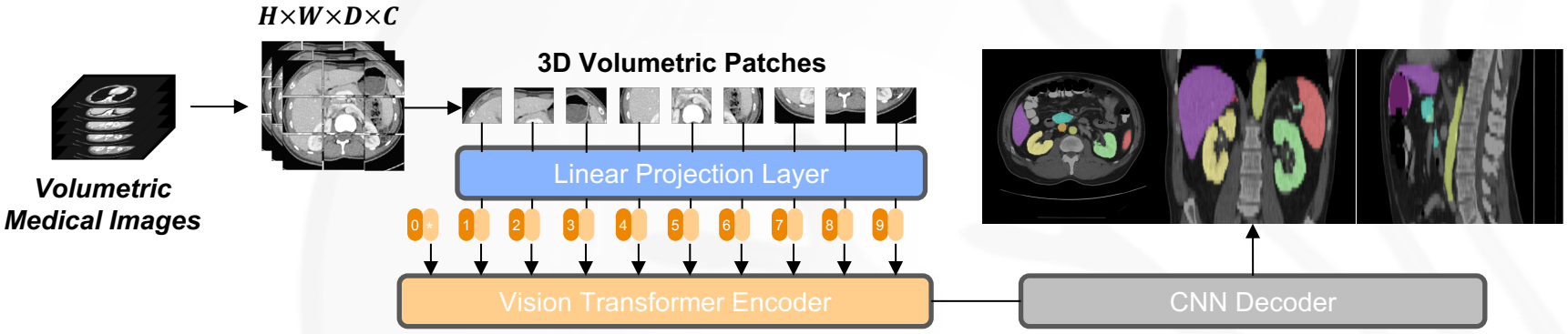
Ho Hin Lee, PhD Candidate

Medical-image Analysis and Statistical Interpretation (MASI)

Department of Computer Science
Vanderbilt University, TN, USA
https://my.vanderbilt.edu/masi/

3D UX-Net: A Large Kernel Volumetric ConvNet Modernizing Hierarchical Transformer for Medical Image Segmentation

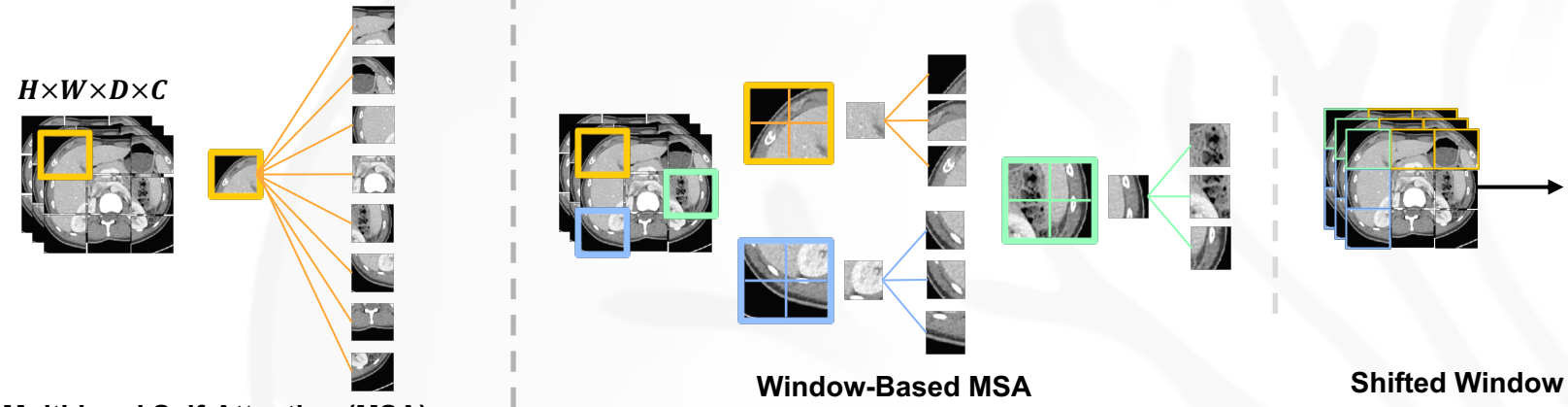
Vision Transformers in Volumetric Segmentation



Hatamizadeh, Ali, et al. "Unetr: Transformers for 3d medical image segmentation.", WACV 2022.

- Significant progress have been demonstrated of integrating vision transformer as a generic backbone in the medical domain
- However, it is challenging to adapt for volumetric segmentation due to the quadratic complexity with respect to the input size, especially for high-resolution images

Hierarchical Transformer



Hatamizadeh, Ali, et al. "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images.", MICCAI 2022 Brainlesion Workshop

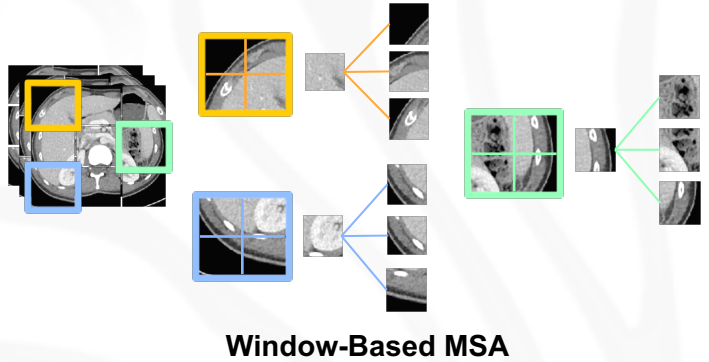
- With Swin Transformer as the generic backbone, we found that the key contribution for significant improvement in performances are attributed to:
  1. **The scaling behavior**
  2. **The self-attention mechanism with large receptive field**

Explorative Insights

We further found that:

- The computation of window-based attention introduce the convolutional prior knowledge such as:
  - Self-Attention in windows with shared weights
    - Correspond to the convolution kernels
  - Hierarchical feature extraction

- Here we raise a question:



Can we leverage convolution modules to enable the capabilities of hierarchical transformers?

Goals

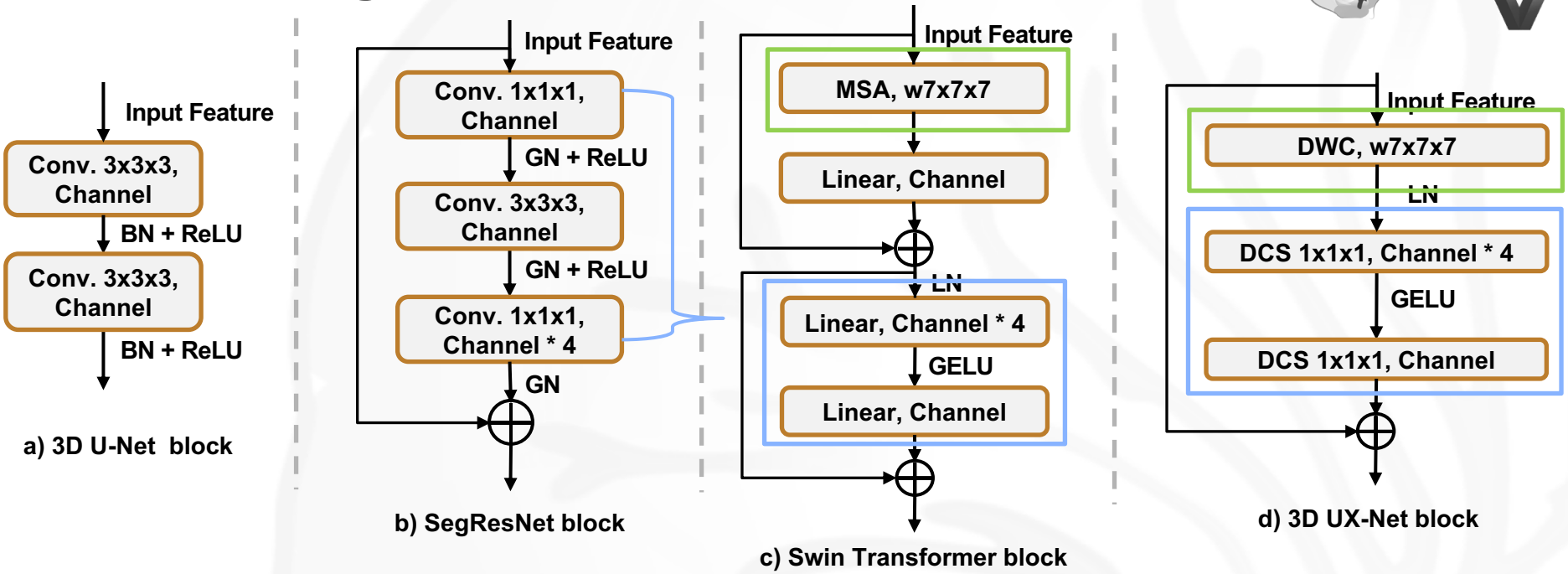
With the recent advances of DepthWise Convolution (DWC),

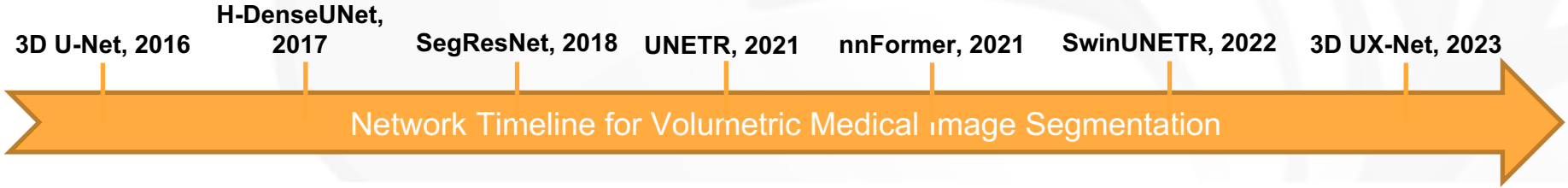
- *Liu et al.* provides a scalable block design for large receptive field by leveraging DWC with large kernel sizes

In this works, we propose **3D UX-Net** to:

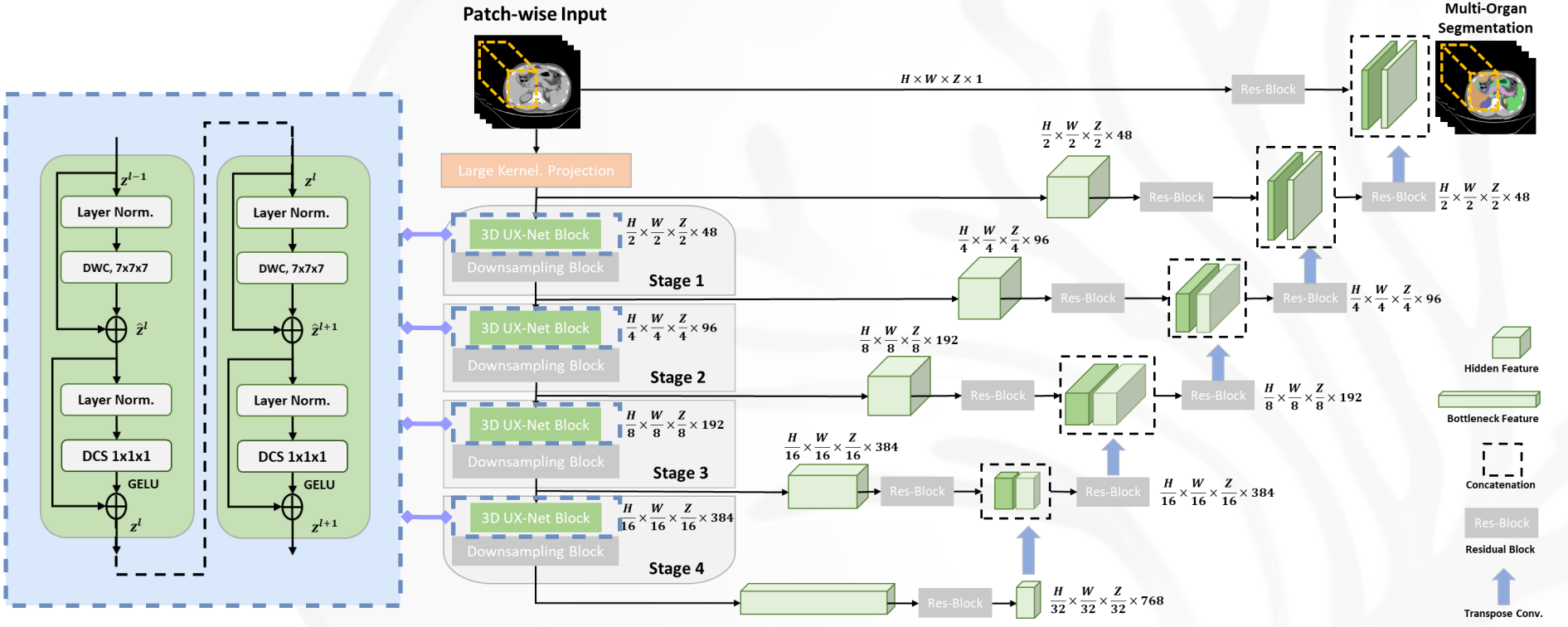
1. To simulate the large receptive field characteristics in hierarchical transformer with **large kernel DWC** design.
2. We introduce **depth pointwise scaling** to distribute each channel-wise feature independently into a wider hidden dimension
3. We evaluate 3D UX-Net with three public volumetric datasets, achieving consistently improvement in 1) direct training and 2) finetuning scenarios with fewer model parameters.

New Block Design for Convolution Neural Network





3D UX-Net: Complete Network Architecture



Quantitative Results

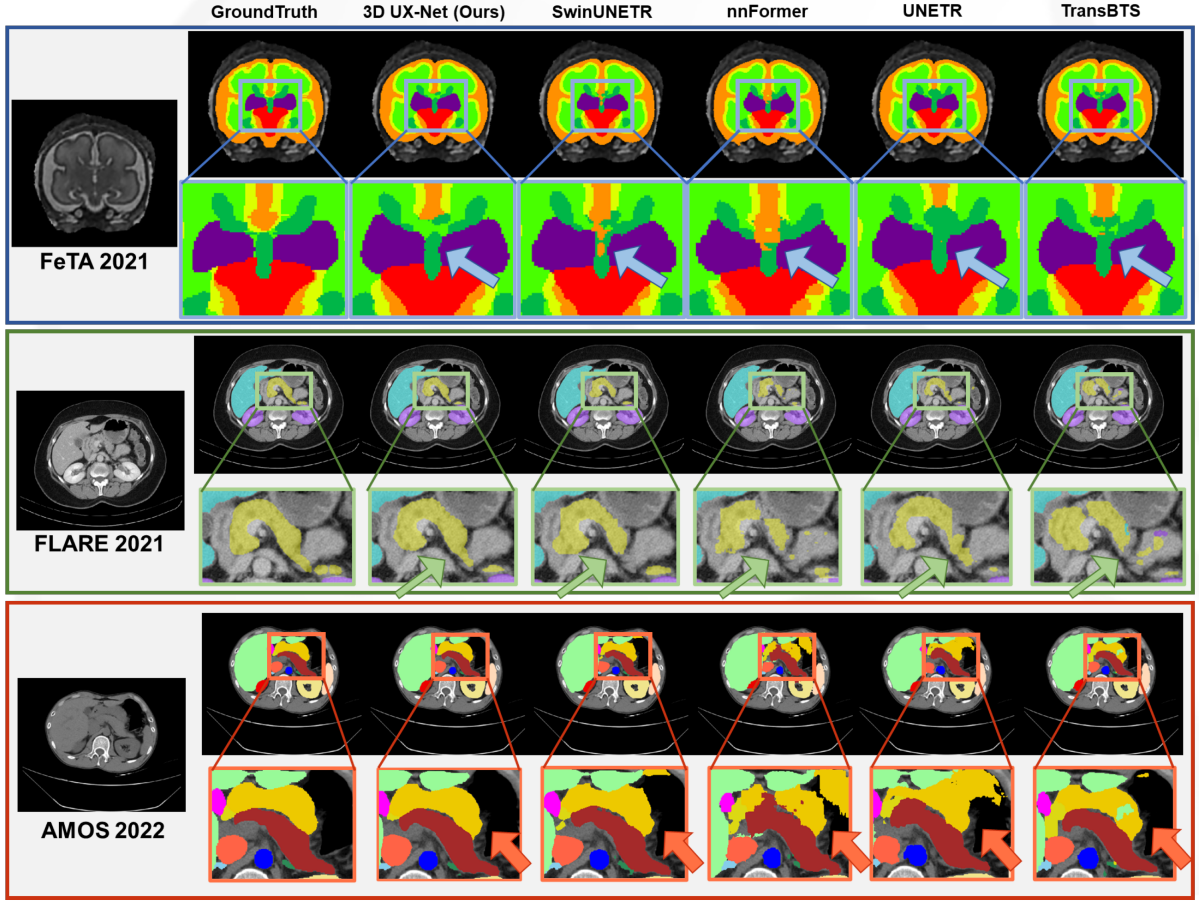
Table 1: Comparison of transformer and ConvNet SOTA approaches on the Feta 2021 and FLARE 2021 testing dataset. (*: $p < 0.01$, with Wilcoxon signed-rank test to all SOTA approaches)

			FeTA 2021								FLARE 2021				
Methods	#Params	FLOPs	ECF	GM	WM	Vent.	Cereb.	DGM	BS	Mean	Spleen	Kidney	Liver	Pancreas	Mean
3D U-Net Çiçek et al. (2016)	4.81M	135.9G	0.867	0.762	0.925	0.861	0.910	0.845	0.827	0.857	0.911	0.962	0.905	0.789	0.892
SegResNet Myronenko (2018)	1.18M	15.6G	0.868	0.770	0.927	0.865	0.911	0.867	0.825	0.862	0.963	0.934	0.965	0.745	0.902
RAP-Net Lee et al. (2021)	38.2M	101.2G	0.880	0.771	0.927	0.862	0.907	0.879	0.832	0.865	0.946	0.967	0.940	0.799	0.913
nn-UNet Isensee et al. (2021)	31.2M	743.3G	0.883	0.775	0.930	0.868	**0.920**	0.880	0.840	0.870	0.971	0.966	0.976	0.792	0.926
TransBTS Wang et al. (2021)	31.6M	110.4G	**0.885**	0.778	0.932	0.861	0.913	0.876	0.837	0.868	0.964	0.959	0.974	0.711	0.902
UNETR Hatamizadeh et al. (2022b)	92.8M	82.6G	0.861	0.762	0.927	0.862	0.908	0.868	0.834	0.860	0.927	0.947	0.960	0.710	0.886
nnFormer Zhou et al. (2021)	149.3M	240.2G	0.880	0.770	0.930	0.857	0.903	0.876	0.828	0.863	0.973	0.960	0.975	0.717	0.906
SwinUNETR Hatamizadeh et al. (2022a)	62.2M	328.4G	0.873	0.772	0.929	0.869	0.914	0.875	0.842	0.867	0.979	0.965	0.980	0.788	0.929
3D UX-Net (Ours)	53.0M	639.4G	0.882	**0.780**	**0.934**	**0.872**	0.917	**0.886**	**0.845**	**0.874***	**0.981**	**0.969**	**0.982**	**0.801**	**0.934***

Table 2: Comparison of Finetuning performance with transformer SOTA approaches on the AMOS 2021 testing dataset.(*: $p < 0.01$, with Wilcoxon signed-rank test to all SOTA approaches)

Methods	Spleen	R. Kid	L. Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	Panc.	RAG	LAG	Duo.	Blad.	Pros.	Avg
nn-UNet	0.965	0.959	0.951	0.889	0.820	0.980	0.890	0.948	0.901	0.821	0.785	0.739	0.806	0.869	0.839	0.878
TransBTS	0.885	0.931	0.916	0.817	0.744	0.969	0.837	0.914	0.855	0.724	0.630	0.566	0.704	0.741	0.650	0.792
UNETR	0.926	0.936	0.918	0.785	0.702	0.969	0.788	0.893	0.828	0.732	0.717	0.554	0.658	0.683	0.722	0.762
nnFormer	0.935	0.904	0.887	0.836	0.712	0.964	0.798	0.901	0.821	0.734	0.665	0.587	0.641	0.744	0.714	0.790
SwinUNETR	0.959	0.960	0.949	0.894	0.827	0.979	0.899	0.944	0.899	0.828	0.791	**0.745**	0.817	0.875	0.841	0.880
3D UX-Net	**0.970**	**0.967**	**0.961**	**0.923**	**0.832**	**0.984**	**0.920**	**0.951**	**0.914**	**0.856**	**0.825**	0.739	**0.853**	**0.906**	**0.876**	**0.900***

Qualitative Results



Summary

In this work, we revisit the 3D ConvNet block design to investigate the feasibility of:

1. Achieved the **state-of-the-art performance** via **a pure ConvNet** architecture
2. Yielded **much less network complexity** compared with 3D vision transformers
3. Provided a **new exploratory direction** of designing 3D ConvNet on volumetric high-resolution tasks

Thank you.

- Made possible by
  - U54DK120058, NSF CAREER 1452485, NIH 2R01EB006136, NIH 1R01EB017230, NIH R01NS09529
  - Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University
  - ImageVU, Research Derivative supported by the VICTR CTSA award (ULTR000445 from NCATS/NIH)
  - PCORI (contract CDRN-1306-04869).

- MASI lab today