



Zeng, C., Yang, X., Smithard, D., Mirmehdi, M., Gambaruto, A. M., & Burghardt, T. (2023). *Video-SwinUNet: Spatio-temporal Deep Learning Framework for VFSS Instance Segmentation*.
<https://doi.org/10.48550/arXiv.2302.11325>

Early version, also known as pre-print

Link to published version (if available):
[10.48550/arXiv.2302.11325](https://doi.org/10.48550/arXiv.2302.11325)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is a pre-print server version of the article. It first appeared online via arXiv at <https://doi.org/10.48550/arXiv.2302.11325>. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

Video-SwinUNet: Spatio-temporal Deep Learning Framework for VFSS Instance Segmentation

Chengxi Zeng*, Xinyu Yang[†], David Smithard[‡], Majid Mirmehdi[†], Alberto M Gambaruto*, Tilo Burghardt[†]

*Department of Mechanical Engineering, University of Bristol, UK

[†]Department of Computer Science, University of Bristol, UK

[‡]Queen Elizabeth Hospital, Woolwich, UK

Abstract—This paper presents a deep learning framework for medical video segmentation. Convolution neural network (CNN) and transformer-based methods have achieved great milestones in medical image segmentation tasks due to their incredible semantic feature encoding and global information comprehension abilities. However, most existing approaches ignore a salient aspect of medical video data - the temporal dimension. Our proposed framework explicitly extracts features from neighbouring frames across the temporal dimension and incorporates them with a temporal feature blender, which then tokenises the high-level spatio-temporal feature to form a strong global feature encoded via a Swin Transformer. The final segmentation results are produced via a UNet-like encoder-decoder architecture. Our model outperforms other approaches by a significant margin and improves the segmentation benchmarks on the VFSS2022 dataset, achieving a dice coefficient of 0.8986 and 0.8186 for the two datasets tested. Our studies also show the efficacy of the temporal feature blending scheme and cross-dataset transferability of learned capabilities. Code and models are fully available at <https://github.com/SimonZeng7108/Video-SwinUNet>.

Index Terms—Deep Learning, Swin Transformer, SwinUNet, Video Tracking, Dysphagia, Swallow difficulty, Videofluoroscopy

I. INTRODUCTION AND RELATED WORK

Dysphagia or swallowing difficulty is a common complication found in 30 - 50% of people following stroke [1]. The prevalence of dysphagia in older people with dementia can be high up to 84%. Risks are identified in people with dysphagia such as malnutrition, development of pneumonia and aspiration. Serious Dysphagia can lead to a strong association with mortality [2], [3]. Hence early detection and treatment of Dysphagia are crucial.

A Videofluoroscopic Swallow Study (VFSS) is accepted as the gold standard assessment for dysphagia. During VFSS, patients are asked to swallow texture-modified foods and liquids that contain barium. It provides visual data on the trajectory of bolus, muscle, and hyoid bone movement and the connection between anatomy and aspiration [4]. However, the clinical assessment requires an extensively experienced speech-language therapist to analyse the visual data on a per-frame basis. The visual data sometimes has both low spatial and temporal quality due to device modalities and radiation noise. Moreover, there can be ambiguity and inconsistency in the judgments of different clinical experts.

Early attempts at automated processing of the data have used traditional methods, such as Hough transforms [5], Sobel Edge detection [6] and Haar classifiers [7] to track lumbar vertebrae,

hyoid bone and epiglottis, which are important anatomical structures in the pharyngeal swallowing reflex. With the more recent impact of deep learning in medical image analysis, others have shown advances in pharyngeal phase detection [8], [9], [10], [11] and hyoid bone detection [12], [13], [14], [15].

Bolus trajectory is one of the main indicators in a VFS study, but there are few studies on the automation of bolus detection or segmentation [16], [17], [18]. CNN-based works, such as [19], demonstrate significant superiority in feature extraction, though they are disadvantaged in computing long-range relations due to their inherent local operations. Vision transformers [20], [21], on the other hand, have exhibited great predominance in modelling global contextual correlations by using attention mechanisms.

Recent works that leverage vision transformers [22], [23] have shown remarkable performance in medical image segmentation. While others have dealt with video dynamics for detection or segmentation in videos, such as Cao et al. [24] and Yang et al. [25], only few has explicitly addressed the use of temporal information in assisting the detection or segmentation of sequential medical data [26]. The dynamics of bolus suggest that an implicit temporal relationship between the frames on a feature level can be exploited in learning detection or segmentation models.

In this paper, we present a deep-learning pipeline that takes account of multi-rater annotations and fuses them into a more consistent and reliable ground truth. Subsequently, an architecture is proposed (see Fig. 1) comprising a ResNet-50 feature extractor, a Temporal Context Module (TCM) feature blender, a non-local attention encoder (Swin Transformer) and a cascaded CNN decoder for detailed segmentation map prediction.

Our main contributions are summarised as follows: i) we provide the VFSS2022 dataset Part 2 in different modalities in contrast to Part 1 annotated with reliable labels for the laryngeal bolus and pharynx. ii) we propose a new architecture enhancing the performance of previous work [18] by extending the vision transformer encoder to a stronger and more generalised Swin Transformer, iii) we perform a detailed ablation study to reveal the importance of temporal feature blending. We also explore the cross-dataset transferability and generalizability of our deep neural networks on data across different modalities.

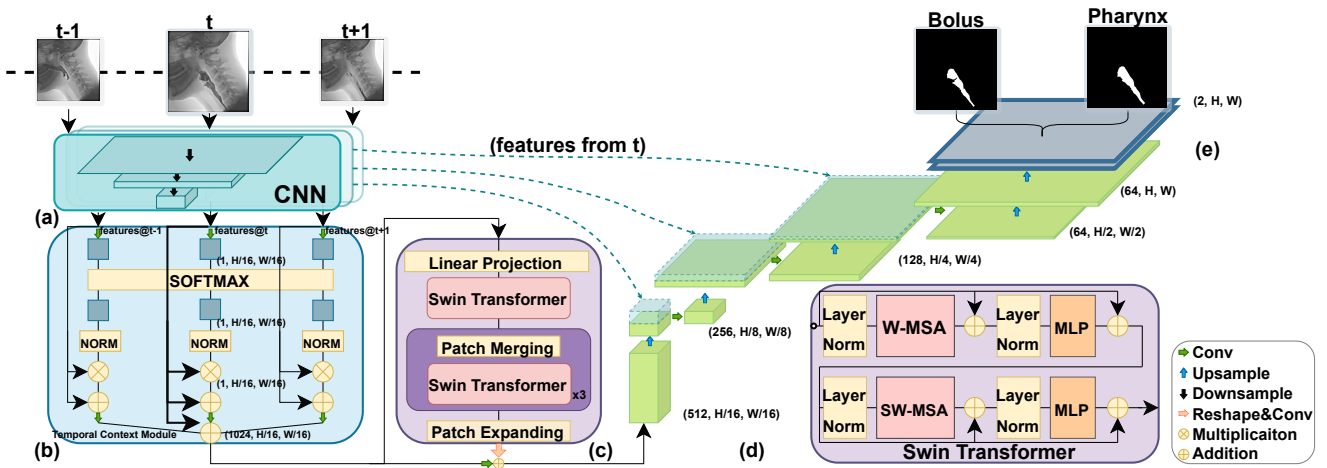


Fig. 1. **Video-SwinUNet Architecture Overview.**(a)A ResNet-50 CNN feature extractor; (b)Temporal Context Module for temporal feature blending; (c)A Swin transformer-based feature encoder; (d)Cascaded CNN up-sampler for segmentation reconstruction; (e)2-layer segmentation head for detailed pixel-wise label mapping. Three skip connections are bridged between the CNN feature extractor and up-sampler as well as from the temporal features.

II. METHODOLOGY

A. Architecture Overview

Inspired by UNet [19], we follow the encoder-decoder structure to build our video instance segmentation network, as shown in Fig. 1. It takes a video snippet as an input that consists of a sequence of frames with dimension $\mathbf{x} \in \mathbb{R}^{t \times H \times W}$, where $H \times W$ represents the spatial resolution of the input and t is a temporal range of the input sequence. The input frames will be successively fed into a ResNet-50 backbone for feature extraction (see Fig. 1(a)). Then, the extracted features are simultaneously passed into a novel Temporal Context Module (TCM) (see Fig. 1(b)) [25], which blends the past and future frame features into the target central frame feature. Thereafter, the output feature that is integrated with high-level spatial and temporal representation is tokenised into image patches by a Swin transformer encoder (see Fig. 1(c)) for global context construction. Finally, an up-sampling decoder (see Fig. 1(d)) reconstructs the segmentation map to the original image size of $H \times W$ with cascaded CNNs and binary segmentation heads (see Fig. 1(e)).

B. Temporal Context Module

The proposed architecture contains a key component Temporal Context Module (TCM) following success in video detection [25]. The design of TCM follows the principled blending framework by [24] where a trainable self-attention module is formulated to a range of frame features from the previous CNN block. The input features $x_t \in \{x_1, \dots, x_i\}$ are separately linear embedded to a feature space by function $e(\cdot)$ and weights w_t in a concurrent manner. After that, a global **Softmax** operation is applied so that the temporal correlation across all the frames in the feature space can be aggregated. The agglomerated features are dispersed again to several stems for further linear embeddings. The normalisation of each stem is necessary to prevent vanishing/exploding gradients and can be done easily by $\hat{x}_{t,i} = \frac{1}{HW} \mathcal{C}(x_{t,i}; w_t) \sum_{j=1}^{HW} \mathcal{C}(x_{t,j}; w_t)$. Identity mapping operations by **multiplication** \otimes and **addition** \oplus

are applied in each stem. In the end, stabilised features are added back to the central frame feature as a final single mixture high-level description of the short-term snippet. In summary, the TCM operation can be formulated by:

$$z_{t,i}^{TCM} = x_{t,i} + \sum_{n \in T} w_n^{**} \left(x_{n,i} \oplus w_n^* \sum_{j=1}^{HW} \hat{x}_{n,j} \otimes x_{n,j} \right), \quad (1)$$

where x_n are the linear embedded features to be combined, w_n^* and w_n^{**} are trainable parameters for identity additions and blending operations.

C. Swin Transformer

Following [20], we tokenise temporal blended features into feature patches x_p and map them into a latent D -dimensional embedding space via learnable linear projection. [21], [23] suggest the unnecessary of employing position embedding \mathbf{E}_{pos} in Swin transformer, hence we omitted it in our work for simplicity. The projected feature can be expressed as $z_0 = [x_p^1 \mathbf{E}_{\text{pat}}; x_p^2 \mathbf{E}_{\text{pat}}; \dots; x_p^N \mathbf{E}_{\text{pat}}]$, where \mathbf{E}_{pat} is the patch embedding linear projection. The conventional vision transformer computes global attention across the vectorized patches. As a result, the computational complexity is quadratically increased along with the increase of the input resolutions. To alleviate the computation overhead in Multihead Self-Attention (MSA), a Window-based Multihead Self-Attention (W-MSA) method is proposed in [21]. The window moves along the feature or image with no over-lapping and conducts self-attention within the local window and makes it more computationally efficient in computer vision tasks. the W-MSA also includes a relative position bias and can be expressed as:

$$\text{Attention}(Q, K, V) = \text{SoftMax} \left(QK^T / \sqrt{d} + B \right) V, \quad (2)$$

where $Q, K, V \in \mathbb{R}^{M^2 \times d}$ stands for the query, key and value matrices respectively; d is the query/key dimension, and M^2 patch numbers in a window. Values in B are taken from the bias matrix \hat{B} .

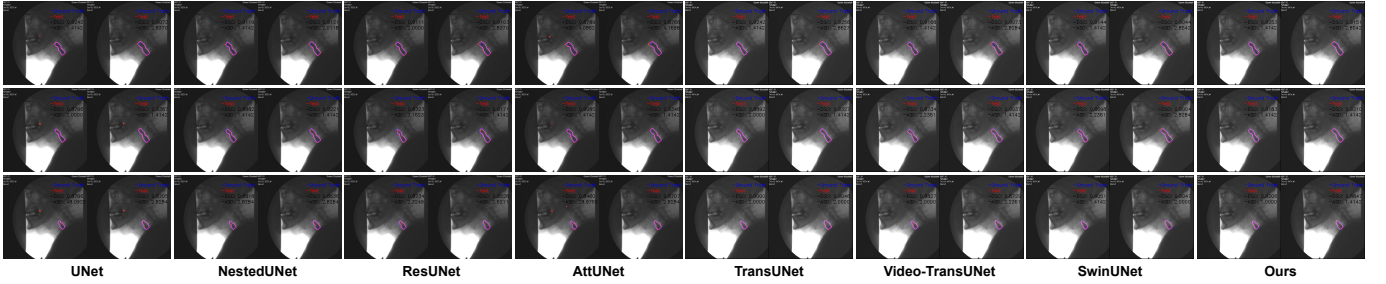


Fig. 2. **Qualitative Results.** Model segmentation results on 3 consecutive frames selected from VFSS Part2 dataset testset. All results are in pairs of bolus and pharynx predictions side by side. The red and blue outlines indicate the output segmentation and ground truth, respectively.(Best viewed zoomed)

To model the relationship between windows, a Shifted-Window MSA (SW-MSA) is proposed in [21], the patches take turns in two consecutive Swin Transformer blocks, each of which contains both a W-MSA and a SW-MSA accompanied with a 2-layer MLP followed a GELU activation function. And LayerNorm(LN) and skip connections are added before the MLP, as illustrated in Fig. 1.

III. EXPERIMENTS AND RESULTS

A. Datasets and Implementation details

Datasets. The VFSS2022 datasets are collected in two major hospitals and the utilisation of the anonymised data is ethically reviewed and approved by the hospitals and our internal institutional Ethics Board. During the VFS studies, the patients carried out modified barium swallow tests under the practitioner’s supervision. VFSS2022 Part 1 produces 3.5 minutes of swallowing videos which result in 440 sampled frames with a spatial resolution of 512×512 pixels. Each frame is annotated by 3 experts and reviewed by 2 speech and language therapists and compromising labels for bolus and pharynx. The final ground truth is fused together with the 3 labels by a common image fusion strategy STAPLE [30]. VFSS2022 Part 2 is annotated by one trained expert consisting of 154 frames and corresponding labels, it appears to have more modal noises and poorer temporal quality, and is used for the model generalisation test.

Implementation details. The bolus and pharynx are concatenated as 2-layer tensors for the end-to-end model co-learning from both. The layers for the frames with no visible bolus are replaced with full-size zeroed tensors. To study the effect of input snippet lengths in our system, the input number of frames is in the range of $t = 3, 5, 7, 9, 11 \& 13$ both for training and testing. All experiments supported online data augmentation such as random limited rotation and flipping. We initialise the weights of the ResNet-50 backbone and Swin-Transformer from the pre-trained models [22], [23]. During training, our system takes in a batch size of 2 and is equipped with an Adam optimizer with an initial learning rate of $1e-3$. For transfer learning, the learning rate is dropped to $1e-4$ at the beginning. A learning rate scheduler is set to drop the learning rate to 80% after 20 epochs of validation loss saturation. The architecture is achieved in Python 3.8.5 and Pytorch 1.9 and trained with an NVIDIA Tesla P100 16GB GPU. We consider

the overall loss of Binary Cross Entropy Loss and Dice Loss as the final training objectives.

B. Comparison with the state of the art

We compare our proposed architecture with major medical image segmentation models including UNet [19], NestedUNet [27], ResUNet [28], AttUNet [29], TransUNet [22], Video-TransUNet [18] and SwinUNet [23] over 5 common evaluation metrics, the Dice Coefficient (DSC), the 95th percentile of the Hausdorff Distance (HD95), the Average Surface Distance (ASD), Sensitivity and Specificity, see Tab. I. Additionally, we also include the total number of parameters of the model and the total floating-point operations(FLOPs) to compare the model size and computing performance. It can be seen that in Tab. I, our method improved segmentation accuracy to 89.86%/81.86%(DSC) and 6.2365/4.5268 pixels(HD95) on VFSS2022 Part1/Part2, the test results dominate the previous SOTA [18] and other methods with a significant margin. The general quality is greatly improved and output noises are less produced, as demonstrated in the qualitative results, see Fig. 2. More importantly, the proposed method achieves a remarkable speed-accuracy trade-off. Although compared with SwinUNet [23] the model size is doubled, it notably improved the segmentation accuracy by 5.09%. And it is observed that our model has reduced the number of parameters to less than half compared with the previous SOTA while not sacrificing computational efficiency due to the design of hierarchy shifting windows. Fig. 3 shows the Grad-CAM output one layer before TCM in Video-TransUNet and Video-SwinUNet. Compared to TransUNet and SwinUNet which don’t include TCM, the attention maps from our method devote great concentrations are computed to task-relevant features. Hence it promotes the efficacy of the TCM and the importance of temporal-relation constructions.

C. Ablation study

We conducted major ablation experiments to reveal the efficacy of the proposed temporal blending framework via a novel TCM component. We modulate 4 main components, CNN extractor(CNN), Swin Transformer Block(Swin), Temporal Context Module(TCM) and CNN up-sampler(CUP) in our experiments. Comparing Tab. II(4) to (3) and (6) to (5), we can see that TCM has increased performance by a margin without an extra expensive cost in computational power. The

TABLE I
QUANTITATIVE RESULTS. SEGMENTATION ACCURACY ON 5 METRICS OF VFSS2022 PART1/PART2 IS SHOWN, AS WELL AS A NUMBER OF PARAMETERS AND FLOPs OF EACH MODEL. PART1/PART2 DATASETS ARE TRAINED SEPARATELY.

Model	DSC	HD95	ASD	Sensitivity	Specificity	FLOPs	#Params
(1) UNet[19]	0.8422/0.7894	14.7530/20.7516	2.1675/4.6458	0.8289/0.7414	0.9988/0.9793	50.1G	34.5M
(2) NestedUNet[27]	0.8335/0.7537	13.7601/6.4952	2.2275/5.1220	0.8305/0.7188	0.9987/0.9682	105.7G	36.6M
(3) ResUNet[28]	0.8465/0.7846	11.982/6.4187	2.0487/2.4218	0.8183/0.7218	0.9991 /0.9994	43.1G	31.5M
(4) AttUNet[29]	0.8501/0.7917	12.9356/16.9552	2.1832/4.2174	0.8328/0.7721	0.9988/0.9985	51.0G	34.8M
(5) TransUNet[22]	0.8586/0.8046	7.4510/4.6291	1.1050/1.9322	0.8486/0.7579	0.9989/0.9929	29.3G	105.3M
(6) Video-TransUNet[18]	0.8796/0.8041	6.9155/4.7775	1.0379 /1.5270	0.8851/0.7423	0.9986/ 0.9996	40.4G	110.5M
(7) SwinUNet[23]	0.8477/0.8001	10.2897/5.9846	2.0817/2.1342	0.8459/0.7336	0.9985/0.9935	6.1G	27.1M
(8) Ours	0.8986/0.8186	6.2365/4.5268	1.3081/ 1.2052	0.9011/0.7756	0.9986/0.9995	25.8G	48.9M

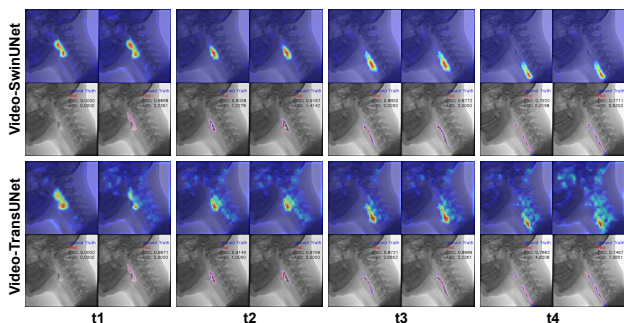


Fig. 3. **Grad-CAM Visualisation.** Comparing the two closest competing architectures, grad-cam maps show where the model pays attention. Note the cleaner focus of our proposed approach.(Best viewed zoomed)

CNN feature extractor and CUP indicate the effectiveness of convolutional operations due to their intrinsic locality characteristic. The use of skip connections is well studied in [19], [22], we attached an additional Temporal feature Skip Connection(TSC) to the decoder path, see Tab. II(7) to (6), it is suggested that the TSC is beneficial in constructing the segmentation map, which further supports the significance of temporal features in the neural network. Grid search over snippet sizes $t = 3, 5, 7, 9, 11 \& 13$ revealed the optimal, application-specific size $t = 5$ both for training and testing.

D. Transfer learning

We also explore the transferability of each component in Fig. 1, CNN(a), TCM(b), Swin Transformer(c), Decoder(d), Segmentation head(e), shown in Tab. III, a* indicates weights are pre-trained on ImageNet from [22], otherwise are trained

TABLE II
 ABLATION STUDY ON IMPACTS OF DIFFERENT ENCODER-DECODER COMBINATIONS TO PERFORMANCES.

Encoder	Decoder	DSC	HD95	#Params
(1)Swin	Swin	0.8477	10.2897	27.1M
(2)CNN+Swin	Swin	0.8483	9.5757	39.1M
(3)CNN+Swin	Swin+CUP	0.8562	8.0544	43.8M
(4)CNN+TCM + Swin	Swin+CUP	0.8634	6.8941	49.1M
(5)CNN+Swin	CUP	0.8592	8.2744	43.2M
(6)CNN+TCM+Swin	CUP	0.8899	5.1234	48.4M
(7)S/A+TSC	S/A	0.8986	6.2365	48.9M

from scratch. We adopt a standard transfer learning approach, fine-tuning, to investigate the generalisation ability of each part in domain shift from VFSS2022 Part 1 to Part 2 and vice versa. It is suggested that fine-tuning the later part after feature extraction is beneficial in domain adaption in both ways, noting row(3) and row(7). It is also shown our model's ability to generalise in part 1/part 2 when the model trains the entire dataset and can even gain performance boosts(DSC 89.94%) in part 1, see row(10) and row(11).

IV. CONCLUSION

We presented an end-to-end framework that exploits multi-frame inputs to segment VFSS2022 data with great success leading to performance gains and model size reduction. Our proposed neural network merits local and global spatial context and leveraged temporal features. Each of the modules can be fine-tuned or exchanged. The final framework achieves superior performance over other designs and provides a new, alternative pipeline for medical video segmentation tasks.

ACKNOWLEDGEMENTS. Data usage and publication are granted by UoB Ethics Approval REF: 11277. We thank project investigators Ian Swaine, Salma Ayis, Aoife Stone-Ghariani, Dharinee Hansjee, Stefan T Kulnik, Peter Kyberd, Elizabeth Lloyd-Dehler, William Oliff, Lydia Morgan and Russel Walker and thank Yuri Lewyckij and Victor Perez for their annotations. Project: CTAR-SwiFt; Funder: NIHR; Grant: PB-PG-1217-20005.

TABLE III
 TRANSFERABILITY TEST ON EACH PART OF THE MODEL.

Pretrained dataset	Training dataset	Frozen weights	Fine-tuning weights	DSC	HD95
(1)Part1	Part2	N/A	All	0.7618	15.1496
(2)Part1	Part2	a	b+c+d+e	0.7979	4.9123
(3)Part1	Part2	a*	b+c+d+e	0.8437	4.6512
(4)Part1	Part2	a + b	c+d+e	0.7295	16.4245
(5)Part1	Part2	a+b+c	d+e	0.7030	18.1302
(6)Part1	Part2	a*+b+c*	d+e	0.8171	5.2039
(7)Part2	Part1	a*	b+c+d+e	0.8920	4.1984
(8)Part2	Part1	a*	b+c*+d+e	0.8850	6.0604
(9)Part1+2	Part1	a*	b+c*+d+e	0.8994	3.8415
(10)Part1+2	Part2	a*	b+c*+d+e	0.8094	4.6473

REFERENCES

- [1] D. G. Smithard, N. Smeeton, and C. D. Wolfe, "Long-term outcome after stroke: does dysphagia matter?" *Age and ageing*, vol. 36 1, pp. 90–4, 2007.
- [2] D. G. Smithard, "Dysphagia: A geriatric giant?" 2016.
- [3] D. G. Smithard, P. A. O'Neill, R. E. England, C. L. Park, R. Wyatt, D. F. Martin, and J. Morris, "The natural history of dysphagia following a stroke," *Dysphagia*, vol. 12, pp. 188–193, 1997.
- [4] D. J. Ramsey, D. G. Smithard, and L. Kalra, "Early assessments of dysphagia and aspiration risk in acute stroke patients," *Stroke: Journal of the American Heart Association*, vol. 34, pp. 1252–1257, 2003.
- [5] Y. Zheng, M. S. Nixon, and R. Allen, "Automated segmentation of lumbar vertebrae in digital videofluoroscopic images," *IEEE Transactions on Medical Imaging*, vol. 23, pp. 45–52, 2004.
- [6] P. M. Kellen, D. Becker, J. M. Reinhardt, and D. J. V. Daele, "Computer-assisted assessment of hyoid bone motion from videofluoroscopic swallow studies," *Dysphagia*, vol. 25, pp. 298–306, 2009.
- [7] S. Noorwali, "Semi-automatic tracking of the hyoid bone and the epiglottis movements in digital videofluoroscopic images," 2013.
- [8] J. T. Lee and E. Park, "Detection of the pharyngeal phase in the videofluoroscopic swallowing study using inflated 3d convolutional networks," in *MLMI@MICCAI*, 2018.
- [9] J. T. Lee, E. Park, and T.-D. Jung, "Automatic detection of the pharyngeal phase in raw videos for the videofluoroscopic swallowing study using efficient data collection and 3d convolutional networks †," *Sensors (Basel, Switzerland)*, vol. 19, 2019.
- [10] K.-S. Lee, E. Lee, B. Choi, and S.-B. Pyun, "Automatic pharyngeal phase recognition in untrimmed videofluoroscopic swallowing study using transfer learning with deep convolutional neural networks," *Diagnostics*, vol. 11, 2021.
- [11] J. T. Lee, E. Park, J.-M. Hwang, T.-D. Jung, and D. Park, "Machine learning analysis to automatically measure response time of pharyngeal swallowing reflex in videofluoroscopic swallowing study," *Scientific Reports*, vol. 10, 2020.
- [12] H. Kim, Y. Kim, B. Kim, D. Y. Shin, S. J. Lee, and S.-I. Choi, "Hyoid bone tracking in a videofluoroscopic swallowing study using a deep-learning-based segmentation network," *Diagnostics*, vol. 11, 2021.
- [13] D. Lee, W. H. Lee, H. G. Seo, B.-M. Oh, J. C. Lee, and H. C. Kim, "Online learning for the hyoid bone tracking during swallowing with neck movement adjustment using semantic segmentation," *IEEE Access*, vol. 8, pp. 157 451–157 461, 2020.
- [14] S. Feng, Q. T.-K. Shea, K.-Y. Ng, C.-N. Tang, E. Kwong, and Y. Zheng, "Automatic hyoid bone tracking in real-time ultrasound swallowing videos using deep learning based and correlation filter based trackers," *Sensors (Basel, Switzerland)*, vol. 21, 2021.
- [15] A. Iyer, M. Thor, R. Haq, J. O. Deasy, and A. P. Apte, "Deep learning-based auto-segmentation of swallowing and chewing structures in ct," *bioRxiv*, 2019.
- [16] H. Caliskan, A. S. Mahoney, J. L. Coyle, and E. Sejdić, "Automated bolus detection in videofluoroscopic images of swallowing using mask-rnn," *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2173–2177, 2020.
- [17] Z. Zhang, E. Lucatorto, J. Coyles, and E. Sejdić, "Deep learning-based auto-segmentation and evaluation of vallecular residue in videofluoroscopy," *SSRN Electronic Journal*, 2021.
- [18] C. Zeng, X. Yang, M. Mirmehdi, A. M. Gambaruto, and T. Burghardt, "Video-transunet: Temporally blended vision transformer for ct vfss instance segmentation," *Proceedings of International Conference of Machine Vision, SPIE*, 2022.
- [19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *ArXiv*, vol. abs/1505.04597, 2015.
- [20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv*, vol. abs/2010.11929, 2020.
- [21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10 002, 2021.
- [22] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *ArXiv*, vol. abs/2102.04306, 2021.
- [23] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," *ArXiv*, vol. abs/2105.05537, 2021.
- [24] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 1971–1980, 2019.
- [25] X. Yang, M. Mirmehdi, and T. Burghardt, "Great ape detection in challenging jungle camera trap footage via attention-based spatial and temporal feature blending," *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 255–262, 2019.
- [26] D. Shi, R. Liu, L. Tao, Z. He, and L. Huo, "Multi-encoder parse-decoder network for sequential medical image segmentation," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 31–35.
- [27] D. Stoyanov, Z. A. Taylor, G. Carneiro, T. F. Syeda-Mahmood, A. L. Martel, L. Maier-Hein, J. M. R. Tavares, A. P. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, and A. Madabhushi, "Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, dlmia 2018, and 8th international workshop, ml-cds 2018, held in conjunction with miccai 2018, granada, spain, september 20, 2018, proceedings," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2018.
- [28] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, pp. 749–753, 2017.
- [29] O. Oktay, J. Schlemper, L. L. Folgoc, M. J. Lee, M. P. Heinrich, K. Misawa, K. Mori, S. G. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," *ArXiv*, vol. abs/1804.03999, 2018.
- [30] S. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, pp. 903–921, 2004.