Article

Sliding-Window Transformer-Convolutional Networks for Robust Longitudinal Prediction of Geographic Atrophy Growth in Low-Data Settings

Peter Racioppo [1], Ziyuan Chris Wang [1], SriniVas R. Sadda [2], Zhihong Jewel Hu [1,*]

1 Doheny Image Analysis Laboratory, **Doheny Eye Institute**, 150 North Orange Grove Blvd, Pasadena, CA 91103
2 Department of Ophthalmology, University of California, Los Angeles, CA
***** Correspondence: Zhihong Jewel Hu, jhu@doheny.org

Abstract:

Age-related macular degeneration (AMD) is the leading cause of central vision loss in aging populations. Geographic atrophy (GA) is the advanced, non-neovascular form of AMD. Predicting the longitudinal progression of GA remains a critical challenge in ophthalmic clinical practice and clinical trial design. Forecasting the trajectory of GA is complicated by highly variable growth rates and the inherent scarcity of long-term, high-quality imaging data. To address these challenges, we introduce the Sliding Window Attention U-Net (SWAU-Net), a hybrid architecture that integrates Transformer-based temporal modeling of GA growth with precise spatial modeling of GA location with a U-Net convolutional neural network (CNN). To ensure generalization in the low-data regime, SWAU-Net embeds explicit temporal and geometric consistency priors via a weight-shared Sliding Window Attention core and feature-level regularization that preserves sparse, high-frequency lesion boundaries across frames. Experimental results demonstrate that these structural constraints prevent the model from overfitting to imaging noise, achieving a Growth Mask Dice Similarity Coefficient (DSC) of 0.66, representing a significant improvement over unregularized Transformer and standard recurrent baseline models. Our framework provides a robust tool for predicting GA lesion trajectories, potentially supporting more efficient clinical trial designs and personalized patient monitoring.

Keywords: Geographic Atrophy, Longitudinal Progression Prediction, Video Prediction, Transformer, Low-Data Regime

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

Citation: To be added by editorial staff during production.

1. Introduction

1.1. Geographic Atrophy and Retinal Imaging

Geographic Atrophy (GA) is the advanced, non-neovascular form of age-related macular degeneration (AMD), representing a leading cause of irreversible central vision loss among elderly populations [1,2]. GA arises from progressive degeneration of the retinal pigment epithelium (RPE), photoreceptors, and the underlying choriocapillaris, producing sharply demarcated, map-like regions of atrophy in the macula. Although peripheral vision is typically preserved, foveal involvement leads to profound loss of reading and face-recognition ability [3,4].

Quantitative characterization of GA progression has become a key endpoint in both natural-history studies and interventional trials [5]. Critical prognostic features include junctional zone hyperautofluorescence, drusen regression, hyperreflective foci, and choroidal thinning—each reflecting local RPE and photoreceptor stress that anticipates lesion expansion [6,7].

Fundus Autofluorescence (FAF) remains the gold-standard non-invasive imaging modality for monitoring GA. FAF visualizes lipofuscin accumulation and loss within the RPE, offering high-contrast delineation of atrophic borders [8]. Complementary to FAF, Optical Coherence Tomography (OCT) provides volumetric cross-sectional views that resolve structural biomarkers [9]. Longitudinal FAF and OCT imaging together enable clinicians to measure both the spatial extent and evolution of GA lesions, supporting visual-function prediction and treatment evaluation in clinical trials [10,11].

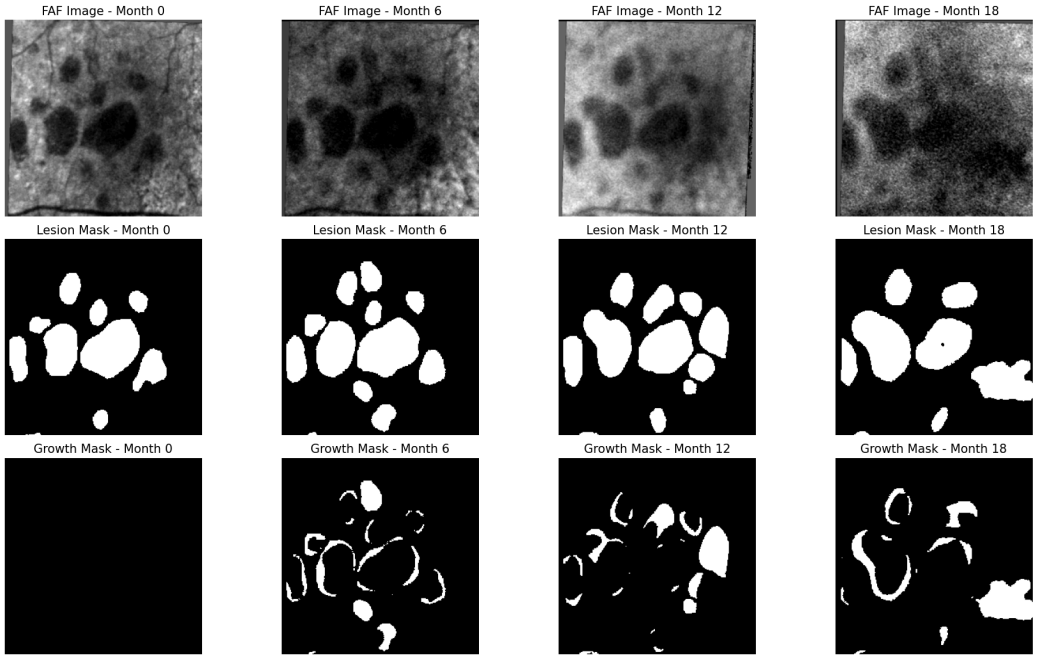

Figure 1. Top: FAF scan, with geographic atrophy (GA) clearly visible as the dark central region; **Center:** Mask of the GA region, annotated by a human expert; **Bottom:** Mask of the growth region since the previous scan (calculated from the difference of adjacent GA masks).

Forecasting GA progression—particularly from limited historical data—remains a critical unmet need for personalizing monitoring schedules and therapeutic decision-making. This challenge arises because GA expansion is a heterogeneous process driven by the local retinal microenvironment and baseline lesion geometry. Progression is rarely uniform; instead, it often manifests through the sudden coalescing of satellite lesions or irregular protrusions into healthy tissue. Because a single clinical snapshot cannot capture this underlying process, accurate forecasting requires models that can interpret the subtle, time-varying shifts at the junctional zone, where metabolic stress precedes visible structural collapse.

The spatial and temporal complexity of the lesion's "growth front" is illustrated in Figure 1, which depicts GA progression over 18 months. The sequence highlights the sparse, irregular nature of the expansion regions, which necessitates high-fidelity spatiotemporal modeling.

1.2. Spatiotemporal Deep Learning

Early sequence-prediction frameworks such as ConvLSTMs and 3D ConvNets [12,13] established spatiotemporal encoder–decoder architectures for video forecasting. These were later extended by attention-based models such as Vision Transformer (ViT) and TimeSformer that capture long-range dependencies [14,15].

While Transformer-based architectures achieve high performance at vision tasks, they typically require large datasets to generalize effectively. Recent works address this limitation through self-supervised pretraining, compact tokenization, and lightweight attention mechanisms that reduce data dependency [16,17]. In medical imaging, where datasets may be small and heterogeneous, directly training pure ViTs often leads to overfitting; thus, hybrid designs that preserve convolutional locality while introducing attention-based long-range reasoning have become standard. Architectures such as TransUNet, MedT (gated axial attention), UTNet, and Swin-UNet are examples of parameter-efficient hybrids that achieve strong transferability in low-data settings [18–21].

While Transformers excel at capturing long-range dependencies, they often overfit in medical imaging where datasets are small and heterogeneous. Consequently, hybrid designs like

TransUNet and Swin-UNet have become standard for their ability to preserve convolutional locality while introducing global reasoning [22–24].

1.3. Deep Learning for GA Detection and Forecasting

Recent deep learning approaches have achieved expert-level segmentation and detection of GA lesions across FAF, NIR, and OCT modalities[25–29]. The research focus is now shifting from static segmentation to temporal forecasting. Prior works have studied applications of CNN–RNN hybrids [30,31] and Neural ODEs [32] to model temporal dependencies across successive imaging visits.

A limitation of standard CNN-RNN models is that they compromise predictive accuracy for spatial expansion tasks by decoupling space and time through the flattening of 2D feature maps, whereas more effective spatio-temporal models maintain spatial coherence by integrating information flow across the feature map dimensions directly into their temporal processing units [33]. Furthermore, FAF and geometric segmentation channels exhibit distinct noise and information profiles. To mitigate signal interference, hybrid multi-path encoders have been shown to improve modality-specific feature extraction and cross-channel fusion [34,35]. Because growth masks represent thin, irregular expansion bands, conventional residual connections tend to over-smooth these regions. Frequency-selective mechanisms such as Gated or Attention-modulated Residual Blocks have proven effective at preserving edge contrast and small-scale structural information [36–38].

Unlike CNN–RNN pipelines, attention mechanisms maintain global spatial and temporal receptive fields and avoid vanishing-gradient bottlenecks, enabling richer modeling of progression dynamics [39]. However, while recurrent networks promote temporal coherence through consistent transition models, attention mechanisms lack such regularity, leading to unstable dynamics in low-data settings. In these regimes, the high expressivity of Transformers can become a liability, as the model may memorize noise or patient-specific artifacts rather than learning meaningful trends [40].

1.4. Contributions

To address these challenges, we introduce the **Sliding Window Attention U-Net (SWAU-Net)**, a hybrid CNN-Transformer architecture that embeds explicit temporal and geometric consistency priors for robust GA progression forecasting. SWAU-Net incorporates the following design principles to improve robustness in the low-data regime:

A regularized U-Net encoder–decoder for spatial feature extraction: To handle the sparsity and thin-band structure of GA growth regions, the model's spatial backbone is constrained to preserve boundary detail while suppressing noise-driven high-frequency artifacts. Lightweight channel mixing and gated residual refinement prevent over-smoothing of junctional zone features, improving robustness in low-data settings.

Sliding Window Attention (SWA): To ensure temporal generalization, we enforce a temporal stationarity prior through architectural weight-sharing across shifted windows. This imposes a structural bottleneck that prevents the attention mechanism from overfitting to noise by forcing it to learn a generalized, time-invariant transition function.

Decoupled Dynamics Network (DynNet): We structurally separate feature estimation from predictive state evolution to mitigate task interference. By decoupling these functions, we simplify the learning objective: the encoder and SWA core focus on stable state estimation, while the DynNet is dedicated purely to modeling feature evolution.

2. Materials and Methods

2.1 Data

The geographic atrophy (GA) dataset consists of longitudinal imaging data of 66 eyes from 66 patients obtained from the Doheny Image Reading and Research Lab (DIRRL) database, with

FAF imaging performed at the initial baseline visit and at six-, twelve-, and eighteen-month follow-ups.

Each FAF image has a 30° field of view with pixel dimensions of 768 × 868. All right-eye images were flipped horizontally to maintain consistency, and each sequence was registered to its baseline image. FAF images were graded using the semi-automated RegionFinder software (Heidelberg Engineering) to delineate areas of atrophy. Each image was initially segmented by a certified reading center grader and subsequently reviewed by a senior grader, with discrepancies resolved by a senior investigator.

2.2 Hybrid Encoder–Decoder Architecture and Feature Regularization

SWAU-Net is built on a four-level U-Net backbone with four downsampling stages, producing five feature resolutions (L1–L5). At each time step t, the input image I_t is a 256 × 256 × 3 tensor consisting of the FAF image, the GA lesion mask ($Mask_t$), and a growth mask defined as $Mask_t - Mask_{t-1}$.

To account for the differing statistical properties of the FAF and masks, the encoder employs a dual-path input design in which the FAF channel and the mask-derived channels are processed separately in the initial layers and concatenated thereafter. This reduces interference from FAF intensity noise while preserving geometric fidelity in the sparse lesion boundaries.

Across the encoder, standard residual blocks are augmented with a gated high-frequency pathway to prevent over-smoothing of thin growth regions. Each block computes:

$$y = f(x) + w \cdot g(x) + x$$

where $f(x)$ denotes the main convolutional pathway, $g(x)$ is a lightweight 1 × 1 detail pathway, and w is a learnable scalar initialized near zero. This formulation allows the network to prioritize stable, low-frequency representations while selectively admitting high-frequency boundary information as supported by the data, improving robustness in low-data regimes.

To further stabilize feature representations across heterogeneous inputs, a lightweight channel-fusion bottleneck (CFB) is applied at each resolution level, consisting of 1 × 1, 3 × 3, and 1 × 1 convolutions with residual connections. This encourages compact, semantically aligned features while limiting parameter growth.

Finally, to mitigate the loss of global context caused by repeated downsampling, single-frame spatial self-attention is introduced at the two deepest encoder levels (L4 and L5). A Transformer encoder layer with Rotary Positional Embeddings (RoPE) enables long-range pixel interactions within each feature map while preserving geometric coherence through relative spatial encoding.

2.3 Sliding Window Attention (SWA) for Temporal Aggregation

The Sliding Window Attention (SWA) module serves as the temporal core, explicitly imposing a *temporal stationarity prior* by applying a single, weight-shared attention operator across multiple shifted windows. This design prevents overfitting to specific time indices in short sequences and forces the model to learn time-invariant relational dynamics. The shared weights and repeated application over time-shifted inputs act as a crucial form of data augmentation and regularization in the temporal domain.

Given refined encoder features $F'_t = CFB(F_t)$ for time steps $t = 0, 1, 2$, the SWA module cyclically applies its core attention block to construct integrated temporal states M_t.
Three distinct 3-frame input tensors are processed sequentially, using F'_0 as the temporal anchor and zero-padding for missing history:

M_0 *(predicts* I_1*):* $(0, 0, F'_0)$
M_1 *(predicts* I_2*):* $(0, F'_0, F'_1)$
M_2 *(predicts* I_3*):* (F'_0, F'_1, F'_2)

Within each window, the Transformer performs unmasked self-attention over all tokens. Causality is enforced externally by the windowing mechanism, allowing the block to exploit spatial and

temporal context without accessing future frames. While this sliding window approach is computationally less efficient than standard global attention due to the redundant processing of overlapping frames, it is well-suited for short clinical sequences, where the regularization effect provided by weight-shared windows is critical for achieving generalization.

The SWA output is upsampled to match the resolution of each U-Net decoder stage and fused with local convolutional features via residual connections. This ensures temporally consistent, globally informed features directly guide GA boundary predictions at every resolution.

To efficiently capture long-range spatiotemporal dependencies, attention is factorized axially into two sequential passes:

- **Time–Width (TW):** interactions across columns and time for each row.
- **Time–Height (TH):** interactions across rows and time for each column.

This factorization preserves global context while scaling linearly with attention dimensions. Relative spatial–temporal information is maintained via Rotary Positional Embeddings (RoPE).

Because simple axial factorization ignores cross-axis spatial interactions, a lightweight **spatio-temporal gated mixer** restores full spatial coherence by fusing information across width, height, and time. The mixer output is combined with the attention residual through a trainable scalar weight (initialized near zero) to balance expressive global attention with a stable local prior.

The final aggregated output is fused with encoder features via a macro-residual connection, preserving high-frequency boundary detail, particularly in deeper U-Net levels (L4 and L5). Shallower levels (L1–L3) use simpler mixers to efficiently enforce temporal smoothness.

2.4 State Evolution and Frame Prediction

The proposed architecture is motivated by classical state-space models, which separate state estimation from temporal evolution. The encoder and Sliding Window Attention (SWA) modules produce a temporally consistent latent state, denoted as M_t, while the Dynamics Network (DynNet), implemented as a smaller 3-level U-Net, predicts the next-step latent state via $E_{t+1} = DynNet(M_t)$. Using a full U-Net for DynNet allows the network to evolve features hierarchically across spatial resolutions, preserving both global structure and high-frequency boundary information. The deepest, most abstract latent state informs predictions at all shallower resolutions, ensuring global temporal context guides local feature evolution.

Deep layers in DynNet refine features by combining attention-based updates with projections of the evolved global state, while shallow layers integrate upsampled global information with local features to capture subtle, high-frequency changes in the growth regions. The evolved features are further refined through channel fusion before being passed to the decoder for frame reconstruction.

By maintaining a separate U-Net for dynamics, the network disentangles spatial representation learning from temporal forecasting, preventing direct entanglement of noisy pixel observations with predictions. This separation stabilizes training, encourages biologically plausible dynamics, and narrows the hypothesis space—critical for modeling sparse, high-frequency growth regions in low-data clinical settings.

A diagram of the full SWAU-Net architecture is shown in Figure 2.

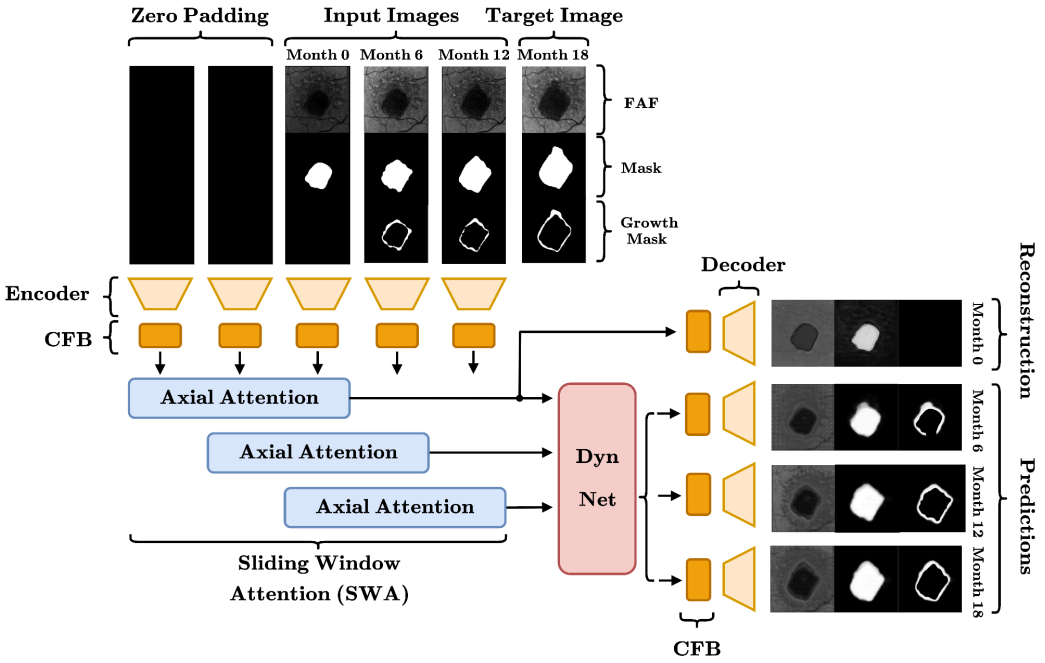

Figure 2. SWAU-Net architecture. Input images have three channels (FAF, GA Mask, Growth Mask). Each is passed through a U-Net Encoder with spatial attention, followed by the CFB block to encourage richer interactions between channels. A shared self-attention block is applied across three windows in parallel, to enforce a temporal stationarity prior. State estimates are passed through the DynNet U-Net to produce next-step predictions, while the current-frame reconstruction bypasses DynNet; this structure separates the tasks of state estimation (SWA) and prediction (DynNet).

2.5 Synthetic Pretraining via Anisotropic Growth Simulation

To address data scarcity and label sparsity, a synthetic pretraining dataset of 2,000 four-frame sequences (8,000 images total) was generated to simulate lesion evolution with realistic image noise and spatial irregularities. The simulator outputs sequences containing the FAF images, full lesion masks, and growth masks.

Key fidelity features include:

- Mask Generation: Lesion masks are initialized by thresholding multi-peak Gaussian fields, then expanded through anisotropic directional dilation combined with stochastic erosion/dilation cycles. This process generates realistic, jagged growth boundaries that specifically mimic the irregular, directional nature of clinical GA progression, enforcing a high-frequency boundary prior.
- FAF Realism: The simulator incorporates clinical artifacts such as vein-like structures and peripheral noise to ensure the model learns features robust to real-world image interference.

Pretraining establishes a strong architectural initialization, enabling faster convergence and enforcing a prior that emphasizes high-frequency boundary fidelity and robustness to noise before fine-tuning on limited clinical data. Example synthetic pretraining data is displayed in Figure 3.

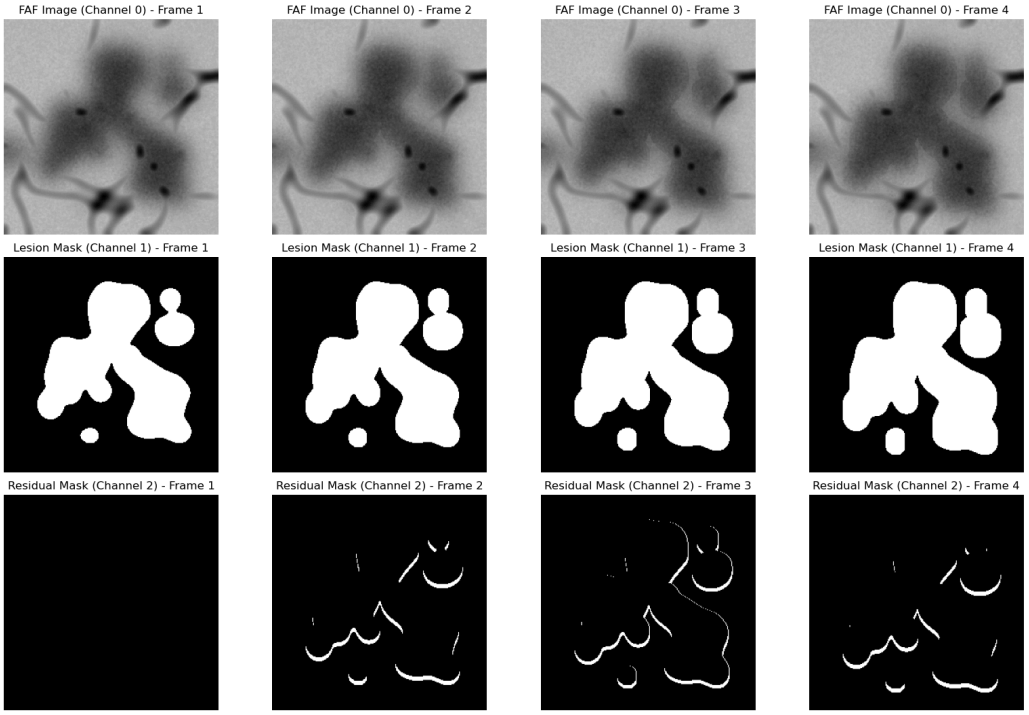

Figure 3. Example synthetic training data. **Top:** Simulated FAF; **Center:** GA Mask; **Bottom:** Growth region.

2.6 Loss Formulation and Training Strategy

We train the model using a hybrid loss:

$$L_{Total} = L_{Pred} + \lambda_{Recon} \cdot L_{Recon}$$

with $\lambda_{Recon} < 0.5$, balancing prediction and reconstruction objectives. The prediction loss L_{Pred} is applied to frames I_1, I_2, and I_3 and combines a soft DICE loss on the GA mask and growth mask—weighted heavily on the sparse growth regions—with a small Binary Cross-Entropy term for stability, and an L1 loss on the FAF channel to ensure accurate reconstruction.

The model is trained end-to-end on the synthetic dataset to establish strong priors for spatial feature recognition and boundary localization. The model is then fine-tuned on real clinical sequences. We use gradient accumulation to simulate larger batch sizes to reduce variance in small datasets. Standard geometric augmentations (rotation, flip, zoom) and targeted random intensity corruption of the FAF channel further encourage the network to prioritize stable geometric features over noisy intensity cues, improving generalization in low-data settings. Each batch of images is dynamically augmented online, applying geometric transformations (rotation, flip, zoom) and targeted random intensity corruption of the FAF channel. This ensures that the network sees a different augmented version of each image every epoch, encouraging it to prioritize stable geometric features over noisy intensity cues and improving generalization in low-data settings.

3. Experiments

3.1 Experimental Setup

The SWAU-Net model uses a U-Net with four down-sampling stages, producing 16 × 16 bottleneck features from a 256 × 256 input (downsampled from the original 768 × 868). With a base channel width of C = 16, the deepest layer contains 256 channels. The SWA block operates on feature maps up to 32 × 32 × 128 at level L4 and 16 × 16 × 256 at level L5. The model contains approximately 8.3 million parameters. All ablated models are built upon this shared backbone, utilizing the same dual-path input, Adaptive Gated Residual Blocks (AGRB), and

two-stage training (synthetic pretraining followed by clinical fine-tuning), unless explicitly ablated below.

To validate the efficacy of our design, particularly the regularization and decomposition scheme required for the low-data regime, we designed the following ablations and benchmarks:

Table 1. Ablations and benchmarks.

Model / Ablation Name	Core Modification	Primary Hypothesis Tested
No Spatial Attention	Removes all Spatial Self-Attention layers within the Encoder and DynNet.	Tests the contribution of non-local pixel interactions vs. purely local convolutional processing in maintaining feature fidelity.
No Channel Fusion Bottleneck	Replaces all Channel Fusion Bottleneck (CFB) blocks with simple residual skips and concatenation.	Tests the importance of explicit semantic alignment of multi-modal input features (FAF, GA Mask, Growth Mask).
SWA Ablation 1 (Standard Attention)	Replaces SWA with Standard Causal Axial Attention (non–weight-shared).	Tests whether SWA's temporal-stationarity prior (via weight-sharing) is needed to prevent highly expressive but unregularized Transformers from overfitting small datasets.
SWA Ablation 2 (Temporal Aggregator)	Replaces SWA with a simple convolutional aggregator (feature concatenation) at L1–L3.	Tests whether the stable CNN backbone (DynNet-based decomposition) alone is sufficient, or if explicit attention-based temporal aggregation is required.
SWA Ablation 3 (ConvLSTM)	Replaces the entire SWA core with a sequence of standard ConvLSTM cells, but retains spatial attention and CFB.	Tests whether our decoupled hybrid architecture (CNN → Attention → DynNet) provides stability or expressivity benefits over conventional coupled ConvLSTM approaches.
No DynNet	Removes the Dynamics Network (DynNet) and forces prediction directly from the estimated state.	Tests whether decoupling feature estimation from state evolution prevents the model from collapsing into a static autoencoder.
No Synthetic Pretraining	Skips phase 1 of training and initializes the model directly on the small clinical dataset.	Tests whether establishing a strong, generalized prior (especially for high-frequency boundaries) is required for Transformer components to converge effectively in the target domain.
No Data Augmentation	Removes online data augmentation, including FAF intensity jitter and noise, and geometric transformations (flips, rotations, etc).	Tests whether data augmentation is necessary to stabilize attention-based layers on the small clinical dataset.

These ablations are designed to isolate which components are genuinely necessary for stable prediction in the low-data regime. Removing spatial attention tests whether non-local pixel interactions matter beyond convolutional context, while removing the CFB measures how much the model relies on explicit fusion of the three input modalities. The three SWA ablations progressively test (1) whether temporal weight-sharing is required to regularize the Transformer, (2) whether SWA's adaptive, non-local temporal reasoning is doing more than a simple causal convolutional mixer, and (3) whether our decoupled CNN→Attention→DynNet design offers advantages over a fully recurrent ConvLSTM, which is sequential and cannot leverage the parallel computation that makes SWAU-Net efficient. The DynNet ablation tests whether explicitly separating state estimation and prediction is an important regularizer. Finally, the pretraining and augmentation ablations assess whether synthetic boundary priors and robustness noise models are required for the attention layers to converge on small clinical datasets.

We pretrain all models for 50 epochs on the synthetic dataset, then finetune on real data for an additional 60 epochs. Optimization is performed using the Adam optimizer with a learning rate of 1E-3 during pretraining and 1E-4 during finetuning. We use a Dropout rate of 0.2.

3.2 Results

Figure 4 compares the ground truth GA masks and growth regions with the predictions generated by the trained SWAU-Net model.

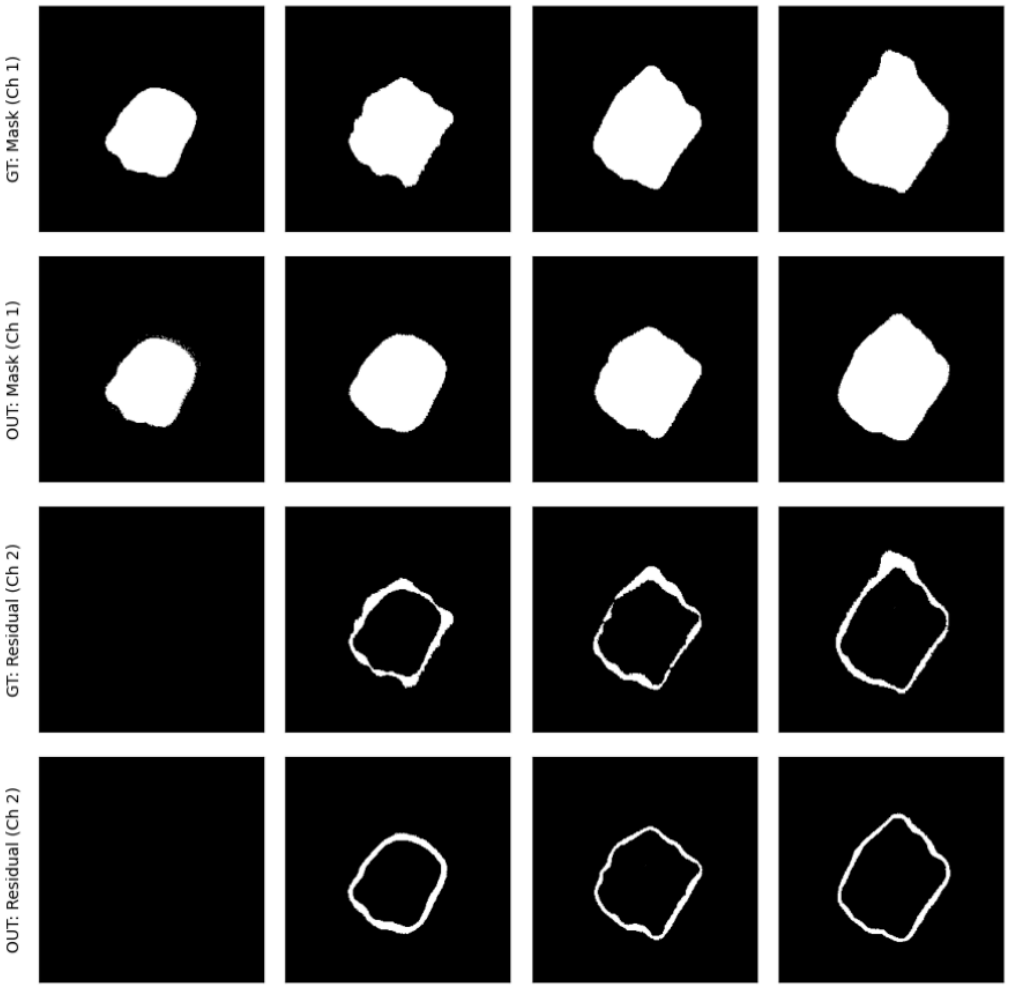

Figure 4. Ground truth masks and predictions for Months 0, 6, 12, 18 (left to right). **First row:** Ground truth GA masks. **Second row:** Predicted masks. **Third row:** Ground truth growth masks. **Fourth row:** Predicted growth masks.

We evaluated the performance of each model by computing the DICE index of the predicted GA and growth masks, thresholded at probabilities of 0.5, as compared to the ground truth growth masks. We performed 5-fold cross validation, using the median DICE score for each test fold, and averaging this over the last 10 epochs of training for additional stability (epochs 50-60). Results are displayed in Table 1. We evaluated statistical significance using a corrected paired *t*-test (Nadeau and Bengio), which accounts for the correlation of train/test folds in *k*-fold cross-validation, as shown in Table 2. Note that metrics for the growth mask are more informative than those for the total GA mask, since even minor errors in the narrow growth region can substantially affect lesion expansion predictions.

Table 2. Five-fold validation accuracy for GA region masks and growth masks for SWAU Net and ablations/benchmarks. Plus/minus signs indicate one standard deviation.

Model	Mask	Growth Mask
SWAU Net	0.94 ± 0.01	**0.66 ± 0.01**
Spatial Attention Ablation	0.94 ± 0.01	0.64 ± 0.01
CFB Ablation	0.92 ± 0.01	0.53 ± 0.02
SWA Ablation 1 (Standard Attention)	0.92 ± 0.02	0.52 ± 0.02
SWA Ablation 2 (Temporal Aggregator)	0.94 ± 0.01	0.63 ± 0.01
SWA Ablation 3 (Attention-ConvLSTM)	0.94 ± 0.01	**0.66 ± 0.01**
DynNet Ablation	0.94 ± 0.01	0.63 ± 0.03
No Synthetic Pretraining	0.94 ± 0.01	0.64 ± 0.02
No Data Augmentation	0.94 ± 0.02	0.63 ± 0.04

SWAU-Net achieved a median Mask DICE score of **0.94 ± 0.01** and Growth Mask DICE score of **0.66 ± 0.01**, a significant improvement over the CNN-LSTM-based ReConNet models [30], which were trained on the same dataset. The latter achieved median Mask DICE scores of **0.87** (CNN-LSTM model, ReConNet1-Initial) and **0.89** ensemble model (ReConNet2-Final), and median Growth Mask DICE scores of **0.60** (ReConNet1-Interval).

The ablation study demonstrated that SWAU-Net's generalization capacity hinges on its explicit consistency priors, achieving a top Growth Mask DICE score of **0.66 ± 0.01**. The most critical components were those enforcing stability in the low-data regime: removing the Sliding Window Attention (SWA) weight-sharing (Standard Attention Ablation, **0.52 ± 0.02**, $p = 0.0002$) and the Channel Fusion Bottleneck (CFB) (**0.53 ± 0.02**, $p = 0.0005$) caused the greatest performance collapse, confirming the necessity of temporal stationarity and multimodal fusion.

Table 3. Corrected Paired *t*-test (Nadeau and Bengio).

Model	*p*-value	Conclusion ($p < 0.05$?)
SWAU Net vs. Spatial Attention Ablation	0.0254	Significant

SWAU Net vs. CFB Ablation	0.0005	Highly Significant
SWAU Net vs. SWA Ablation 1 (Standard Attention)	0.0002	Highly Significant
SWAU Net vs. SWA Ablation 2 (Temporal Aggregator)	0.0116	Significant
SWAU Net vs. SWA Ablation 3 (Attention-ConvLSTM)	0.9899	Not Significant
SWAU Net vs. DynNet Ablation	0.1187	Not Significant
SWAU Net vs. No Synthetic Pretraining	0.0109	Significant
SWAU Net vs. No Augmentation	0.1415	Not Significant

Ablation 1, the primary test of the SWA mechanism's efficacy, achieved a significantly lower median Mask DICE score of **0.92 ± 0.02** and a catastrophic decline in performance on the growth prediction tasks, with a median Growth Mask DICE score of **0.52 ± 0.02**, ($p = 0.0002$)) demonstrating the ablated model's inability to generalize in this data-scarce setting. This result validates our central architectural thesis: by injecting strong temporal priors into the Transformer framework, SWAU-Net successfully stabilizes the expressive attention mechanism, enabling it to achieve recurrent-level robustness without sacrificing the parallel computation and adaptive context modeling unique to Transformers.

Furthermore, the model benefited from the non-local reasoning of the Transformer over a simple convolutional mixer (SWA Ablation 2, **0.63 ± 0.01**, $p = 0.0116$), and benefited significantly from Synthetic Pretraining (**0.64 ± 0.02**, $p = 0.0109$).

SWAU-Net achieves the same high robustness and generalization performance (**0.66 ± 0.01**, $p = 0.9899$) as the best recurrent model (SWA Ablation 3) while offering the benefits of parallel processing, superior expressivity, and adaptive context modeling.

In a clinical context, this enhanced stability translates directly to a more reliable estimate of lesion expansion volume and timing, enabling clinicians to optimize patient monitoring schedules and better time emerging interventional therapies that rely on predicting the boundary of future atrophy.

4. Conclusion

We introduced the Sliding Window Attention U-Net (SWAU-Net), a hybrid CNN–Transformer architecture that embeds explicit temporal and spatial consistency priors for robust prediction under limited data. By separating estimation, temporal aggregation, and state evolution, the model preserves the inductive strengths of classical filtering while leveraging the expressivity of attention. The Sliding Window Attention module enforces temporal stationarity through weight sharing, and the separate dynamics network isolates feature evolution from reconstruction, encouraging interpretable and stable temporal learning.

This work illustrates that combining explicit causal structure with data-driven attention improves generalization in low-data regimes. Future efforts will extend this framework to longer sequences and continuous-time dynamics, advancing toward data-efficient, physically grounded Transformer architectures for general spatiotemporal modeling.

References

1. Fleckenstein, M.; Schmitz-Valckenberg, S.; Chakravarthy, U. Age-Related Macular Degeneration: A Review. *JAMA* **2024**, *331*, 147–157, doi:10.1001/jama.2023.26074.
2. Bakri, S.J.; Bektas, M.; Sharp, D.; Luo, R.; Sarda, S.P.; Khan, S. Geographic Atrophy: Mechanism of Disease, Pathophysiology, and Role of the Complement System. *J. Manag. Care Spec. Pharm.* **2023**, *29*, S2–S11, doi:10.18553/jmcp.2023.29.5-a.s2.
3. Li, M.; Huisingh, C.; Messinger, J.; Dolz-Marco, R.; Ferrara, D.; Freund, K.B.; Curcio, C.A. HISTOLOGY OF GEOGRAPHIC ATROPHY SECONDARY TO AGE-RELATED MACULAR DEGENERATION: A Multilayer Approach. *Retina Phila. Pa* **2018**, *38*, 1937–1953, doi:10.1097/IAE.0000000000002182.
4. Sohn, E.H.; Flamme-Wiese, M.J.; Whitmore, S.S.; Workalemahu, G.; Marneros, A.G.; Boese, E.A.; Kwon, Y.H.; Wang, K.; Abramoff, M.D.; Tucker, B.A.; et al. Choriocapillaris Degeneration in Geographic Atrophy. *Am. J. Pathol.* **2019**, *189*, 1473–1480, doi:10.1016/j.ajpath.2019.04.005.
5. Fleckenstein, M.; Mitchell, P.; Freund, K.B.; Sadda, S.; Holz, F.G.; Brittain, C.; Henry, E.C.; Ferrara, D. The Progression of Geographic Atrophy Secondary to Age-Related Macular Degeneration. *Ophthalmology* **2018**, *125*, 369–390, doi:10.1016/j.ophtha.2017.08.038.
6. Schmitz-Valckenberg, S.; Nadal, J.; Fimmers, R.; Lindner, M.; Holz, F.G.; Schmid, M.; Fleckenstein, M.; FAM Study Group Modeling Visual Acuity in Geographic Atrophy Secondary to Age-Related Macular Degeneration. *Ophthalmol. J. Int. Ophtalmol. Int. J. Ophthalmol. Z. Augenheilkd.* **2016**, *235*, 215–224, doi:10.1159/000445217.
7. Huang, A.; Wu, Z.; Ansari, G.; Von Der Emde, L.; Pfau, M.; Schmitz-Valckenberg, S.; Fleckenstein, M.; Keenan, T.D.L.; Sadda, S.R.; Guymer, R.H.; et al. Geographic Atrophy: Understanding the Relationship between Structure and Function. *Asia-Pac. J. Ophthalmol. Phila. Pa* **2025**, *14*, 100207, doi:10.1016/j.apjo.2025.100207.
8. Sparrow, J.R. Bisretinoids of RPE Lipofuscin: Trigger for Complement Activation in Age-Related Macular Degeneration. *Adv. Exp. Med. Biol.* **2010**, *703*, 63–74, doi:10.1007/978-1-4419-5635-4_5.
9. Ferrara, D.; Silver, R.E.; Louzada, R.N.; Novais, E.A.; Collins, G.K.; Seddon, J.M. Optical Coherence Tomography Features Preceding the Onset of Advanced Age-Related Macular Degeneration. *Invest. Ophthalmol. Vis. Sci.* **2017**, *58*, 3519–3529, doi:10.1167/iovs.17-21696.
10. Papaioannou, C. Advancements in the Treatment of Age-Related Macular Degeneration: A Comprehensive Review. *Postgrad. Med. J.* **2024**, *100*, 445–450, doi:10.1093/postmj/qgae016.
11. Ferrara, D.; Silver, R.E.; Louzada, R.N.; Novais, E.A.; Collins, G.K.; Seddon, J.M. Optical Coherence Tomography Features Preceding the Onset of Advanced Age-Related Macular Degeneration. *Invest. Ophthalmol. Vis. Sci.* **2017**, *58*, 3519–3529, doi:10.1167/iovs.17-21696.
12. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.; Woo, W. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting 2015.
13. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks 2015.
14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale 2021.
15. Bertasius, G.; Wang, H.; Torresani, L. Is Space-Time Attention All You Need for Video Understanding? 2021.
16. Cao, Y.-H.; Yu, H.; Wu, J. Training Vision Transformers with Only 2040 Images 2022.
17. Hassani, A.; Walton, S.; Shah, N.; Abuduweili, A.; Li, J.; Shi, H. Escaping the Big Data Paradigm with Compact Transformers 2022.
18. 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 2021.
19. Valanarasu, J.M.J.; Oza, P.; Hacihaliloglu, I.; Patel, V.M. Medical Transformer: Gated Axial-Attention for Medical Image Segmentation 2021.
20. Gao, Y.; Zhou, M.; Metaxas, D. UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation 2021.
21. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation 2021.
22. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows 2021.
23. Li, K.; Wang, Y.; Zhang, J.; Gao, P.; Song, G.; Liu, Y.; Li, H.; Qiao, Y. UniFormer: Unifying Convolution and Self-Attention for Visual Recognition 2023.
24. Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; Feichtenhofer, C. Multiscale Vision Transformers 2021.

25. Shi, X.; Keenan, T.D.L.; Chen, Q.; Silva, T.D.; Thavikulwat, A.T.; Broadhead, G.; Bhandari, S.; Cukras, C.; Chew, E.Y.; Lu, Z. Improving Interpretability in Machine Diagnosis: Detection of Geographic Atrophy in OCT Scans. *Ophthalmol. Sci.* **2021**, *1*, doi:10.1016/j.xops.2021.100038.
26. Yao, H.; Wu, Z.; Gao, S.S.; Guymer, R.H.; Steffen, V.; Chen, H.; Hejrati, M.; Zhang, M. Deep Learning Approaches for Detecting of Nascent Geographic Atrophy in Age-Related Macular Degeneration. *Ophthalmol. Sci.* **2024**, *4*, doi:10.1016/j.xops.2023.100428.
27. Spaide, T.; Jiang, J.; Patil, J.; Anegondi, N.; Steffen, V.; Kawczynski, M.G.; Newton, E.M.; Rabe, C.; Gao, S.S.; Lee, A.Y.; et al. Geographic Atrophy Segmentation Using Multimodal Deep Learning. *Transl. Vis. Sci. Technol.* **2023**, *12*, 10, doi:10.1167/tvst.12.7.10.
28. Dow, E.R.; Jeong, H.K.; Katz, E.A.; Toth, C.A.; Wang, D.; Lee, T.; Kuo, D.; Allingham, M.J.; Hadziahmetovic, M.; Mettu, P.S.; et al. A Deep-Learning Algorithm to Predict Short-Term Progression to Geographic Atrophy on Spectral-Domain Optical Coherence Tomography. *JAMA Ophthalmol.* **2023**, *141*, 1052–1061, doi:10.1001/jamaophthalmol.2023.4659.
29. Elsawy, A.; Keenan, T.D.L.; Chen, Q.; Shi, X.; Thavikulwat, A.T.; Bhandari, S.; Chew, E.Y.; Lu, Z. Deep-GA-Net for Accurate and Explainable Detection of Geographic Atrophy on OCT Scans. *Ophthalmol. Sci.* **2023**, *3*, doi:10.1016/j.xops.2023.100311.
30. Mishra, Z.; Wang, Z.; Xu, E.; Xu, S.; Majid, I.; Sadda, S.R.; Hu, Z.J. Recurrent and Concurrent Prediction of Longitudinal Progression of Stargardt Atrophy and Geographic Atrophy. *MedRxiv Prepr. Serv. Health Sci.* **2024**, 2024.02.11.24302670, doi:10.1101/2024.02.11.24302670.
31. Yoshida, K.; Anegondi, N.; Pely, A.; Zhang, M.; Debraine, F.; Ramesh, K.; Steffen, V.; Gao, S.S.; Cukras, C.; Rabe, C.; et al. Deep Learning Approaches to Predict Geographic Atrophy Progression Using Three-Dimensional OCT Imaging. *Transl. Vis. Sci. Technol.* **2025**, *14*, 11, doi:10.1167/tvst.14.2.11.
32. Mai, J.; Lachinov, D.; Reiter, G.S.; Riedl, S.; Grechenig, C.; Bogunovic, H.; Schmidt-Erfurth, U. Deep Learning-Based Prediction of Individual Geographic Atrophy Progression from a Single Baseline OCT. *Ophthalmol. Sci.* **2024**, *4*, 100466, doi:10.1016/j.xops.2024.100466.
33. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.; Woo, W. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting 2015.
34. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation 2018.
35. Li, S.; Tang, H. Multimodal Alignment and Fusion: A Survey 2025.
36. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas 2018.
37. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module 2018.
38. Schlemper, J.; Oktay, O.; Schaap, M.; Heinrich, M.; Kainz, B.; Glocker, B.; Rueckert, D. Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images. *Med. Image Anal.* **2019**, *53*, 197–207, doi:10.1016/j.media.2019.01.012.
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems; Curran Assoc. Inc., 2017; Vol. 30.
40. Tay, Y.; Dehghani, M.; Bahri, D.; Metzler, D. Efficient Transformers: A Survey 2022.