2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-7358-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISBI53787.2023.10230625

A ROBUST AGATSTON SCORE FOR CORONARY ARTERY CALCIUM SCORING FROM NON-ECG-GATED CT WITH DIFFERENT RECONSTRUCTION KERNELS

Shuo Zhang[1,2*], *Xinmei Su*[2,4*], *Quanlong Feng*[2,3], *Zifeng Wu*[2 †], *Zhenghua Xu*[1]

[1]State Key Laboratory of Reliability and Intelligence of Electrical Equipment,
School of Health Sciences and Biomedical Engineering, Hebei University of Technology, China
[2] DeepWise AI Lab, Beijing, China
[3]College of Land Science and Technology, China Agricultural University, Beijing, China
[4]School of Information and Electronics, Beijing Institute of Technology, Beijing, China

ABSTRACT

A coronary artery calcium score (CACS) is a vital measure to screen individuals at risk for early coronary heart disease. However, as the main evaluation system of CACS, the Agatston score computed from CT images via HU-thresholding may vary significantly even for the same individual as the protocol of image acquisition changes (e.g., reconstruction kernels). This may harm the compatibility of CACS, when evaluated in different health facilities at different times. To tackle this issue, we propose the robust Agatston score (RAS), wherein we predict the calcification level per pixel via deep learning, rather than directly thresholding the HU value from CT images, as we do for the classic Agatston score. In this way, we make the CACS more robust to the change of acquisition protocols, and let the comparison among CACS from various sources easier. Experimental results show that our method can improve the CACS level accuracy from 64.21% to 95.78%. Code is available at https://github.com/lucas-dw/ras.

***Index Terms*—** Coronary Artery Calcium Score, reconstruction kernel, deep learning, CT image, Agatston score

1. INTRODUCTION

The prevalence of the Cardiovascular Disease (CVD) is 49.2% overall in recent years and it increases with age [1]. This kind of disease often increases the incidence of angina pectoris, myocardial infarction, heart failure and sudden cardiac death, which is a potential threat in daily life. The coronary artery calcium score (CACS) obtained by computed tomography (CT) scanning is a common indicator for assessing the risk of cardiovascular disease, wherein the Agatston score (AS) is the main evaluation system [2, 3].

However, various acquisition methods of the CT images, such as different reconstruction kernels, can affect the CACS drastically[4], where soft kernels (i.e., B20f) improved CACS accuracy while sharp kernels (i.e., B46f) degraded the precision [5]. This is because in the classic Agatston score, calcification level of a certain pixel is determined via HU-thresholding, therefore, the change of CT image acquisition protocols would lead to the variability of HU value even for the same patient, leading to incomparability of Agatston scores (Fig. 1). In a word, there is an obvious need for a more reliable method than the classic Agatston score, which is more robust against the change of acquisition protocols, especially for the variability of reconstruction kernels. This fact inspires our proposed robust Agatston score (**RAS**).

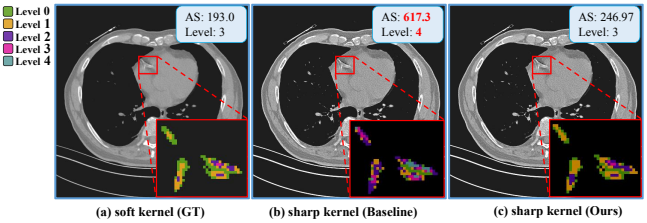

Fig. 1. Overview of proposed method (c) from sharp kernel, and comparison using the classic Agatston score with (a) from a soft kernel (GT), and (b) from sharp kernel (Baseline). AS refers to Agatston score and Level refers to CACS level, both of which are calculated from a total CT series.

Deep learning has been popular for semantic segmentation in medical images these years, such as the well-known U-Net [6]. Several deep models are used to locate the coronary calcification regions in non-ECG-gated CT imagess [7, 8, 9], and naturally, to compute the CACS [10]. Generally, calcification regions are predicted by a U-Net, then the calcification levels (from level 0 to level 4) are obtained by HU threshold-based method. These methods could yield promising results when acquisition protocols are stable, in the sense of directly comparing one CACS to another, e.g. in a follow-up study, but will probably fail when acquisition protocols vary largely. To this end, one possible pipeline would be generating soft and thin series (ideal for computing CACS) from any original series using deep-learning models such as GAN [11], and

*Authors contributed equally.
†Corresponding author: wuzifeng@deepwise.com.

then calculate the classic Agatston score. Although this work reduces CACS errors, it deciphers the original data structure information and increases unreliability, while still being affected by the reconstructed kernels. However, we find that it is possible to directly predict the pixel-wise calcification level from original CT series, which is the very approach that we adopt to realize RAS against various CT image acquisition protocols.

The main contribution of this work is three-fold: 1) We proposed directly predicting the calcification level using deep semantic segmentation models with designed multi-losses to realize the robust Agatston score (RAS) against any reconstruction kernels. 2) We performed a detailed ablation study of our method to verify the contribution of each component, and verify the generalization for different segmentation networks. 3) We conducted extensive experiments and showed that our method can greatly improve the calcification level accuracy compared to the HU-based classic Agatston score.

2. METHODS

2.1. Method Overview

We propose RAS (Fig. 2) against the change of different reconstruction kernels in CT image acquisition and compare with the HU threshold-based method for AS acquisition. As shown in Fig. 2, first, we utilize the CT series reconstructed by soft kernel and the well-labeled masks that locate the calcification regions to create the ground-truth masks (GT masks), which contain the calcification level for each pixel. Second, we adopt a series of deep semantic segmentation models to predict the calcification level of each pixel in CT scans of sharp kernel. Third, we adopt various loss functions to further improve the performance. Finally, we use the predicted calcification level to calculate the Agatston score as RAS and compare with the classical Agatston score (as our baseline).

2.2. Robust Calcification Level Prediction

The classical Agatston score from the HU-threshed CT images usually does not perform well in the face of changes in reconstruction kernels (especially sharp kernels). To improve the robustness of the Agatston score, we adopt the deep model to directly predict pixel-wise calcification level. Specifically, a total of five-class calcifications are considered ranging from level 0 to level 4. Therefore, the calcification level estimation has been transferred to a semantic segmentation task, where the variations among different CT image acquisition protocols could be eliminated by deep learning. We assume that the input $x \in \mathbb{R}^{H\times W}$ is a CT image. Then, we define a segmentation model $F(\cdot)$ to obtain a predicted calcification level map $\hat{z}$ followed by a softmax activation function of the same resolution as x. Further, we optimize the model using cross-entropy loss and Dice loss as follows:

$$\mathcal{L} = \mathcal{L}_{\text{ce}}\left(\hat{z}, z\right) + \mathcal{L}_{\text{dice}}\left(\hat{z}, z\right), \tag{1}$$

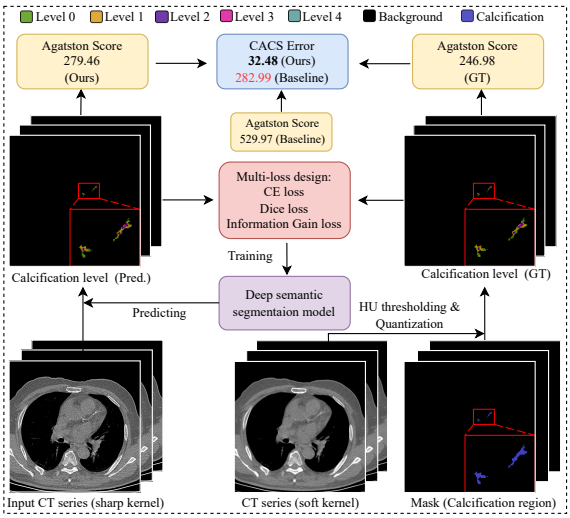

Fig. 2. Overview of proposed method and comparison with the classic CACS directly from sharp series (baseline).

where $\hat{z} = F(x)$ and $z \in \{0,1\}^{5\times H\times W}$ is the ground-truth with 5 classes. In addition, U-Net is mainly adopted due to its high-performance and robustness to noise in medical image analysis. Besides, other segmentation models such as SegResNet [12] and AttentionUnet [13] are also considered to further justify the generalization capability of RAS.

2.3. Information Gain Loss

Since we treat the calcification level prediction as a task of five-class quantization estimation, the loss from depth estimation could be another alternative for this very task. To justify this hypothesis, we also use the information gain loss (IGL) [14] that is widely used in depth estimation to achieve better results. The rational of IGL lies in that predictions close to ground-truth labels can also help in updating network parameters. IGL is defined as:

$$\mathcal{L}_{IG} = -\frac{1}{V}\sum_{i=0}^{V-1}\sum_{D=0}^{N-1} M(D_i^*, D) log(P(D|z_i)), \tag{2}$$

where V is the total amount of the pixels of CT images, N is the number of calcification levels and D is the calcification level. D_i^* is the ground-truth labels of each pixel. z_i is the output of the last convolutional layer in the network. The metric $P(D|z_i)$ refers to the probability of each pixel labeled with D. The M metric is defined as: $M(D_i^*, D) = e^{-\alpha(D_i^* - D)^2}$, where α is a constant parameter. By using IGL, it encourages the predicted calcification levels that are closer to ground-truths have higher contributions in updating network parameters, which provides a better way to utilize the quantized calcification labels. Finally, we use CE loss, Dice loss and IG loss to optimize the segmentation model.

Table 1. Comparison of our method with the baseline for sharp reconstruction kernels on the test set.

Methods	Dice	Prec.	Sens.	CACS Acc	CACSE	CACSRE
baseline	0.2945	0.3986	0.5793	0.6421	119.8728	9.7569
Ours	**0.8764**	**0.8993**	**0.9044**	**0.9578**	**13.1270**	**0.1208**

Table 2. Comparison of our method with the baseline for each CACS level (L0, L1, L2, L3 and L4) on the test set.

Methods	CACSE					CACSRE				
	L0	L1	L2	L3	L4	L0	L1	L2	L3	L4
baseline	3.87	19.04	59.14	115.98	401.30	-	7.45	1.59	0.61	0.33
Ours	**0.0**	**1.14**	**4.51**	**12.87**	**47.09**	-	**0.38**	**0.11**	**0.07**	**0.03**

3. EXPERIMENTAL SETTING

3.1. Data Preparation

The image data used in our experiments are retrospectively collected from multiple medical facilities, taken by CT scanners made by various manufacturers. To evaluate the performance of the robust Agatston score (RAS) against the change of reconstruction kernels, we carefully build up one dataset that considers as various conditions as possible. These factors include the age, sex and calcification level (i.e., CACS level) for an individual. For all the CT scans, we only use the studies having multiple thin-slice series (with slice thicknesses no greater than 3mm) reconstructed using soft and sharp kernels respectively. We manually label the regions of coronary calcification on soft series, and reuse them for the sharp ones. The total dataset consists of 951 CT scans and we hold about 10% to make a test set, which is not accessible during model training and for final accuracy assessment. Similarly, we hold another 10% to make a validation set for online accuracy evaluation and best model selection. In addition, both patient's age, sex and calcification level are considered during the split of train, test and validation to make a balanced dataset.

As for the ground-truth of the calcification level prediction task, we firstly selected a calcification sub-region inside the manually labeled coronary calcification masks according to the HU value of CT image pixels (CT_i) and the preset threshold, and then quantify them into five pixel-wise calcification levels, including level 0 ($CT_i \leq 130\ HU$), level 1 ($130\ HU \leq CT_i < 200\ HU$), level 2 ($200\ HU \leq CT_i < 300\ HU$), level 3 ($300\ HU \leq CT_i < 400\ HU$) and level 4 ($CT_i \geq 400\ HU$). Afterwards, we treat the calcification level per pixel in soft series as the ground-truth, which is stable and has a high-resolution. In contrast, those in sharp series are noisy, which are not preferred when computing CACS. To bridging this gap, we feed a network with sharp series, and let it learn to predict the calcification level masks obtained from soft series to make the calcification level prediction more robust against various CT image acquisition protocols.

Besides, CT kernel together with slice spacing, pixel spacing and CT manufacturer witness a great variety in this study. CT kernels include sharp kernels (e.g., B_SHARP_C, B_VSHARP_C, FC51, FC52, LUNG) and soft kernels (e.g., B_SOFT_B, FC02, FC03, FC18, STANDARD), while CT manufacturers include GE, SIEMENS, TOSHIBA and UIH. Slice spacing ranges from 0.7 mm to 3 mm while pixel spacing ranges from 0.46 mm to 0.94 mm.

3.2. Evaluation Metrics

The adopted evaluation metrics include both segmentation-based and Agatston score-based metrics. Segmentation-based metrics include the widely used Dice coefficient (**Dice**), Precision (**Prec.**) and Sensitivity (**Sens.**). Agatston score-based metrics include: a) the accuracy of the CACS level (**CACS Acc**), which is the proportion of correctly predicted CACS levels in the whole dataset, b) the error of CACS (**CACSE**), which is the absolute error between the predicted CACS and the ground-truth one, c) the relative error of CACS (**CACSRE**), which is the CACS error divided by the ground-truth CACS. As the main evaluation system of CACS, the Agatston score [15] per slice is computed as: $Agatston\ score = f \times p \times s$, where f is the density factor (calcification level), p is the pixel number and s is the area per pixel, given that slice thickness is 3mm. In other case, we can re-weight it by a factor of (slice thickness) / 3.

3.3. Implementation Details

In this work, all experiments are implemented using PyTorch 1.7 on two TITAN Xp GPUs. We build up the whole pipeline on top of the fastai [16] toolkit. A fixed random seed is set to mantain the evaluation results reproducible. We use three segmentation networks including U-Net [6], SegResNet [12] and AttentionUnet [13] to verify the robustness and generalization of our method, which are implemented from MONAI[1]. AdamW is chosen as the optimizer with an initial learning rate of 1e-3. For data augmentation, random rotation and random scaling are used in the training phase. We discard color-based data augmentation, which is not suitable for this task and hurts calcification level prediction accuracy. More experimental setups can be found in the public code repository.

4. EXPERIMENTAL RESULTS

4.1. Quantitative Results

We evaluate the proposed deep segmentation model-based method and the classic HU threshold-based Agatston method (baseline) for the change of reconstruction kernels on the test set. We show the results of the mean values of all CACS levels for each metric in Table 1 and CACSE, CACSRE for each CACS level in Table 2, as well as confusion matrices in Fig. 4, which all demonstrate significant and consistent improvements compared to the baseline. From Table 1, due to a large amount of noise in the sharp kernel, the prediction

[1]https://monai.io/

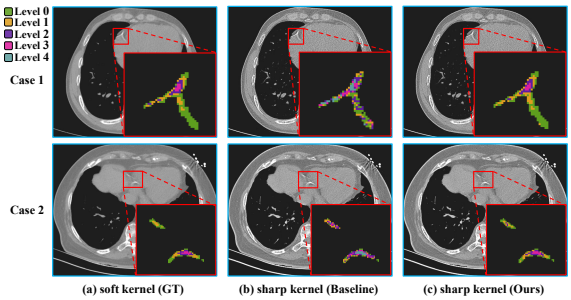

Fig. 3. The Visual comparison of the CT images and masks among different kernels for baseline and our method.

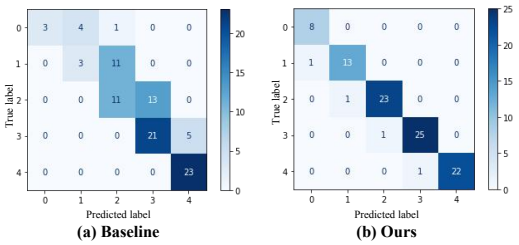

Fig. 4. Confusion matrices of the CACS level prediction for sharp reconstruction kernels.

accuracy of the calcification level of baseline is unacceptable, and the Dice metric is only 0.2945. However, our method can accurately predict the calcification level, with Dice reaching 0.8764. Meanwhile, our CACS level accuracy is as high as 0.9578, which is 0.3157 higher than baseline. For each CACS level error in Table 2, our method achieves extremely small errors, while the baseline errors are usually huge. Similarly, in the confusion matrices in Fig. 4, our method has high prediction accuracy for each level for 95 cases in the test set. These results illustrate the robustness of our method in the face of noisy reconstruction kernels.

4.2. Qualitative Results

As shown in Fig. 3, (a) and (b) are images and masks among the change of different kernels using HU threshold-based method. (b) and (c) are images and masks for HU threshold-based and our method under the same sharp kernel. The image of the soft kernel (a) is smooth but the image of the sharp kernel (b) is rough, and the masks (GT and baseline) obtained from their images vary significantly. We can clearly see that the calcification levels obtained by sharp kernel using HU threshold-based method have serious noise, which greatly impairs the prediction of CACS. However, the mask obtained by our proposed model can better predict the calcification level per pixel. This fully demonstrates that our model can well resist the change of the reconstruction kernel, greatly improving the robustness of the Agatston score.

Table 3. Comparison with the classic Agatston score (baseline) and ablation study on our method on the test set.

Method	Dice	Prec.	Sens.
baseline	0.2945	0.3986	0.5793
U-Net	0.8659	0.8881	0.8988
U-Net + IGL + LDA	0.8444	0.8688	0.8848
U-Net + IGL (Ours)	**0.8764**	**0.8993**	**0.9044**

4.3. Ablation Study

To evaluate the impact of each component in our method, we show the result of ablation study in Table 3. Our method with all the considered configurations outperforms the baseline significantly, showing the effectiveness of our method against the change of reconstruction kernels. First, the U-Net segmentation model can accurately predict the calcification level of each pixel compared to the baseline despite under the sharp kernel. Second, with information gain loss (IGL), our method performs even better. This shows the importance of encouraging the calcification levels that are closer to their ground-truths. Finally, although the lighting data-augmentation (LDA) also improves the performance, it has a significant performance drop compared to U-Net + IGL. This illustrates that color-based data augmentation is ineffective or even harmful for such image brightness-related tasks in the training of deep models.

Furthermore, to demonstrate that our method also generalizes well under different segmentation models, we additionally evaluate two widely used segmentation networks, SegResNet [12] and AttentionUnet [13] on the test set in Table 4. We can see that our method, regardless of the segmentation network used, can significantly improve the performance of the baseline. In addition, different segmentation networks have little effect on our method, which shows that our proposed method has good generalization.

Table 4. Comparison with different segmentation models.

Method	Dice	Prec.	Sens.
baseline	0.2945	0.3986	0.5793
Ours (U-Net [6])	0.8764	0.8993	0.9044
Ours (SegResNet [12])	0.8806	0.9044	0.9057
Ours (AttentionUnet [13])	0.8798	0.8956	0.9144

5. CONCLUSIONS

We have proposed the robust Agatston score (RAS), and shown its stability against the change of reconstruction kernels. We have built up the whole pipeline on top of U-Net, wherein we have studied the respective impact of data-augmentation and training loss. In practice, this technique can enhance the compatibility of CACS, which could probably increase its reliability and would possibly be in clinical applications and researches such as multi-center studies that involve various CT image acquisition protocols.

Discussion: The method proposed in this paper can only model the CT of the same shot, and cannot predict the calcification score for CT of different shots and changes in the heart position. In addition, we only consider CT series with different kernels, however, the slice spacing also impacts the robustness of CACS values. Therefore, a future study should consider the effect of slice spacing variances.

6. COMPLIANCE WITH ETHICAL STANDARDS

We state that our work is a retrospective study for which no ethical approval was required. This study has obtained the local committee approval, and all data are retrospectively collected and have been anonymized to ensure the privacy of patients before study.

7. ACKNOWLEDGMENTS

This work is funded by the National Natural Science Foundation of China (No. 81971616, 62076218, 82072005, 82171934, 82272085) and the Beijing Municipal Science and Technology Planning Project (No. Z201100005620008, Z211100003521009).

8. REFERENCES

[1] S.S. Virani, A. Alonso, H.J. Aparicio, E.J. Benjamin, M. S. Bittencourt, C.W. Callaway, A.P. Carson, et al., "Heart disease and stroke statistics—2021 update: a report from the american heart association," *Circulation*, vol. 143, pp. 254–743, 2021.

[2] M. Petretta, S. Daniele, W. Acampa, M. Imbriaco, T. Pellegrino, G. Messalli, E. Xhoxhi, G. Del Prete, C. Nappi, D. Accardo, et al., "Prognostic value of coronary artery calcium score and coronary ct angiography in patients with intermediate risk of coronary artery disease," *The international journal of cardiovascular imaging*, vol. 28, pp. 1547–1556, 2012.

[3] M.J. Blaha, J. Yeboah, M. Al Rifai, K. Liu, R. Kronmal, and P. Greenland, "Providing evidence for subclinical cvd in risk assessment," *Global heart*, vol. 11, pp. 275–285, 2016.

[4] S. An, R. Fan, B. Zhao, Q. Yi, S. Yao, X. Shi, Y. Zhu, X. Yi, and S. Liu, "Evaluating coronary artery calcification with low-dose chest ct reconstructed by different kernels," *Clinical Imaging*, 2022.

[5] S. Achenbach, K. Boehmer, T. Pflederer, D. Ropers, M. Seltmann, M. Lell, K. Anders, et al., "Influence of slice thickness and reconstruction kernel on the computed tomographic attenuation of coronary atherosclerotic plaque," *Journal of cardiovascular computed tomography*, vol. 4, pp. 110–115, 2010.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[7] C. Cano-Espinosa, G. González, G.R. Washko, M. Cazorla, and R.S.J. Estépar, "Automated agatston score computation in non-ecg gated ct scans using deep learning," in *Medical Imaging 2018: Image Processing*. International Society for Optics and Photonics, 2018, vol. 10574, p. 105742K.

[8] H. Lee, S. Martin, J.R. Burt, P.S. Bagherzadeh, S. Rapaka, H.N.Gray, et al., "Machine learning and coronary artery calcium scoring," *Current Cardiology Reports*, vol. 22, pp. 1–6, 2020.

[9] Nicolas Gogin, Mario Viti, Luc Nicodème, Mickaël Ohana, Hugues Talbot, Umit Gencer, Magloire Mekukosokeng, Thomas Caramella, Yann Diascorn, Jean-Yves Airaud, et al., "Automatic coronary artery calcium scoring from unenhanced-ecg-gated ct using deep learning," *Diagnostic and Interventional Imaging*, vol. 102, no. 11, pp. 683–690, 2021.

[10] W. Wang, H. Wang, Q. Chen, Z. Zhou, R. Wang, N. Zhang, Y. Chen, Z. Sun, and L. Xu, "Coronary artery calcium score quantification using a deep-learning algorithm," *Clinical Radiology*, vol. 75, pp. 237.e11–237.e16, 2020.

[11] Q. Liu, Z. Zhou, F. Liu, X. Fang, Y. Yu, and Y. Wang, "Multi-stream progressive up-sampling network for dense ct image reconstruction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 518–528.

[12] Andriy Myronenko, "3d mri brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 311–320.

[13] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al., "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[14] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, pp. 3174–3182, 2017.

[15] M.J. Blaha, M.B. Mortensen, S. Kianoush, R. Tota-Maharaj, and M. Cainzos-Achirica, "Coronary artery calcium scoring: is it time for a change in methodology?," *JACC: Cardiovascular Imaging*, vol. 10, pp. 923–937, 2017.

[16] J. Howard and S. Gugger, "Fastai: a layered api for deep learning," *Information*, vol. 11, pp. 108, 2020.