality prediction.The full process of the l^{th} transformer layer is as follows, in which LN is layernorm and MLP includes two fully-connected layers with a GELU non-linearity:

$$\begin{aligned} \boldsymbol{z_l^{'}} &= MHSA(LN(\boldsymbol{z_{l-1}})), \quad l = 1, 2, ...L \\ \boldsymbol{z_l} &= MLP(LN(\boldsymbol{z_l^{'}})) + \boldsymbol{z_l^{'}}, \quad l = 1, 2...L \end{aligned} \tag{3}$$

After feeding input token embedding into L transformer layers, we can obtain the output token embeddings $\boldsymbol{z_L} \in \mathbb{R}^{(n+1) \times D}$, in which D is the dimension of the token embedding. The first quality token embedding $\boldsymbol{c_L}(\mathcal{B}) = \boldsymbol{z_L}[0]$ is used to quality classification and following instance embedding $\boldsymbol{b_L} = \boldsymbol{z_L}[1, 2, ..., n]$ can be used as the states of PRID, in which the instance embeddings $\boldsymbol{b_L}$ can be understood as features of n instances in one vessel extracted by T-MIL.

2.2 PRID

To reduce the intervention of negative instances (*e.g.*, the instance that do not match vessel-level labels or the instance contains only background), we propose

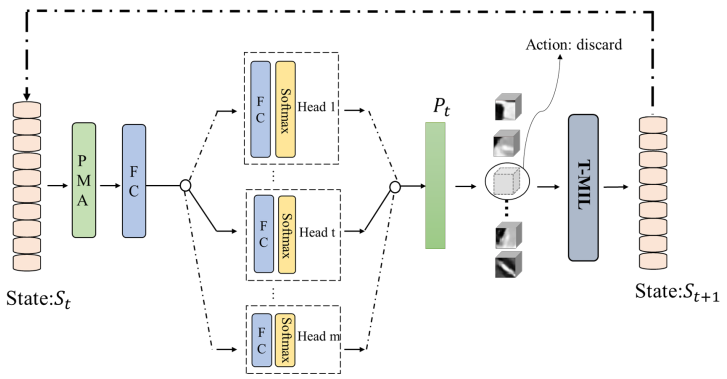

Fig. 2. The agent network of PRID, contains common Pooling by Multi-Head Attention(PMA) module and various MLP layers.

to utilize reinforcement learning (RL) agent to adaptively identify them and discard them progressively [20]. Specifically, we model the process of progressively instance discarding as a Markov Decision Process (MDP) [1,12] and introduce a RL agent to obtain the optimal solution for it. The state, action, reward and agent in RL are clarified clearly as follows.

States. As shown in Fig. 2, in the t^{th} iteration, the state $\boldsymbol{s_t}$ is defined as the output instance embedding $\boldsymbol{b_L}(t-1) \in \mathbb{R}^{(n-t+1)\times D}$ of the $(t-1)^{th}$ iteration's transformer layer, since the features captured by transformer are more representative for quality prediction.

Action. Action $\boldsymbol{a_t}$ is the instance index that is discarded within the scope of the instance set. In the t-th iteration, the action search space $\mathcal{A} = \{1, 2, ..k, ..., n-t+1\}$ is the current instances' index list. The agent's output probability vector $\boldsymbol{p_t} \in \mathbb{R}^{n-t+1}$can be regarded as the selected distribution of current instance set. Thus we can encode action as multinomial distribution sampling when training and top one sampling when testing, the selected k is equal to action: $k = sample(\boldsymbol{p_t})$. The state $\boldsymbol{s_t}$ transforms to $\boldsymbol{s_{t+1}}$ through the action because of changes in the instance set: $\{\boldsymbol{x_i}\}_{i=1}^{n-t+1} \longrightarrow \{\boldsymbol{x_i}\}_{i=1,i\neq k}^{n-t+1}$.

Reward. The reward $\boldsymbol{r_t}$ need to reflect the effect of transforming from the state $\boldsymbol{s_t}$ to $\boldsymbol{s_{t+1}}$ due to the action. After discarding one instance, we feed new instance set into the pre-trained T-MIL and compare the new predicted result with the label to calculate the reward $(t > 1)$:

$$\boldsymbol{r_t} = \begin{cases} 2, & if\ y_t = label \quad and \quad y_{t-1} = label \\ 1, & if\ y_t = label \quad and \quad y_{t-1} \neq label \\ -1, & if\ y_t \neq label \quad and \quad y_{t-1} = label \\ -2, & if\ y_t \neq label \quad and \quad y_{t-1} \neq label \end{cases} \tag{4}$$

In the first selection, the predict result y_1 need to compare with label. If the prediction is correct, give a positive reward (+1), otherwise give a negative reward (−1). In the next choice, as the Eq. 4 shown, the value of reward is not only related to the accuracy of the current selection's prediction result, but also to the last selection's result. This is because in the MDP problem, the current selection (iteration) is related to the last selection (iteration).

Agent. As shown in Fig. 2, the agent in PRID receives the states from T-MIL. We first aggregate the n tokens of states into one token through PMA module $PMA(.)$ [7]. In the t^{th} iteration, this module sets a learnable embedding $I \in \mathbb{R}^{1\times D}$ as a query, and directly regards the instance embedding $\boldsymbol{b_L}(t-1) \in \mathbb{R}^{(n-t+1)\times D}$ as key and value to calculate a attention matrix of $1 \times (n-t+1)$ dimension to gather these feature embedding. Similarly, the cross attention here is also implemented in the form of multi-head. Then we feed the fused token into the MLP head $g_t(.)$ to obtain the probability vector $\boldsymbol{p_t} \in \mathbb{R}^{n-t+1}$. Note that in t^{th} iteration, we will use t^{th} MLP head $g_t(.)$:

$$\boldsymbol{p_t} = g_t(PMA(\boldsymbol{b}_L(t-1))) \tag{5}$$

Instance Discarding Strategy. The implementation requires above two modules: PRID and T-MIL. In the first stage, we need to pre-train the T-MIL by randomly selecting $n-m$ instances from n instances. Secondly, fix the parameters of T-MIL and update the agent's parameter through m progressive selections through interaction with T-MIL. At each iteration, we can obtain the selected index (k) probability from distribution $\boldsymbol{p_t}$ and reward $\boldsymbol{r_t}$, so the training loss is

$$loss = -\sum_{t=1}^{m} log(\boldsymbol{p_t}[k]) \times \boldsymbol{r_t} \tag{6}$$

Table 1. Performance comparisons with state-of-the-arts on the CCTA dataset.

MIL methods	Accuracy	AUC
AttentionMIL [6]	0.7574	0.7576
MIL-RNN [2]	0.7322	0.6842
CLAM [15]	0.7761	0.7161
DSMIL [10]	0.6917	0.5378
T-MIL (ours)	**0.8036**	**0.7658**
RTN(PRID+T-MIL) (ours)	**0.8546**	**0.8461**

3 Experiment

3.1 Implementation Details

Our CCTA VIQA dataset is collected with the help of a partner hospital, where the vessel-level quality labels of each CCTA image are provided by experienced

imaging doctors and the resolution (*i.e.*, 512×512) of CCTA slices along axis is commonly used in the hospital. There are two quality levels in our dataset *i.e.*, "1" and "0". "1" means the CCTA image is high-quality and accepted by doctors, while "0" represents the CCTA image is low-quality and cannot be used for diagnosis. The dataset consists of 80 CCTA scans from 40 patients in both systole and diastole, which can generate 210 coronary branches by the centerline tracking algorithm [21]. The centerline algorithm realizes the rough detection of coronary region, and the rough detection accuracy can reach 94%, which has little influence on the subsequent VIQA. Therefore, our dataset contains 210 pairs of coronary branches and vessel-level quality labels, where the ratio of label "1" and label "0" is 114/96. And we possibly plan to make this CCTA VIQA dataset public later.

We adopt the numbers of instances (*i.e.*, cubes) n in MIL as 19, which are uniformly cropped along the vessel centerline. All cubes are with the size of $20 \times 20 \times 20$ and cover the whole coronary artery branch.We also augment the data by moving the cube's center point randomly to three voxels in any direction along 6 neighborhoods as in [16]. We follow the 5-fold cross validation setting with 80% of data for training and 20% for testing in each split. Both T-MIL and PRID are implemented with Pytorch and trained on one NIVDIA 1080Ti GPU. In the training process, we first train T-MIL for 200 epoches with the batchsize 2. Then, we optimize the PRID module for 400 epoches with batchsize 2. We utilize two metrics of quality classification at vessel level to measure the effectiveness of the proposed framework: Accuracy and Area Under the Curve (AUC) scores. Moreover, the best selection of instance discarding number m is based on experiments. And m pick 14 as baseline.

Table 2. Performance comparison with different discarding numbers and discarding strategies in RTN on the CCTA dataset.

Discarding Number	PRID(Accuracy/AUC)	Random(Accuracy/AUC)
4	0.7964/0.7777	0.7682/0.7409
9	0.8253/0.7674	0.8007/0.7567
14	0.8546/0.8461	0.8107/0.7994

Table 3. Performance comparison with different pooling module in agent network of RTN on the CCTA dataset and different cube size on one vessel.

Pooling Module	Accuracy	AUC	Crop Size	Accuracy	AUC
PMA	**0.8546**	**0.8461**	15	0.8042	0.7459
Avg Pooling	0.8443	0.8257	20	**0.8546**	0.8461
Max Pooling	0.8273	0.8198	30	0.8510	**0.8668**

3.2 Comparisons with State-of-the-arts

We compare our methods with the state-of-the-art MIL methods on our dataset, including attention-based MIL [6], RNN-based MIL [2], attention-based and cluster-based MIL [15], non-local attention based MIL [10]. In order to ensure fairness, the feature extraction process of the above methods shares the same two layers of 3D residential blocks. As shown in Table 1, transformer-based MIL exceeds the second best method CLAM [15] by 2.75%, thanks to its better long-range relationship modeling capability. Furthermore, our proposed RTN achieves the best performance, outperforming previous MIL-based methods by 7.85%, which reveals the effectiveness of our PRID. In other words, discarding quality-irrelevant instances is vital for CCTA VIQA. See supplementary material, the visualization of index distribution of discarded instance and remained instance shows that only limited instances will play a role in CCTA VIQA.

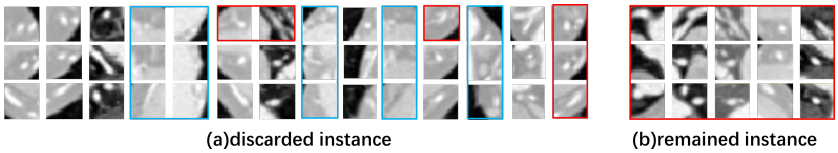

Fig. 3. Example of instance discarding with label "1". Three rows represent views in axial, sagittal, and coronal orientations from all 3D cubes on the coronary artery, among them, the blue box is the case where the coronary in instances is not identifiable, and the red box is the case with obvious distortion.

3.3 Ablation Study

In this section, we verify the effectiveness of our proposed PRID from four aspects: the number of discarding instances, discarding strategy, pooling operations and cube size. Table 2 shows the comparison results of different discarding numbers and different discarding strategies. According the results, the discarding number $m = 14$ is the best solution. This is because after iterative discarding, the five instances with the most information are retained at last, which will make it easier for network to classify, as shown in Fig. 3. We also compare the PRID with random discarding strategy in Table 2. Our PRID exceeds random discarding strategy by a large margin regardless of the discarding number, which reveals the effectiveness of our PRID on instance selection. In Table 3, we compare the different pooling operations for RL agent in PRID. We can draw a conclusion that PMA has a stronger aggregation ability to input instance embedding. This also shows that it is more explanatory to aggregate tokens through cross attention [7]. The comparison of different cube size in Table 3 shows that the cubes with small size cannot cover the whole vessel and the cubes with larger size will contain a little more quality-unrelated content.

4 Conclusion

In this paper, we present a novel Reinforced Transformer Network(RTN) model for CCTA VIQA, which contains two modules: Transformer-based MIL backbone (T-MIL) and Progressive Reinforcement learning based Instance Discarding module (PRID). T-MIL can solve the challenge that local part of coronary that determines final quality is hard to locate. Moreover, PRID can overcome the intervention from quality-irrelevant/negative instances. Compared with previous MIL methods, our RTN has achieved great improvement. In the future, we plan to adaptively select the number of discarded instances, which will continue to be improved in the later work and put into clinical use.

Acknowledgement. This work was supported in part by NSFC under Grant U1908209, 62021001 and the National Key Research and Development Program of China 2018AAA0101400.

References

1. Bellman, R.: A markovian decision process. Journal of mathematics and mechanics pp. 679–684 (1957)
2. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)
3. Chen, S., Ma, K., Zheng, Y.: Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Ghekiere, O., Salgado, R., Buls, N., Leiner, T., Mancini, I., Vanhoenacker, P., Dendale, P., Nchimi, A.: Image quality in coronary ct angiography: challenges and technical solutions. The British journal of radiology **90**(1072), 20160567 (2017)
6. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
7. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: International Conference on Machine Learning. pp. 3744–3753. PMLR (2019)
8. Leipsic, J., Labounty, T.M., Heilbron, B., Min, J.K., Mancini, G.J., Lin, F.Y., Taylor, C., Dunning, A., Earls, J.P.: Adaptive statistical iterative reconstruction: assessment of image noise and image quality in coronary ct angiography. American Journal of Roentgenology **195**(3), 649–654 (2010)
9. Lerousseau, M., Vakalopoulou, M., Deutsch, E., Paragios, N.: Sparseconvmil: Sparse convolutional context-aware multiple instance learning for whole slide image classification. In: MICCAI Workshop on Computational Pathology. pp. 129–139. PMLR (2021)

10. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14318–14328 (2021)
11. Li, X., Jin, X., Lin, J., Liu, S., Wu, Y., Yu, T., Zhou, W., Chen, Z.: Learning disentangled feature representation for hybrid-distorted image restoration. In: European Conference on Computer Vision. pp. 313–329. Springer (2020)
12. Littman, M.L.: Reinforcement learning improves behaviour from evaluative feedback. Nature **521**(7553), 445–451 (2015)
13. Liu, J., Li, X., Peng, Y., Yu, T., Chen, Z.: Swiniqa: Learned swin distance for compressed image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1795–1799 (2022)
14. Liu, J., Zhou, W., Xu, J., Li, X., An, S., Chen, Z.: Liqa: Lifelong blind image quality assessment. arXiv preprint arXiv:2104.14115 (2021)
15. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
16. Ma, X., Luo, G., Wang, W., Wang, K.: Transformer network for significant stenosis detection in ccta of coronary arteries. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 516–525. Springer (2021)
17. Myronenko, A., Xu, Z., Yang, D., Roth, H.R., Xu, D.: Accounting for dependencies in deep learning based multiple instance learning for whole slide imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 329–338. Springer (2021)
18. Nakanishi, R., Sankaran, S., Grady, L., Malpeso, J., Yousfi, R., Osawa, K., Ceponiene, I., Nazarat, N., Rahmani, S., Kissel, K., et al.: Automated estimation of image quality for coronary computed tomographic angiography using machine learning. European radiology **28**(9), 4018–4026 (2018)
19. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image li2021dsmilclassification. Advances in Neural Information Processing Systems **34** (2021)
20. Tang, Y., Tian, Y., Lu, J., Li, P., Zhou, J.: Deep progressive reinforcement learning for skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5323–5332 (2018)
21. Wolterink, J.M., van Hamersvelt, R.W., Viergever, M.A., Leiner, T., Išgum, I.: Coronary artery centerline extraction in cardiac ct angiography using a cnn-based orientation classifier. Medical image analysis **51**, 46–60 (2019)
22. Yu, S., Ma, K., Bi, Q., Bian, C., Ning, M., He, N., Li, Y., Liu, H., Zheng, Y.: Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 45–54. Springer (2021)
23. Zhao, Y., Yang, F., Fang, Y., Liu, H., Zhou, N., Zhang, J., Sun, J., Yang, S., Menze, B., Fan, X., et al.: Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4837–4846 (2020)
24. Zhou, Z.H.: A brief introduction to weakly supervised learning. National science review **5**(1), 44–53 (2018)