3D UX-NET: VOLUMETRIC LARGE KERNEL CONVNET MODERNIZING TRANSFORMER FOR MEDICAL IMAGE SEGEMNTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The introduction of vision transformers (ViTs) quickly superseded convolutional networks (ConvNets) as the current state-of-the-art (SOTA), especially for medical image segmentation. Instead of only leveraging vanilla ViT, the hierarchical transformers (e.g., Swin Transformers) reintroduced several ConvNet priors and further enhanced the practical viability of adapting volumetric segmentation in 3D medical datasets. However, the effectiveness of such hybrid approaches is largely credited to the large receptive field for non-local self-attention and the large capacity of model parameters. We hypothesis that the ConvNet can simulate the large receptive field behavior in transformer with less model capacities using depth-wise convolution. In this work, we propose a lightweight volumetric ConvNets termed 3D UX-Net, which adapts the capability of hierarchical transformer using ConvNet modules for robust volumetric segmentation. Specifically, we revisit the volumetric depth-wise convolution with large kernel size (e.g. $7 \times 7 \times 7$) to enable the global receptive field similar to the swin transformer. We further substitute the multi-layer perceptron in swin transformer blocks with pointwise group convolutions and enhance model performances with fewer normalization and activation layers, thus to reduce the overall model parameters. 3D UX-Net is competed favorably with current SOTA transformers (e.g. SwinUNETR) using three challenging public datasets on volumetric brain and abdominal imaging: 1) MICCAI Challenge 2021 FLARE, 2) MICCAI Challenge 2021 Feta datasets and 3) MICCAI Challenge 2022 AMOS. Significant improvement is consistently achieved and outperforms SwinUNETR from 0.909 to 0.920 Dice (FLARE2021) and 0.867 to 0.874 Dice (Feta2021). We further evaluate the transfer learning capability of 3D UX-Net with AMOS2022 and demonstrates another significant improvement of 2.27% Dice (from 0.880 to 0.900). The GitHub and model are available in `https://github.com/MASILab/3DUX-Net`.

1 INTRODUCTION

Significant progress has been made with the introduction of vision transformers (ViTs) over the last years. It demonstrates superior performance over the convolutional networks (ConvNets) on many downstream tasks . ViTs have further extended to extract sequential representation with volumetric patches and generate non-local self-attentions correspondence to enhance volumetric downstream tasks, especially for medical image segmentation. Another characteristics of ViTs is the zero image-specific inductive bias and the scaling behaviour with the help of large model and dataset sizes. Such specialities contributes to the significant improvement compared to ConvNets. However, without the inductive biases from ConvNets, it is challenging to adapt ViT model as the generic backbone due to the high complexity of computing global self-attention with respect to the input size. Therefore, hierarchical transformers are proposed to bridge these gaps with its intrinsic hybrid structure. The "sliding window" strategy is introduced into ViTs and proposes swin transformer to behave similarily with the ConvNets. SwinUNETR adapts the swin transformer blocks as the generic vision encoder backbone and achieve state-of-the-art performance for volumetric segmentation. With such significant improvement by leveraging the ConvNet priors into transformers, we further question

about the essence and the capability of convolution for volumentric segmentation: **Why we cannot leverage convolutions to enable the capabilities of hierarchical transformers?**

As the advancement of ViTs starts to bring back the basis of convolutions, the key component for such large performance differences are attributed to the **scaling behavior** and **the global self-attention with large receptive field**. Therefore, with the motivation of Liu et al. (2022), we hypothesis that the key attributes for the performance differences can be tackled with ConvNets modules and the newly designed convolution blocks can further enhance the model generalizability for volumetric segmentation. In this work, (3)

2 Related Work

Accurate 3D Medical image segmentation aims to analyze body anatomy that plays an essential role in computer-aided diagnosis. The popular segmentation tasks focus on structures and tumor on brain, liver, lung, abdominal, leg, using X-ray, CT, MRI and ultrasound imaging etc Wang et al. (2022).

2.1 Transformer based

In recent years, with the introduction of Vision Transformers (ViT) from natural language processing into visual recognition, it quickly replaced ConvNet in various tasks of computer vision, including segmentation Hatamizadeh et al. (2022b); Chen et al. (2021). To improve the competence of capturing long-range dependencies information for 3D Medical Image Segmentation than a pure transformer, Swin Transformer contains a multi-stage hierarchical architecture that shifts the window partition between consecutive self-attention to capture local overlapping regions Liu et al. (2021). Swin UNETR equips the encoder with the Swin transformers to compute self-attention for enhancing brain tumor segmentation accuracy in 3D MRI Images Hatamizadeh et al. (2022a). Tang et al. extend the Swin UNETR by adding a self-supervised learning pre-training strategy for fine-tuning segmentation tasks. Swin-Unet is an Unet-like architecture that adds on Swin Transformer on both the encoder and decoder side with skip-connections to learn local and global semantic features for multi-abdominal CT segmentation Cao et al. (2021). Similarly, SwinBTS adds the Swin transformer on both encoder–decoder sides with an enhanced transformer module for detailed feature extraction Jiang et al. (2022).

The transformer-based 3D medical image segmentation frameworks usually require lengthy training time and are accompanied by high computational complexity associated with extracting features at multi-scale levels Xie et al. (2021); Shamshad et al. (2022). Such limitations motivate us to rethink if ConvNets can emulate transformer behavior with the benefit of depthwise and large kernel convolution.

2.2 Depthwise and large kernel convolution based

Depthwise separable convolution performs a depthwise convolution followed by a pointwise convolution. It has been proved to be a powerful variation of standard convolution that helps reduce the number of parameters and transfer learning Guo et al. (2019). Zunair et al. introduced an encoder-decoder depthwise fully convolutional network architecture for lung CT segmentation that can be easily scaled for improved performance and outperforming baselines with three-fold learnable parameters Zunair & Hamza (2021). 3D U^2-Net aims to solve multi-modality segmentation in one model fed with depth-wise, depth-wise separate convolution, and shared cross-channel correlations that achieved comparable performance with traditional models but required only about 1% parameters. 3D U^2-Net can be effortlessly adapted to a new domain without sacrificing performance in the modalities used to train the model Huang et al. (2019).

The tradeoff of utilizing a single convolution with a larger kernel size would lead to search-efficient space in 3D CNNs and reduce the number of parameters but high computation and time cost when compared with smaller kernel sizes on GPUs Tang et al. (2020). Researchers have investigated the usefulness of large kernel convolution for medical image segmentation. For instance, GCN employs large kernels (5x5, 7x7) for semantic segmentation, which simultaneously addresses the classification and localization problems Peng et al. (2017); Huo et al. integrated large convolutional kernel

(7x7) layers and the conditional generative adversarial networks to address the spatial and anatomical variations in splenomegaly segmentation task Huo et al. (2018); ConvUNeXt aims to reduce the training parameters of the convolution block of UNet while retaining segmentation superiority utilizing large convolution kernels and depth-wise separable convolution for the convolution block Han et al. (2022); Li et al. implemented large kernel convolution to assemble the attention map of long-range relationships on U-Net for optimizing automatic MRI Brain tumor segmentation Li et al. (2022).

Attribute to the potential of depthwise and large kernel convolution, we hypothesize that combining depthwise convolution and large kernel size can emulate the transformer behavior and meanwhile preserve the segmentation downstream 3D medical image segmentation performance.

3 3D UX-NET: INTUITION

Inspired by Liu et al. (2022), we introduce 3D UX-Net, a simple volumetric ConvNets that adapts the capability of hierarchical transformers and preserves the advantages of using CNN modules such as inductive biases. The basic idea of designing the encoder block in 3D UX-Net can be divided into 1) block-wise and 2) layer-wise perspectives. First, we discuss the block-wise perspective into three views:

- **Patch-wise Features Projection**: Comparing the similarities between ConvNets and ViTs, we found that a common block is introduced to aggressively downscale the feature representations into particular patch size. Here, we adapt a large kernel projection layer to extract patch-wise features as encoder's inputs.
- **Volumetric Depth-wise Convolution with Large Kernels**: One of the intrinsic specialities of swin transformer is the window-shifting non-local self-attention, which allows to generate global receptive field across layers. Inspired by the idea of depth-wise convolution, we found that there are similarities between the weight sum approach in self attention and the convolution per-channel basis. We argue that using depth-wise convolution with large kernel size can provide large receptive field in extracting features similar to the MSA blocks. Therefore, we propose to compress the window shifting characteristics in swin transformer with a volumetric depth-wise convolution using large kernel size (e.g. starting from 7×7). Each channel of the kernel is convolved with the corresponding input channel, so that the output feature can have the same channel dimension as the input. Furthermore, we observe that the multi-self-attention (MSA) block is applied prior to the multi-layer perceptron (MLP). We also move up the position of our re-designed convolutional layer and attempt to have the similar intrinsic structure as swin transformer block.
- **Inverted Bottleneck**: Another intrinsic structure in ViTs is designed with the hidden dimension of the MLP block four times wider than the input dimension. Such design is interestingly correlated to the expansion ratio in ResNet block. Therefore, we leverage the ResNet block design as the basis and convert the 1×1 convolutional layers as linear layers to reduce the model parameters. To further adapt the depth-wise convolution into the ResNet block, the depth-wise convolutional layer is moved up and followed with two consecutive linear layer, as shown in Figure 1.

The macro-design in convolution blocks demonstrates the possibility of adapting the large receptive field and use similar operation of extracting features comparing with swin transformer. We want to further investigate the variation between between CNN and swin transformer in layer-wise setting and refine the model architecture . Here, we further define and adapt the layer-wise difference into another three perspectives:

- **Applying Residual Connections**: With the layer structure comparison of different network state-of-the-arts in Figure 2, 3D U-Net block demonstrates the naive approach of using small kernels to extract local representations with the increase of channels, while the segresnet block applies the residual as similar to the transformer block. Here, we also apply the residual connection between the input and the extract features after the last linear layer. However, we haven't applied any normalization and activation layers before and after the summation of residual, as to be equivalent with the swin transformer structure.

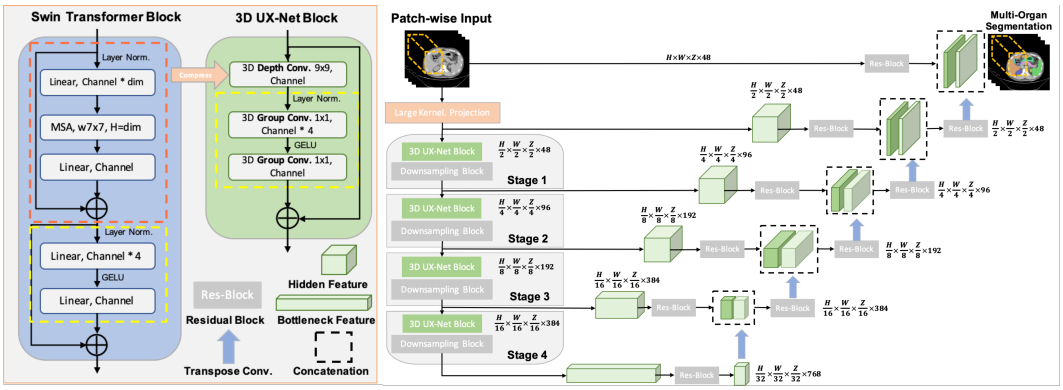

Figure 1: Caption

- **Adapting Layer Normalization (LN)**: In ConvNets, batch normalization (BN) is the common strategy that normalize the convolved representations to enhance convergence and reduce overfitting. However, previous works have been demonstrated to have detrimental effect in model generalizability. Although several approaches are proposed to have an alternative normalization techniques, BN still remains as the optimal choice in volumetric vision task. With the motivation of vision transformer and other NLP models, we directly substitute BN with LN in the encoder block and demonstrate similar operations in ViTs. Also, a significant improvement in segmentation performance is also observed by leveraging LN into encoding the features.
- **Using GELU as Activation Layer**: Previous works have demonstrated the extensive ability of using rectified linear unit (ReLU), which provide the non-linearity in both ConvNets and ViTs. However, for transformer models in NLP domain, the guassian error linear unit (GELU) is shown to be a smoother variant and tackle the limitation of sudden zero in the negative input range in ReLU. Therefore, we further subsitute the ReLU with the GELU activation function.

4 3D UX-NET: COMPLETE NETWORK DESCRIPTION

3D UX-Net comprises multiple re-designed volumetric convolution blocks that directly utilizes 3D patches. Skip connections are further leveraged to connect the multi-resolution features to a convolutional-based decoder network. Figure 1 illustrates the complete architecture of 3D UX-Net. We further describe the details of encoder and decoder in this section.

4.1 DEPTH-WISE CONVOLUTION ENCODER

Given a set of 3D image volumes $V_i = X_i, Y_{i i=1,\ldots,L}$, random sub-volume $P_i \in \mathcal{R}^{H\times W\times D\times C}$ is extracted to be the input for the encoder network. Instead of flattening the patches and project it with linear layer Hatamizadeh et al. (2022b), we leverage a large kernel convolutional layer to compute partitioned feature map with size $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}$ that are projected into a $C = 48$-dimensional space. To adapt the characteristics of compute local self-attention, we use the group convolution with kernel size $7\times7\times7\times7$ (GC-7) with padding of 3 to act as a "shifted window" and evenly divide the feature map. As global self-attention is generally not affordable with large number of patches extracted in swin transformer Liu et al. (2021), we hypothesis that performing group convolution with large kernel size can effectively extract features with global receptive field. Therefore, we define the output of encoder blocks in layers l and $l+1$ as follows:

$$\begin{aligned}
\hat{z}^{l} &= \text{GC-7}(\text{LN}(z^{l-1})) + z^{l-1} \\
z^{l} &= \text{MLP}(\text{LN}(\hat{z}^{l})) + \hat{z}^{l} \\
\hat{z}^{l+1} &= \text{GC-7}(\text{LN}(z^{l})) + z^{l} \\
z^{l+1} &= \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1}
\end{aligned} \tag{1}$$

Where $\hat{z}_l$ and $\hat{z}_{l+1}$ are the outputs from the GC-7 layer in different depth level; LN and MLP denote as layer normlaization and multi-layer perceptron (see. Figure 1). Comparing to the swin transformer, we subsitute the regular and window partiioning multi-head self-attention modules, W-MSA and SW-MSA respectively, with two GC-7 layers.

With the motivation of SwinUNETR Tang et al. (2022), the complete architecture of the encoder consists of 4 stages comprising of 2 large kernel convolution blocks at each stage (*i.e.* L=8 total layers). Between every stage, instead of having a patch merching layer in swin transformer, we leverage a convolution block with kernel size $2 \times 2 \times 2$ with stride 2 to reduce the feature resolution by a factor of 2. Furthermore, a MLP network is followed with the large kernel group convolution in each block. The linear embedding layer helps to maintain the dimension of feature map and extract depth-wise feature with 4 times of the input channel size. The same procedure continues in stage 2, stage 3 and stage 4 with the resolutions of $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}$, $\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}$ and $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$ respectively. Such hierarchical representations in multi-scale setting are extracted in each stage and further leverage for learning dense volumetric segmentation.

4.2 Decoder

The multi-scale output from each stage in the encoder is connected to a ConvNet-based decoder via skip connections and form a "U-shaped" like network for downstream segmentation task. Specifically, we extract the output feature mapping of each stage $i (i \in 0, 1, 2, 3, 4)$ in the encoder and further leverage a residual block comprising with two post-normalized $3 \times 3 \times 3$ convolutional layers with instance normalization to stablize the extracted features. The processed features from each stage are then upsampled with a transpose convolutional layer and concatentate the features from the preceding stage. For downstream volumetric segmentation, we also concatenate the residual features from the input patches with the upsampled features and input it into a residual block with $1 \times 1 \times 1$ convolutional layer with a softmax activation to predict the segmentation probabilities.

5 Experimental Setup

Datasets We conduct experiments on three public multi-modality datasets for volumetric segmentation, which comprising with 1) MICCAI 2021 FeTA Challenge dataset (FeTA2021), 2) MICCAI 2021 FLARE Challenge dataset (FLARE2021) and 3) MICCAI 2022 AMOS Challenge dataset (AMOS2022). For FETA2021 dataset, we employ 80 T2-weighted infant brain MRI from the University Children's Hospital with 1.5 T and 3T clinical whole-body scanners for brain tissue segmentation. Seven specific tissues are manually well-annotated in FeTA 2021 including external cerebrospinal fluid (ESF), grey matter, white matter, ventricles, cerebellum, deep grey matter (DGM) and brainstem. For FLARE2021 and AMOS2022, we employ 511 multi-contrast abdominal CT from FLARE2021 with four anatomies manually annotated (spleen, kidney, liver and pancreas), and 200 multi-contrast abdominal CT from AMOS 2022 with sixteen anatomies manually annotated (spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, prostate/uterus) for abdominal organs segmentation.

Preprocessing & Model Training We apply hierarchical steps for data preprocessing: 1) intensity clipping is applied to further enhance the contrast of soft tissue (FeTA2021:{min:0, max:1000}, FLARE2021 & AMOS2022:{min:-175, max:250}). 2) Intensity normalization is performed after clipping for each volume and use min-max normalization: $(X - X_1)/(X_{99} - X_1)$ to normalize the intensity value between 0 and 1, where X_p denote as the p_{th} percentile of intensity in X. We then randomly crop sub-volumes with size $96 \times 96 \times 96$ at foreground and perform data augmentations including rotations, intensity shifting, and scaling (scaling factor: 0.1). All training process with 3D UX-Net are optimized with an Adam optimizer (weight decay: $1 \times 10^{-}4$; momentum: 0.9; batch size: 1). We trained all models for 40000 steps using a learning rate of 0.0001 on a NVIDIA-Quadro RTX 5000. 1 epoch takes approximately about 1 minutes for FeTA2021, 10 minutes for FLARE2021 and 7 minutes for AMOS2022, respectively.

Implementation Details Following with prior works, we perform evaluations on two scenarios: 1) direct supervised training and 2) transfer learning with pretrained weights. For the supervised training scenario, FeTA2021 and FLARE2021 datasets are used to evaluate the segmentation per-

formance of 3D UX-Net with different modalities and anatomies. We perform five-fold cross validations with 80% (train)/10% (validation)/10% (test) split to both FeTA2021 and FLARE2021 datasets. For the transfer learning scenario, we leverage the pretrained weights from the best fold model trained with FLARE2021, and finetune the model weights on AMOS2022 to evaluate the fintuning capability of 3D UX-Net . We also perform the same train/validation test split in the direct training scenario for finetuning. Overall, we evaluate 3D UX-Net performance by comparing with current transformer and ConvNet state-of-the-art approaches for volumetric segmentation in fully-supervised setting. We use Dice similarity coefficient as an evaluation metric to compare the overlapping regions between predictions and ground-truth label. Furthermore, we performed ablation studies to investigate the effect on different kernel size and the variability of substituting linear layers with group-wise convolution for feature extraction.

6 RESULTS

7 DISCUSSION

8 CONCLUSION

REFERENCES

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Yunhui Guo, Yandong Li, Liqiang Wang, and Tajana Rosing. Depthwise convolution is all you need for learning multiple visual domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8368–8375, 2019.

Zhimeng Han, Muwei Jian, and Gai-Ge Wang. Convunext: An efficient convolution neural network for medical image segmentation. *Knowledge-Based Systems*, pp. 109512, 2022.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pp. 272–284. Springer, 2022a.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 574–584, 2022b.

Chao Huang, Hu Han, Qingsong Yao, Shankuan Zhu, and S Kevin Zhou. 3d u

ˆ

-net: A 3d universal u-net for multi-domain medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 291–299. Springer, 2019.

Yuankai Huo, Zhoubing Xu, Shunxing Bao, Camilo Bermudez, Hyeonsoo Moon, Prasanna Parvathaneni, Tamara K Moyo, Michael R Savona, Albert Assad, Richard G Abramson, et al. Splenomegaly segmentation on multi-modal mri using deep convolutional networks. *IEEE transactions on medical imaging*, 38(5):1185–1196, 2018.

Yun Jiang, Yuan Zhang, Xin Lin, Jinkun Dong, Tongtong Cheng, and Jing Liang. Swinbts: a method for 3d multimodal brain tumor segmentation using swin transformer. *Brain Sciences*, 12(6):797, 2022.

Hao Li, Yang Nan, and Guang Yang. Lkau-net: 3d large-kernel attention-based u-net for automatic mri brain tumor segmentation. In *Annual Conference on Medical Image Understanding and Analysis*, pp. 313–327. Springer, 2022.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4353–4361, 2017.

Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *arXiv preprint arXiv:2201.09873*, 2022.

Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European conference on computer vision*, pp. 685–702. Springer, 2020.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20730–20740, 2022.

Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K Nandi. Medical image segmentation using deep learning: A survey. *IET Image Processing*, 16(5):1243–1267, 2022.

Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 171–180. Springer, 2021.

Hasib Zunair and A Ben Hamza. Sharp u-net: depthwise convolutional network for biomedical image segmentation. *Computers in Biology and Medicine*, 136:104699, 2021.

A Appendix

You may include other additional sections here.