Dual-Encoder Transfer Learning for Surgical Image Segmentation

Dennis Barnes✉[1], Antonio Rodriguez-Sanchez[2], Ana Lamas Rodríguez[2], Florian Ponholzer[3]

[1]Universität Innsbruck, Austria.
✉dennis.barnes@uibk.ac.at.
[2]CiTIUS, Universidad de Santiago de Compostela, Spain.
[3]Medizinische Universität Innsbruck, Austria.

Abstract

Correct identification of anatomical structures is important for safe oncologic thoracic surgery, but inflammation and tissue changes after neoadjuvant therapy can obscure the operative field. Since video-assisted thoracoscopic surgery (VATS) is performed through a camera system, AI-based identification of anatomy can be integrated without altering the surgical setup. The final success in such a task is hindered by scarce expert annotations, target structures covering only a small fraction of the image, and real-time requirements. In addition, AI-based systems rely on network configuration and on pretrained weights that greatly affect the performance of identifying anatomical structure and tissue.

We present FTSpec-UNet, which combines a pretrained Foundation Encoder (ResNet or DINOv2) with a task-specific encoder trained from scratch, allowing the configuration to be adapted to a new procedure without repeating pretraining. On SurgiMind, an in-house dataset of 1539 frames from 81 VATS resections videos, FTSpec-UNet reaches an average Dice of 0.68 across eight anatomical structures, outperforming pretrained U-Net, DeepLabV3+, and SegFormer, and runs in real time on a single consumer GPU. On the public CholecSeg8K benchmark, it provides comparable results to pretrained state-of-the-art baselines. We also propose the Segmentation True Negative Rate (S-TNR), which measures whether absent structures are correctly ignored. This measure complements the background Dice score to provide a better performance picture of AI-based systems, evaluating both segmentation quality and whether absent classes are correctly ignored.

Keywords: Surgical AI, Intraoperative guidance, Transfer learning, Anatomical structure segmentation

1 Introduction

Surgical treatment is essential for the management of oncologic disease, and the accurate identification of anatomical structures is critical for a safe procedure. In thoracic surgery, misidentification of pulmonary vessels, bronchi, or the phrenic nerve can lead to intraoperative complications. Surgeons bring extensive anatomical knowledge and long experience with anatomical variation to this task. Nevertheless, acute inflammation or tissue changes following neoadjuvant immunotherapy can obscure the surgical field, which makes confident identification difficult. Video-assisted thoracoscopic surgery (VATS) is performed through a camera system that already produces a continuous view of the operative field and is routinely recorded for documentation. An AI model can be trained using this existing video data, requiring no change to the surgical setup, and no additional intraoperative burden on the team. The trained AI model can then overlay identified structures on the monitor the surgeon is already watching. The same model applied retrospectively to recorded procedures could be used for a systematic study of which conditions enable or impede the visual identification of anatomical structures. Such a system has to run in real time on the video stream, and it has to remain reliable under the visual conditions of thoracoscopic surgery, including blood and instrument occlusion. For intraoperative use, the system must produce accurate segmentation and must not predict structures that are absent as even a small false-positive outline of a vessel or nerve could have a negative impact on decisions made during the procedure.

Developing such a system for a new surgical procedure is difficult in practice. Annotated data is limited because every frame has to be labeled by experienced surgeons, and the anatomical structures of interest cover only a small part of the image. In our own experiments, established segmentation networks did not reach the accuracy we require for intraoperative use during VATS resection of lung parenchyma. U-Net-like models are widely used because they combine semantic information with spatial localization [1, 2]. Other methods, such as DeepLabV3+ and hybrid CNN Transformer models [3–5], improve the receptive field and global context modeling. More recently, vision foundation models such as SAM or DINO have provided transferable representations that can be reused for downstream dense prediction tasks [6, 7].

We also observed that the achieved performance depends strongly on the chosen network configuration, and that pretrained weights are essential when only few annotated frames are available. Transferring a method to a new procedure is therefore not only a question of retraining, but also of finding a suitable configuration for the new task. However, if the network depends on pretrained weights, the architecture becomes less flexible and optimizing hyperparameters may require repeating the expensive pretraining process. Methods such as neural architecture search and nnU-Net show that finding a suitable architecture is difficult [2, 8]. The process becomes even more time-consuming if every candidate architecture also depends on large-scale pretraining. The motivation of our work is to provide a strategy where we can use architectures that can reuse large-scale training without requiring the network to be pretrained for every new configuration.

We introduce FTSpec-UNet, a dual-encoder architecture in which a pretrained Foundation Encoder provides transferable representations while a task-specific encoder

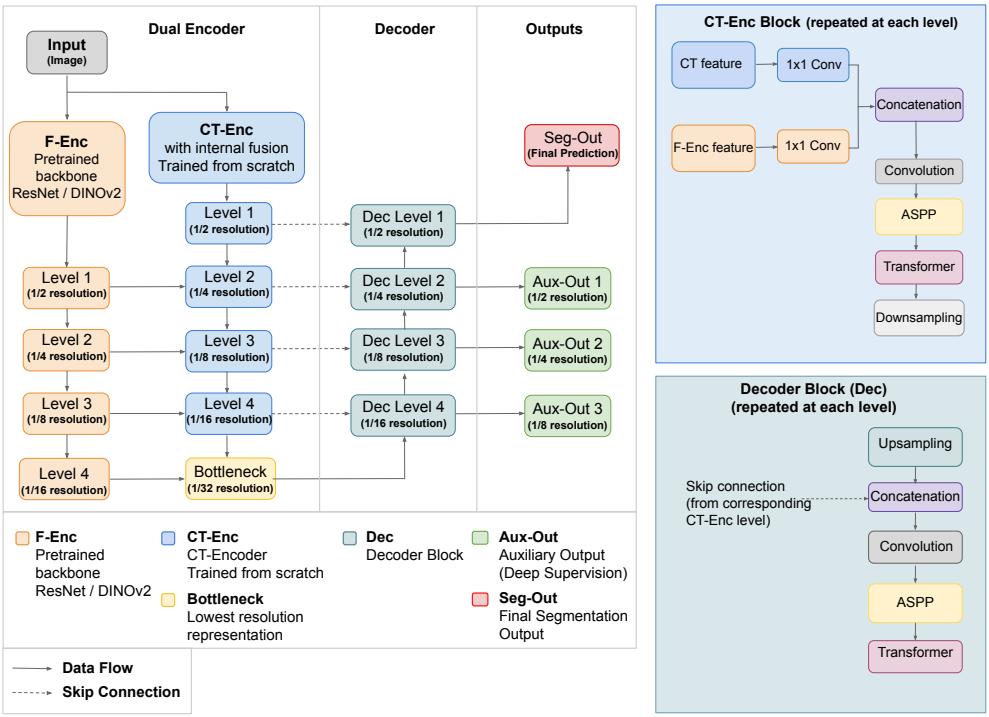
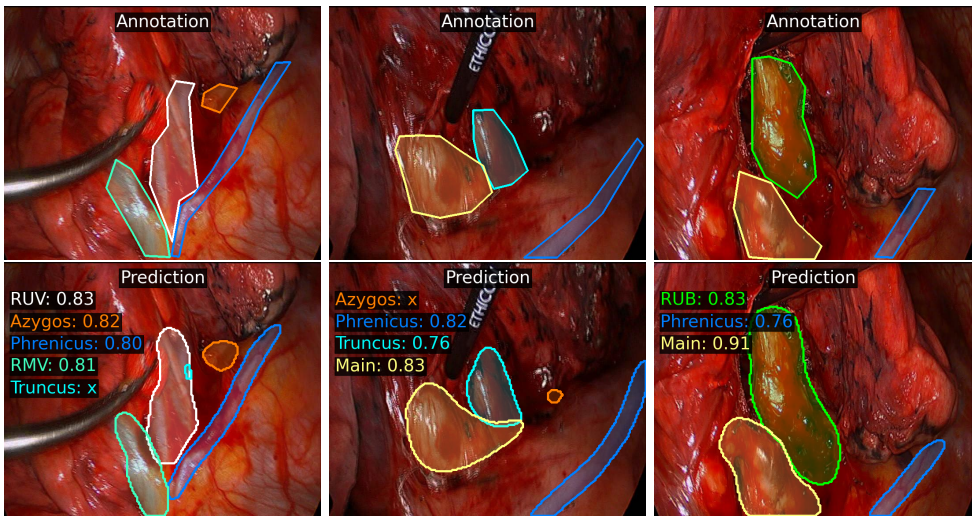


Fig. 1: Overview of the proposed Foundation Task-specific U-Net (FTSpec-UNet) architecture.

is trained from scratch and can be reconfigured freely for a new procedure. Pretraining and network optimization are thereby decoupled. The backbone can be exchanged (ResNet or DINOv2) according to the target task, and the task-specific branch can be adapted without any additional large-scale pretraining. We evaluate the resulting system on the public CholecSeg8K benchmark and on SurgiMind, an in-house dataset of 1539 frames from 81 VATS resections videos annotated for eight anatomical structures. Our contributions are two-fold:

- A deployable AI-based assistance system for thoracoscopic anatomy identification. FTSpec-UNet operates on the recorded intraoperative video stream at real-time rates (25 FPS on a single NVIDIA RTX 4090) and achieves the best average Dice across anatomical structures on our clinical VATS dataset.
- A clinically motivated evaluation criterion: the Segmentation True Negative Rate (S-TNR) complements the Dice score, by measuring whether a model correctly refrains from predicting structures that are absent from the field of view.



(a) (b) (c)

Fig. 2: Example predictions on the in-house SurgiMind dataset. Top row: ground-truth annotations; bottom row: predictions by FTSpec-UNet DINO with Dice. Small false-positive regions for Truncus and Azygos are marked with x in (a) and (b); despite near-perfect background Dice, these errors motivate the complementary S-TNR.

2 Methodology

2.1 Deep Neural Network

We propose **Foundation Task-specific U-Net** (FTSpec-UNet), a dual-encoder U-shaped segmentation network that combines two complementary feature sources, shown in fig. 1. The first is a task-specific Convolution-Transformer encoder, called the **CT-Encoder**, which is trained from scratch. The second is a pretrained encoder, called the **Foundation Encoder**, which provides features learned from large-scale pretraining. The decoder mirrors the CT-Encoder and constructs the segmentation map from the fused encoder features. The main idea is to combine pretrained representations with newly trained task-specific layers. This allows the task-specific part to be reconfigured without repeating the pretraining, a practical bottleneck when extending intraoperative guidance to a further operation.

Components. The CT-Encoder and the decoder are built from the same block, following U-Net-style architectures [1] and later U-Net variants [5, 9]. Each block applies convolution layers, an Atrous Spatial Pyramid Pooling (ASPP) module for multi-scale context [10], and a Transformer block for long-range dependencies [11]. The attention projections are implemented with convolutions to stay efficient for dense prediction. The network has four levels, from 1/2 to 1/16 of the input resolution, followed by a bottleneck at 1/32. Encoder levels end with a downsampling by a factor of two; decoder levels start with a bilinear upsampling by the same factor, followed by

a skip connection from the corresponding encoder level. An auxiliary output head is attached after each decoder level for deep supervision [12], with stage-wise loss weights given in section 3.2.

Foundation Encoder and fusion. The Foundation Encoder is initialized with pretrained weights and can be exchanged without changing the rest of the network, which is the property that makes the configuration adaptable to a new task. We use a convolutional model such as ResNet [13] or a foundation model such as DINOv2 [7] and freeze the weights during training. Backbones like ResNet already provide feature maps at progressively lower resolutions, so each output is used at the corresponding depth. For DINOv2, which produces all patch tokens at the same spatial resolution, the output is scaled to match the resolution of each depth level. At the beginning of each CT-Encoder block, the incoming CT-Encoder and Foundation Encoder features are projected to a common channel width with a 1×1 convolution and fused by concatenation. Only during training, the Foundation Encoder is randomly disabled using branch dropout. This prevents the network from relying too strongly on the pretrained features and forces the CT-Encoder to learn representations that are useful on their own.

2.2 Evaluation

For evaluation, we report the Dice score for each class including the background. The Dice score measures the overlap between the predicted mask and the ground-truth mask. It is widely used in medical segmentation because it measures region overlap and is robust to foreground-background class imbalance.

For an intraoperative guidance system, however, a second aspect matters in addition to overlap. The system should highlight a structure when it is visible, but it should also stay silent when the structure is not in the current field of view. Even a small false-positive outline of a pulmonary vessel that is not present can be misleading to the surgeon. This behavior is only partially reflected by the Dice score. If a model predicts a small false-positive region for a class that is absent from an image, the background Dice can remain high because the background still contains many true positives. In this case, the model appears to perform well, even though it incorrectly indicates that a class is present. Two such cases can be seen in fig. 2 at predictions (a) and (b) for the classes Truncus and Azygos.

To make this behavior visible in the evaluation, we additionally report the **Segmentation True Negative Rate** (S-TNR). It measures whether the model correctly ignores classes that are not part of an image. For each image i, we look at every class c separately. We set $g_c^i = 1$ iff the ground-truth mask contains at least one pixel of class c, and $p_c^i = 1$ iff the predicted mask contains at least one pixel of class c. Otherwise, they are set to 0. We then compute the image-level specificity over all images i and foreground classes c:

$$\mathrm{TNR}_c = \frac{\sum_i [g_c^i = 0 \wedge p_c^i = 0]}{\sum_i [g_c^i = 0]}, \qquad \text{S-TNR} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{TNR}_c. \tag{1}$$

An S-TNR value close to 1 means that the model rarely indicates classes that are not present in an image. Unlike the background Dice, S-TNR penalizes small and large false-positive predictions equally.

S-TNR is not intended to replace Dice, and it does not capture segmentation quality. We therefore report both metrics together. Dice describes how well present classes are segmented, while S-TNR describes how consistently absent classes are ignored. Reporting them jointly makes the trade-off between sensitivity and specificity explicit, which can be useful when a guidance system is configured for clinical use.

3 Experiments and Results

3.1 Datasets

Our in-house SurgiMind dataset consists of 1539 non-consecutive frames extracted from videos of 81 recorded VATS resections of lung parenchyma. Frames were sampled to cover different phases of the procedure and varying visual conditions. Annotations were created by a medical student and reviewed by a medical expert. They cover eight anatomical structures that were defined as most relevant for the procedure (classes are listed in table 1). Surgical instruments are visible in the images but are not annotated.

Annotated structures cover only 1 to 8% of an image, and not every structure is visible in every frame. So a model must both segment sparse targets and recognize when a structure is outside the current field of view. We use five patient-level splits, with 80% of patients for training, 10% for validation, and 10% for testing. Splitting by patient rather than by frame prevents leakage between frames of the same procedure and gives a realistic estimate of generalization to unseen patients. Ethics approval for the use of the in-house data was obtained from the local ethics committee (Nr.: 1121/2024).

We also evaluate on the public CholecSeg8K dataset to test a complementary evaluation setting and to enable a comparison on an established surgical benchmark. CholecSeg8K provides pixel-level annotations for laparoscopic cholecystectomy. It contains surgical scenes from 17 videos and includes 13 annotated classes. Following previous work, we use the same test protocol and report results on video IDs 12 and 52 using the eight foreground classes visible in these videos [14].

3.2 Implementation Details

All models were trained with the same standardized procedure. Training used the soft Dice loss $\mathcal{L}_{\text{Dice}}$ and binary cross-entropy $\mathcal{L}_{\text{BCE}}$ in equal weighting (eqs. (2) and (3)), with deep-supervision weights $\boldsymbol{\omega}_{\text{deep}} = [0.05, 0.15, 0.15, 0.65]$, so that the final decoder output is the main optimization target. Here $p_{i,c}$ is the predicted probability and $y_{i,c}$ is the ground-truth label of pixel i for class c, and ϵ ensures numerical stability. Models were trained at an input resolution of 192×192. The complete training configuration, per-dataset settings, and split definitions are available online.[1]

For comparison we train U-Net and DeepLabV3+ with ImageNet-pretrained encoders [15, 16], SegFormer pretrained on ADE20K [4], and FTSpec-UNet with

[1] https://github.com/DennisBarnesUIBK/FTSpec-UNet-for-Surgical-Image-Segmentation

Method	RUV	Az	RUB	Ph	RMV	A2	Tr	Ma	Avg	Bkg	S-TNR
U-Net°	0.68	0.61	0.66	0.63	0.52	0.46	0.55	0.72	0.60	0.93	0.28
U-Net⋆	0.66	0.66	0.65	0.66	0.52	0.41	0.56	0.73	0.61	0.95	0.57
DeepV3+°	0.61	0.56	0.53	0.53	0.37	0.29	0.48	0.65	0.50	0.93	0.46
DeepV3+⋆	0.73	0.67	0.64	0.68	0.44	0.47	0.61	0.75	0.62	0.95	0.63
SegFormer°	0.63	0.58	0.59	0.53	0.43	0.39	0.52	0.68	0.54	0.93	0.32
SegFormer⋆	0.74	**0.71**	0.67	0.69	0.56	0.53	**0.63**	0.80	0.67	**0.96**	**0.65**
FTSpec-UNet ResNet	**0.78**	0.70	**0.71**	**0.72**	**0.58**	0.54	0.61	0.79	**0.68**	0.95	0.60
FTSpec-UNet DINO	0.77	**0.71**	0.70	0.68	0.55	**0.55**	**0.63**	**0.81**	**0.68**	**0.96**	0.61

Table 1: Per-class average Dice and S-TNR across 5 splits on our SurgiMind dataset. °: random initialization; ⋆: pretrained weights; RUV: right upper pulmonary vein; Az: azygos vein; RUB: right upper lobe bronchus; Ph: phrenic nerve; RMV: middle lobe vein; A2: segmental artery; Tr: truncus of the right upper lobe; Ma: right pulmonary main artery; Avg: Average foreground class Dice; Bkg: Background

Method	AW	Liver	GT	Fat	Grasper	CT	L-hook	Gb	Bkg	S-TNR
U-Net°	0.85	0.82	0.22	0.91	0.56	0.23	0.48	0.63	0.97	0.45
U-Net⋆	0.88	0.85	0.15	**0.93**	0.70	**0.50**	0.78	0.70	0.97	0.57
DeepV3+°	0.83	0.80	0.00	0.90	0.57	0.00	0.63	0.58	0.95	0.54
DeepV3+⋆	0.89	0.84	0.12	0.92	0.66	0.19	**0.79**	0.70	0.95	0.57
SegFormer°	0.83	0.81	0.00	0.91	0.61	0.00	0.61	0.63	0.95	0.61
SegFormer⋆	**0.92**	**0.89**	0.48	**0.93**	0.70	0.37	0.77	**0.73**	0.97	**0.64**
FTSpec-UNet ResNet	0.86	0.88	0.38	**0.93**	**0.71**	0.03	0.49	0.67	**0.98**	0.52
FTSpec-UNet DINO	0.90	0.86	**0.58**	0.91	0.70	0.19	0.78	0.70	**0.98**	**0.64**

Table 2: Per-class Dice and S-TNR on CholecSeg8K. °: random initialization; ⋆: pretrained weights; AW: abdominal wall; GT: gastrointestinal tract; CT: connective tissue; Gb: gallbladder; Bkg: background

ResNet pretrained on ImageNet or with DINOv2 [15, 17]. Each baseline is additionally trained from random initialization to isolate the contribution of pretraining. Although model sizes differ, every evaluated model processes 25 FPS video in real time on a single NVIDIA RTX 4090, so inference speed does not distinguish the methods and accuracy is the deciding factor for intraoperative use.

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{1}{C}\sum_{c=1}^{C}\frac{2\sum_i p_{i,c}y_{i,c} + \epsilon}{\sum_i p_{i,c} + \sum_i y_{i,c} + \epsilon}, \tag{2}$$

$$\mathcal{L} = \omega_{\text{Dice}}\mathcal{L}_{\text{Dice}} + \omega_{\text{BCE}}\mathcal{L}_{\text{BCE}}, \qquad \omega_{\text{Dice}} = \omega_{\text{BCE}} = 0.5. \tag{3}$$

3.3 Results and Discussion

Performance on the target clinical task. Table 1 reports per-class Dice averaged over five patient-level splits on SurgiMind. Both FTSpec-UNet variants reach the highest average foreground Dice (0.68), ahead of pretrained SegFormer (0.67), DeepLabV3+ (0.62), and U-Net (0.61). The margin over the strongest baseline is small, but the improvement of FTSpec-UNet achieves the best or equal-best Dice for every structure. The target structures occupy a small fraction of the image and are

frequently partially occluded, which is where the combination of a well-configured task-specific encoder with pretrained features is expected to help. The two FTSpec-UNet variants reach the same average Dice but differ per class, which indicates that the choice of foundation model matters for individual structures. Since the Foundation Encoder can be exchanged without retraining the rest of the network, this choice can be made for a given procedure at low cost.

CholecSeg8K differs from SurgiMind in anatomy and includes surgical instruments as classes. Table 2 shows that FTSpec-UNet DINO performs competitively with pretrained baselines, with the best Dice for the gastrointestinal tract (0.58) and background (0.98) and the highest S-TNR (0.64) together with pretrained SegFormer. Both variants outperform the randomly initialized models.

The role of pretraining. Pretrained initialization (marked with *) improves over random initialization (marked with $^\circ$) for every method on both datasets. On SurgiMind the average Dice increases by 0.12 for DeepLabV3+ and 0.13 for SegFormer. On CholecSeg8K, SegFormer improves from 0.00 to 0.48 for the gastrointestinal tract and from 0.00 to 0.37 for connective tissue, i.e. classes it fails to segment without pretraining. The effect is even more pronounced for S-TNR, which rises from 0.28 to 0.57 for U-Net and from 0.32 to 0.65 for SegFormer on SurgiMind. Pretraining therefore does not only improve overlap on visible structures, it also substantially reduces spurious predictions of absent structures. With annotation budgets of the size available in a single-centre surgical dataset, pretrained representations are thus not an optimization detail but a prerequisite, which is what makes the ability to adapt the configuration without repeating pretraining practically relevant.

Benefit of S-TNR. Background Dice is nearly identical across methods on SurgiMind, varying by only 0.03, while S-TNR varies by up to 0.37. On CholecSeg8K, U-Net$^\circ$ and U-Net* both reach a background Dice of 0.97 but differ in S-TNR (0.45 versus 0.57). A high background Dice therefore does not imply that a model reliably refrains from predicting absent structures, and the two metrics rank methods differently. On SurgiMind, pretrained SegFormer attains the highest S-TNR (0.65) while FTSpec-UNet reaches the highest Dice at a somewhat lower S-TNR (0.61 and 0.60). For an intraoperative overlay this is an explicit trade-off between segmenting visible structures accurately and remaining silent when a structure is absent, and reporting both metrics makes it visible to whoever configures the system for clinical use.

4 Conclusion

We presented FTSpec-UNet, an AI-based application for the intraoperative identification and segmentation of anatomical structures in video-assisted thoracoscopic resections. Effective learning from small medical datasets requires both strong pretrained representations and a configuration tailored to the target procedure. However, tightly coupling these two aspects limits flexibility if changing the network configuration requires additional pretraining. By combining a pretrained Foundation Encoder with a task-specific encoder trained from scratch, FTSpec-UNet keeps the benefit of pretraining while leaving the task-specific part free to be adapted.

On SurgiMind, a dataset of 1539 frames from 81 VATS resection videos, the method achieved the highest average Dice across eight anatomical structures while running in real time on a single consumer GPU, and it remained competitive on the public CholecSeg8K benchmark. Across all evaluated methods, pretraining improved not only segmentation quality but, more markedly, the ability to refrain from predicting absent structures. Measuring that behavior required a metric beyond Dice. The Segmentation True Negative Rate makes visible whether a model stays silent when a structure is outside the field of view. Background Dice cannot distinguish this, so both metrics together give a more complete picture of how a network performs.

Prospects of application: This work developed an AI/ML-based identification of anatomical structures in thoracic surgery to reduce intraoperative complications. The flexible architecture presented here allows easier adaptation to different procedures. Applied retrospectively to recorded procedures, it further enables systematic study of the conditions under which anatomical structures can be reliably identified.

Disclosure of Interests

This work is supported by the SurgiMind project (Land Tirol, 443393) and approved by the local ethics committee (Nr.: 1121/2024). The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] 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 (2015). Springer

[2] Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)

[3] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)

[4] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021)

[5] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

[6] Zhang, Y., Shen, Z., Jiao, R.: Segment anything model for medical image segmentation: Current applications and future directions. Computers in Biology and

Medicine **171**, 108238 (2024)

[7] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

[8] Elsken, T., Metzen, J.H., Hutter, F.: Neural architecture search: A survey. Journal of Machine Learning Research **20**(55), 1–21 (2019)

[9] Tragakis, A., Kaul, C., Murray-Smith, R., Husmeier, D.: The fully convolutional transformer for medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3660–3669 (2023)

[10] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[12] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Transactions on Medical Imaging **39**(6), 1856–1867 (2020) https://doi.org/10.1109/TMI.2019.2959609

[13] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[14] Zhang, L., Hayashi, Y., Oda, M., Mori, K.: Towards better laparoscopic video segmentation: A class-wise contrastive learning approach with multi-scale feature extraction. Healthcare Technology Letters **11**(2-3), 126–136 (2024)

[15] TorchVision Contributors: ResNet. https://docs.pytorch.org/vision/main/models/resnet.html. Accessed: 2026-06-17 (2026)

[16] Yakubovskiy, P.: Segmentation Models PyTorch. https://github.com/qubvel-org/segmentation_models.pytorch. Accessed: 2026-06-17 (2020)

[17] Meta AI Research: DINOv2: PyTorch Code and Models. https://github.com/facebookresearch/dinov2. Accessed: 2026-06-17 (2023)