Simultaneous Bone and Shadow Segmentation Network Using Task Correspondence Consistency

Aimon Rahman[1(✉)], Jeya Maria Jose Valanarasu[1], Ilker Hacihaliloglu[2], and Vishal M. Patel[1]

[1] Johns Hopkins University, Baltimore, USA
arahma30@jhu.edu
[2] University of British Columbia, Vancouver, Canada

Abstract. Segmenting both bone surface and the corresponding acoustic shadow are fundamental tasks in ultrasound (US) guided orthopedic procedures. However, these tasks are challenging due to minimal and blurred bone surface response in US images, cross-machine discrepancy, imaging artifacts, and low signal-to-noise ratio. Notably, bone shadows are caused by a significant acoustic impedance mismatch between the soft tissue and bone surfaces. To leverage these complementary features between these highly related tasks, we propose a single end-to-end network with a shared transformer-based encoder and task independent decoders for simultaneous bone and shadow segmentation. To share complementary features, we propose a cross task feature transfer block which learns to transfer meaningful features from decoder of shadow segmentation to that of bone segmentation and vice-versa. We also introduce a correspondence consistency loss which makes sure that network utilizes the inter-dependency between the bone surface and its corresponding shadow to refine the segmentation. Validation against expert annotations shows that the method outperforms the previous state-of-the-art for both bone surface and shadow segmentation.

Keywords: Multi-task · Ultrasound · Bone segmentation · Shadow segmentation

1 Introduction

There has been a significant interest in incorporating ultrasound (US) imaging for computer assisted orthopedic surgery (CAOS) procedures owing to its non-invasive, radiation-free, and cost-effective nature. However, due to bone surfaces appearing only several millimeters (mm) in thickness along with noisy artifacts, researchers have been focusing on developing automated bone segmentation and enhancement methods [7]. These bone surfaces generally have the highest intensity in US images which is then followed by a low-intensity region, namely bone shadows. Bone shadow is the result of a high acoustic impedance mismatch

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
L. Wang et al. (Eds.): MICCAI 2022, LNCS 13434, pp. 330–339, 2022.
https://doi.org/10.1007/978-3-031-16440-8_32

between the bone surface and the adjacent soft tissue, which reflects the US signal to the transducer. The bone shadow information is essential to guide the orthopedic surgeon to a standardized viewing plane with minimal noise and artifacts. Hence, both bone surface and shadow segmentation are crucial to CAOS procedures.

Recent literature on bone and shadow segmentation focus on learning individual networks for each problem separately [1–3,13]. However, in [11], Wang et al. [11] proposed a pre-enhancement network that leverages bone shadow information for bone surface segmentation. The bone shadow was obtained using a bone shadow enhancement method where a signal transmission map is constructed from the local phase bone image features [6]. The enhanced bone shadow information has also been used in [12] where a multi-task learning-based method to segment bone shadow region is proposed.

It should be noted that bone shadow is a signal void that indicates the loss of energy as US waves propagate through bone tissues. Thus, the quality of bone surface segmentation can have major impact on shadow segmentation accuracy and vice-versa. However, existing works do not fully exploit the structure of these highly related tasks. Despite being closely-related, existing top networks for bone and shadow segmentation have significantly different and specialized architectures. Our proposed method explores the idea of exploiting shared features for a more compact network and taking advantage of interactions between the two tasks to generate a better feature representation. We hypothesize that the interrelation between bone and shadow response in US images can be leveraged to significantly improve the quality of both learned networks. In summary, we present the following contributions in this paper:

- We are the first to integrate two highly-related homogeneous tasks into a single framework for unified bone surface and shadow segmentation. The common encoder brings powerful synergy across both tasks when extracting shared deep features for the two tightly-coupled problems.
- We propose a cross task feature transfer block to extract complementary features at decoders to improve the quality of performance in the multi-task learning framework.
- We propose a task correspondence consistency loss to further regularize the network by ensuring the transitivity between the two related predictions.
- We conduct extensive experiments using the in vivo US scans of knee, femur, distal radius, spine, and tibia bones collected using two US machines and demonstrate that the proposed method is competitive with other individual specialized state-of-the-art methods.

2 Method

2.1 Preliminaries

Instead of using only B-mode US scan as input, the proposed network takes the concatenation of three filtered images along with the original B-mode US scan

$(US(x, y))$. The filtered images are shown in Fig. 1(a)–(d). This has been done to reduce the domain discrepancy between the images obtained using different US machine settings or different orientations of the transducer. During the extraction of filtered images we have used the original parameters and constant values described in [6,8]. The Local Phase Tensor Image $(LPT(x, y))$ is computed by defining odd and even filter responses using [8]. Local Phase Bone Image $LP(x, y)$ is computed using: $LP(x, y) = LPT(x, y) \times LPE(x, y) \times \text{LwP}\, A(x, y)$, where $LPE(x, y)$ and $LwPA(x, y)$ represent the local phase energy and local weighted mean phase angle image features, respectively. These two features are computed using monogenic signal theory as [6]. Bone Shadow Enhanced image $BSE(x, y)$ is obtained by modeling the interaction of Ultrasound signal at position (x,y) within the tissues as scattering and attenuation information using the method proposed in [6],

$$BSE(x, y) = [(CM_{LP}(x, y) - \rho)/[max(US_A(x, y), \epsilon)]^{\delta}] + \rho$$

Here the confidence map is denoted by $CM_{LP}(x, y)$ which is obtained by modeling the US signal propagation inside the tissue considering bone feature in local phase bone image LP(x, y). $US_A(x, y)$ maximizes the visibility of bone features with high intensity inside a local region. δ represents the tissue attenuation coefficient. ρ is related to echogenicity confining the bone surface and ϵ is a small constant to avoid division by zero.

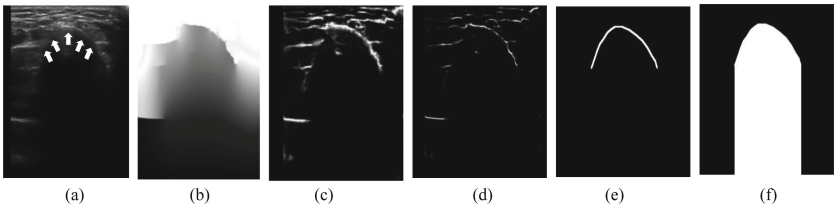

Fig. 1. (a) B-mode US scan. Thick white arrows point to the bone response in US image. (b) LPT (c) LP (d) BSE (e) Bone Surface Segmentation and (f) Bone Shadow Segmentation.

2.2 Network Architecture

We propose Shadow and Surface Segmentation Network (SSNet) for simultaneous bone surface and shadow segmentation from US images which is illustrated in Fig. 2. SSNet is composed of a shared LeViT-based encoder to extract global and long-range spatial features and two CNN-based decoders with a cross task feature transfer block to leverage complementary features between the two tasks.

(i) LeViT-based Shared Encoder: The shared encoder for bone and shadow surface segmentation is built based on the LeViT architecture [5]. The encoder

part consists of four 3×3 convolution layers with stride 2 initially followed by three transformer blocks. Features from the convolution layers are forwarded to the LeViT transformer blocks which require fewer floating-point operations (FLOPs) than ViTs [4]. The local and global features at different scales are exploited by concatenating the features from both transformer and convolution layers.

(ii) CNN-based Decoders: The decoder part of the network consists of two separate branches for bone surface and shadow segmentation. Inspired by UNet [10], the features from decoders are concatenated with skip connection to effectively reuse spatial information of feature maps. The resolution from the previous layers is recovered using the cascaded upsampling technique similar to UNet. The decoder blocks consist of a 3×3 convolution, batch normalization layer followed by a ReLU layer.

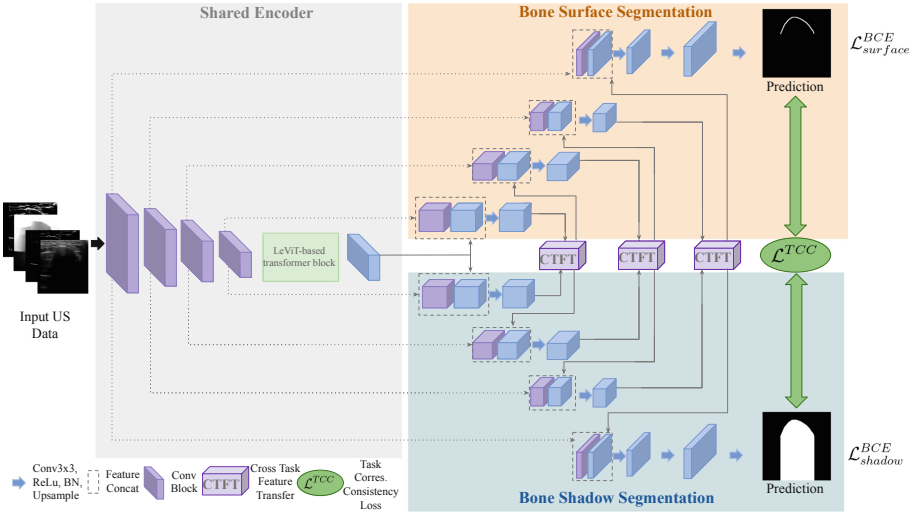

Fig. 2. An overview of the proposed SSNet for simultaneous bone surface and shadow segmentation from US images.

2.3 Cross Task Feature Transfer Block

To leverage the joint-learning capabilities of these two highly-related tasks, we propose a cross task feature transfer (CTFT) block used in between the two decoders. CTFT extracts complementary features from the two decoder branches using a squeeze and excitation block [9] and forwards them to the next decoder blocks of respective branches. We use squeeze and excitation block to learn which features of the surface segmentation decoder would help in segmenting bone shadow and vice-versa. Squeeze and excite enables dynamic channel-wise feature re-calibration thus help extract features that contributes to the complementary task. The details of CTFT are illustrated in Fig. 3. It takes in two inputs: $F_{surface}$

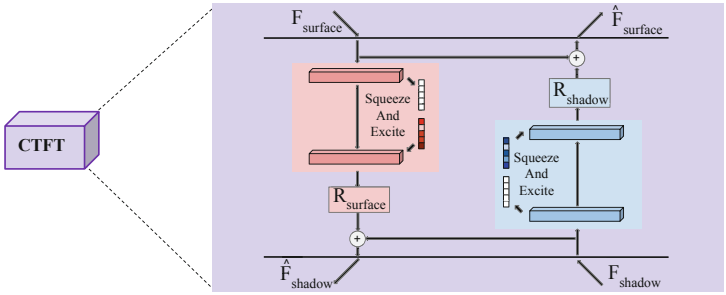

Fig. 3. An overview of the Cross Task Feature Transfer Block.

and F_{shadow} corresponding to the feature maps of bone surface and shadow decoders. $F_{surface}$ is passed through a squeeze and excite layer to obtain the residual $R_{surface}$ which is added to F_{shadow} to obtain $\hat{F}_{shadow}$. $\hat{F}_{shadow}$ is then passed to the next block of the shadow segmentation decoder. Similarly, F_{shadow} is passed through a squeeze and excite layer to obtain the residual R_{shadow} which is added to $F_{surface}$ to obtain $\hat{F}_{surface}$.

2.4 Task Correspondence Consistency Loss

To guarantee both networks capture the inter-dependency between bone surface and its corresponding shadow, we introduce two additional loss terms called Task Correspondence Consistency Loss. For an US image $X \in \mathcal{X}$, the annotations $Y = (y_1, y_2)$ is a set of labels containing bone surface and shadow segmentation masks, respectively. Let, $\hat{Y} = (\hat{y_1}, \hat{y_2})$ be the predictions of the decoder networks. Our additional loss term includes two mapping $F_1 : y_1 \rightarrow y_2$ and $F_2 : y_2 \rightarrow y_1$. For any US image X, each loss term ensure consistency by translating in between bone surface and shadows, i.e., $y_1 \rightarrow F_1(y_1) \approx y_2$. The task corresponding consistency loss further regularizes the network to produce robust segmentation masks for both task and prevent them to contradict each other. The proposed Task Correspondence Consistency Loss $\mathcal{L}^{TCC}(X, Y)$ is defined as:

$$\mathcal{L}^{TCC}(X, Y) = \mathcal{L}^{BCE}(y_1, F_2(\hat{y_2})) + \mathcal{L}^{BCE}(y_2, F_1(\hat{y_1})).$$

3 Experiments and Results

Dataset: The study includes 25 healthy volunteers with the approval of the institutional review board (IRB). Total 1042 different US images have been collected using SonixTouch US machine (Analogic Corporation, Peabody, MA, USA) with 2D C5-2/60 curvilinear and L14-5 linear transducer. For independent testing, 3 new subjects have been included in the study. Using handheld wireless US scans (Clarius C3, ClariusMobile Health Corporation, BC, Canada), a total of 185 scans have been collected. Depending on the depth setting, scan resolution varies between 0.1 mm to 0.15 mm. As both transducer and reconstruction

pipelines are different, Clarius have low image quality. The scans include knee, femur, radius, and spine data and all of them are manually segmented by an expert ultrasonographer. For the Sonix dataset, a random 80:20 split has been applied based on the subject, making the final training set with 834 samples and the test set with 208 samples.

Implementation Details: SSNet is trained using a batch size of 32. For training both branches, a two-step training phase is adapted. Each of these steps are trained until convergence. The weights and bias of the network are optimized using Adam optimizer with a learning rate of 10^{-4}. All US scans and their corresponding masks are resized to 224×224 pixels and rescaled between 0 to 1. All transformer blocks in the LeViT architecture were pre-trained on ImageNet-1k. The overall loss function we use to train the multi-task network is,

$$\mathcal{L}^{total}(X, Y) = \mathcal{L}^{BCE}(y_1, \hat{y_1}) + \mathcal{L}^{BCE}(y_2, \hat{y_2}) + \mathcal{L}^{TCC}(X, Y).$$

Binary-cross entropy loss has been used between the prediction and the ground truth, which is expressed as,

$$\mathcal{L}_{CE(p,\hat{p})} = -\left(\frac{1}{wh}\sum_{x=0}^{w-1}\sum_{y=0}^{h-1}(p(x,y)\log(\hat{p}(x,y))) + (1-p(x,y))\log(1-\hat{p}(x,y))\right).$$

Here, w and h represents the dimension of ultrasound scan, $p(x, y)$ denotes the pixel in scan and $\hat{p}(x, y)$ denotes the output prediction at a specific location (x, y). Test images can be forwarded through the network for both tasks in one shot. The experiments are carried out on a Linux workstation with Intel 3.50 GHz CPU and a 12GB NVidia Titan Xp GPU using the PyTorch framework. Dice coefficients are used to measure the segmentation performance of different methods.

Quantitative Comparison: For bone shadow segmentation, we compare the performance of our proposed method with that of UNet [10], MFG-CNN [11], and PSPGAN MTL [12]. PSPGAN MTL is the current state-of-the-art for bone shadow segmentation. For bone surface segmentation, we compare with UNet [10], MFG-CNN [11] without the classification labels, and LPT+GCT [13]. All the methods are trained using the same training dataset as used to train the proposed method. PSPGAN-MTL uses a conditional shape discriminator to enforce bone interval boundaries which provides more accurate and robust bone segmentation. Instead of using bone interval boundaries during the training, we enforce the boundary from the bone surface segmentation mask during inference instead. Average test results are shown in Table 1. It can be observed that the shared network SSNet outperforms the current state-of-the-art [12] and individual networks for both bone and shadow segmentation (paired t-test < 0.05).

Qualitative Comparison: We present sample qualitative results in Fig. 4 for both bone surface and shadow segmentation. It can be observed that the current state-of-the-art methods result in either missed shadow regions or disjoint bone segmentation maps. As our proposed method uses the inter-dependency between

Table 1. Results averaged over 5 folds. Numbers correspond to dice score with standard deviation. Boldface numbers indicate the best segmentation performance.

	SonixTouch		Clarius	
Method	Surface (%)	Shadow (%)	Surface (%)	Shadow (%)
UNet [10]	76.01 ± 0.20	88.33 ± 0.06	75.11 ± 0.31	84.03 ± 0.14
MFG-CNN [11]	81.05 ± 0.06	–	82.23 ± 0.14	–
LPT + GCT [13]	81.65 ± 0.10	–	83.05 ± 0.21	–
PSPGAN-MTL [12]	–	93.49 ± 0.06	–	91.01 ± 0.18
SSNet + CTFT + TCC loss (ours)	**87.03 ± 0.21**	**96.18 ± 0.43**	**83.33 ± 0.31**	**93.01 ± 0.23**

these tasks, we see a significant improvement with less discrepancies compared to the ground truth annotations.

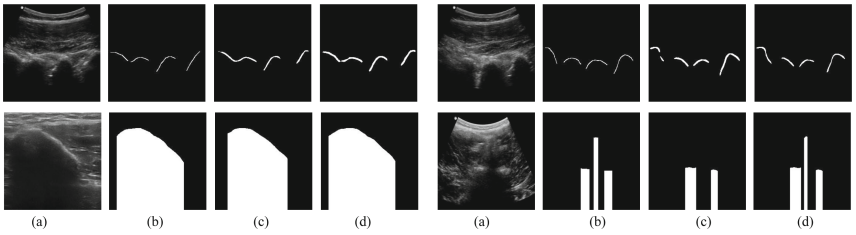

Fig. 4. Top Row - Bone surface segmentation. Bottom Row - Bone shadow segmentation. (a) Input US scan (b) Ground Truth (c) Output from current state-of-the-art [13] (surface), [12] (shadow) (d) Ours.

4 Discussion

Ablation Study: To understand the contribution of each individual module in the proposed SSNet, we conduct an ablation study and report it in Table 2. It can be observed that addition of CTFT helps improve the performance of both surface and shadow segmentation by injecting complementary features to the respective decoders. Also, using the propose task consistency ($\mathcal{L}^{TCC}$) further regularizes the network and boosts the segmentation performance.

Importance of Joint Learning: Qualitative results in Fig. 5 shows the importance of the joint learning framework. The result from cascaded network demonstrates that the faulty output from either of the network can produce wrong corresponding prediction. Cascaded network corresponds to using a deep network to predict the bone shadow map from bone surface segmentation map and vice-versa. For example, missing or joint boundaries in bone surface segmentation may result in wrong bone intervals in shadow network as demonstrated in the top row of Fig. 5. Similarly, over or under-segmented bone shadow predictions may produce faulty surface estimations. However, as each of the decoders

Table 2. Ablation study. Numbers correspond to dice score.

	SonixTouch		Clarius	
Method	Surface (%)	Shadow (%)	Surface (%)	Shadow (%)
SSNet (Base)	82.95 ± 0.13	93.34 ± 0.06	81.71 ± 0.20	90.94 ± 0.22
SSNet + CTFT	84.03 ± 0.11	94.88 ± 0.16	81.13 ± 0.19	92.43 ± 0.18
SSNet + CTFT + $\mathcal{L}^{TCC}$ (ours)	$\mathbf{87.03 \pm 0.21}$	$\mathbf{96.18 \pm 0.43}$	$\mathbf{83.33 \pm 0.31}$	$\mathbf{93.01 \pm 0.23}$

in our network is specialized for their respective task and further regularized by ensuring cross-task consistency, our network produces more consistent results.

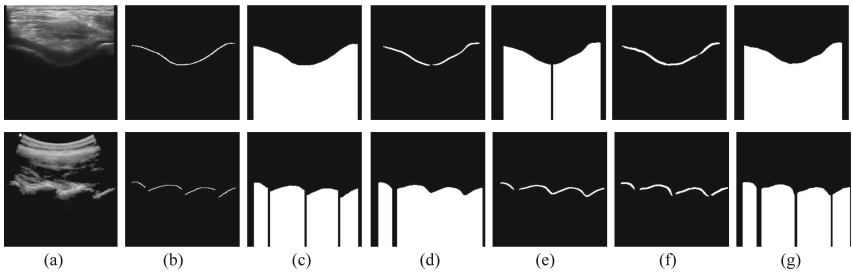

Fig. 5. (a) Input US scan (b) Surface ground truth (c) Shadow ground truth (d) Top row corresponds to output from an individual bone surface segmentation network and bottom row corresponds to output from an individual bone shadow segmentation network (e) Top row corresponds to cascaded shadow segmentation output generated using the segmentation from individual network and bottom row corresponds to cascaded surface segmentation output generated using the segmentation from individual network (f) Surface output from ours (g) Shadow output from ours.

Effectiveness of CTFT: In Table 3, we show that adding CTFT to the base network improves the segmentation performance. To further validate the claim, we conduct more experiments as seen in Table 3. It can be observed that adding CTFT to a joint-UNet architecture results in a boost in performance.

Table 3. Ablation study. All results are reported in Dice score.

	SonixTouch		Clarius	
Method	Surface (%)	Shadow (%)	Surface (%)	Shadow (%)
Joint-UNet	76.45 ± 0.03	86.06 ± 0.15	75.11 ± 0.33	84.01 ± 0.17
Joint-UNet + CTFT	77.19 ± 0.17	89.01 ± 0.15	75.81 ± 0.21	84.71 ± 0.11

5 Conclusion

Accurate, complete, and robust bone and shadow segmentation are important to make ultrasound an essential imaging modality in clinically acceptable orthopedics procedures. In this paper, we propose an end-to-end network to simultaneously perform robust and accurate bone and shadow segmentation by leveraging complementary features between the two tasks. The main novelty of our work lies in (1) the first systematic design of exploiting interrelation between two tasks to improve both bone and shadow segmentation, and (2) the design of fusion method of CNN and vision transformer to leverage multi-task learning while optimizing accuracy-efficiency trade-off. We believe the multi-task learning framework is an important contribution to the field of US-based orthopedic procedures.

References

1. Alsinan, A., Vives, M., Patel, V., Hacihaliloglu, I.: Spine surface segmentation from ultrasound using multi-feature guided CNN. CAOS **3**, 6–10 (2019)
2. Alsinan, A.Z., Patel, V.M., Hacihaliloglu, I.: Automatic segmentation of bone surfaces from ultrasound using a filter-layer-guided CNN. Int. J. Comput. Assist. Radiol. Surg. **14**(5), 775–783 (2019)
3. Alsinan, A.Z., Patel, V.M., Hacihaliloglu, I.: Bone shadow segmentation from ultrasound data for orthopedic surgery using GAN. Int. J. Comput. Assist. Radiol. Surg. **15**(9), 1477–1485 (2020)
4. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Graham, B., et al.: LeViT: a vision transformer in convnet's clothing for faster inference. arXiv preprint arXiv:2104.01136 (2021)
6. Hacihaliloglu, I.: Enhancement of bone shadow region using local phase-based ultrasound transmission maps. Int. J. Comput. Assisted Radiol. Surg. **12**(6), 951–960 (2017)
7. Hacihaliloglu, I.: Ultrasound imaging and segmentation of bone surfaces: a review. Technology **5**(02), 74–80 (2017)
8. Hacihaliloglu, I., Rasoulian, A., Rohling, R.N., Abolmaesumi, P.: Local phase tensor features for 3-D ultrasound to statistical shape+ pose spine model registration. IEEE Trans. Med. Imaging **33**(11), 2167–2179 (2014)
9. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28
11. Wang, P., Patel, V.M., Hacihaliloglu, I.: Simultaneous segmentation and classification of bone surfaces from ultrasound using a multi-feature guided CNN. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 134–142. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_16

12. Wang, P., Vives, M., Patel, V.M., Hacihaliloglu, I.: Robust bone shadow segmentation from 2D ultrasound through task decomposition. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 805–814. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_78
13. Wang, P., Vives, M., Patel, V.M., Hacihaliloglu, I.: Robust real-time bone surfaces segmentation from ultrasound using a local phase tensor-guided CNN. Int. J. Comput. Assist. Radiol. Surg. **15**, 1127–1135 (2020)