





EchoCoTr: Estimation of the Left Ventricular Ejection Fraction from Spatiotemporal Echocardiography

Rand Muhtaseb^(✉)  and Mohammad Yaqub 

Mohamed Bin Zayed University of Artificial Intelligence,
Abu Dhabi, United Arab Emirates
{rand.muhtaseb,mohammad.yaqub}@mbzuai.ac.ae

Abstract. Learning spatiotemporal features is an important task for efficient video understanding especially in medical images such as echocardiograms. Convolutional neural networks (CNNs) and more recent vision transformers (ViTs) are the most commonly used methods with limitations per each. CNNs are good at capturing local context but fail to learn global information across video frames. On the other hand, vision transformers can incorporate global details and long sequences but are computationally expensive and typically require more data to train. In this paper, we propose a method that addresses the limitations we typically face when training on medical video data such as echocardiographic scans. The algorithm we propose (EchoCoTr) utilizes the strength of vision transformers and CNNs to tackle the problem of estimating the left ventricular ejection fraction (LVEF) on ultrasound videos. We demonstrate how the proposed method outperforms state-of-the-art work to-date on the EchoNet-Dynamic dataset with MAE of 3.95 and R^2 of 0.82. These results show noticeable improvement compared to all published research. In addition, we show extensive ablations and comparisons with several algorithms, including ViT and BERT. The code is available at <https://github.com/BioMedIA-MBZUAI/EchoCoTr>.

Keywords: Transformers · Deep learning · Echocardiography · Ejection fraction · Heart failure

1 Introduction

In medical imaging, there are different imaging modalities that are crucial to real-time clinical assessment and visualization. An example of this is echocardiography, which produces spatiotemporal data made of a sequence of two-dimensional (2D) images. When dealing with spatiotemporal data, it is essential to learn the spatial information as well as take into account the temporal factor in these sequences for an accurate diagnosis. In order to detect abnormalities and certain diseases, cardiologists also tend to take into consideration the temporal information when measuring the left ventricular ejection fraction (LVEF) or while

assessing heart wall motion [5]. LVEF can be measured as the difference in the left ventricle volume at end-diastole and end-systole divided by the end-diastolic volume estimated from the apical four-chamber (a4c) or apical-two chamber (a2c) views of the heart. LVEF is an important biomarker that can predict heart failure (HF), which is a serious condition that can be caused when the heart cannot pump enough blood and consequently, oxygen to other parts of the body. In 2018, heart failure contributed to 13.4% of the recorded deaths in the United States [16]. Early diagnosis of HF will help cardiologists prescribe medications and encourage patients to have effective lifestyles [18]. Heart failure is typically diagnosed if LVEF is less than the normal range (50–80%). Echocardiography is the most common imaging modality used to assess cardiac function by measuring the left ventricle volume, wall thickness and LVEF since it is real-time, low-cost, ionizing radiation free, portable and a highly sensitive tool compared to other modalities. However, ultrasound technology has many drawbacks, such as operator-dependence, noise, artifacts and decreased contrast that may affect its quality which could lead to a high inter- and intra- observer variability in the diagnosis [17].

In this paper, we study the impact of different CNNs and transformer models to estimate left ventricle ejection fraction (LVEF) from ultrasound videos. Convolutional neural networks (CNNs) have shown great success when training the models to tackle problems in medical or natural images. However, vision transformers have shown that they may be good contenders to CNNs when solving certain image analysis problems. There are major differences between the two approaches. CNNs have limited receptive fields in the initial layers, but can progressively enlarge the field of view through convolution operations. In contrast, vision transformers (ViTs), can have the entire field of view starting from the initial layers through the self-attention process. However, unlike CNNs, ViTs do not have inductive bias and hence typically require a large amount of data to train on which is not always available especially in medical imaging. A research study shows that the initial layers of a ViT cannot acquire local information if the dataset is small, which highly impacts the model accuracy [9]. Hence, having a method that combines the strengths of both CNNs and ViTs, to work efficiently with spatiotemporal data in medical imaging assessment, is of great value.

Our contribution in this work is three fold:

- We propose EchoCoTr (**E**cho **C**onvolutional **T**ransformer) which is a method that is able to analyze echocardiography video sequences by combining the strength of CNNs and vision transformers to accurately estimate the heart’s ejection fraction. Even though EchoCoTr is adapted from UniFormer [6] which worked on natural video datasets, some changes were made to address the challenging problems we face such as proper frame sampling.
- We show how our proposed method outperforms all published work to-date on a large scale public dataset [8,10], which does not require: 1) information regarding the position of end-systolic (ES) and end-diastolic (ED) frames, 2) segmentation masks as EchoNet-Dynamic’s beat-to-beat pipeline [8], and 3) a pre-defined length of the cardiac scan.

- We compare our proposed method with several existing deep learning algorithms and perform thorough ablation studies to provide a deep discussion of the results.

2 Related Works

Many research papers [11,13,14,19] were introduced to improve the segmentation of the left ventricle to accurately estimate ejection fraction. Silva et al. [12] used a 3D CNN with residual learning blocks to estimate ejection fraction from transthoracic echocardiography (TTE) exams. Ouyang et al. [8] proposed a deep learning approach to estimate the beat-to-beat ejection fraction and predict heart failure with reduced ejection fraction (HFrEF) by combining the semantic segmentation results and the clip-level ejection fraction prediction using spatiotemporal CNN [15]. Recently, Reynaud et al. [10] proposed a transformer model based on residual auto-encoder to reduce the dimensions followed by Bidirectional Encoder Representations from Transformers (BERT) for end-systolic (ES) and end-diastolic (ED) frame detection and ejection fraction estimation. Understanding spatiotemporal data using transformers can also be found in other medical imaging domains. Latest research areas have been focusing on using transformers to diagnose COVID-19 [4,20,21] and perform 3D image segmentation of multi-organ and on brain tumor datasets [2].

A recent work was proposed by Li et al. [6] in a modified transformer version that combines the strengths of 3D CNNs and spatiotemporal transformers. The UniFormer has three main components. The first component is Dynamic Position Embeddings (DPE) which maintains the spatiotemporal positions of the video tokens by applying 3D depthwise convolution without padding. The second component is Multi-Head Relation Aggregator (MHRA) which learns the local token relations to ignore the redundancy due to the small differences found in adjacent frames in the initial layers. However, in the last two stages, MHRA learns the global token affinity, which is similar to the self-attention scheme. The last component is Feed Forward Network (FFN) which has two linear layers.

3 Methods

In this section, we describe the frames sampling approach, model architecture and the proposed method when estimating ejection fraction from echocardiographic videos.

3.1 Frames Sampling

Deep learning networks require a fixed number of video frames from each scan. However, EchoNet-Dynamic videos contain one or more cardiac cycles, which also vary in the number of frames per cycle (approximately 20-30 frames). Moreover, the differences between the adjacent video frames are small. Because of

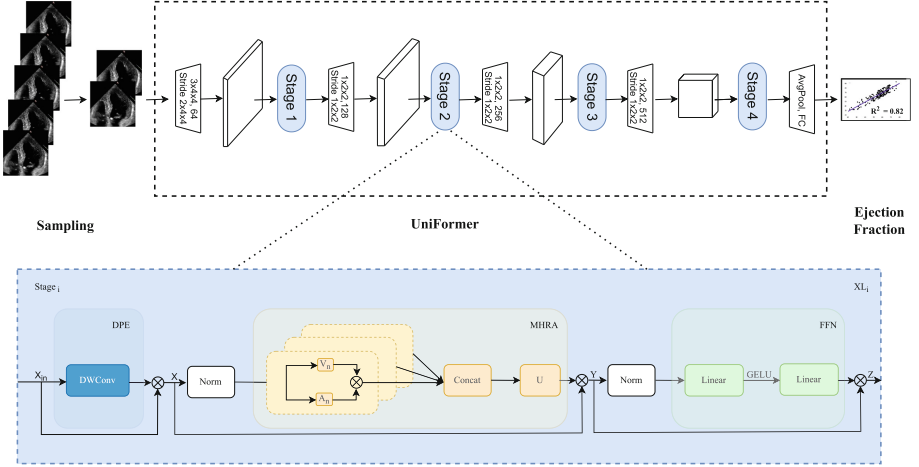


Fig. 1. The overall architecture of EchoCoTr is based on UniFormer [6]. Echocardiographic videos will be first sampled, to introduce dissimilarity between the frames, then fed to the UniFormer model to predict the LVEF for the entire video sequence.

that, we had to perform a video frame sampling by experimenting with different number of frames $\{32, 36, 40\}$ and uniform frequencies $\{2, 4, 6\}$ adapted by [8]. The sampling operation starts with a random clip within the range of $[0 - (\text{Number of original video frames} - (\text{Number of sampling video frames} - 1) * \text{Sampling frequency})]$. Prior to that, in the case of short videos, frames filled with zeros will be added to the end of the video. The strength of using video sampling techniques replaces the traditional methods that clinicians do, which requires knowing the location of ES and ED frames before calculating the LVEF. In addition to that, as the location of ES and ED frames are already known beforehand, we also experimented with only selecting ES and ED frames from the video to check if these are sufficient to give an adequate LVEF prediction. A summary of some experiments related to video sampling is found in Table 2.

3.2 Architecture Overview

EchoCoTr builds on UniFormer [6] to address both the challenges of the local redundant features and the complex dependency among the video frames in the cardiac echo scans. Subtle differences between adjacent frames make it important that the network selects the most representative frames when estimating LVEF. Therefore, we had to adapt an architecture that effectively learns the local features without redundancy in the adjacent frames while capturing the global information along the video. An illustration of the overall architecture is found in Fig. 1. Before feeding the ultrasound videos to the UniFormer model to generate LVEF prediction, we sample the video frames to introduce dissimilarity and make sure that there is no redundancy between the neighboring frames. Before each stage in the UniFormer model, $1 \times 2 \times 2$ convolution with stride of

$1 \times 2 \times 2$ is applied. However, to downsample the spatiotemporal dimensions of the input video, $3 \times 4 \times 4$ convolution with stride of $2 \times 4 \times 4$ is used instead in the first stage. As a method for echocardiography, we experimented with two different UniFormer variants: UniFormer-S and UniFormer-B with the aim of investigating the impact of the number of UniFormer blocks on the LVEF estimation. The number of UniFormer blocks used for EchoCoTr-S (small model) and EchoCoTr-B (baseline version) are $\{3, 4, 8, 3\}$ and $\{5, 8, 20, 7\}$, respectively. The drop rates are set to 0.1 for EchoCoTr-S and 0.3 for EchoCoTr-B.

3.3 Existing Methods for LVEF Estimation

In this subsection, for the sake of ablations and comparisons, we present recent published methods that addressed LVEF prediction. The work of [10] has shown that using a BERT model could be used to estimate LVEF. First, the dimensions of the input videos are reduced to a vector of size (Batch Size \times Number of Frames) \times 1024 using a ResNetAE [3] encoder. Two sampling strategies were introduced by [10]. The first is mirroring (M), which places the repeated sequence between the ES and ED frames after the last annotated frame. The second strategy is random sampling (R), which adds up 10-70% of the distance between the two annotated frames before and after the sampled frames from a heart cycle. However, the result that was reported did not outperform [8] that used a spatiotemporal convolution based ResNet (ResNet (2+1)D) [15]. Therefore, we compare our proposed method with the BERT method [10] and with other transformer models, such as DistilBERT and ViT.

4 Experiments

In this section, we aim to give a brief summary of the dataset used and experimental setup that we had for our experiments.

4.1 Datasets

EchoNet-Dynamic. [7] is the largest publicly available dataset of echocardiographic scans for the apical four-chamber (a4c) view of the heart acquired from the Stanford University Hospital. It consists of 10,030 videos in total. Each video consists of a sequence of 112×112 grayscale images and traces for the left ventricle end-systole (ES) and end-diastole (ED) frames. In addition, every video is labelled with the corresponding end-systolic volume (ESV), end-diastolic volume (EDV) and ejection fraction (EF).

4.2 Experimental Setup

The data split sizes for training, validation and testing are 7460, 1288 and 1277, respectively. This is the same split chosen by [7]. All selected hyperparameters are optimized experimentally. The evaluation metrics used are mean absolute

error (MAE), root mean squared error (RMSE) and R-squared (R^2). In addition to that, we also compare the floating point operations (FLOPs) values for the different models using fvcore package [1].

EchoCoTr Experiments: EchoCoTr models are trained on an NVIDIA A100 GPU for 45 epochs. The batch sizes used for EchoCoTr-S and EchoCoTr-B are 25 and 16, respectively. AdamW is used as an optimizer with a value of $1e-4$ for both the learning rate and weight decay. Both models were pretrained on the Kinetics-400 dataset with different pretraining strategies. EchoCoTr-S is pretrained on $16 \times 1 \times 4$ frames with sampling stride of 8. However, the weights used for EchoCoTr-B is $32 \times 1 \times 4$ frames with sampling stride of 4. Frame resolutions are kept as same as in the original public dataset (112×112).

Other Experiments. BERT, DistilBERT and ViT models are trained for 5 epochs with batch size of 2, which is small because of the large model size. AdamW is used as an optimizer with a learning rate of $1e-5$ and weight decay of $1e-2$. Images are padded to be 128×128 in size to facilitate fair comparison and easy integration for the three models. The Hugging Face Python library is used for the transformer experiments.

Table 1. Comparison with the state-of-the-art results on EchoNet-Dynamic dataset. "R." and "M" are the sampling methods proposed by [10], which refer to random and mirroring sampling. EchoNet-Dynamic (1) predicts the clip-level LVEF using 32 frames. EchoNet-Dynamic (2) uses the segmentation and clip-level LVEF outputs to evaluate the beat-to-beat LVEF estimation for the entire video sequence. One sample from the testing dataset is used to calculate the FLOPs.

Model	No. of frames	FLOPs	MAE ↓	RMSE ↓	R^2 ↑
UVT R. [10]	128	130.00G	6.77	8.70	0.48
UVT M. [10]	128	130.00G	5.95	8.38	0.52
R3D [8]	32	92.273G	4.22	5.62	0.79
MC3 [8]	32	97.656G	4.54	5.97	0.77
EchoNet-Dynamic [8] (1)	32	91.974G	4.22	5.56	0.79
EchoNet-Dynamic [8] (2)	beat-to-beat	-	4.05	5.32	0.81
EchoCoTr-B	36	44.907G	3.98	5.34	0.81
EchoCoTr-S	36	19.611G	3.95	5.17	0.82

5 Results

As Table 1 shows, our EchoCoTr-S model, which was trained on only 36 frames with sampling frequency of 4 (3.95 MAE), outperforms the state-of-the-art results reported by [8, 10]. It is also noticeable from the results that the

EchoCoTr-S experiment (3.95 MAE) performed slightly better than EchoCoTr-B (3.98 MAE). We test the effect of various sampling frequencies and sizes on the LVEF prediction. Results in Table 2 show that a sampling frequency of 4 frames achieves the best result for both small and baseline models. In addition, the optimal number of frames is found to be 36 for both models. Surprisingly, training both EchoCoTr-S and EchoCoTr-B models on only two frames (ES and ED) from each video achieves lower yet satisfactory results (4.432 and 4.494 MAE).

Table 2 also displays the results of our experiments that we performed using BERT, DistilBERT and ViT. We only report the experiments for the mirroring sampling strategy, as it achieved better results than the random one in [10]. Results suggest that the BERT model with the mirroring sampling on 36 and 128 frames (5.788 and 5.950 MAE, respectively) [10] performs better than DistilBERT and ViT when estimating LVEF. Moreover, reducing the number of frames to 36 was negatively impacting DistilBERT’s MAE score the most (6.689).

Table 2. Ablation study: Summary of experiments performed on the EchoNet-Dynamic Dataset using EchoCoTr and transformer models. The sampling strategy used for BERT, DistilBERT and ViT experiments is mirroring [10]. 2* refers to the two video frames used, which are ES and ED.

Model	Frequency	No. of frames	Batch size	MAE ↓	RMSE ↓	R ² ↑
BERT	-	36	2	5.788	8.137	0.545
BERT [10]	-	128	2	5.950	8.380	0.520
DistilBERT	-	36	2	6.689	9.234	0.414
DistilBERT	-	128	2	6.430	8.940	0.451
ViT	-	36	2	6.454	8.955	0.448
ViT	-	128	2	6.527	9.053	0.436
EchoCoTr-S	-	2*	25	4.432	5.998	0.759
EchoCoTr-S	2	36	25	4.168	5.541	0.795
EchoCoTr-S	4	32	25	3.966	5.290	0.813
EchoCoTr-S	4	36	25	3.947	5.174	0.821
EchoCoTr-S	4	40	25	4.010	5.326	0.810
EchoCoTr-S	6	36	25	4.135	5.434	0.803
EchoCoTr-B	-	2*	16	4.494	6.205	0.743
EchoCoTr-B	2	36	16	4.184	5.590	0.791
EchoCoTr-B	4	36	16	3.980	5.342	0.809
EchoCoTr-B	6	36	16	4.068	5.410	0.804

6 Discussion

In this paper, we propose EchoCoTr which is a method that combines the strengths of 3D CNNs and vision transformers for spatiotemporal echocardiography assessment in order to estimate LVEF on ultrasound videos.

The results in Table 1 show that the model trained using EchoCoTr-S on only 36 frames with a uniform sampling frequency of 4 (3.95 MAE), outperforms the state-of-the-art results reported by EchoNet-Dynamic on the beat-to-beat pipeline for LVEF prediction (4.05 MAE). In addition, unlike EchoNet-Dynamic, our method does not require the segmentation masks. Furthermore, our score is also better than the result that EchoNet-Dynamic stated for 32 frames and a sampling frequency of 2 frames (4.22 MAE). As illustrated in Table 2, a proper video sampling strategy plays a role in improving the results when using EchoCoTr models. This might be due to the different details that the model attends to spatially and temporally. For instance, not all adjacent frames might be needed during training and frames from multiple heart cycles are likely needed to provide a better temporal representation. Furthermore, we think that 36 frames with sampling frequency of 4 is found to be an ideal configuration to the problem at hand, because it covered multiple cardiac cycles (4-5 cycles) while skipping redundant and similar frames in most of the videos found in the EchoNet-Dynamic dataset. Hence, this has led to a more accurate estimation of LVEF prediction for the entire video. In fact, the frame sampling strategy we propose is aligned with the clinical guidelines that suggest estimating LVEF from up to 5 cardiac cycles.

Another remarkable result found is that EchoCoTr achieves satisfactory LVEF estimations when trained on only two frames (ES and ED). Due to its design, it ignores the local redundant features but learns the long-range dependencies. This follows the same methodology that clinicians do when calculating the EDV and ESV values to estimate LVEF.

It is also clearly seen from Table 2 that training on 36 frames achieves comparable results to 128 frames for BERT, DistilBERT and ViT models. However, all these experiments did not perform as well as our proposed method. We hypothesize that these models could not capture the temporal information as effectively as our proposed method while learning the local features within different frames. We believe that EchoCoTr-B performed marginally less than EchoCoTr-S due to its large architectural size that might be an overkill for the LVEF estimation problem.

7 Conclusion

We propose EchoCoTr which utilizes CNNs' discriminative spatial ability with transformers' temporal perception to estimate LVEF from a set of sampled

frames from multiple heart cycles. The method outperforms other recent work when estimating ejection fraction on the EchoNet-Dynamic dataset. The goal of this paper is not to comprehensively study the performance of different transformer models, but to compare their performances with our CNN-Transformer method on spatiotemporal image analysis. For future work, it is valuable to study the effect of self-supervision on EchoCoTr's performance by using the unlabelled frames from each video. EchoNet-Dynamic dataset size proved to be enough to produce good results using EchoCoTr and spatiotemporal convolutional neural networks. Furthermore, it is also worth experimenting with the impact of performance on smaller datasets and datasets with abnormal motion of the heart.

Acknowledgments. We thank Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI) for providing funding for this study, and Mohamed Saeed for providing his support.

References

1. Facebookresearch: fvcore: flop counter for pyTorch models. https://github.com/facebookresearch/fvcore/blob/main/docs/flop_count.md
2. Hatamizadeh, A., et al.: UNETR: transformers for 3D medical image segmentation (2021)
3. Hou, B.: ResNetAE-<https://github.com/farrell236/resnetae> (2019). <https://github.com/farrell236/ResNetAE>
4. Hsu, C.C., Chen, G.L., Wu, M.H.: Visual transformer with statistical test for COVID-19 classification (2021)
5. Lara Hernandez, K.A., Rienmüller, T., Baumgartner, D., Baumgartner, C.: Deep learning in spatiotemporal cardiac imaging: a review of methodologies and clinical usability. *Comput. Biol. Med.* **130**, 104200 (2021). <https://doi.org/10.1016/j.complbiomed.2020.104200>. <https://www.sciencedirect.com/science/article/pii/S001048252030531X>
6. Li, K., et al.: UNIFORMER: unified transformer for efficient spatiotemporal representation learning (2022)
7. Ouyang, D., et al.: EchoNet-Dynamic: a large new cardiac motion video data resource for medical machine learning. In: *NeurIPS ML4H Workshop*, Vancouver, BC, Canada (2019)
8. Ouyang, D., et al.: Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**(7802), 252–256 (2020). <https://doi.org/10.1038/s41586-020-2145-8>
9. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? (2021)
10. Reynaud, H., Vlontzos, A., Hou, B., Beqiri, A., Leeson, P., Kainz, B.: Ultrasound video transformers for cardiac ejection fraction estimation (2021)
11. Saeed, M., Muhtaseb, R., Yaqub, M.: Contrastive pretraining for echocardiography segmentation with limited data (2022). <https://doi.org/10.48550/ARXIV.2201.07219>. <https://arxiv.org/abs/2201.07219>
12. Silva, J.F., Silva, J.M., Guerra, A., Matos, S., Costa, C.: Ejection fraction classification in transthoracic echocardiography using a deep learning approach. In: *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 123–128 (2018). <https://doi.org/10.1109/CBMS.2018.00029>

13. Smistad, E., et al.: Real-time automatic ejection fraction and foreshortening detection using deep learning. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **67**(12), 2595–2604 (2020). <https://doi.org/10.1109/TUFFC.2020.2981037>
14. Smistad, E., Østvik, A., Salte, I.M., Leclerc, S., Bernard, O., Lovstakken, L.: Fully automatic real-time ejection fraction and mapse measurements in 2D echocardiography using deep neural networks. In: 2018 IEEE International Ultrasonics Symposium (IUS), pp. 1–4 (2018). <https://doi.org/10.1109/ULTSYM.2018.8579886>
15. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6450–6459 (2018)
16. Virani, S.S., et al.: Heart disease and stroke statistics—2020 update: a report from the american heart association. *Circulation* **141**(9), 139–596 (2020). <https://doi.org/10.1161/cir.0000000000000757>. <https://doi.org/10.1161/cir.0000000000000757>
17. Voorhees, A., Han, H.C.: Biomechanics of cardiac function. *Compr. Physiol.* **5**(4), 1623–1644 (2015). <https://doi.org/10.1002/cphy.c140070>
18. Wang, Y., et al.: Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records. In: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2015, Milan, Italy, 25–29 August 2015, pp. 2530–2533. IEEE (2015). <https://doi.org/10.1109/EMBC.2015.7318907>. <https://doi.org/10.1109/EMBC.2015.7318907>
19. Zhang, J., et al.: Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation* **138**, 1623–1635 (10 2018). <https://doi.org/10.1161/CIRCULATIONAHA.118.034338>
20. Zhang, L., Wen, Y.: MIA-COV19D: a transformer-based framework for COVID19 classification in chest CTs (2021). <https://doi.org/10.13140/RG.2.2.12992.05125>
21. Zhang, L., Wen, Y.: A transformer-based framework for automatic COVID19 diagnosis in chest CTs. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 513–518 (2021). <https://doi.org/10.1109/ICCVW54120.2021.00063>