TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation

Jieneng Chen[1], Yongyi Lu[1], Qihang Yu[1], Xiangde Luo[2],
Ehsan Adeli[3], Yan Wang[4], Le Lu[5], Alan L. Yuille[1], and Yuyin Zhou[3]

[1]Johns Hopkins University
[2]University of Electronic Science and Technology of China
[3]Stanford University
[4] East China Normal University
[5]PAII Inc.

Introduction

Background

- U-Net, de-facto choice in medical image segmentation
- Limitations in long-range relation due to locality of convolution operations.
- Other studies apply self-attention to CNNs or use Transformers instead to capture global contexts.

UNet

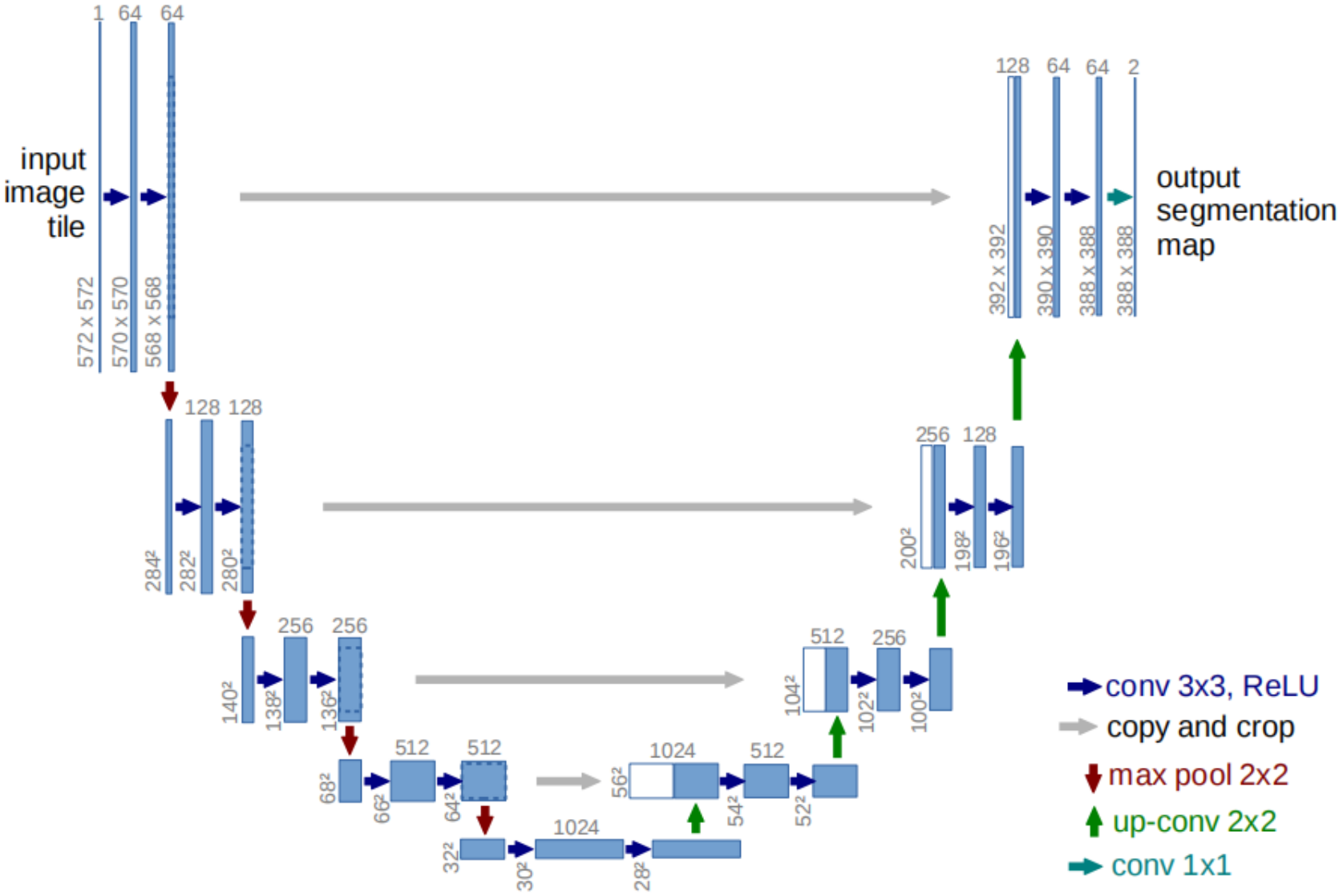


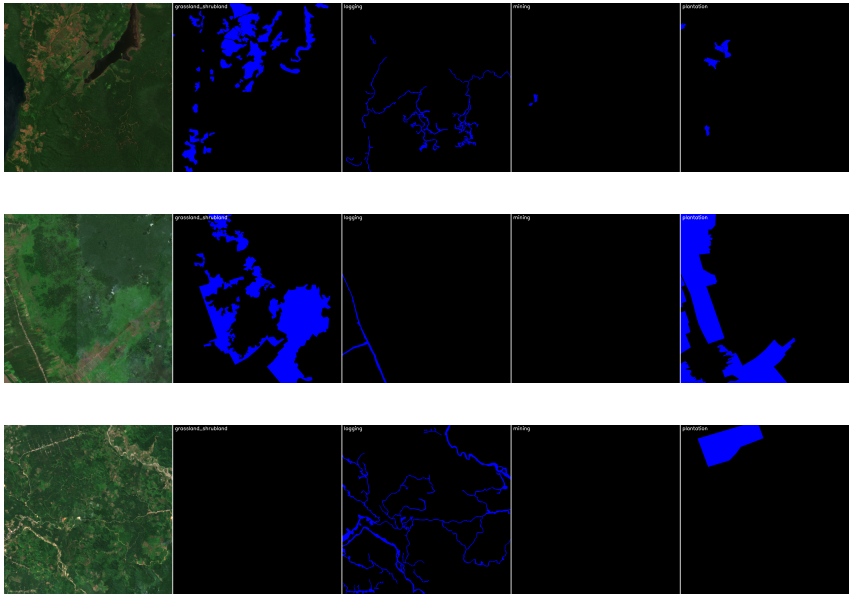
Figure 1: UNet Architecture

In this paper

- Study the potential of transformers in medical image segmentation.
- Transformers capture global contexts.
- UNet (CNNs) extracts low-level details.
- TransUNet: a hybrid CNN-Transformer architecture
- Compare with other architectures on medical images segmentation.
- Ablation studies on model parameters

Relevance for competition (Deforestation Drivers)

- UNet as baseline
- Both are segmentation tasks
- Does TransUNet perform better?



Method and Architecture

Transformer as Encoder

- Input image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$
- Image Sequentialization
- Patch Embeddings
- Naive Upsampling
- Output a $H \times W \times S$ pixel-wise segmentation.

Image Sequentialization

- Reshape input $\mathbf{x}$ into a sequence of 2D patches.
- $\{\mathbf{x}_p^i \in \mathbb{R}^{P^2} \cdot C | i = 1, \ldots, N\}$
- Each patch is $P \times P$.
- $N = \frac{HW}{P^2}$ is the number of patches, i.e. the sequence length.

Patch Embedding

- Pass patches into a latent D-dimensional embedding space with a trainable linear projection.
- $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}$ is the patch embedding projection.
- $\mathbf{E}_{pos} \in \mathbb{R}^{N \times D}$ is the position embedding.

$$z_0 = [x_p^1 \mathbf{E}; x_p^2 \mathbf{E}; \ldots; x_p^N \mathbf{E}] + \mathbf{E}_{pos}$$

Transformer Encoder

- L layers of Multihead Self-Attention and Multi-Layer Perceptron
- $z'_\ell = MSA(LN(z_{\ell-1})) + z_{\ell-1}$
- $z'_\ell = MLP(LN(z'_\ell)) + z'_\ell$
- $LN(\cdot)$ is the layer normalization

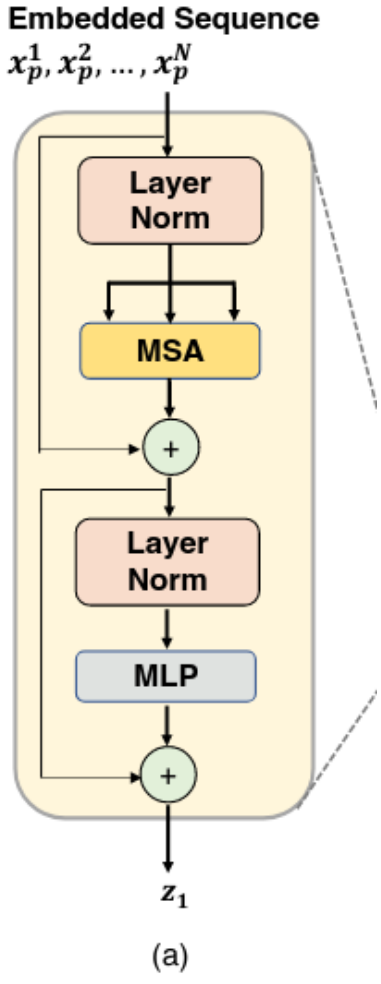


Figure 2: Transformer Encoder

TransUNet

- Hybrid CNN and Transformer encoder
- Cascaded Upsampler (CUP)
- Skip connections from CNN

TransUNet

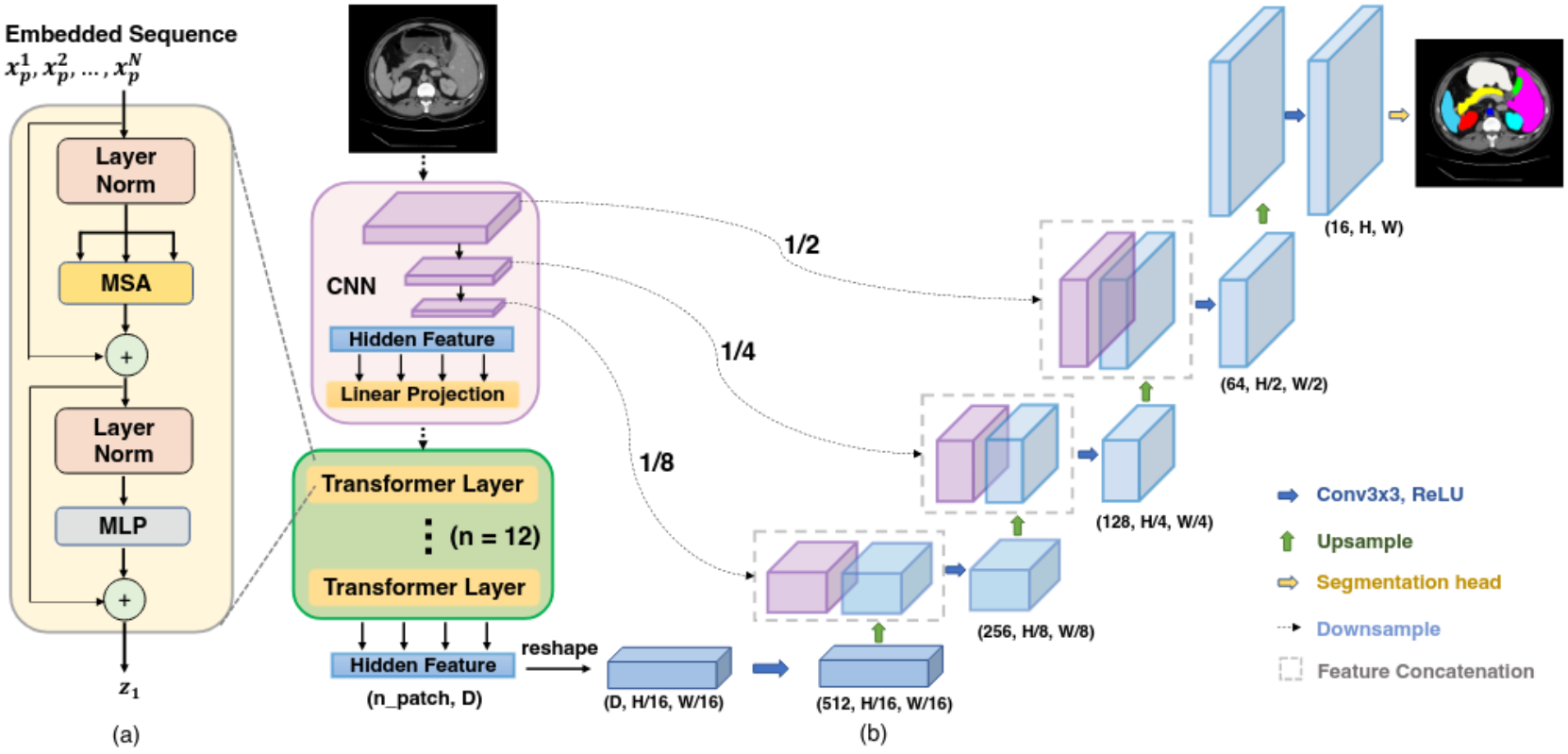


Figure 3: TransUNet architecture

Experiments and Results

Dataset and Evaluation

- Synapse multi-organ segmentation dataset
- CT scans images
- Metrics: Dice Similarity Score (DSC) and Hausdorff Distance (HD)

Comparison with State-of-the-arts

Framework		**Average**		Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
Encoder	Decoder	DSC ↑	HD ↓								
V-Net [9]		68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
DARR [5]		69.77	-	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
R50	U-Net [12]	74.68	36.87	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
R50	AttnUNet [13]	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
ViT [4]	None	61.50	39.61	44.38	39.59	67.46	62.94	89.21	43.14	75.45	69.78
ViT [4]	CUP	67.86	36.11	70.19	45.10	74.70	67.40	91.32	42.00	81.75	70.44
R50-ViT [4]	CUP	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUNet		**77.48**	**31.69**	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62

Figure 4: Synapse dataset results

Segmentation comparison

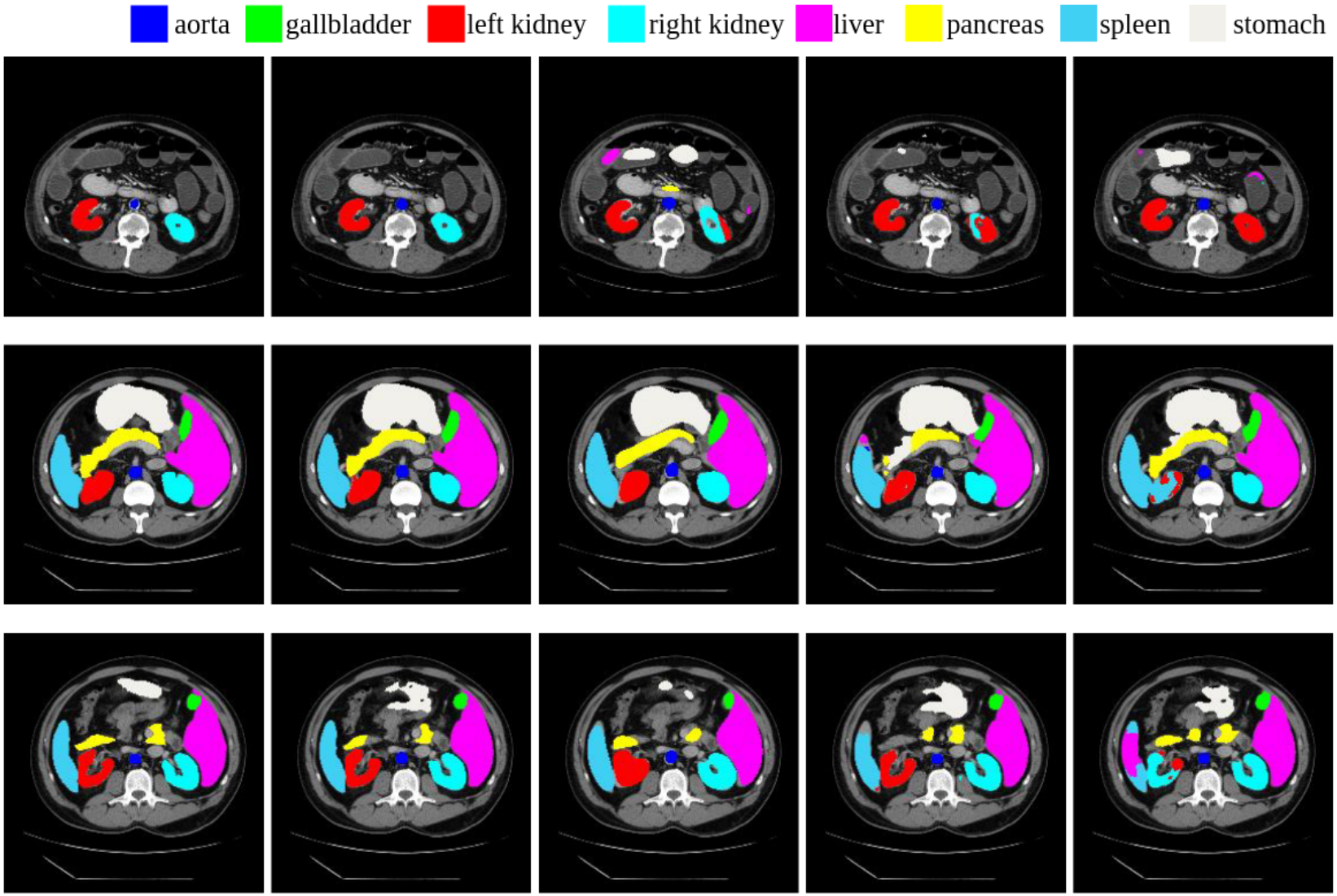


Figure 5: TransUNet predicts less false positive and keep finer information

Ablation Studies

Number of Skip-connections

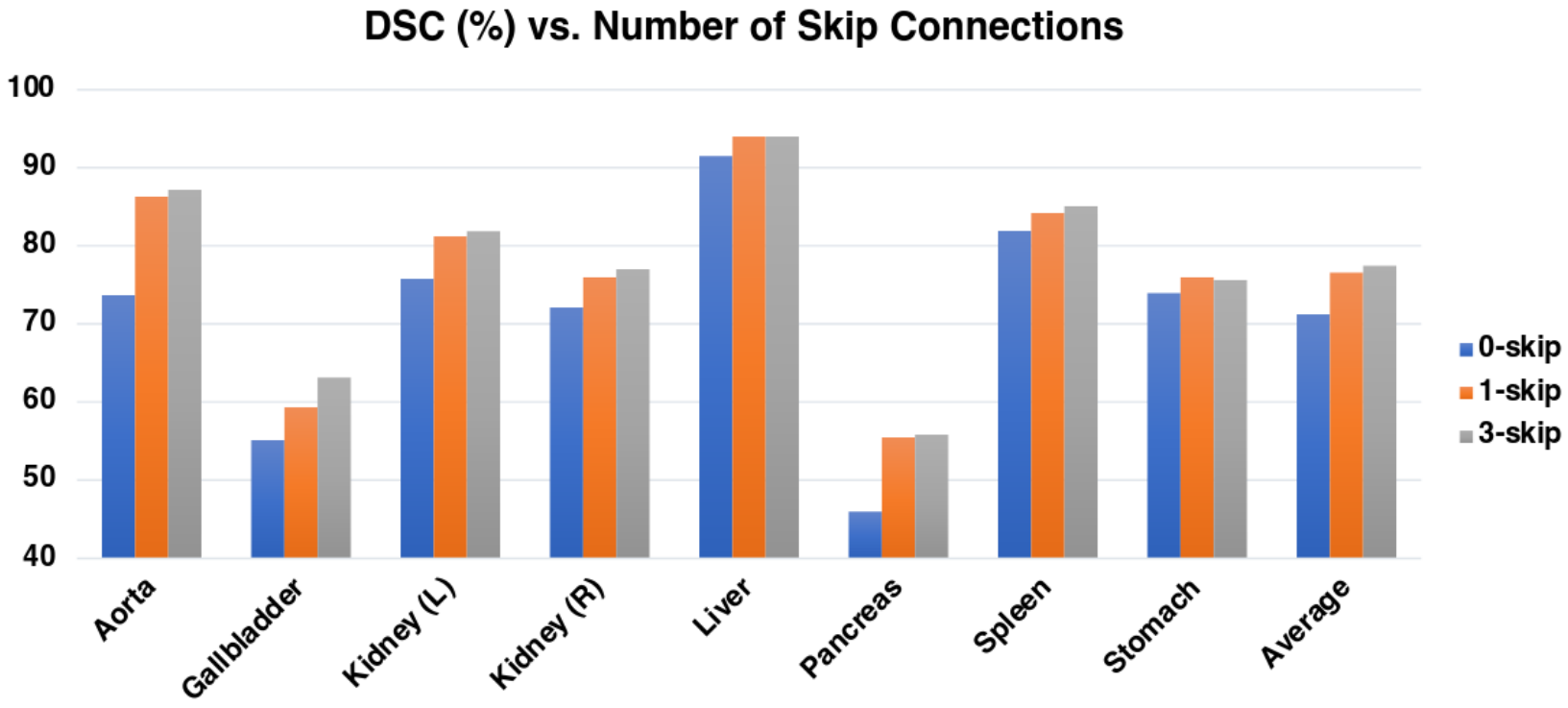


Figure 6: Ablation study on the number of skip-connections

Influence of Input Resolution

Resolution	Average DSC	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
224	**77.48**	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
512	**84.36**	90.68	71.99	86.04	83.71	95.54	73.96	88.80	84.20

Figure 7: Ablation study on the influence of input resolution

Influence of Patch Size/Sequence Length

Patch size	Seq_length	Average DSC	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
32	49	76.99	86.66	63.06	81.61	79.18	94.21	51.66	85.38	74.17
16	196	77.48	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
8	784	**77.83**	86.92	58.31	81.51	76.40	93.81	58.09	87.92	79.68

Figure 8: Ablation study on patch size and sequence length

Model Scaling

- Parameters for the hidden size D, number of layers, MLP size, and number of heads.
- Base: 12, 768, 3072, 12
- Large: 24, 1024, 4096, 16

Model scale	Average DSC	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
Base	77.48	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Large	78.52	87.42	63.92	82.17	80.19	94.47	57.64	87.42	74.90

Figure 9: Ablation study on model scale

Conclusion

- First usage of Transformers in (medical) image segmentation.
- Global context with Transformers and low-level details with CNNs.
- TransUNet performs better than other competing methods.