Special Paper
방송공학회논문지 제27권 제7호, 2022년 12월 (JBE Vol. 27, No. 7, December 2022)
https://doi.org/10.5909/JBE.2022.27.7.1011
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

High-Speed Transformer for Panoptic Segmentation

Jong-Hyeon Baek[a)], Dae-Hyun Kim[a)], Hee-Kyung Lee[b)], Hyon-Gon Choo[b)], and Yeong Jun Koh[a)‡]

Abstract

Recent high-performance panoptic segmentation models are based on transformer architectures. However, transformer-based panoptic segmentation methods are basically slower than convolution-based methods, since the attention mechanism in the transformer requires quadratic complexity w.r.t. image resolution. Also, sine and cosine computation for positional embedding in the transformer also yields a bottleneck for computation time. To address these problems, we adopt three modules to speed up the inference runtime of the transformer-based panoptic segmentation. First, we perform channel-level reduction using depth-wise separable convolution for inputs of the transformer decoder. Second, we replace sine and cosine-based positional encoding with convolution operations, called conv-embedding. We also apply a separable self-attention to the transformer encoder to lower quadratic complexity to linear one for numbers of image pixels. As result, the proposed model achieves 44% faster frame per second than baseline on ADE20K panoptic validation dataset, when we use all three modules.

Keywords : Panoptic segmentation, Transformer

I. Introduction

The transformer[1] has been widely adopted in many state-of-the-art models. The core of the transformer, the attention mechanism, allows neural networks to exploit global information effectively and improves many vision tasks, such as image classification[2], object detection[3], and semantic segmentation[4] on various benchmark datasets. Recently, the transformer is also used in panoptic segmentation[5] and has achieved remarkable improvements in the panoptic segmentation field. Despite such advances, the attention mechanism in the transformer demands the high computational complexity, and this yields the bottleneck for inference. There are two main issues for the high computational complexity. First, the self-attention in the transformer has the quadratic complexity with respect to the number of image pixels. Thus, the transformer-based pan optic segmentation methods are weak to high-resolution

a) Department of Computer Engineering, ChungNam National University
b) Electronics and Telecommunications Research Institute
‡ Corresponding Author : Yeong-Jun Koh
E-mail: yjkoh@cnu.ac.kr
Tel: +82-42-821-7442
ORCID: https://orcid.org/0000-0003-1805-2960

※ This work was supported partly by Institute of Information & communications Technology Planning & evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00011, Video Coding for Machine) and partly by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2022 R1I1A3069113).
· Manuscript October 17, 2022; Revised December 2, 2022; Accepted December 8, 2022.

images in terms of inference speed. Second, the positional encoding[1] based on sine and cosine computation in the original transformer is also the problem for slow inference speed. To address this problem, some works tried to lower the complexity of the attention mechanism. For example, Axial-Deeplab[6] computes attention along the height and width axis, respectively. Also, MobileViTv2[7] reduces the computation cost of attention. Moreover, it does not even use the positional encoding.

To utilize transformer-based panoptic segmentation in real-world applications such as MPEG video coding for machine, efficient models are essential. In this work, we introduce three models to achieve high-speed panoptic segmentation. First, we use a depth-wise separable convolution[8] to reduce the channel dimensions of image features for the transformer decoder. The depth-wise separable convolution can effectively reduce the computational complexity in the transformer decoder without any performance degradation. Second, we replace sine and cosine-based positional embedding in the original transformers with a convolution layer, called conv-embedding, in the transformer encoder and decoder. The simple convolution layer can extract the positional information, and thus it substitutes the traditional positional embedding. We also further reduce the complexity of the attention using the separable attention[7] for transformer encoder. Experimental results demonstrate that these three modules efficiently improve inference speed of MaskFormer[9] on the ADE20K panoptic segmentation dataset.

II. Related works

1. Efficient Vision Transformer

Recently, there are some efficient vision transformer models to improve the inference speed. For the classification task, EfficientFormer[10] applied vision transformers to the light-weight backbone models such as MobileNet[8]. However, EfficientFormer employed the original attention module, and thus it suffered from the quadratic complexity in the attention process. MobileViT[11] used inter-patch attention, unfolding, and folding structure to connect local information and global information, but it was still slower than MobileNetv2[12]. MobileViTv2[7], which is an extension of MobileViT, developed a separable self-attention to decrease the computational complexity of the traditional attention mechanism. For semantic segmentation, Top-Former[13] used the average pooling operator to reduce the computational costs for various scale inputs.

2. Panoptic segmentation

Segmentation is one of the most important areas in the field of computer vision. Segmentation has various tasks such as semantic segmentation, instance segmentation, and panoptic segmentation. Semantic segmentation is a method of assigning classes to all pixels of an image, and instance segmentation is a method of determining whether each pixel in an image is an object. Panoptic segmentation unifies the above two tasks and defines ideal outputs for thing classes as instance segmentations and for stuff classes as semantic segmentation.

First, Kirillov et al.[14] proposed the heuristic method, which combines independently obtained semantic segmentation and instance segmentation results, to assign class labels to each pixel. Bowen Cheng et al.[15] presented an end-to-end model, called Panoptic DeepLab, which consists of a semantic head, instance center head, and regression head. It determines stuff classes based on the semantic head while finding instance locations by selecting instance centers. Recently, transformer-based approaches have been developed in panoptic segmentation. DETR[16], which is the first end-to-end model based on the transformer in object detection, constructed queries for thing and stuff classes and it performed object detection through the attention mechanism of the learnable queries and image features. Then, DETR extracted segmentation results from the detected boundingboxes. Unlike DETR requires two-stage process to obtain panoptic

segmentation, MaskFormer proposed a new approach to convert any per-pixel classification model into mask classification. The model also adopted the transformer decoder to extract class prediction and mask embedding vectors for thing and stuff queries. However, transformers in MaskFormer require the high computational complexity that reduces the inference speed. In this work, we introduce three modules to improve the inference runtime of MaskFormer.

III. Proposed Methods

In this section, we introduce three modules, a depth-wise separable convolution module, a convolution embedding module, and a separable attention module[7], which reduce inference runtime of transformer-based panoptic segmentation. In this work, we employ MaskFormer as the baseline and add the three modules to the baseline. The depth-wise separable convolution module applies channel-level reduction, the separable attention module reduces the computational complexity in the transformer encoder, and the convolution embedding module improves the computation speed as compared with the positional encoding. Figure. 1 illustrates an overview of the proposed network.

1. Depth-wise Separable Convolution

The depth-wise separable convolution focuses on reducing the channel dimension of inputs (query, key, and value) of multi-head cross attention in the transformer decoder. For this purpose, the image feature $F' \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C}$, extracted from the transformer encoder, passes through the depth-wise separable convolution to form a reduced feature $\mathrm{R} \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times \frac{C}{2}}$. Here, $H \times W$ is a spatial resolution and C is a feature dimension. The depth-wise separable convolution consists of two layers: the first layer is depth-wise convolution and the second layer is point-wise convolution. In depth-wise convolution, one filter is applied to on only one channel. This operation reduces the computational complexity. Point-wise convolution is a convolution with a 1×1 kernel size to reduce the number of channel dimensions and combine the information on image features, extracted from depth-wise convolution. Then, the feature R is transformed into the

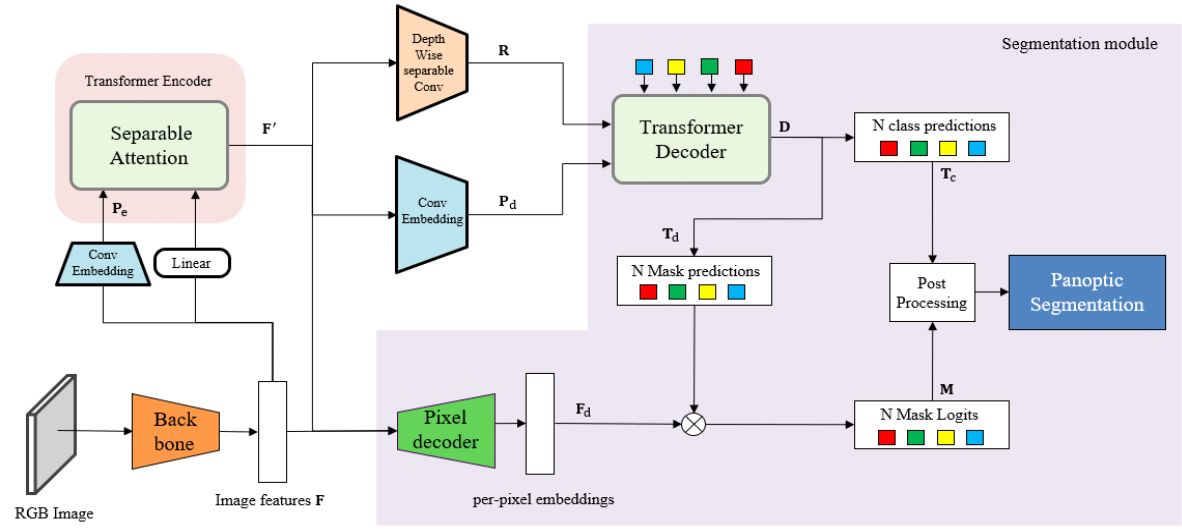

Fig. 1. Overview of the proposed network

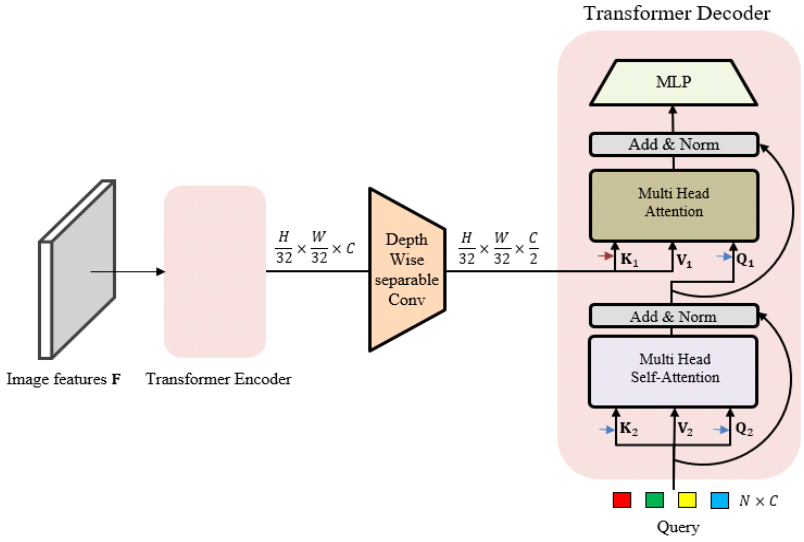

Fig. 2. Diagram of the depth-wise separable convolution

key $K_1 \in \mathbb{R}^{\frac{H}{32}\frac{W}{32}\times\frac{C}{2}}$ and value $V_1 \in \mathbb{R}^{\frac{H}{32}\frac{W}{32}\times\frac{C}{2}}$ that have reduced channel dimension $\frac{C}{2}$. Also, N per-segment embeddings $Q \in \mathbb{R}^{N\times C}$ are projected into $K_2 \in \mathbb{R}^{N\times\frac{C}{2}}$, $V_2 \in \mathbb{R}^{N\times\frac{C}{2}}$, and $Q_2 \in \mathbb{R}^{N\times\frac{C}{2}}$ for multi-head self-attention, and then the query $Q_1 \in \mathbb{R}^{N\times\frac{C}{2}}$ is obtained to perform cross attention with K_1 and V_1.

2. Convolution Embedding

In general, positional encoding is employed in the transformer encoder and decoder to inject relative position information in images. In the baseline, positional encoding adopts sine and cosine functions to obtain positional embeddings $P \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times C}$ for x and y coordinates

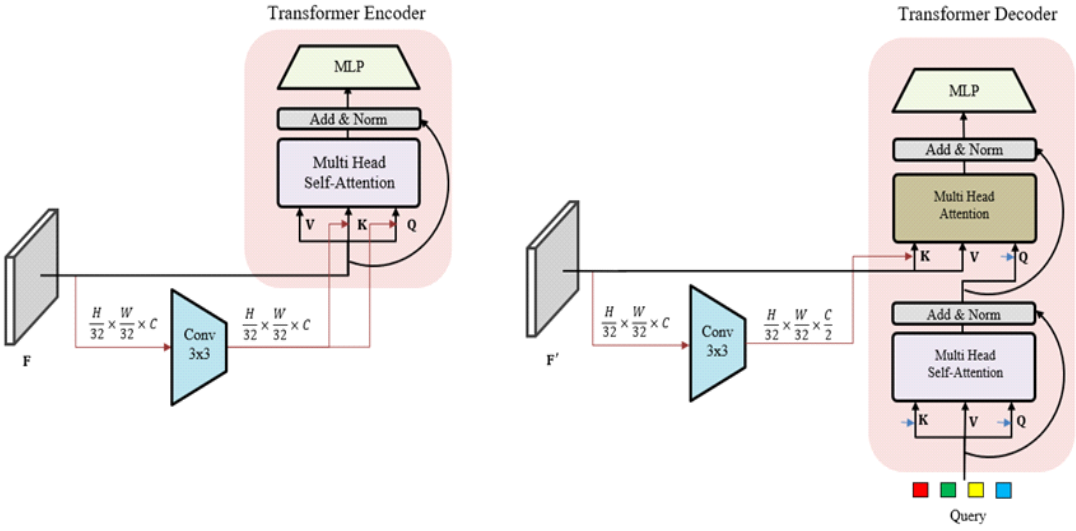

Fig. 3. The convolution embedding in the transformer encoder and transformer decoder

$$\begin{aligned} &\mathrm{P}(x,y,2i) \\ &= \sin(x/10000^{2i/C}), \mathrm{P}(x,y,2i+1) \\ &= \cos(x/10000^{2i/C}) \end{aligned} \tag{1}$$

$$\begin{aligned} &\mathrm{P}(x,y,2i+C/2) \\ &= \sin(x/10000^{2i/C}), \mathrm{P}(x,y,2i+1+C/2) \\ &= \cos(x/10000^{2i/C}) \end{aligned}$$

where x and y are normalized absolute coordinates in the input feature. P can be simply computed, but we experimentally observe that the sine and cosine computation require a high computational complexity as compared with convolution. Therefore, we replace positional encoding with the convolution embedding that contains a 3×3 convolution layer, as shown in Figure 3. The convolution embedding takes image features F and F′ as inputs for transformer encoder and decoder, respectively. Thus, using the convolution embedding, we can obtain two positional embeddings $\mathrm{P_e}$ and $\mathrm{P_d}$ for transformer encoder and decoder, respectively.

3. Separable Attention

The transformer encoder in the baseline is designed to enhance image features. The original transformer encoder in the baseline consists of self-attention, layer normalization, and multi-layer-perceptron. Since the input of self-attention contains image feature information extracted from the backbone, the size of the input greatly affects the inference speed. Therefore, we replaced the self-attention in the original transformer encoder with the separable attention[7] in Figure 4, which can improve the inference speed by reducing the computation in attention processes.

As in Equation (2), the general attention requires quadratic complexity with regard to image resolutions:

$$\mathrm{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{softmax}\left(\frac{\mathrm{QK}^T}{\sqrt{d_k}}\right)\mathrm{V}. \tag{2}$$

In contrast, the separable attention makes feature di-

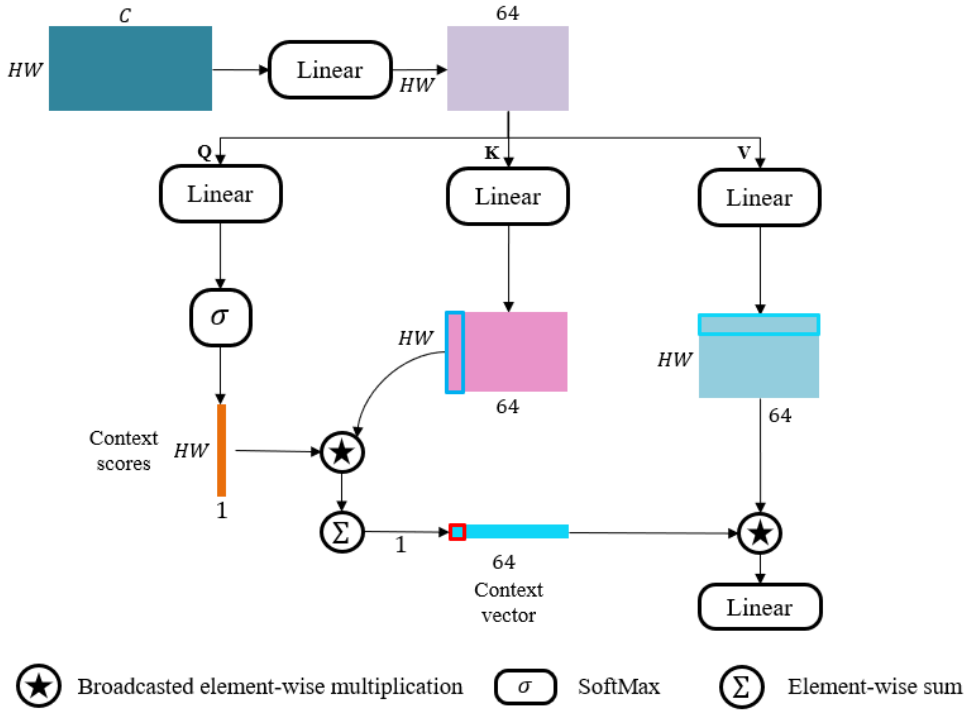

Fig. 4. The process of the separable attention

mension for each pixel in query Q to scalar 1 to obtain context scores $c_s \in \mathbb{R}^{HW\times 1}$. Then, using c_s, a context vector c_v is computed by a weighed sum of key K as

$$c_v = \sum_{i=1}^{HW} c_s(i)\mathrm{K}(i) \quad (3)$$

Finally, value V is updated by broadcasted element-wise multiplication with c_v. By reducing dimension to scalar 1 for c_s, the computation cost for attention is reduced from $O(H^2W^2)$ to $O(HW)$.

4. Segmentation module

Using learnable parameters $\mathrm{D} \in \mathbb{R}^{N\times C}$, extracted from the transformer decoder, the segmentation module predicts N class prediction features $\mathrm{T}_c \in \mathbb{R}^{N\times C_{class}}$ containing the class information and N Mask prediction features $\mathrm{T}_d \in \mathbb{R}^{N\times C_{class}}$. To generate N mask logits $\mathrm{M} \in \mathbb{R}^{H\times W\times N}$, it performs a dot product operation between per-pixel embeddings $\mathrm{F}_d \in \mathbb{R}^{H\times W\times C}$, which is obtained from the pixel decoder, and N mask prediction features T_d. Finally, the panoptic segmentation is predicted through the post-processing process that takes N class predictions and N mask logits as inputs and generates a mask with the highest probability value without duplication.

Ⅳ. Experimental results

1. Dataset and Metric

To validate the proposed network, we use the ADE20K panoptic dataset. ADE20K contains 20,210 images for training and 2,000 images for validation. It has 100 classes for things and 50 classes for stuff. Figure 5 shows examples of pairs of images and labels. For the metric, we use the standard PQ[14] metric. PQ is composed of segmentation quality (SQ) and recognition quality (RQ). SQ is the average IoU of matched segments and RQ is F1 score for quality estimation in class and mask. Thus, PQ is defined as $\mathrm{PQ} = \mathrm{SQ} \times \mathrm{RQ}$

2. Implementation Details and Training Settings

For the baseline, we use the same structure and hyperparameters to MaskFormer[9]. Also, we use the same loss to the baseline, which is composed of cross-entropy loss, mask loss, and dice loss. For the backbone, we employ ResNet50[17] with pre-trained weights from the baseline. For training, input images are randomly cropped and flipped horizontally and vertically. We train the network during 200 epochs with a batch size of 8 using four RTX A6000 GPUs. We employ the AdamW[18] optimizer and the

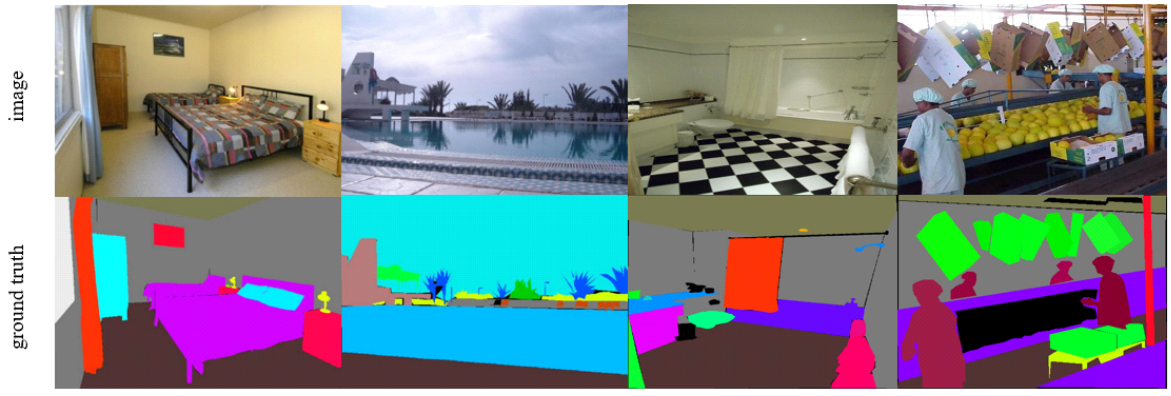

Fig. 5. ADE20K panoptic dataset

poly learning rate schedule with an initial learning rate of 10^{-4} and a weight decay of 10^{-4}.

3. Experimental Results on ADE20K

We perform experiments by adding the depth-wise separable convolution, convolution embedding, and separable attention modules to the baseline sequentially. Table 1 shows PQ scores and frame per second (fps) according to those three modules on ADE20K. Fps is measured on a RTX A6000 GPU with a batch size of 1 by computing the average runtime on the entire validation set. Also, the runtime includes the time to perform the post-processing. First, as compared with the baseline, the depth-wise separable convolution (S_1) improves the inference runtime about 11% without any performance degradation. Also, when the convolution embedding is applied to the transformer decoder only, setting S_2 provides the best PQ score and the higher fps than the baseline. Next, when we use the convolution embedding to both transformer encoder and decoder, we observe that setting S_3 yields the best trade-off between the performance and runtime. This case outperforms the baseline by 0.7 PQ and 9.7fps. Finally, in setting S_4, the separable attention extremely boosts the inference speed and requires the smallest number of parameters, but it degrades the PQ performance.

Figure 6 shows the trade-off between performance, speed, and network complexity. As compared with the baseline, settings S_2 and S_3 reduce runtime, while increasing PQ scores. This indicates that the proposed convolution embedding demands the lower computational complexity than the positional embedding, but it yields more effective features for panoptic segmenta- tion. By replacing sign and cosine operations with the simple convolution operation, we achieve the remarkable trade-off between performance and runtime. Setting S_4 significantly reduces the number of network parameters and operations, but it produces the lowest performance due to the restricted network capacity.

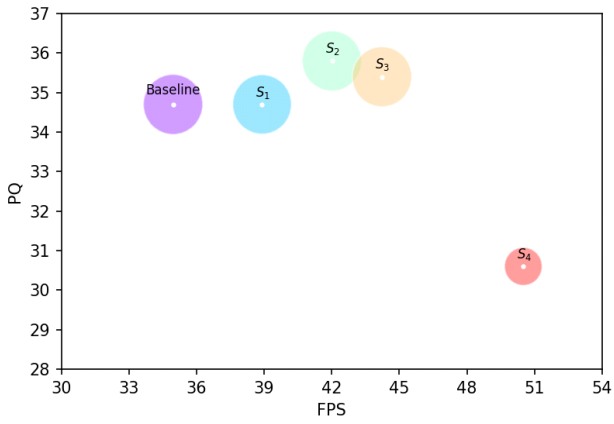

Fig. 6. A ball chart for PQs vs. fps of various proposed settings, in which a ball size indicates the number of network parameters. The PQs and fps are measured on the ADE20K dataset

Table 1. Performance according to three modules on ADE20K Dataset

		PQ	fps	#Params
	Baseline	34.7	34.96	45M
S_1	+ Depth-wise separable convolution	34.7	38.91	44M
S_2	++ Conv-embedding in the transformer decoder	35.8	42.01	45M
S_3	+++ Conv-embedding in the transformer encoder	35.4	44.23	45M
S_4	++++ Separable attention	30.6	50.5	38M

(a) (b) (c)

Fig. 7. (a): images in ADE20K, Panoptic segmentation results of (b) the baseline and (c): the proposed model with the depth-wise separable convolution and Conv-embedding

Table 2 compares the performance of setting with the existing methods: MaskFormer[9], BGRNet[19], Auto- Panoptic[20], and Mask2Former[21]. Setting outperforms MaskFormer in terms of both accuracy and runtime. Also, provides the higher PQ score than BGRNet and Auto-Panoptic. Although yields the lower PQ than Msk2Former, it achieves the best running speed.

Table 2. Performance comparison with the existing methods on the ADE20K dataset.

	PQ	fps	#Params
MaskFormer[9]	34.7	34.96	45M
BGRNet[19]	31.8	-	-
Auto-Panoptic[20]	32.4	-	-
Mask2Former[21]	39.7	22.00	44M
S_3	35.4	44.23	45M

Figure 7 qualitatively compares the proposed model with the baseline. For the proposed model, we select the model with the depth-wise separable convolution and Conv-embedding, which shows the best trade-off between the performance and the runtime. In Figure 7, we can observe that the proposed model provides good panoptic segmentation results as compared with the baseline.

Ⅴ. Conclusion

In this paper, we proposed the high-speed transformer for panoptic segmentation. First, we applied the depth-wise separable convolution to decrease feature dimensions of the transformer encoder output. Second, we substituted sine and cosine computation for positional encoding in the original transformer with the conv-embedding in the transformer encoder and decoder. Finally, we adopted the separable attention instead of the self-attention in the transformer encoder. Experimental results demonstrated that these three modules efficiently improve the inference speed of MaskFormer on the ADE20K panoptic segmentation dataset. In experimental results, we observed that the depth-wise separable convolution and conv-embedding on the transformer encoder and decoder provide the best trade-off between the performance and inference runtime, while the usage of three modules achieves the best inference speed.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. NIPS, 30, 2017.

[2] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Contrastive Captioners are Image-Text Foundation Models," arXiv:2205.01917, 2022.
doi: https://doi.org/10.48550/arXiv.2205.01917

[3] Y. Wei, H. Hu, Z. Xie, Z. Zhang, Y. Cao, J. Bao, D. Chen, and B. Guo, "Contrastive Learning Rivals Masked Image Modeling in Fine-tuning via Feature Distillation," arXiv:2205.14141, 2022.
doi: https://doi.org/10.48550/arXiv.2205.14141

[4] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei, "BEiT Pretraining for All Vision and Vision-Language Tasks," arXiv:2208.10442, 2022.
doi:https://doi.org/10.48550/arXiv.2208.10442

[5] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Towards A Unified Transformer-based Framework for Object Detection and Segmentation," arXiv:2206.02777, 2022.
doi: https://doi.org/10.48550/arXiv.2206.02777

[6] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-DeepLab:Stand-Alone Axial-Attention for Panoptic Segmentation," in Proc. ECCV, pp.108-126, 2020.
doi: https://doi.org/10.1007/978-3-030-58548-8_7

[7] S. Mehta and M. Rastegari, "Separable Self-attention for Mobile Vision Transformers," arXiv:2206.02680, 2022.
doi: https://doi.org/10.48550/arXiv.2206.02680

[8] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arxiv: 1704.04861, 2017.
doi: https://doi.org/10.48550/arXiv.1704.04861

[9] B. Cheng, A. Schwing, and A. Kirillov, "Per-Pixel Classification is Not All You Need for Semantic Segmentation," in Proc. NIPS, 34, 2021.

[10] Y. Li, G. Yuan, Y. Wen, E. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, and J. Ren "EfficientFormer: Vision Transformers at MobileNet Speed," arxiv:2206.01191, 2022.
doi: https://doi.org/10.48550/arXiv.2206.01191

[11] S. Mehta and M. Rastegari, "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer," arxiv:2110.02178, 2022.
doi: https://doi.org/10.48550/arXiv.2110.02178

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L.-C. Chen, "MobileNetV2 Inverted Residuals and Linear Bottlenecks," in Proc. CVPR, Salt Lake City, USA, pp.4510-4520, 2018.
doi: https://doi.org/10.1109/CVPR.2018.00474

[13] W. Zhang, Z. Huang, G. Luo, T. Chen, X. Wang, W. Liu, G. Yu, and C. Shen, "TopFormer: Token Pyramid Transformer for Mobile Semantic Segmentation," in Proc. CVPR, New Orleans, USA, pp.12083-12093, 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.01177

[14] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollar, "Panoptic segmentation," in Proc. CVPR, California, USA, pp.9404-9413, 2019.
doi: https://doi.org/10.1109/CVPR.2019.00963

[15] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. Chen, "Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation," in Proc. CVPR, pp. 12475-12485, 2020.
doi: https://doi.org/10.1109/CVPR42600.2020.01249

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in Proc. ECCV, pp.213-229, 2020.
doi: https://doi.org/10.1007/978-3-030-58452-8_13

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. CVPR, Nevada, USA, pp.770-778, 2016.
doi: https://doi.org/10.1109/CVPR.2016.90

[18] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularizatio n," in ICLR, 2019.

[19] Y. Wu, G. Zhang, Y. Gao, X. Deng, K. Gong, X. Liang, and, L. Lin, "Bidirectional Graph Reasoning Network for Panoptic Segmentation," in CVPR, pp.9080-9089, 2020.
doi: https://doi.org/10.1109/CVPR42600.2020.00910

[20] Y. Wu, G. Zhang, H. Xu, X. Liang, L. Lin, "Auto-Panoptic: Cooperative Multi-Component Architecture Search for Panoptic Segmentation," in NeurIPS, 2020.

[21] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, R. Girdhar, "Masked-Attention Mask Transformer for Universal Image Segmentation," in CVPR, New Orleans, USA, pp.1290-1299 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.00135

Introduction Authors

Jong-Hyeon Baek

- received the B.S. degree computer engineering from Chungnam National University, Daejeon, South Korea, in 2022 As an master degree in Mar. 2022, he joined the Department of Computer Science & Engineering, Chungnam National University.
- ORCID : https://orcid.org/0000-0003-3993-8648
- Research interests : include computer vision and machine learning, especially in the problems of image panoptic

Dae-Hyun Kim

- received the B.S. degree in Computer Science & Engineering from Chungnam National University, Daejeon, South Korea, in 2022. As an MS student in September, 2022, he joined the Department of Computer Science & Engineering, Chungnam National University.
- ORCID : https://orcid.org/0000-0002-3773-1224
- Research interests : computer vision and machine learning, especially in the problems of panoptic segmentation

Hee-Kyung Lee

- received her BS degree in computer engineering from Yeungnam University, Daegu, Rep. of Korea, in 1999, and her MS degree in engineering from the Information and Communication University, Daejeon, Rep. of Korea, in 2002. Since 2002, she has worked for Electronics and Telecommunications Research Institute Daejeon, Rep. of Korea, where she is now serving as a senior member of engineering staff. She participated in “TV-Anytime” standardization and IPTV Metadata standardization. She also involved in the development of gaze tracking technology.
- ORCID : https://orcid.org/0000-0002-1502-561X
- Research interests : personalized service via metadata, HCI, Gaze Tracking, Bi-directional advertisement and video content analysis, and VR/AR/MR

Hyon-Gon Choo

- received his B.S. and M.S. degree in electronic engineering in 1998 and 2000 respectively, and his Ph.D degree in electronic communication engineering in 2005 from Hanyang University, Korea.
  He is currently working as a Principal Researcher in Electronics and Telecommunications Research Institute, Daejeon, Korea. He was a director of Digital holographic research section
- ORCID : https://orcid.org/0000-0002-0742-5429
- Research interests : video coding for machines, holography, multimedia protection and 3D broadcasting technolgies

Yeong Jun Koh

- received the B.S. degree and the Ph.D. degree in electrical engineering from Korea University, Seoul, South Korea, in 2011 and 2018, respectively. As an assistant professor in Mar. 2019, he joined the Department of Computer Science \& Engineering, Chungnam National University.
- ORCID : https://orcid.org/0000-0003-1805-2960
- Research interests : computer vision and machine learning, especially in the problems of video object segmentation and image enhancement