Deep learning for automated analysis of cardiac imaging: applications in Cine and 4D flow MRI

Sun, X.

Chapter 2 Combination special data augmentation and sampling inspection network for cardiac magnetic resonance imaging quality classification

This chapter was adapted from:

Xiaowu Sun, Li-Hsin Cheng, Rob J. van der Geest. **Combination Special Data Augmentation and Sampling Inspection Network for Cardiac Magnetic Resonance Imaging Quality Classification.** International Workshop on Statistical Atlases and Computational Models of the Heart (STACOM). Springer, Cham, 2022.

Abstract

Cardiac magnetic resonance imaging (MRI) may suffer from motion-related artifacts resulting in non-diagnostic quality images. Therefore, image quality assessment (IQA) is essential for the cardiac MRI analysis. The CMRxMotion challenge aims to develop automatic methods for IQA. In this paper, given the limited amount of training data, we designed three special data augmentation techniques to enlarge the dataset and to balance the class ratio. The generated dataset was used to pre-train the model. We then randomly selected two multi-channel 2D images from one 3D volume to mimic sample inspection and introduced ResNet as the backbone to extract features from those two 2D images. Meanwhile, a channel-based attention module was used to fuse the features for the classification. Our method achieved a mean accuracy of 0.75 and 0.725 in 4-fold cross validation and the held-out validation dataset, respectively. The code can be found here (https://github.com/xsunn/CMRxMotion).

2.1. Introduction

Cardiac magnetic resonance imaging (MRI) is considered as the standard reference for the evaluation of cardiac function due to its excellent image resolution and soft-tissue contrast. However, the MR scanner's hardware itself or the interaction of patient with hardware can result in artifacts in MRI, yielding a low quality imaging, which is often detrimental to the analysis of cardiac function especially in the large-scale imaging studies [1]. Although the artifacts can be minimized by carefully designed image protocols, they still cannot be fully eliminated [2]. Visual inspection of imaging quality is time-consuming and high-cost labor, and also relies on experienced radiologists. Therefore, an automatic method is needed to classify the MR image quality.

In the field of natural images, the approaches to image quality assessment (IQA) can be divided into two categories: full-reference and no-reference, depending on the availability of the original reference image. Meanwhile, recent Convolutional Neural Network (CNN) based methods, such as ResNet [3] and VGG [4], demonstrate promising performance in the automatic image classification task. Bosse et al. [5] employed a Siamese network to extract the features from the distorted and reference patch respectively and fused the difference of those features for IQA. Su et al. [6] proposed a self-adaptive hyper network to blindly assess image quality in the wild without any reference.

However, unlike IQA in natural images, in medical imaging it is particularly challenging for several reasons. There is no large-scale publicly available medical image dataset for IQA. In addition, the distinction between the diagnostic and non-diagnostic imaging is not always evident. Therefore, the labels annotated by radiologists are often subjective [7]. Previously, Fu tried to integrate the information from different color-spaces at feature-level and prediction-level to assess retinal image quality [8]. Oksuz et al. proposed a CNN model to automatically detect and correct motion-related artifacts in cardiac MRI using the K-space lines [9]. Lyu et al. used a recurrent generative adversarial network to reduce motion artifacts in cardiac MRI [10].

The CMRxMotion challenge aims to encourage the participants to develop an IQA model and a segmentation method for the extreme cardiac MRI dataset. In this paper we focuse only on the task of image quality assessment. Our contributions are as follows: (1) We designed specific data augmentation methods to enlarge the given limited data. (2) We proposed a two-branch network and combined a channel-based attention mechanism to fuse features from two random samples of the 3D volume, improving the IQA performance.

2.2. Dataset

The challenge provides short-axis cardiac MR images of 45 healthy volunteers (20 for training, 5 for validation and 20 for testing), obtained through the same 3T MR system (Siemens MAGNETOM Vida) under four different levels of respiratory motion, including full breath-hold, half breath-hold, free breath and intensive breath. Only the images of the end-diastolic (ED) and end-systolic (ES) phase are available. Therefore, there are 160 (20 volunteers × 4 scans × 2 phases), 40 and 160 3D volumes for training, validation and testing. The number of slices in one phase ranges from 9 to 13. The image resolution varies from 0.66×0.66×9.6 mm^3 to 0.76×0.76×10 mm^3, and the range of field of view (FOV) varies from 400×512 mm^2 to 512×512 mm^2. Independent from motion levels, all images were reviewed and scored by multiple radiologists using a standard 5-point Likert scale which can be found in https://www.synapse.org/#!Synapse:syn32407769/wiki/618241. For better reproducibility, the organizer divided those images into three classes based on the 5-point scores: mild motion, intermediate motion, and severe motion.

During data preprocessing, we excluded slices outside of the heart region and selected the 9 slices in the center to make each processed case having the same number of slices. Afterwards, all the cases were cropped or zero-padded into a uniform matrix size of 192×192×9 and the image intensity was normalized to [0,1] using the min-max method.

2.3. Methods

2.3.1. Data augmentation

In this section, we describe the specially designed data augmentation method for IAQ in detail. The first two strategies are based on weighted interpolation of images from the same subject, while the third strategy employs histogram matching plus interpolation to generate new images. All of the data augmentation methods are based on the 3D volume.

Generating transition phases between ED and ES. The ED and ES phases capture the two extreme scenarios in a cardiac cycle. The transition phases between ED and ES in the same cardiac cycle have almost identical intensity distribution [12]. Therefore, given the available ED and ES phases, we first generate new transition phases between ED and ES using weighted interpolation defined as following:

$$wp = wI_1 + (1-w)I_2 \ , \ n\_label \approx wL_1 + (1-w)L_2 \tag{2.1}$$

where *wp* is the generated volume and *n_label* is its corresponding label, I_1, I_2, and L_1, L_2 are the 3D volume and labels of ED and ES phases, and w is the weight. In this work, we used three values for w, namely 0.2, 0.5 and 0.8, to generate transition phases. Figure.2.1 shows an example of the generated images using this approach.

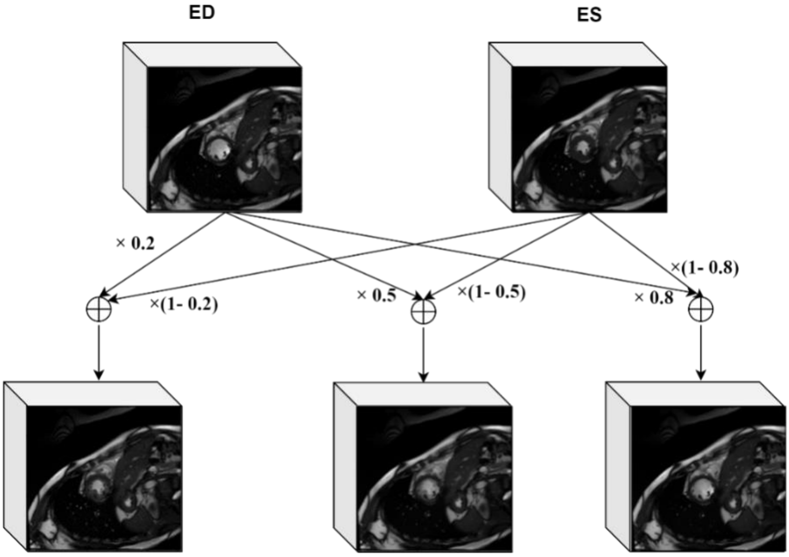

Figure.2.1. An example of generated new phases by weighting ED and ES phases from the same case. The first row implies the 3D volume selected from ED and ES phase. The second row presents 3D volume selected from the generated phases using the weights of 0.2, 0.5 and 0.8, respectively.

Generating intermediate images from different levels of respiratory motion. Interestingly, each volunteer was scanned four times under different levels of respiratory motion. Those paired cases from the same volunteer have the same anatomy structures but with different image qualities. Therefore, we used the paired images at the same phase but from different respiratory motion to generate new images. As illustrated in Figure.2.2, two paired images (P004-1-ED, P004-4-ED for example) both from the ED phase of the same volunteer, but possibly with different image qualities, are selected randomly as the source images. After an intensity-based registration, the method described in formula (2.1) is used to generate the new image and its corresponding label. Similar as the previous augmentation strategy, the new images are generated using weights of 0.2, 0.5 and 0.8.

Generating degraded images with histogram matching and linear interpolation. Within the 160 training cases, the numbers of cases with mild, intermediate, and severe motion artifacts are 70, 69, 21, respectively. To enlarge the subset with severe motion artifacts, the cases with intermediate artifacts were degraded into a lower-quality ones using the linear interpolation approach. As shown in Figure.2.3 a 3D volume with severe artifacts is randomly selected as the reference and another one with intermediate artifacts is considered as the source image. The pixel intensity distribution of the source image is matched to that of the reference image. We then randomly choose 5% of the pixels from the matched image, and apply the linear interpolation approach on those selected pixels to expand into a new image. The label of the generated image is assigned as severe.

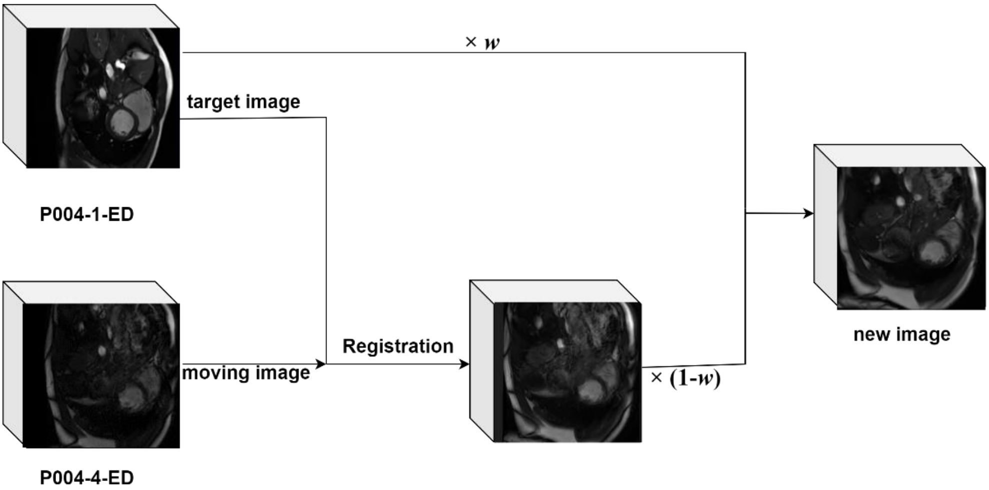

Figure.2.2. An illustration of using two ED phases under different respiratory motion levels of the same volunteer to generate a new image.

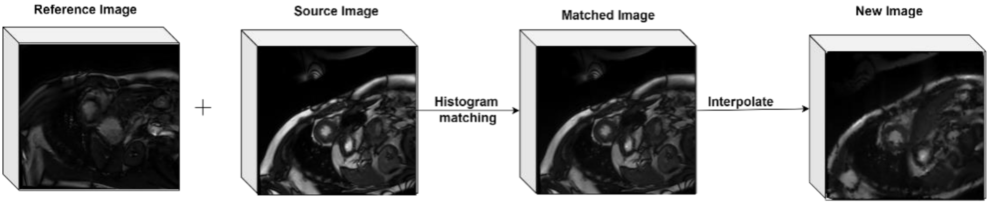

Figure.2.3. The procedure of image degradation includes histogram matching and linear interpolation. The generated result is considered as a new image with lower quality.

2.3.2. Sampling Inspection Network Architecture

To mimic the sampling inspection, the quality of a 3D volume is determined by estimating the quality of random samples drawn from the volume. The advantage of the random selection strategy is that it can generate more data from a single volume to train a model. However, because the selected sample occasionally missed certain critical slices, we introduced another sample with different combinations of 2D slices, mimicking ensembling two times of sampling inspection. The model architecture is shown in Figure.2.4.

The two samples were regarded as 2D multi-channel images, and were each processed by a convolution layer. In addition, according to our intuition, the slices from the apical, middle and basal regions contribute differently for IQA. The slices in the middle section, with a relatively larger size of the left ventricle than those in apex and base, have a significant impact on the quality assessment. Therefore, the channel attention module (CAM) proposed in [11] was introduced to explore the intra-channel relationship of the input. After that, ResNet was introduced as a backbone to extract the features for each input. The features from those two branches were concatenated along the channel dimension, and a Feature Fusion Module (FFM) was introduced to fuse those features. The FFM block contains one channel attention module to explore the inter-channel relationship and one averaged pooling layer to extract the global information. Lastly, a fully connected layer was used to predict the result.

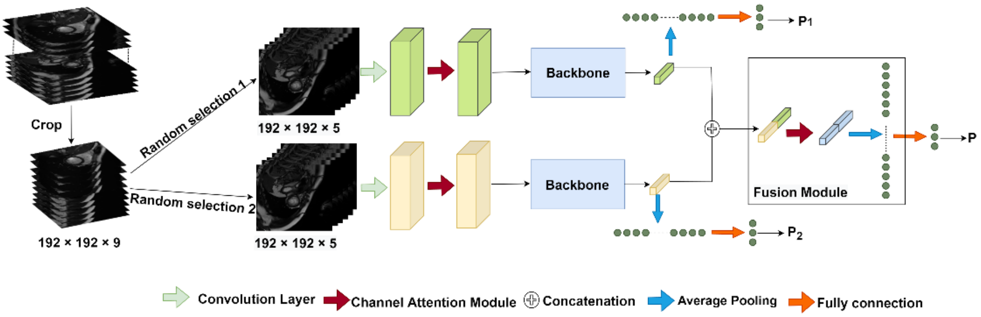

Figure.2.4. Sampling inspection network architecture. The input of each branch is a multi-channel 2D image.

Inspired by the idea of deep supervision [13], besides the final prediction P, each branch also has one prediction denoted as P_1 and P_2. Therefore, the total loss function can be expressed as:

$$Loss = CE(P, L) + \sum_{i=1}^{2} CE(P_i, L) \tag{2.2}$$

where CE is cross-entropy loss. Only the prediction P was used for the validation and testing. During the validation and testing, the sampling inspection was repeated 50 times for one 3D volume, the averaged result was regarded as the final prediction.

2.4. Experiments and results

The training data was divided into 4-fold for cross validation. The metrics, including accuracy, precision, recall, F1-Score and Cohen's Kappa were used to evaluate the performance. All the results were reported as the mean value of four folds. All the experiments were implemented with Pytorch trained on a machine with a NVIDIA Quadro RTX 6000 GPU with 24 GB memory. Adam was employed as the optimizer with 0.00001 as the learning rate.

The ResNet was employed as the baseline, and it took a multi-channel 2D image with a size of 192×192×9 as the input. We first compared our method with the baseline. Due to the small and imbalanced dataset, the baseline failed to predict the severe class, yielding a relatively poor result with accuracy being 0.41, as reported in Table.2.1. The performance indicated that a larger and balanced dataset is needed.

Table.2.1. The 4-fold cross validation performance. Over-Acc: the overall accuracy based on all classes. DA: data augmentation. P: Precision. R: Recall. F: F1-Score.

Model	DA	Over-Acc	Cohen's Kappa	Severity Level	Acc	P	R	F
ResNet	No	0.41	-0.04	Mild	0.76	0.42	0.76	0.54
				Intermediate	0.19	0.38	0.19	0.25
				Severe	0.00	0.00	0.00	0.00
Ours	Yes	**0.75**	**0.58**	Mild	0.77	**0.83**	0.77	**0.80**
				Intermediate	0.81	0.69	0.81	0.75
				Severe	0.48	0.71	0.48	0.57
	No	0.68	0.45	Mild	**0.86**	0.74	**0.86**	0.79
				Intermediate	0.68	0.64	0.68	0.66
				Severe	0.10	0.40	0.10	0.15

We also evaluated the performance of the proposed network and data augmentation (DA) techniques. The new data generated from the offline data augmentation approach was used to pre-train the model and the pre-trained model was fine-tuned using the original training data. Table.2.1 also reports the classification results derived from the proposed methods. It shows that the overall accuracy increased from 0.68 to 0.75 after using DA. Although on the class of mild, the accuracy using DA is a little lower than that without DA, the accuracy for the other two classes is better. The method using DA achieved the best performance on the metric of F1-Score in all classes. The confusion matrix in Figure.2.5 further reveals that the number of false negatives for the severe class reduced after introducing DA. Therefore, the performance confirmed that the carefully designed DA works well for the IQA task.

For the validation part, the labels were hidden by the organizer, we submitted our predicted results and evaluated the performance online. Our method achieved a competitive results, yielding accuracy of 0.725 and Cohen's Kappa 0.645. The best model in the validation data was chosen as the final model, and we submitted it to the organizer and evaluated the performance in the testing dataset with 120 image volumes [14], achieving accuracy of 0.6417 and Cohen's Kappa 0.456.

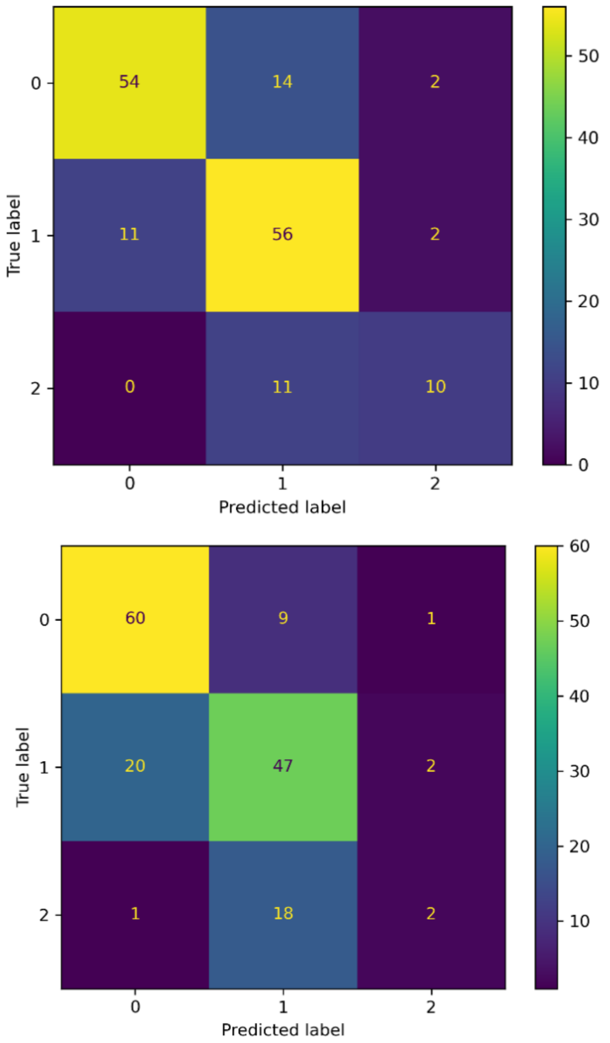

Figure.2.5. Confusion matrix derived from the proposed network. 0, 1, 2 represent the classes of mild, intermediate and severe. The upper one is the result using data augmentation and the bottom one is the result without data augmentation.

Ablation. In the proposed classification network, a module named FFM was used to fuse the features from two branches. To reveal the effectiveness of FFM, we evaluated the accuracy and confusion matrix derived from the three predictions P, P_1, P_2 as reported in Table.2.2, Figure.2.5 and Figure.2.6. P_1, P_2 were derived from two individual braches, while P was generated using the FFM. Compared with the other two predictions, P achieved the best performance on all those classes and the overall.

Table.2.2. Comparison of the accuracy for each class derived from different branches.

	P	P_1	P_2
Mild	**0.77**	0.71	0.73
Intermediate	**0.81**	0.79	0.80
Severe	**0.48**	0.29	0.29
Overall	**0.75**	0.69	0.71

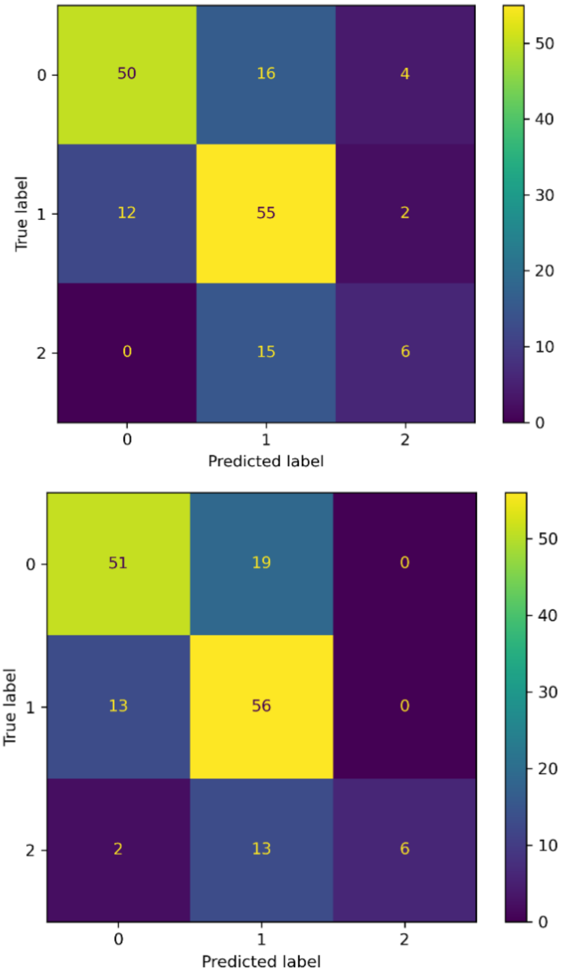

Figure.2.6. Confusion Matrix of two predictions P_1, P_2. The upper one is from the result P_1, the bottom one is derived from P_2.

2.5. Conclusion

In this paper, we designed three data augmentation methods to enlarge the dataset and balance the classes for the cardiac MR image quality assessment task. Inspired by the idea of sample inspection, to enlarge the training data and to extract sufficient features, we randomly selected different combinations of 2D slices as the input of each branch of the network. The proposed method was trained and evaluated using four-fold cross validation. The results of the classification accuracy, precision, recall and F1-Score demonstrate that our method performed better than the baseline, and the results on the validation dataset shows a competitive performance against the other participants' methods.

Declaration. The authors of this paper declare that they did not use any additional medical image datasets other than those provided by the organizers. They also would like to acknowledge the organizer of the CMRxMotion challenge for collecting and sharing the dataset.

References

1. Krupa, Katarzyna, and Monika Bekiesińska-Figatowska. "Artifacts in magnetic resonance imaging." Polish journal of radiology 80 (2015): 93.
2. Zhang, Le, et al. "Automated quality assessment of cardiac MR images using convolutional neural networks." International Workshop on Simulation and Synthesis in Medical Imaging. Springer, Cham, 2016.
3. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
4. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
5. Bosse, Sebastian, et al. "Deep neural networks for no-reference and full-reference image quality assessment." IEEE Transactions on image processing 27.1 (2017): 206-219.
6. Su, Shaolin, et al. "Blindly assess image quality in the wild guided by a self-adaptive hyper network." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
7. Ma, Jeffrey J., et al. "Diagnostic image quality assessment and classification in medical imaging: Opportunities and challenges." 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, 2020.
8. Fu, Huazhu, et al. "Evaluation of retinal image quality assessment networks in different color-spaces." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019.
9. Oksuz, Ilkay, et al. "Detection and correction of cardiac MRI motion artefacts during reconstruction from k-space." International conference on medical image computing and computer-assisted intervention. Springer, Cham, 2019.
10. Lyu, Qing, et al. "Cine cardiac MRI motion artifact reduction using a recurrent neural network." IEEE Transactions on Medical Imaging 40.8 (2021): 2170-2181.
11. Wang, Qilong, et al. " ECA-Net: Efficient channel attention for deep convolutional neural networks." Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Seattle, WA, USA. 2020.
12. Zhang, Yao, et al. "Semi-supervised cardiac image segmentation via label propagation and style transfer." International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, Cham, 2020.
13. Wang L, Lee CY, Tu Z, Lazebnik S. Training deeper convolutional networks with deep supervision. arXiv preprint arXiv:1505.02496. 2015.
14. Wang S, Qin C, Wang C, Wang K, Wang H, Chen C, et al. “The Extreme Cardiac MRI Analysis Challenge under Respiratory Motion (CMRxMotion)”. arXiv preprint arXIv: 2210.06385 (2022)