Available online at **www.joac.info**

ISSN: 2278-1862

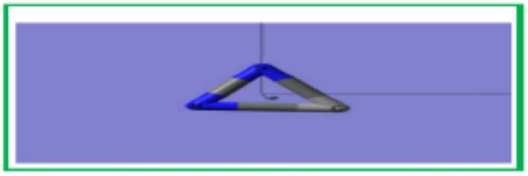
Journal of Applicable Chemistry

2025, 14 (2): 471-547

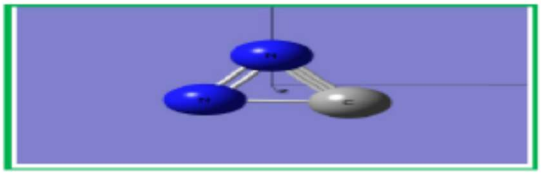
(International Peer Reviewed Journal)

New Chemistry News

$^{-}N=C=N^{-}$



New News of Chem (NNC)



ChemNewsNew (CNN)

CNN – 66D

Iam (Intelligence Augmented /Assisted Method(s))

Transformers - -architectures & Fits (2025)

Information Source	sciencedirect.com;	
S. Narasinga Rao M D *Associate Professor,* *Emergency Medicine dept.,* *Andhra Medical College,* *King George Hospital* *Visakhapatnam, A.P., India*	*K. Somasekhara Rao, Ph D* *Dept. of Chemistry,* *Acharya Nagarjuna Univ.,* *Dr. M.R.Appa Rao Campus,* *Nuzvid-521 201, India*	*R. Sambasiva Rao, Ph D* *Dept. of Chemistry,* *Andhra University,* *Visakhapatnam 530 003,* *India*
snrnaveen007@gmail.com (+91 98 48 13 67 04)	sr_kaza1947@yahoo.com (+91 98 48 94 26 18)	rsr.chem@gmail.com (+91 99 85 86 01 82)

Conspectus: Ashish Vaswani et al. published a paper entitled "Attention is All You Need" **in** the year 2017. It brought renaissance not only in sequence data processing, but also in computational paradigm with other data structures. The new approach won the favour of data scientists as a whole. This new model gained popularity as Transformer net (TransF Net) or Transformer neural network (TransF NN).

TransF NN contains two important modules, viz., attention layer and MLP-NN which help to carry out Natural Language processing (NLP).

Attention: In 2014, Bahdanau et al. proposed the idea of attention in the context of sequence-to-sequence models, used for neural machine translation (NMT). The attention mechanism targets at improving the performance of sequence-to-sequence models by allowing the model to focus on different parts of the input sequence during calculation of each output token. It overcomes limitation of encoding the entire input sequence into a single fixed-size vector which was earlier practice.

Luong et al. (2015) proposed different scoring mechanisms viz. Dot-Product (multiplicative attention), Generalized (Additive) attention and Concatenation-based attention. Another way of looking at it is global attention (where all encoder states considered) and local attention (focusing on only a subset of encoder states). The multiplicative attention is simple, efficient, and scalable for large datasets. It became a standard approach in later models and is it was a direct precursor to the self-attention mechanism in the Transformer model. The local attention is instrumental in mitigating the computational bottleneck for long input sequences, as it focuses on a smaller subset of tokens.

Attention: The attention in Transformer NN is calculated as

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q is Query matrix (current token's information), K: Key matrix (context of all tokens in the sequence), V: Value matrix (values to be passed along after attention weighting) and dk: Dimension of the key vectors (used for scaling).

Cross-Attention: It is essentially same except using the query from one sequence with the key and value from another sequence.

Mix-attention: It is computed as

$$\text{MixAttention} = \alpha \cdot \text{SelfAttention} + (1 - \alpha) \cdot \text{CrossAttention}$$

This is also widely employed in sequence text modelling. ALiBi (Attention with Linear Biases) allows more efficient handling of long sequences in transformer models.

Multi-Head Attention: The multiple attention heads are concatenated and passed through a final linear transformation

$$\text{MultiHead}(\text{Q}, \text{K}, \text{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_\text{h})\text{W}^\text{O}$$

and

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

Where W_i^Q, W_i^K, W_i^V are learned weight matrices for each attention head i.

FlashAttention: It is an optimized CUDA-based approach (on frameworks like PyTorch and TensorFlow) designed to efficiently compute the attention with GPUs for both training and inference of LLMs((GPT, BERT). It results in higher speed and increased scalability.

The evolution of architecture of TransF NN, attention mechanism, and hybridization with other approaches, during these few years, revolutionized computational modelling. This approach is in the front-line in dealing with multi-modal data (viz.Text, numerical time-series, sound (speech), image/video sequence, and tactile-sense-output) with local and global inter-dependencies

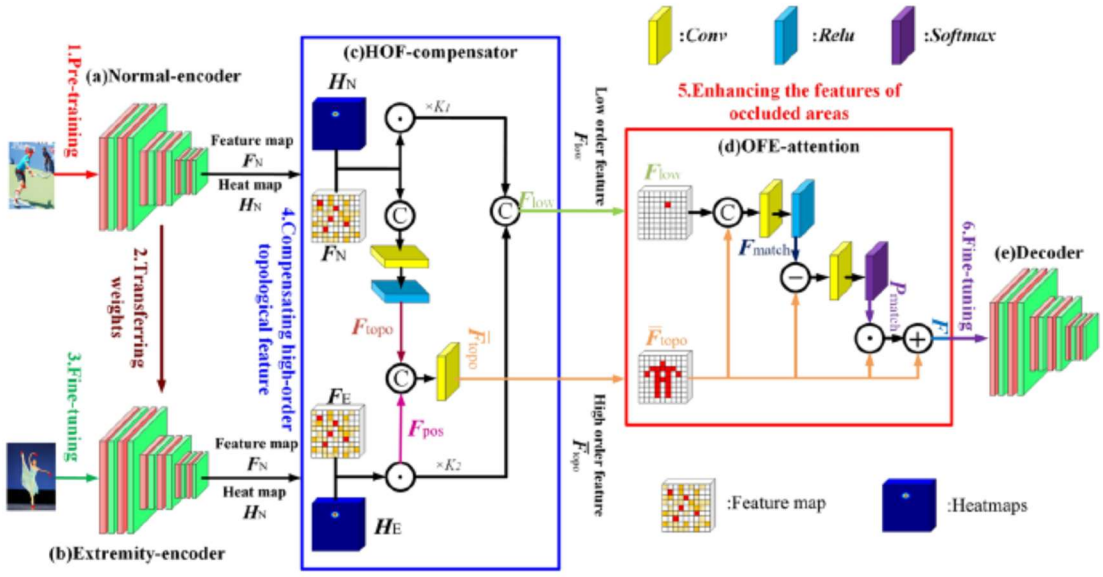
A few of Transformer net (TransF-N) architectures documented in this state-of-knowledge-methods-module for dataTOknowledge transformation are

- PCNet,
- Multi-modal multi-task hierarchical feature fusion (MM-HiFuse) ,
- Estimating energy expenditure based on video (E3V) using E3V-K5 dataset,
- Performer with Graph Self-attention Mechanism,
- Grouped Attention and Cross-Layer Fusion Network (GACLFNet),
- Neural Ordinary Differential Equation (N-ODE) Transformer,
- BioMechanically Accurate Neural Inverse KINematics solver (MANIKIN),
- Gradient Origin Embeddings (GOEmbed),
- Transformer-based VGQA model,
- Triplet convolutional twin transformer,
- Hierarchical Multi-Task Learning (HirMTL),
- Memory-augmented Deformable Detection TRansformer (MD-DETR),
- **G**eolocalization with **A**dapters and Auto-**Re**gressive **T**ransformers (GAReT),
- Vision Transformer (ViT),
- RealViformer.,
- AuraLLM,
- 3D Transformers,
- Multi-Relational Graph Contrastive Learning architecture (MRGCL) using a Multi-relational Graph Hierarchical Attention Networks (MGHAN),
- TWiX, C-TWiX,
- Squeeze and Excitation based UNet TRansformers (SE-UNETR),
- Squeeze and Excitation based High-Quality Resolution Swin Transformer Network (SE-HQRSTNet)
- BiMKANsDformer,
- Hybrid Transformer with Multi-level Fusion for Multimodal Knowledge Graph Completion (MKGformer),
- Transformer Choice Net (TCNet) [Transformer Neural Network for Choice of Transformer],
- Transformer-Based Framework
- Physics-Informed Neural Networks (PhysIinfNNSFormer), and
- EfficientNet and Vision Transformer (ViT)-based Swin Ttransformer (SwinE-Net)

Keywords: Artificial intelligence (AI); Convolution Neural Nets-- Capsule Neural Nets— MLP-Attention Mechanism-TransFormer Nets—Hybrid TransFormer Networks--

CNN : [C [Computations; Computer; Chemistry, Cell, Cellestial, Cerebrum] NN [New News; News New; Neural Nets; Nature News; News of Nature;]]
Fits : [Figure Image Table Script;]

K(nowledge)Lab
rsr.chem1979



Structure diagram. PCNET

- ✓ In HOF-compensator and OFE-attention, the colored arrows represent the output variables.
- ✓ ⊗, _, ⊕, and c_ represent matrix multiplication, element-wise matrix multiplication, element-wise matrix summation, and feature map concatenation

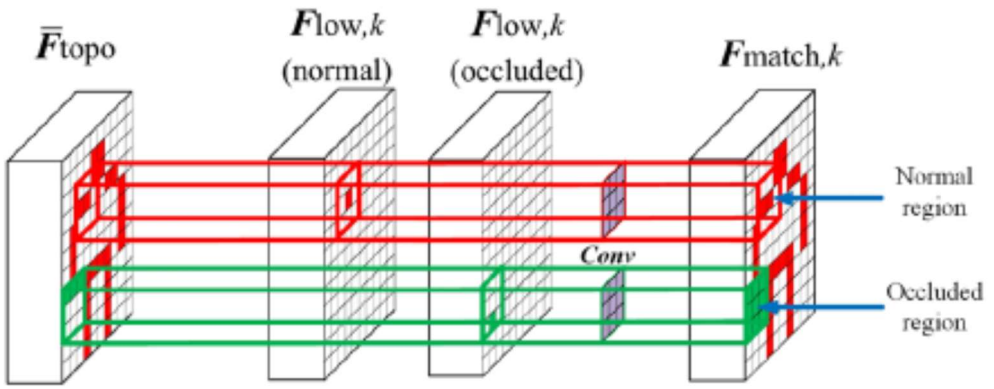
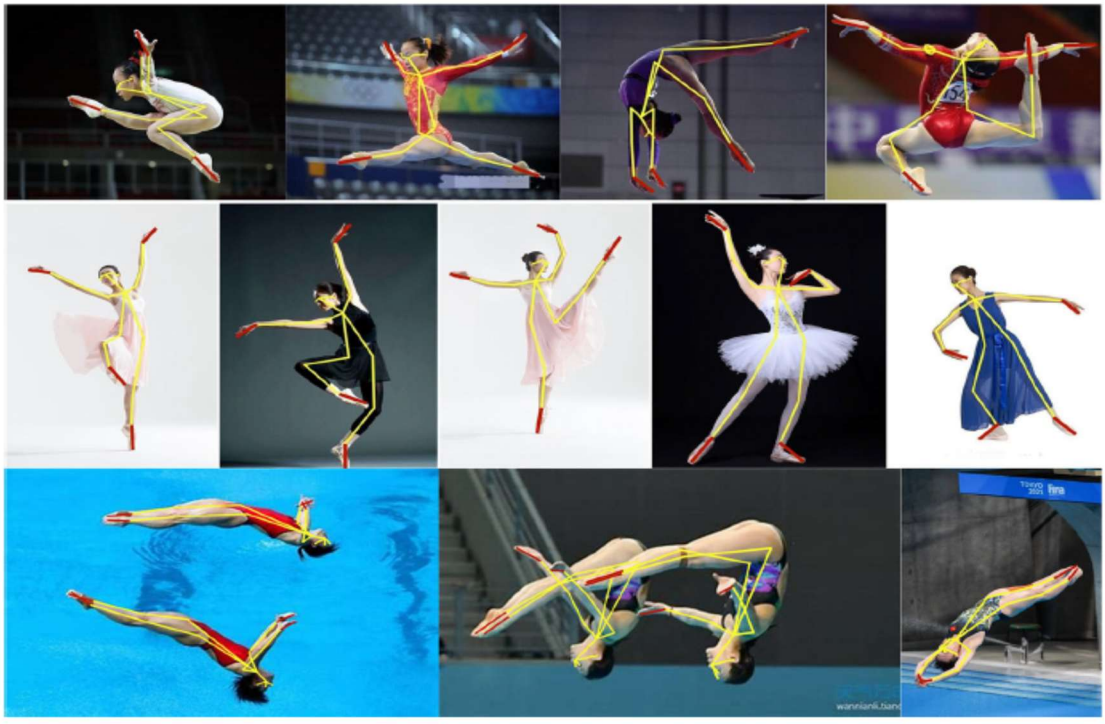


Fig. 3 Diagram of new connection establishment: After convolution, no new relationships are established for normal points (red), while new relationships are established for occluded points (green)

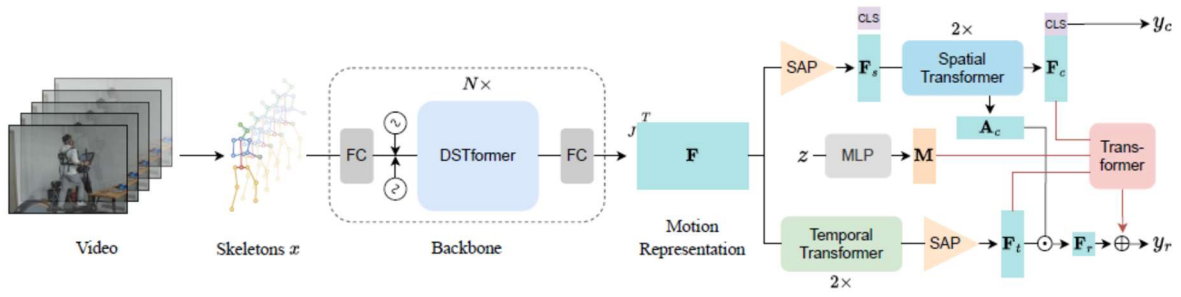
Key point detection results

- Blue points: torso and limb keypoints
- Red points: supplementary extremity keypoints



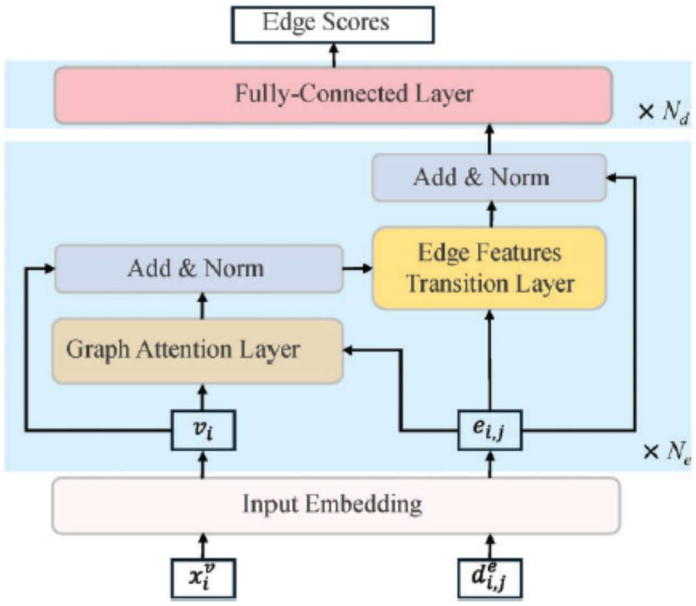
Pose estimation results in sports

- Red lines: Compensation pose estimation of extremities



Framework. E3SFormer

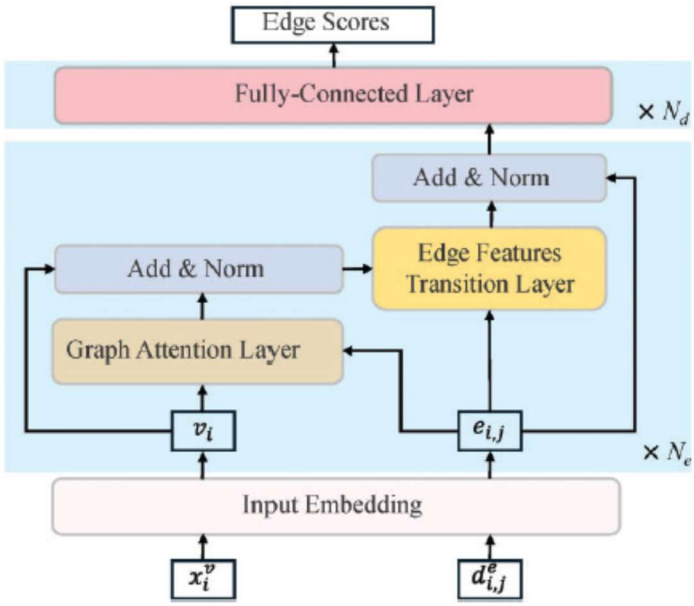
- Human skeleton sequence x is extracted using a pose estimator from the video
- Then fed into a backbone to obtain motion representation F.
- Then sent to an action recognition branch (upper) and an energy estimation regression branch (lower). The category-related joint-specific attention Ac from the action recognition branch is transferred to the energy estimation regression branch to boost its performance.
- The multi-modal data z are used for more personalized energy estimation estimation

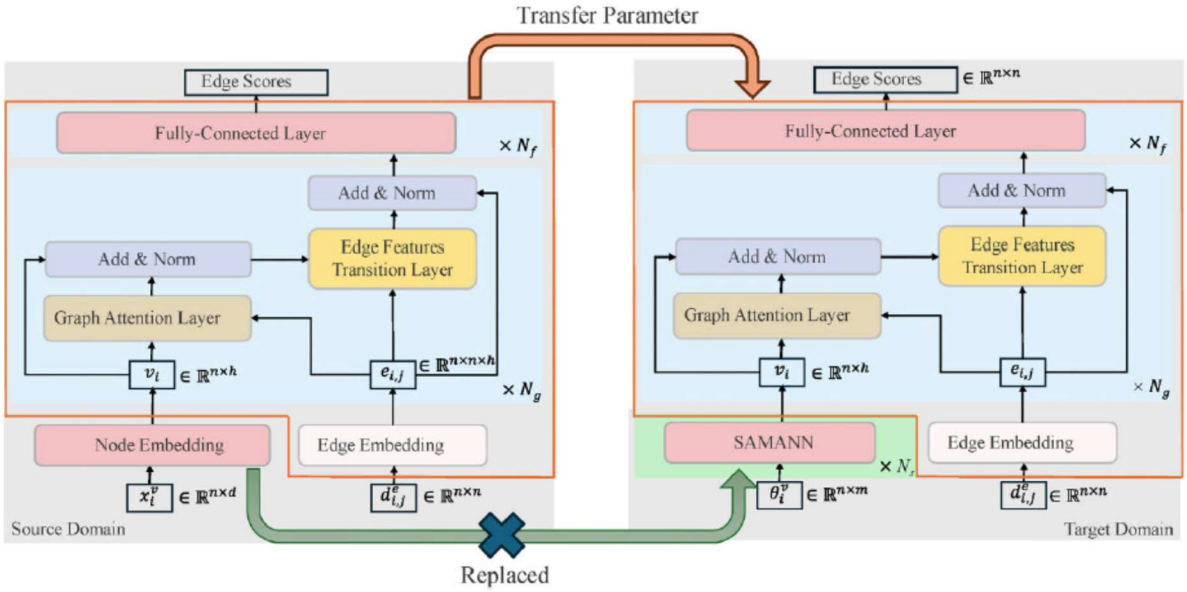


Architecture. GNN-based model

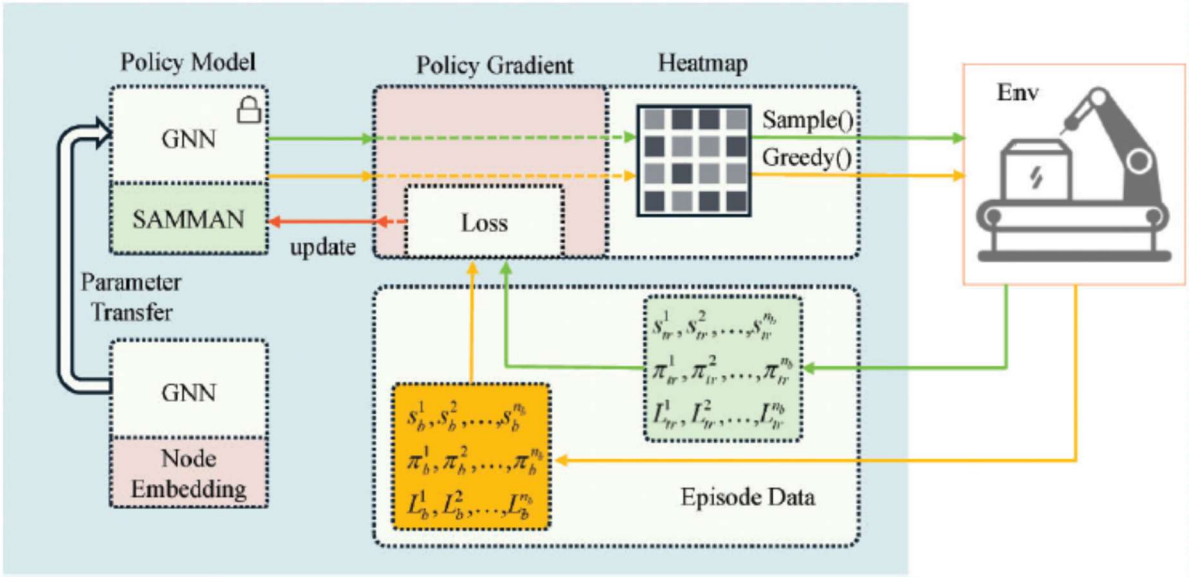
Table 1: Survey of deep/RL methods in solving TSPs.

Mothods	Learning type	Network structure
Vinyals *et al.* (2015)	SL + AR	LSTM
Bello *et al.* (2016)	RL + AR	LSTM
Khalil *et al.* (2017)	RL + AR	GNN
Nazari *et al.* (2018)	RL + AR	LSTM
Kool *et al.* (2018)	RL + AR	Transformer
Deudon *et al.* (2018)	RL + AR	Transformer
Ma *et al.* (2019)	RL + AR	GNN
Bresson & Laurent. (2021)	RL + AR	Transformer
Wu *et al.* (2022)	RL + AR	GNN
Lei *et al.* (2022)	RL + AR	GAT
Yang *et al.* (2023)	RL + AR	Transformer
Zhu *et al.* (2023)	RL + AR	CosFormer
Wang *et al.* (2023)	RL + AR	BERT
Jung *et al.* (2024)	RL + AR	CNN-Transformer
Zhu *et al.* (2024)	RL + AR	GEIAM
Zhao & Gu (2024)	RL + AR	GPN
Nowak *et al.* (2016)	SL + NAR	GNN
Joshi *et al.* (2019)	SL + NAR	GCN
Xiao *et al.* (2023)	RL + NAR	GNN





Transferring model parameter



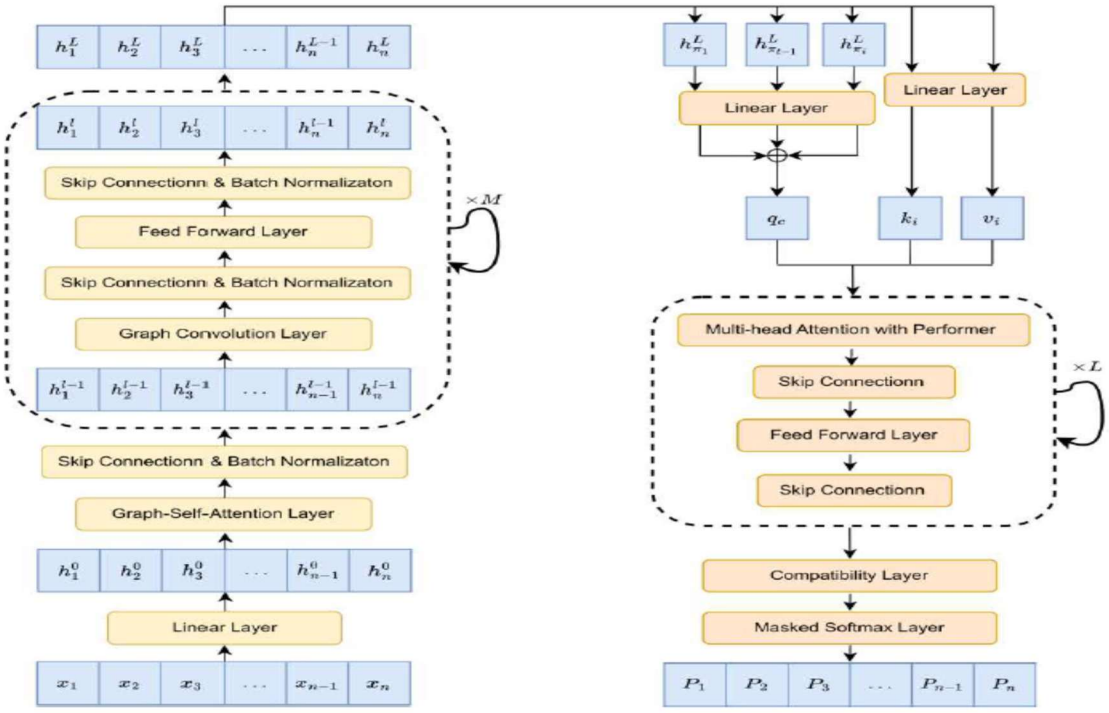
Architecture. REINFORCE transfer learning algorithm

Table 4: Performance comparisons of our method with benchmarking methods.

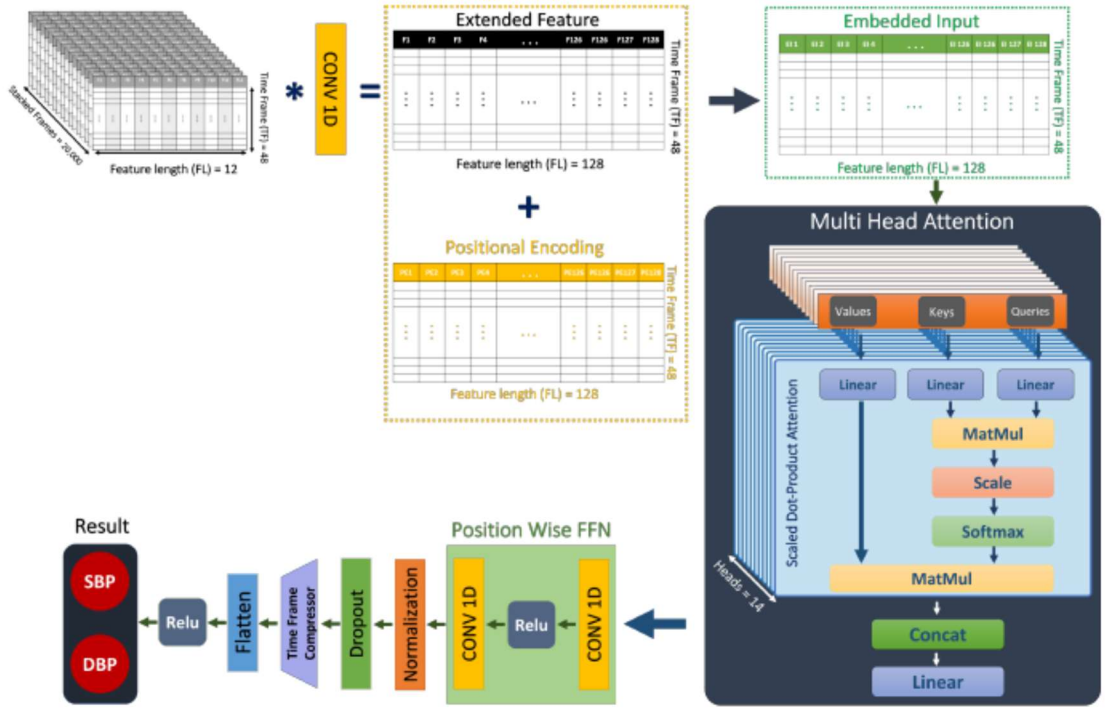
Method	Type	TSP20		
		Length	Gap (%)	Time (ms)
Gurobi	Exact	24.25	0.00	1.77e + 01
Greedy algorithm (Dijkstra, 1959)	Heuristic	26.51	9.30	6.00e−02
Nearest insertion (Held & Karp, 1961)	Heuristic	29.00	19.58	1.80e−01
2-OPT (Lin & Kernighan, 1973)	Heuristic	26.16	7.87	8.45e + 00
Fastest insertion (Chvátal *et al.*, 2010)	Heuristic	27.84	14.79	1.83e + 00
Christofides (Christofides, 2022)	Heuristic	26.51	9.32	3.50e + 00
SA (Aarts & Korst, 1989)	Meta-heuristic	26.34	8.61	7.77e + 02
GA (Chung & Xu, 2012)	Meta-heuristic	25.16	3.73	1.77e + 04
Ant colony optimization (Zhao *et al.*, 2016)	Meta-heuristic	26.33	8.56	4.55e + 02
Hill climbing (Yelmewad & Talawar, 2019)	Meta-heuristic	28.13	16.01	1.86e + 03
NAR4TSP(GS)	RL, NAR	25.75	6.19	3.61e−01
NAR4TSP(BS), B=100	RL, NAR	24.59	1.40	3.54e−01
NAR4TSP(BS), B=1000	RL, NAR	24.53	1.15	3.82e−01
Ours (GS)	TRL, NAR	25.20	3.90	3.51e−01
Ours (BS), B=100	TRL, NAR	24.50	1.05	3.65e−01
Ours (BS), B=1000	TRL, NAR	24.47	0.92	3.77e−01

- 'TRL' stands for tr ansfer r einforcement learning;
- '2-OPT' stands for 2-OPT Local Search;
- 'GS' stands for greedy search;
- 'BS' represents beam search with a width B .
- 'Time' indicates the inference time of solving a TSP instance on average
- batch sizes:TSP20, TSP50, and TSP100 in searching solutions are 1024, 128, and 16,

Transformer Net | **2025-105**



Architecture. Network
Performer Graph Self-Attention Mechanism



The block diagram depicts the streamlined process for estimating systolic and diastolic blood pressures.

Figure 3. A comparison of the signal before and after preprocessing. The photo demonstrates the exceptional noise filtering achieved, resulting in a clean and smooth signal. Furthermore, the peaks and feet are accurately preserved, showcasing the effectiveness of the preprocessing techniques employed.

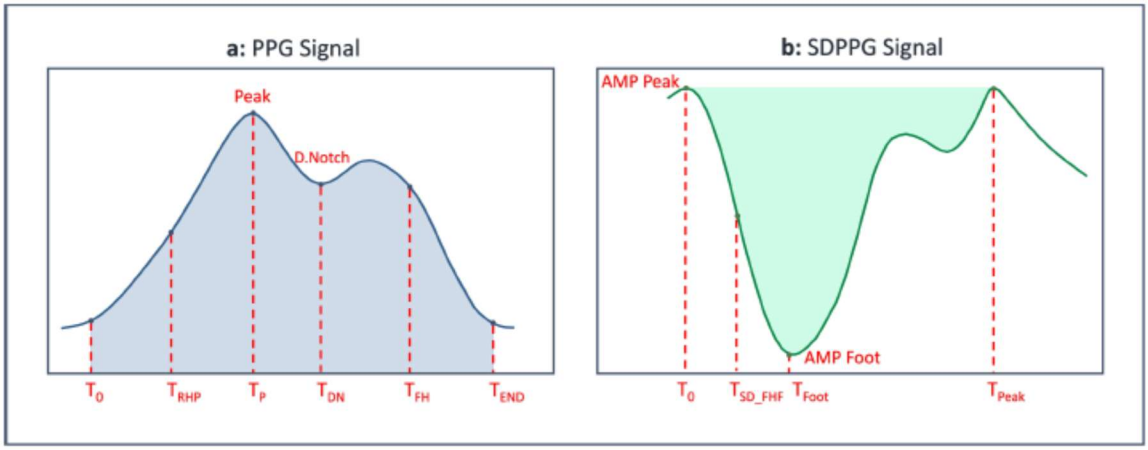
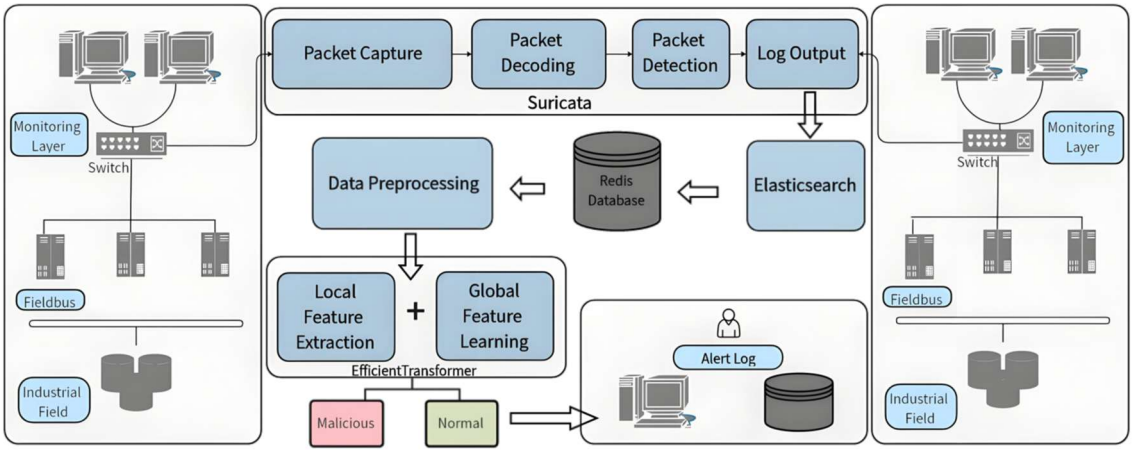
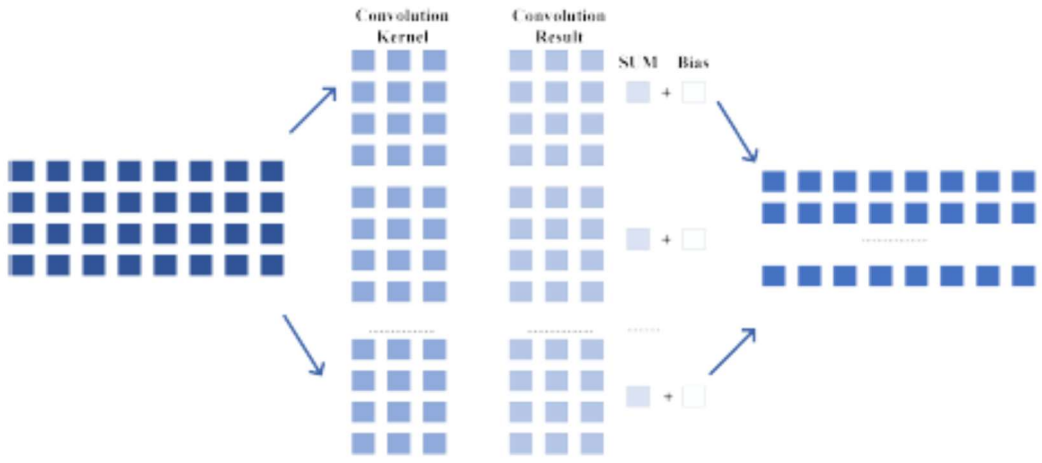


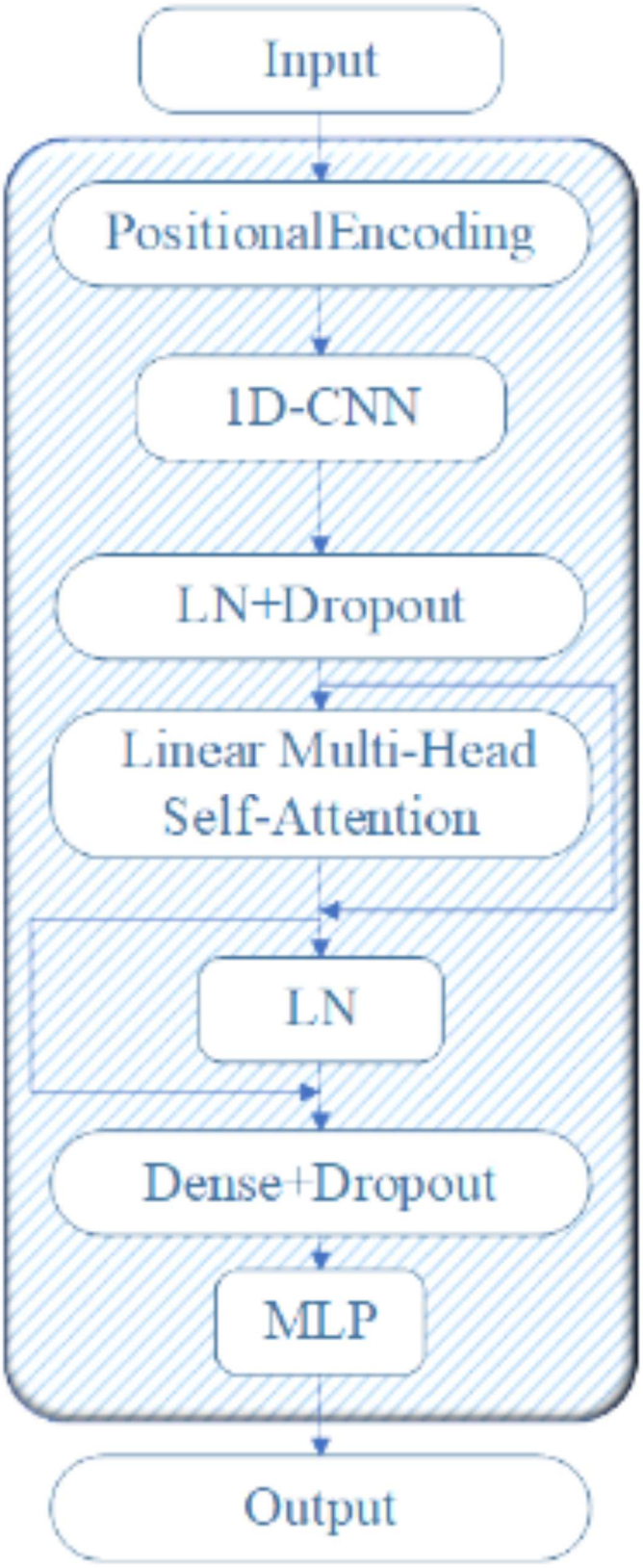
Figure 4. Characterization of features on the signal. In the PPG signal, the peak is defined as the maximum value of a PPG wave, while d.notch is defined as the minimum value between two consecutive peaks. In the SD PPG signal, amp peak refers to the maximum amplitude, while amp foot refers to the minimum amplitude between two consecutive waves.

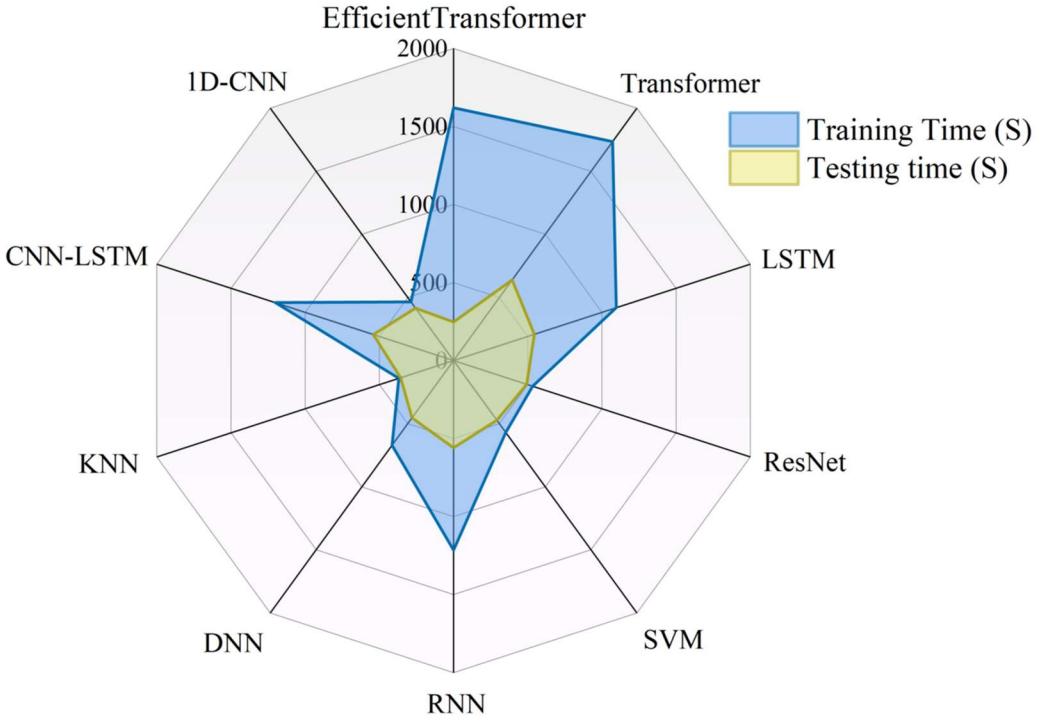
Transformer Net	2025-109	



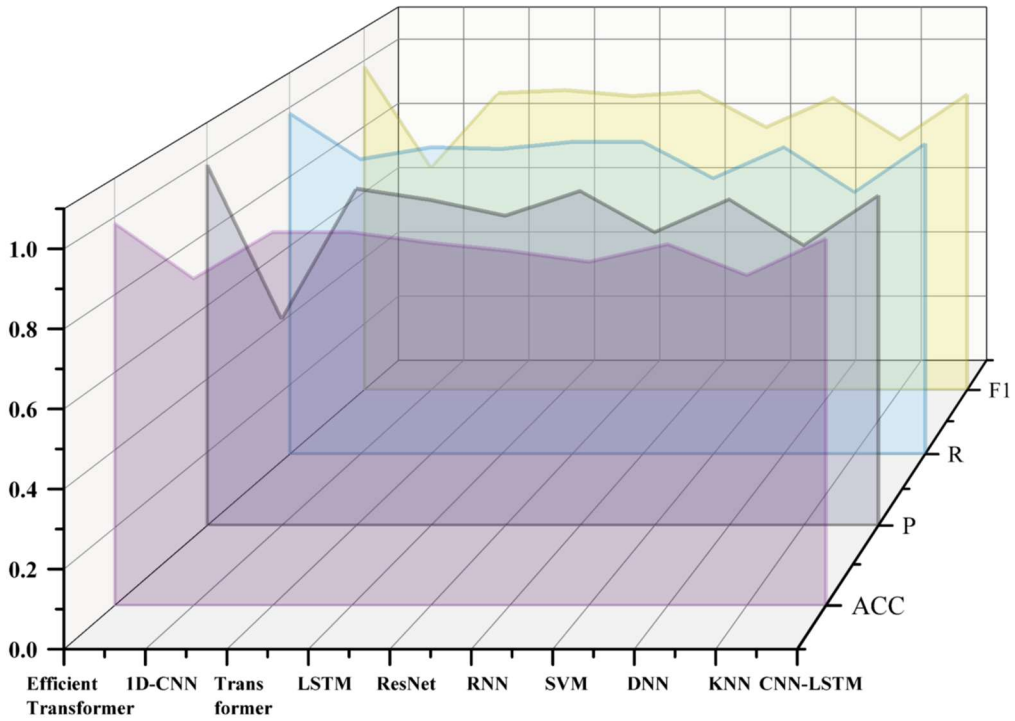
Research Path of the Dynamic Anomaly Detection Model of EfficientTransformer




Input
PositionalEncoding
1D-CNN
LN+Dropout
Linear Multi-Head Self-Attention
LN
Dense+Dropout
MLP
Output

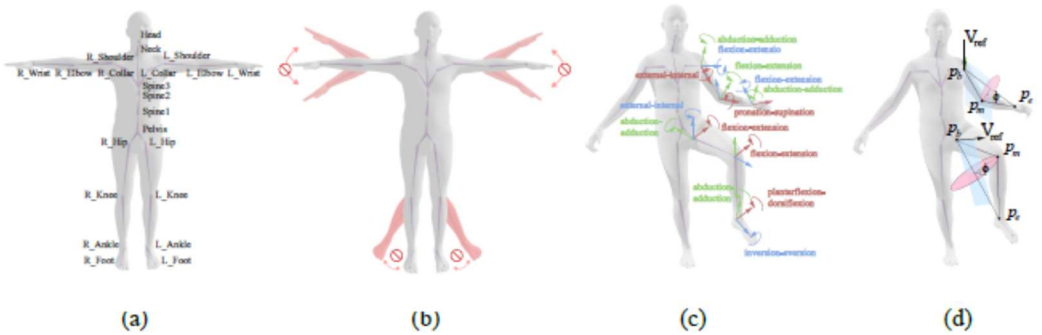


Radar Chart of Training and Testing Time for Multiple Models on the Oil and Gas Gathering and Transportation Full-Process Industrial Platform Attack-Defense Field



3D Plot of Evaluation Metrics for Multiple Models on the

Transformer Net	2025-110	



- ✓ Comparison between SMPL and biomechanically constraint limb motion
  - (a) skeleton of the SMPL model
  - (b) unrealistic SMPL joint configurations in the AMASS dataset
  - (c) biomechanically plausible joint configurations
  - (d) swivel angle parametrization of arm and leg

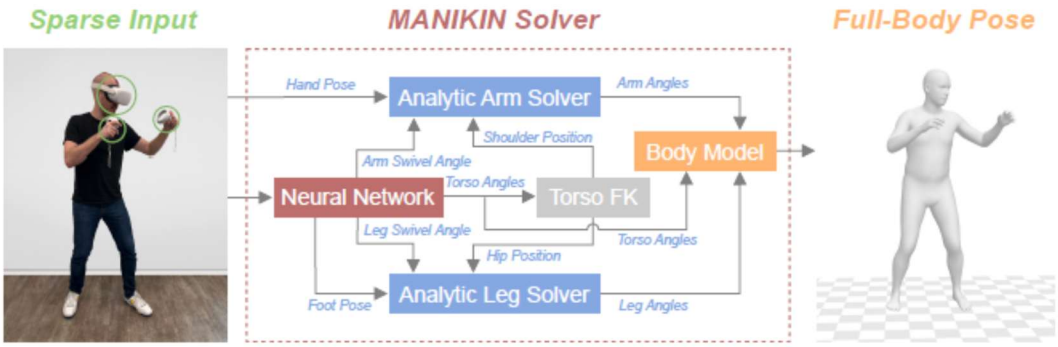
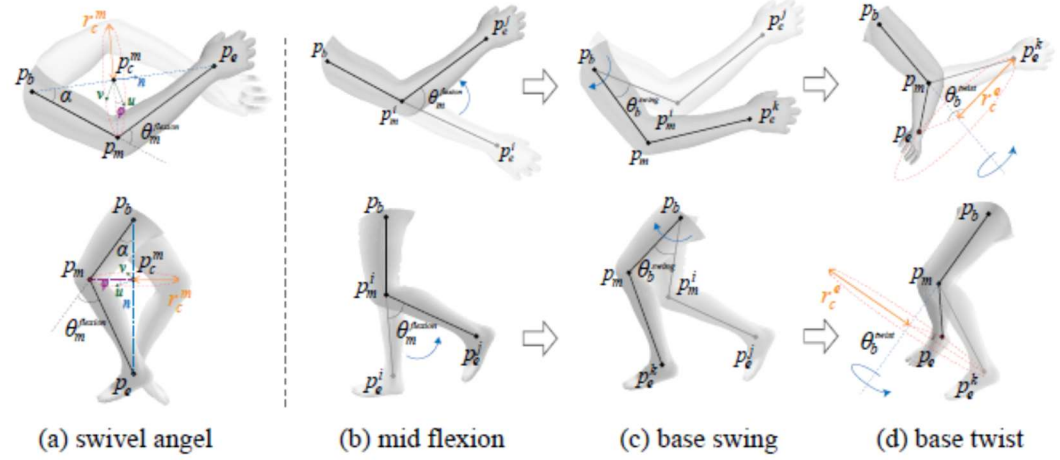


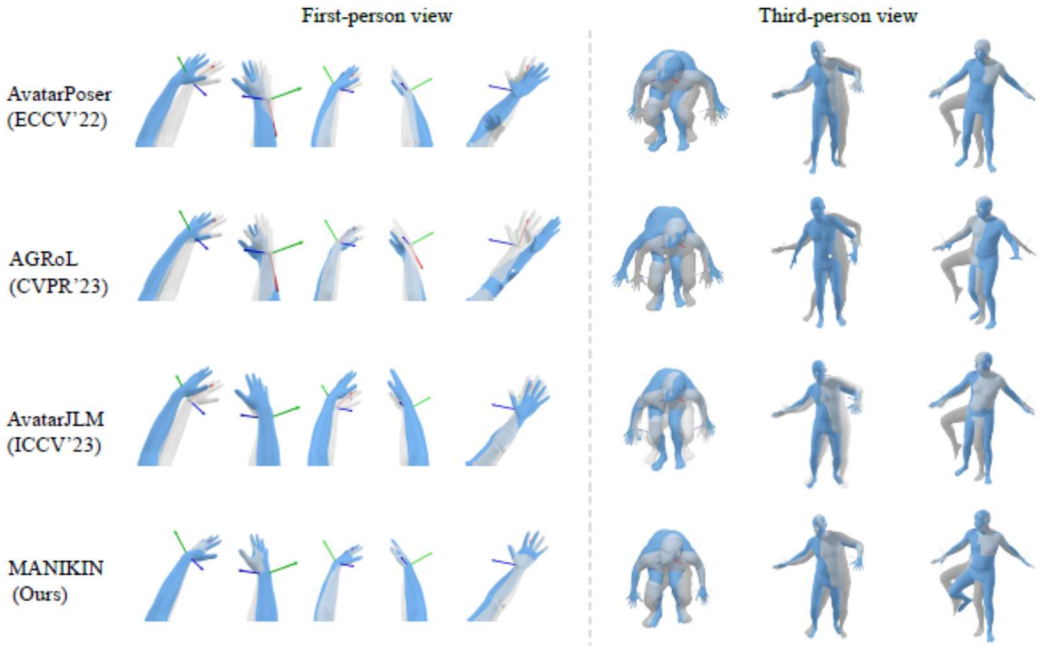
Fig. 2: Overview of MANIKIN for full-body pose estimation from sparse tracking signals. Given the 6 DoF of the head and two hands, the neural network predicts the foot pose, swivel angles for arms and legs, as well as joint angles in the torso. Next, the shoulder and hip position can be obtained through forward kinematics on the torso. Given the poses of an end effector (hand and foot), predicted swivel angles, and base joint (shoulder and hip) positions, we use analytic geometry to get all limb joint configurations. Combined with the predicted torso joints, we obtain a full-body pose.



(a) swivel angel (b) mid flexion (c) base swing (d) base twist

Illustrations of the triangular geometry of the human limbs.

- ✓ (a) shows relationship between swivel angel and mid joint position
- ✓ (b) to (d) procedure to rotate the limb from T-pose to desired **positions.**



Visual comparisons of different methods under the first and third person views

- ✓ The ground truth pose is colored in transparent gray.
- ✓ This method can perfectly match the hand observation and has better full-body prediction results than other methods

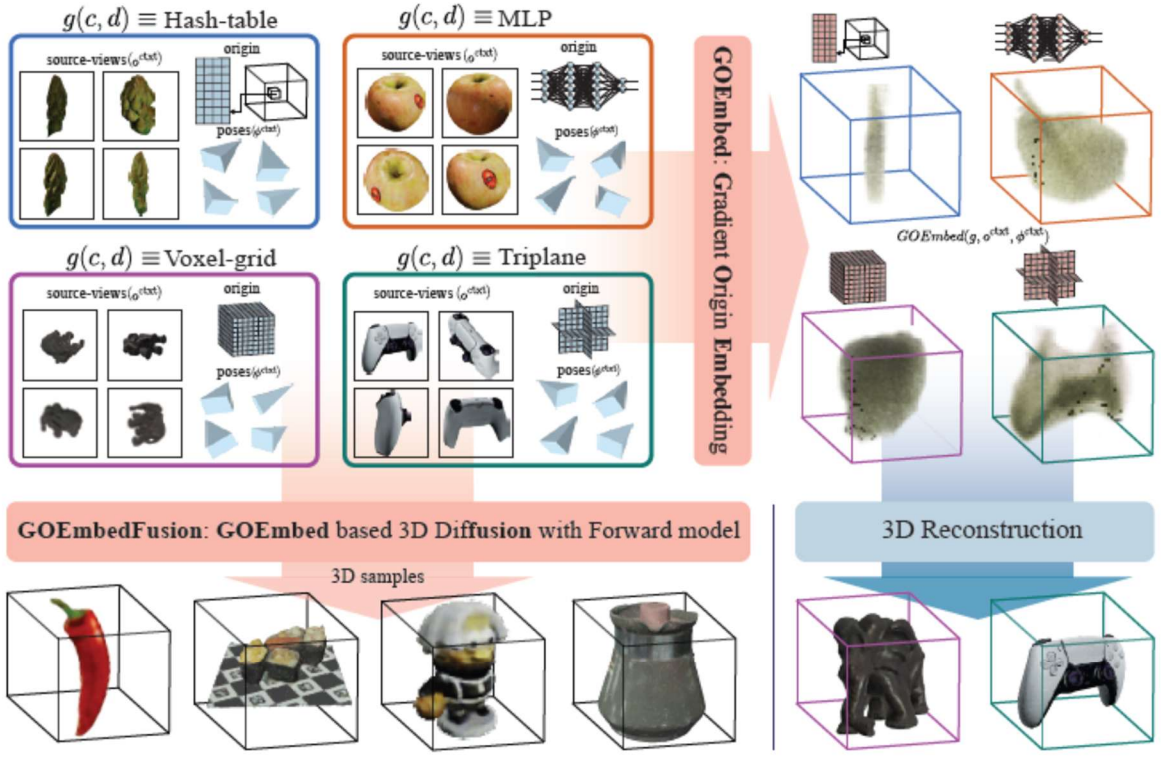


Fig. 1: We propose the **GOEmbed** (**G**radient **O**rigin **Embed**ding) mechanism that encodes source views (o^{ctxt}) and camera parameters (ϕ^{ctxt}) into arbitrary 3D Radiance-Field representations $g(c,d)$ (sec. 3). We show how these general-purpose GOEmbeddings can be used in the context of 3D DFMs (Diffusion with Forward Models) (sec. 5) and for sparse-view 3D reconstruction (sec. 6).

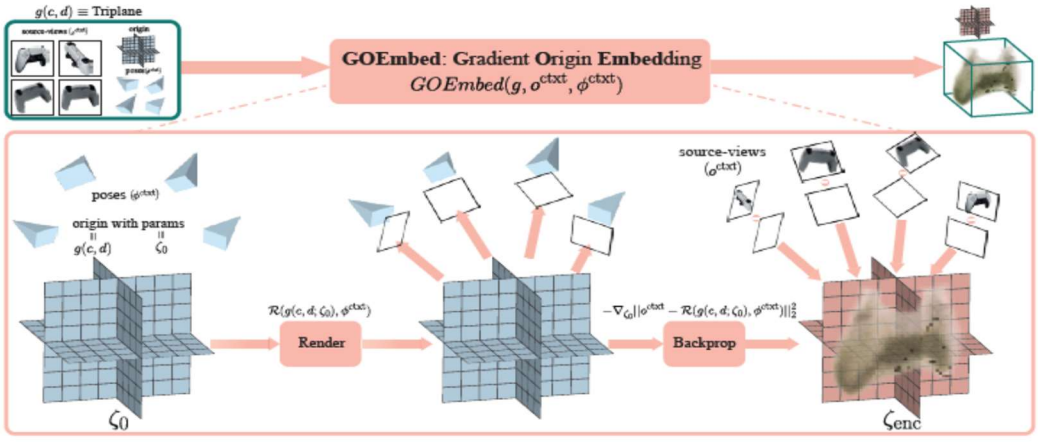
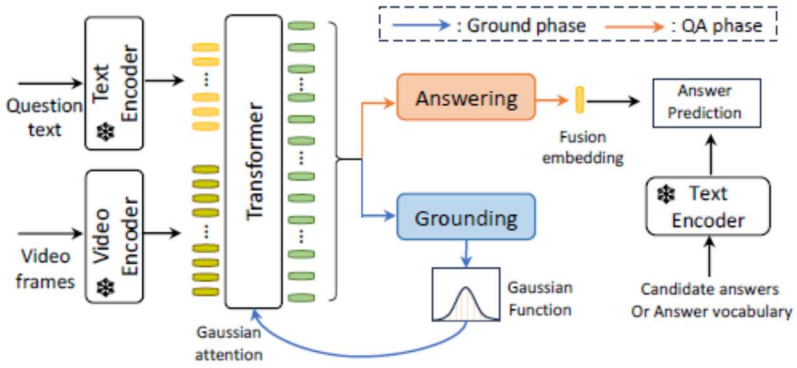
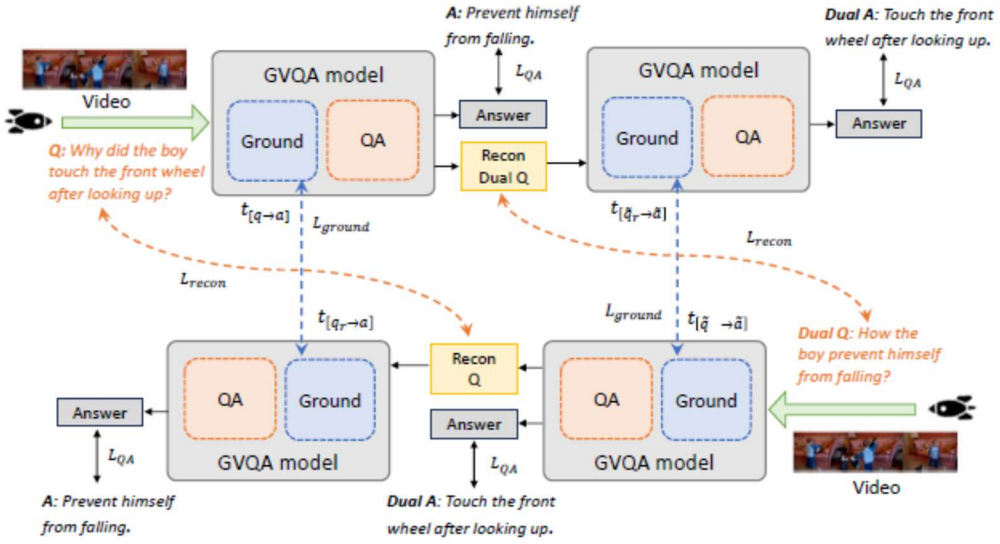


Fig. 2: GOEmbed illustration. We demonstrate the mechanism here using the Tri-plane representation for $g(c, d)$, but note that this can be applied to other representations as well. The GOEmbed mechanism (eq. 1) consists of two steps. First we render the origin ζ_0 from the context-poses ϕ^{ctxt}; then we compute the gradient of the MSE between the renders and the source-views o^{ctxt} wrt. the origin ζ_0 which gives us the GOEmbed encoding ζ_{enc}.

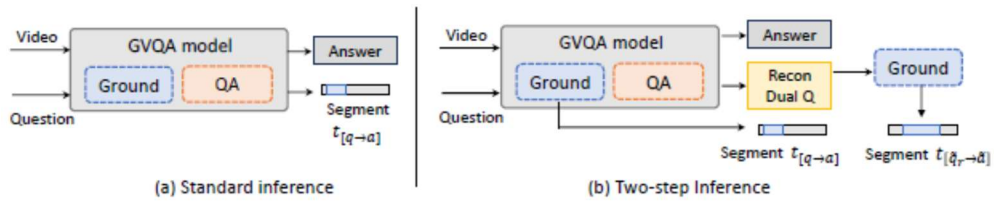
Transformer Net	2025-112	



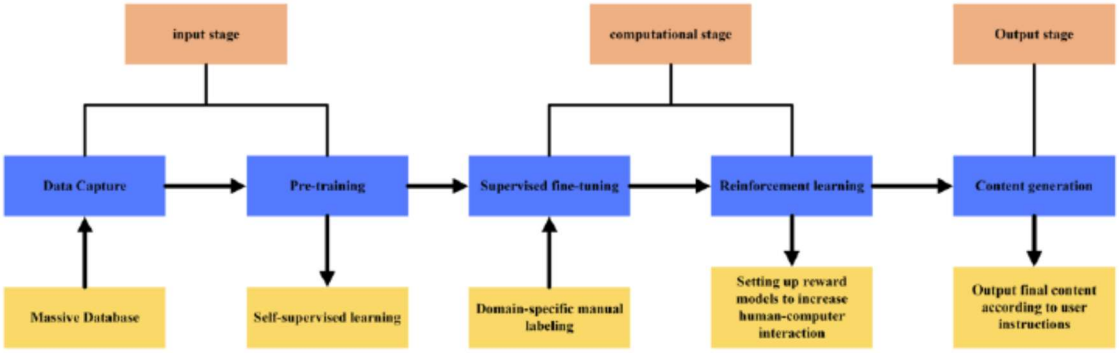
Architecture. Transformer-based VGQA model.



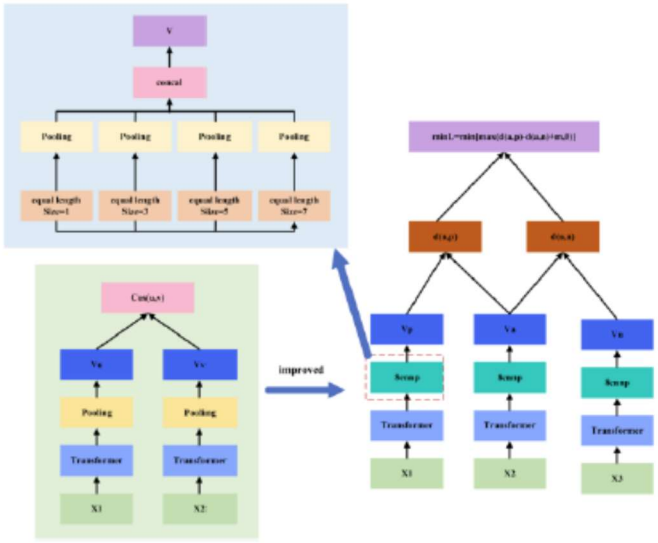
Training pipeline of bi-directional reasoning framework



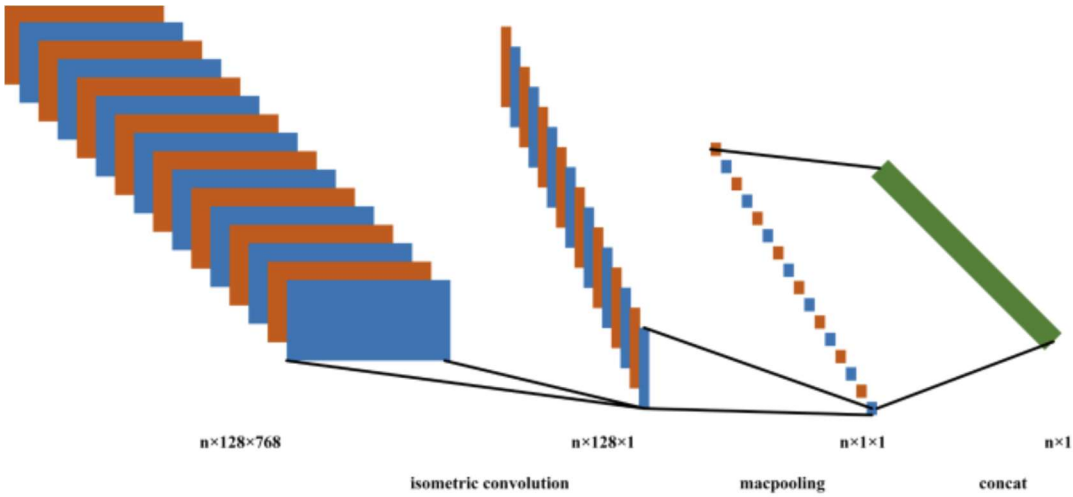
Framework. inference phase



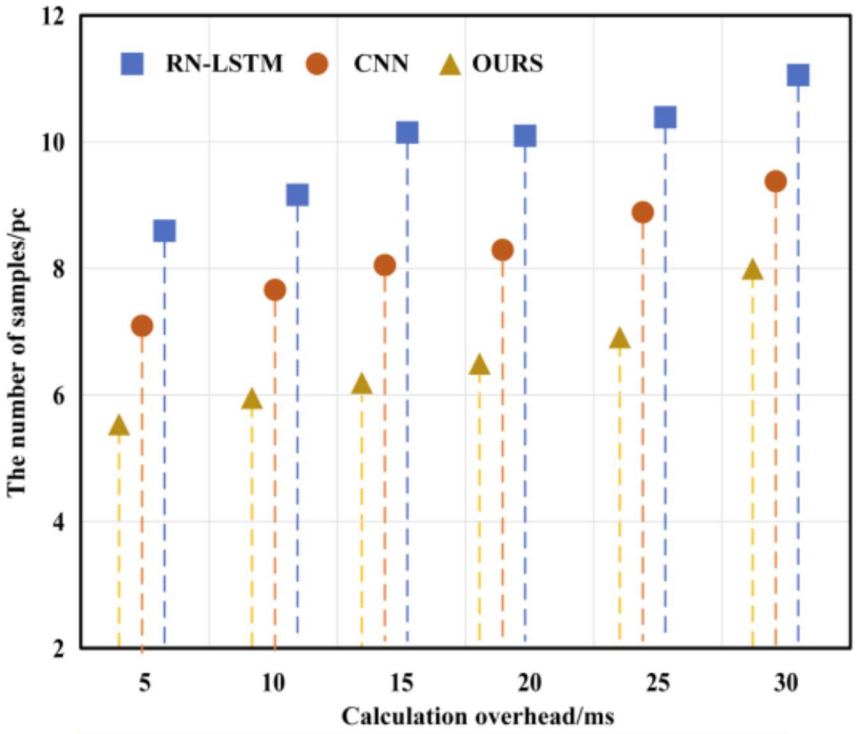
Schematic diagram of operational mechanism of generative artificial intelligence technology based on reinforcement learning



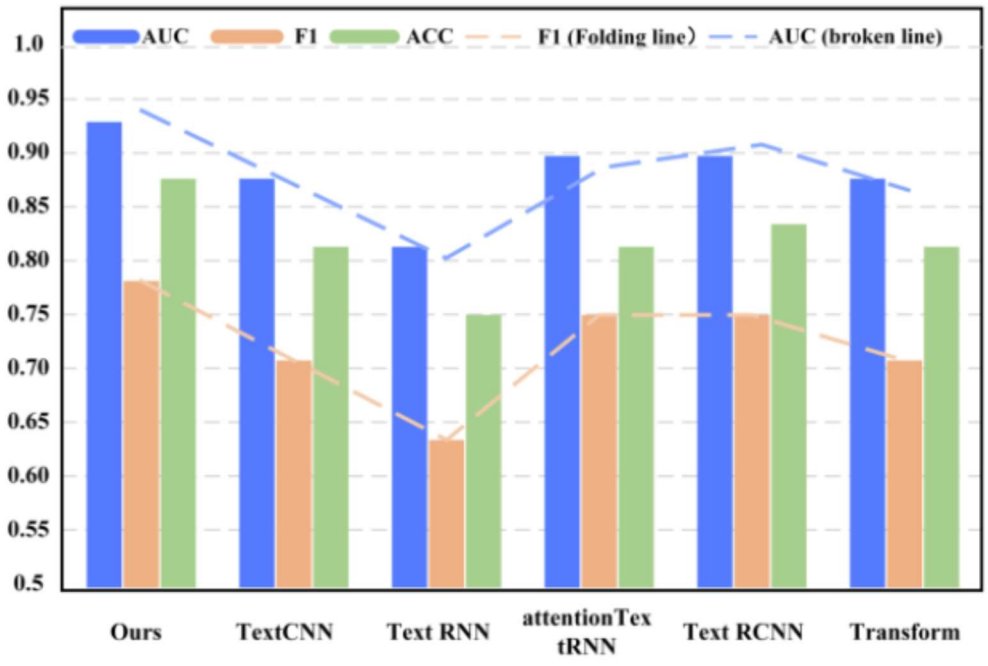
Network structure diagram. Triplet convolutional twin transformer



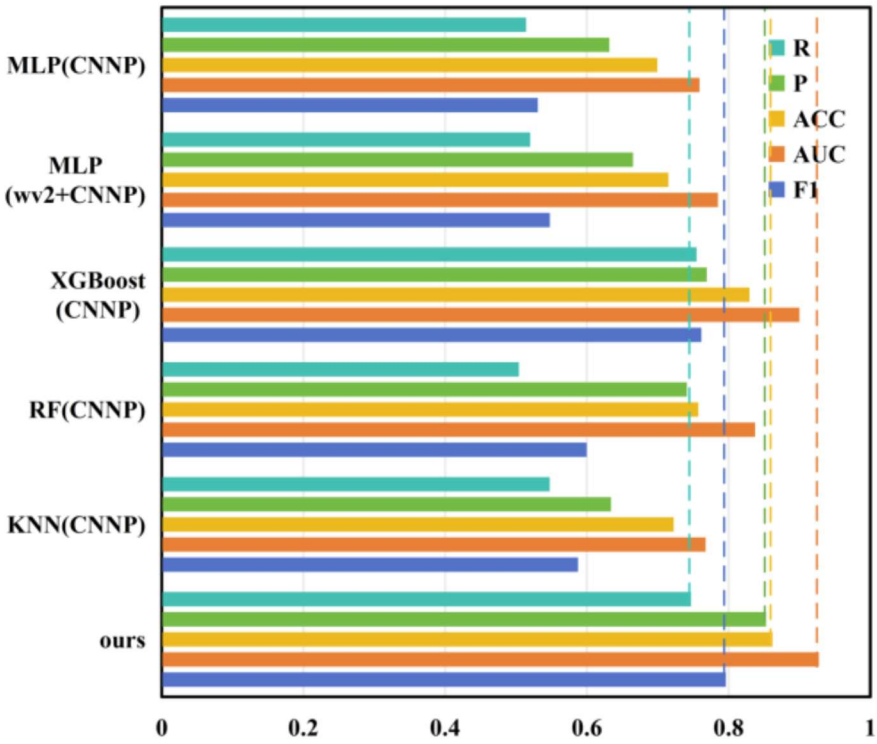
Triplet convolution process



Three kinds of methods to calculate overhead test results

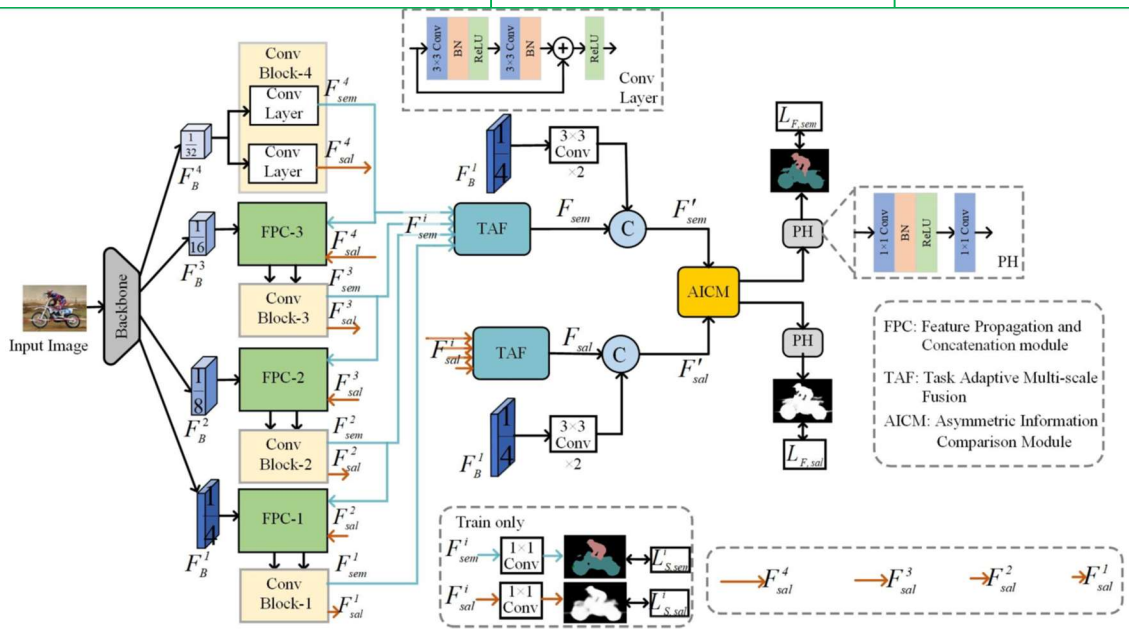


Performance comparison results of different models



Comparison results of structural performance of different models

Transformer Net | **2025-114**



Framework. HirMTL

- ✓ Highlighting its hierarchical multi-task learning strategy for dense scene prediction tasks.
- ✓ Integration of FPC, TAF and AICM modules,

- which together facilitate a sophisticated feature fusion and interchange process, tailored to enhance the network's performance across multiple tasks.

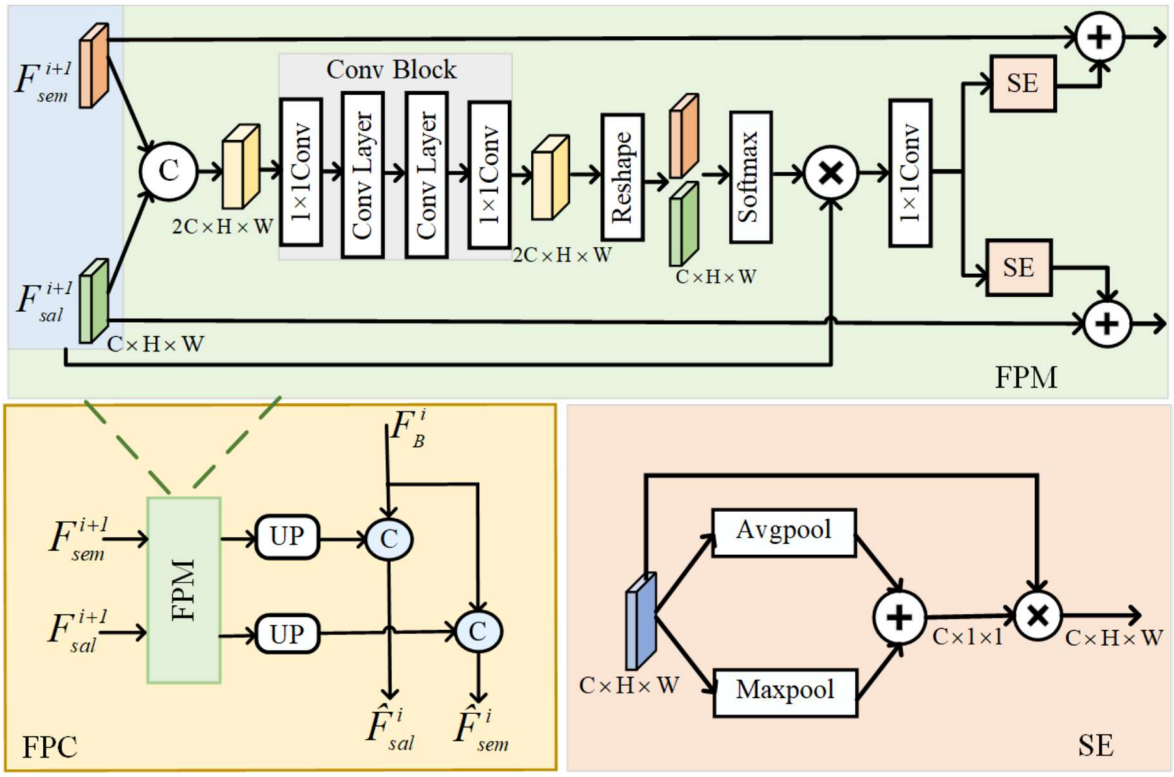
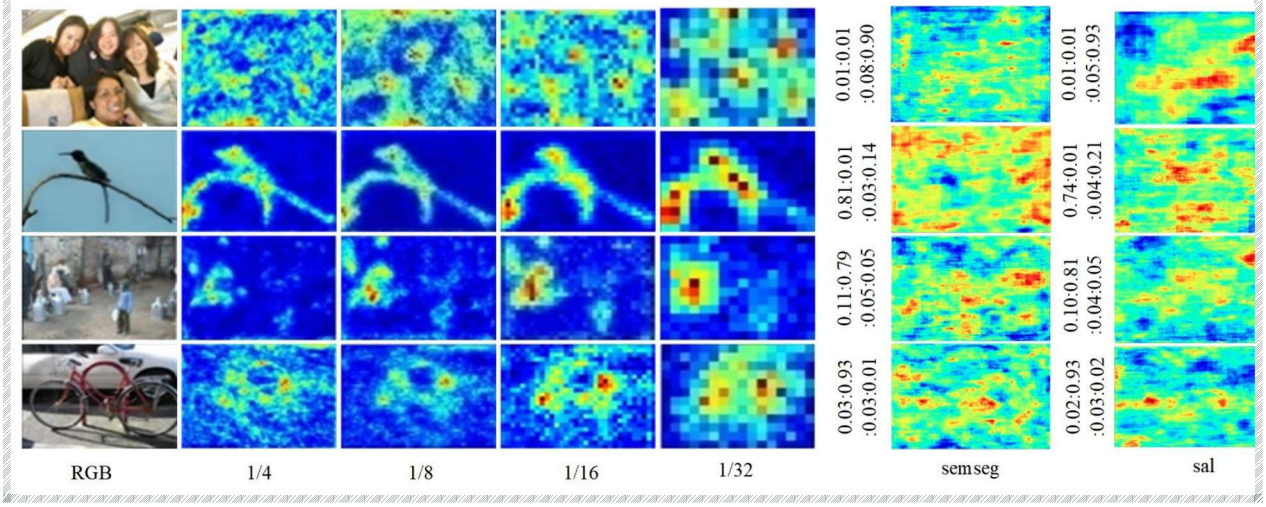


Fig. 2: Schematic of the FPC-i module, where $i \in \{1,2,3\}$, highlighting its role in the hierarchical multi-task learning architecture. The FPC-i module is integral for promoting single-scale level interactions by propagating and enhancing features from higher-level semantic and saliency detection tasks into lower-level but larger resolution representations, thereby establishing a foundational layer for subsequent multi-scale and task-level feature fusion within the HirMTL framework.

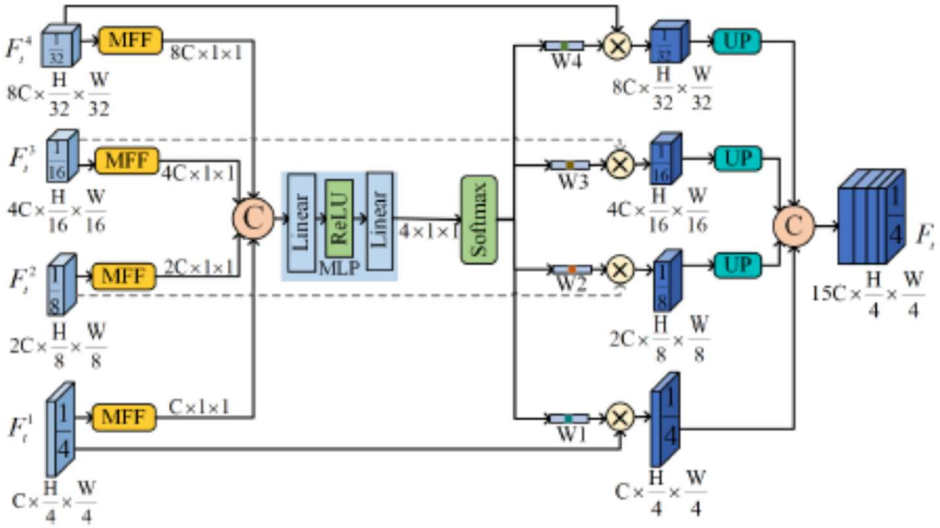


Visual illustration of the TAF module effectiveness

+ Demonstrating its role in adaptively fusing features across multiple scales.

- Leftmost column: input images
- Columns two to five: four scales of features
- Final two columns: taf's sophisticated multi-scale fusion results,
  - showcasing the heatmaps of fusion features for semantic segmentation (semseg) and salient object detection (sal),

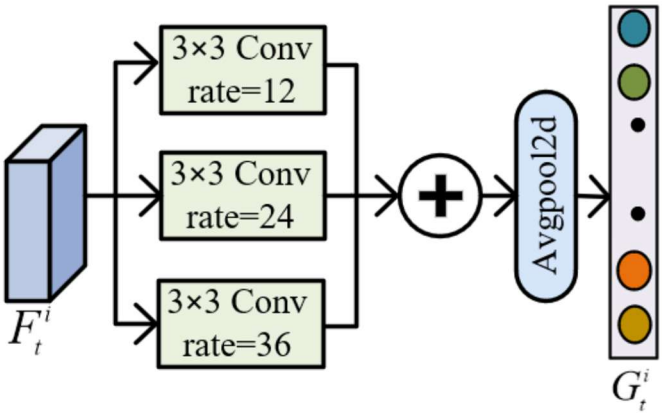
✓ The fusion ratios across scales are shown on the left side of each heatmap,
  - showcasing the module's adeptness in customizing feature fusion for different task requirements: focusing more on the boundaries of objects for semantic segmentation, while more on the centers of objects for salient object detection



Detailed depiction of the TAF module, highlighting adaptive routing mechanism across different feature scales tailored for the "t" task

✓ This illustrates the module's capability to intelligently weigh
✓ and combine multi-scale features, ensuring a robust
✓ and nuanced feature integration for improved performance

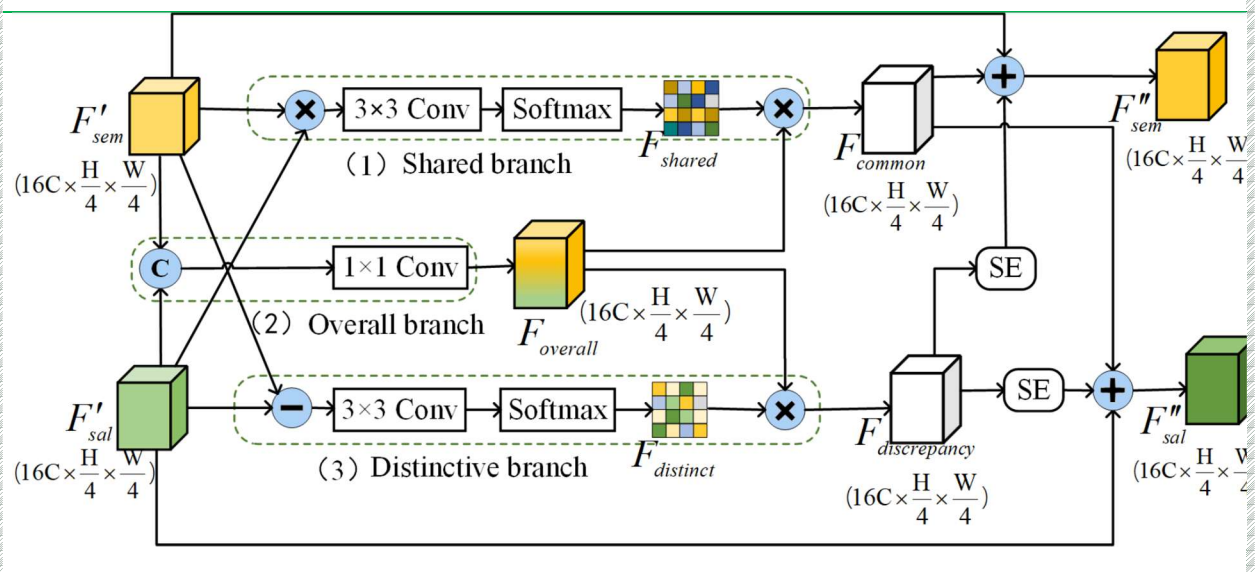
✓



✓

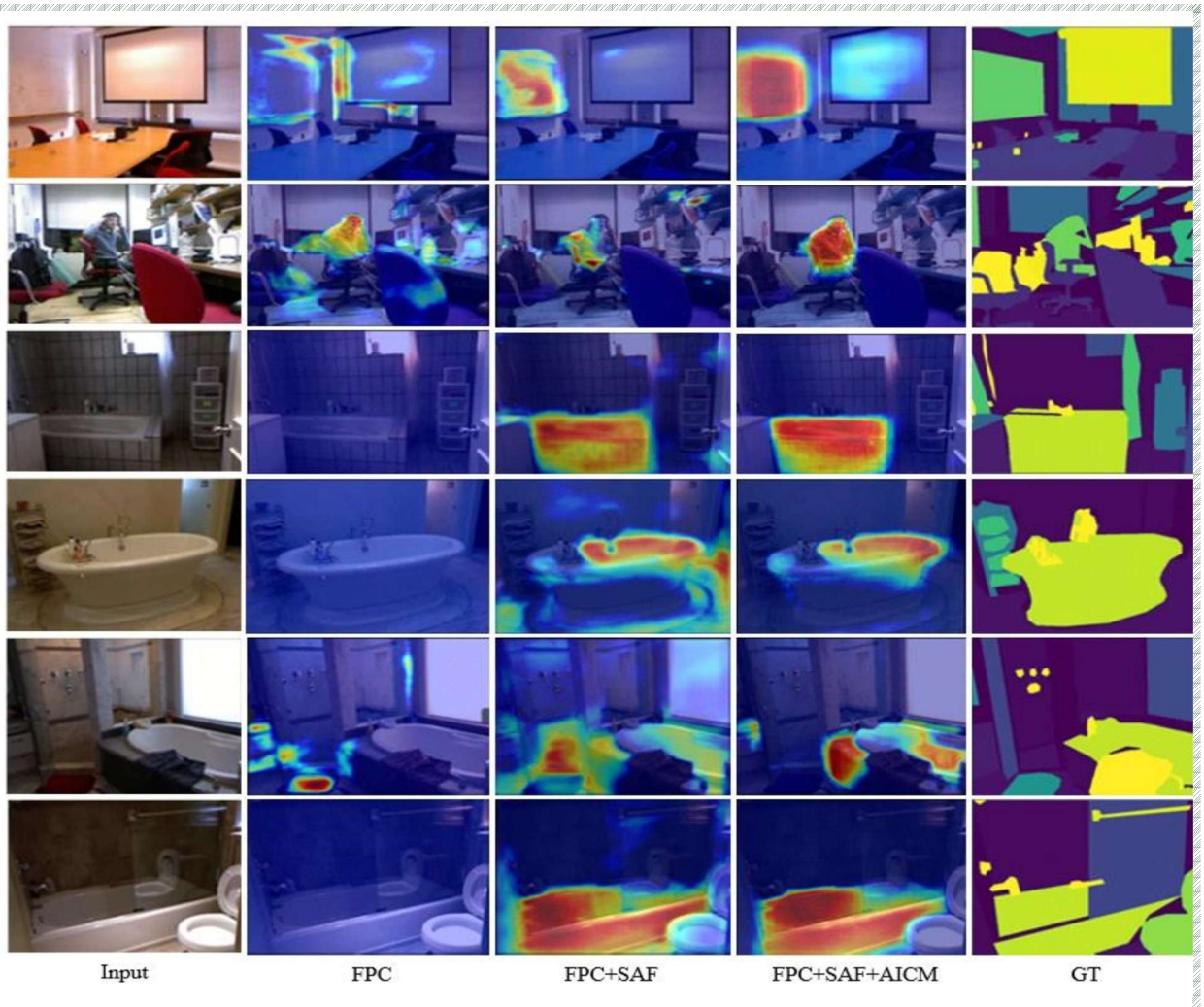
✓ **Schematic representation of the Multi-Receptive Field Feature (MFF) module**

✓ **highlighting its role in expanding the receptive field of features to capture global contextual information at each scale**



Specific implementation details of AICM

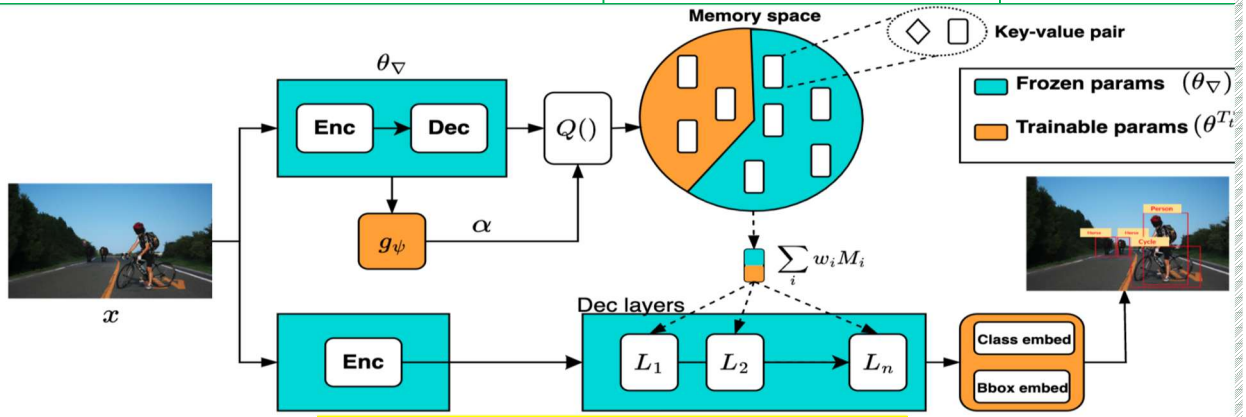
✓ It models the shared and distinctive characteristics between two tasks, alleviating the negative transfer phenomenon in the network.



Heat map visualizations for one class

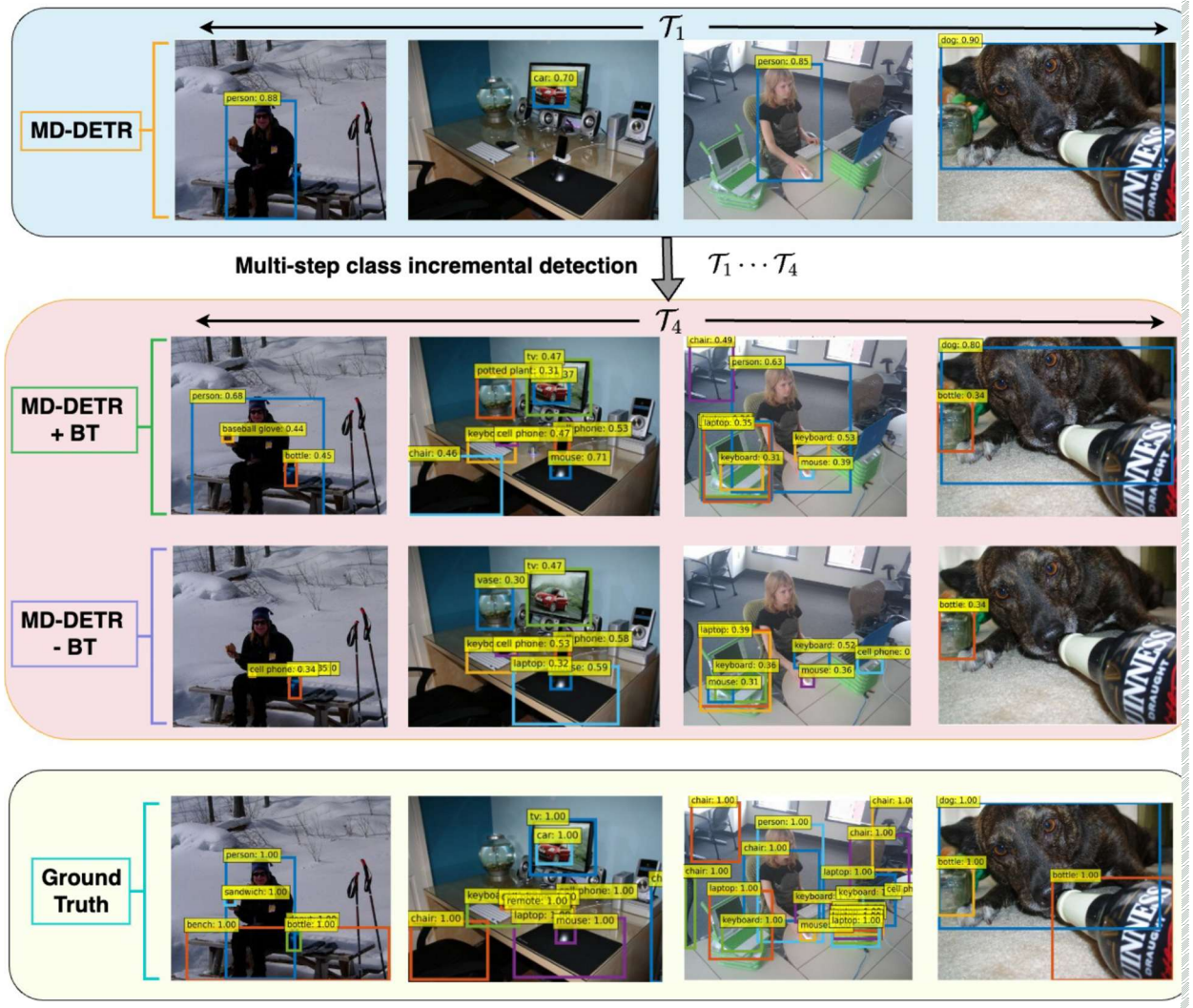
- ✓ Reflecting the incremental impact of the FPC, TAF, and AICM modules on the semantic segmentation predictions of HirMTL on NYUD-v2
- ✓ The gradation from left to right corresponds to the successive addition of modules, vividly illustrating the refinement in predictions.

Transformer Net | **2025-115**



Architecture. MD-DETR at a given time-step t.

- ✓ Given an input image x,
- ✓ query function Q(x, $\theta \nabla$, α) proposed to retrieve relevant memory units as a linear combination.
- ✓ The obtained information from the memory is utilized by the decoder across various decoding layers.
- ✓ The majority of the architecture remains frozen, encompassing the encoder and decoder; the trainable modules consist of memory units M, class embedding, bounding box embedding, and ranking function g ϕ .



Progression of MD-DETR performance from T1 → T4 when trained in a multi-step class incremental setting on MS-COCO

- ✓ To illustrate the effectiveness of MD-DETR in addressing background relegation of previously encountered classes, a comparison is presented between two architecture designs:
- ✓ with (MD-DETR + BT) and without (MD-DETR - BT) background thresholding
- ✓ In the images shown in
  - First column (1st image for each block), the class {person} is relegated by MD-DETR - BT;

- For the second column, the category {car} is relegated,
- Third column displays relegation of classes {person, chair)
- Fourth column shows the relegation of the category {dog}.
- Ground truth block shows images with all annotations across all four tasks {T1 · · · T4 }

Transformer Net	2025-116	

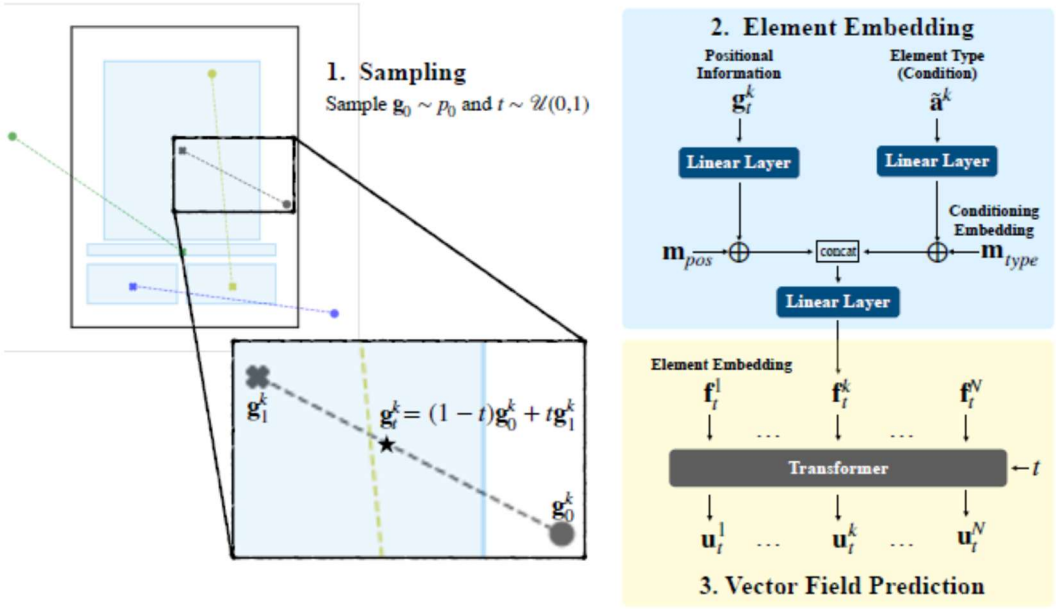


Fig. 2: Overview of the training procedure of LayoutFlow for the type-conditioned scenario. First, we sample an initial layout from a base distribution and a time t. Then, an intermediate sample $\mathbf{g}_t$ is calculated by linearly interpolating between the initial sample and the ground truth layout. Each intermediate element is embedded jointly with the given element condition $\tilde{\mathbf{a}}^k$. Lastly, the Transformer architecture takes all the element embeddings to predict a vector field.

Transformer Net	2025-117	

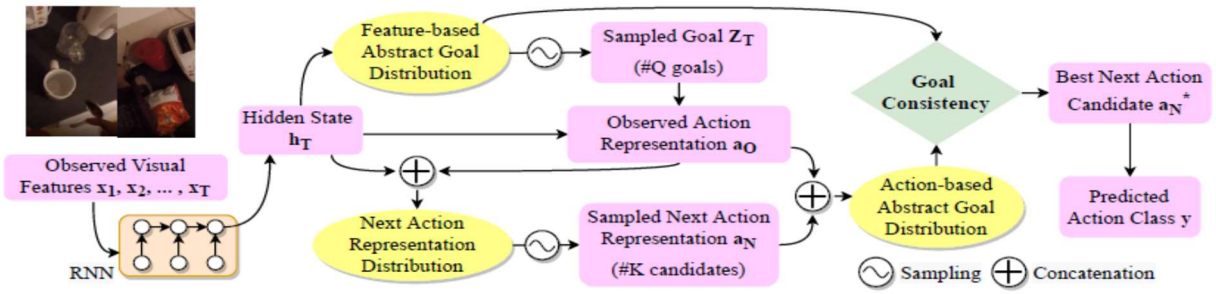
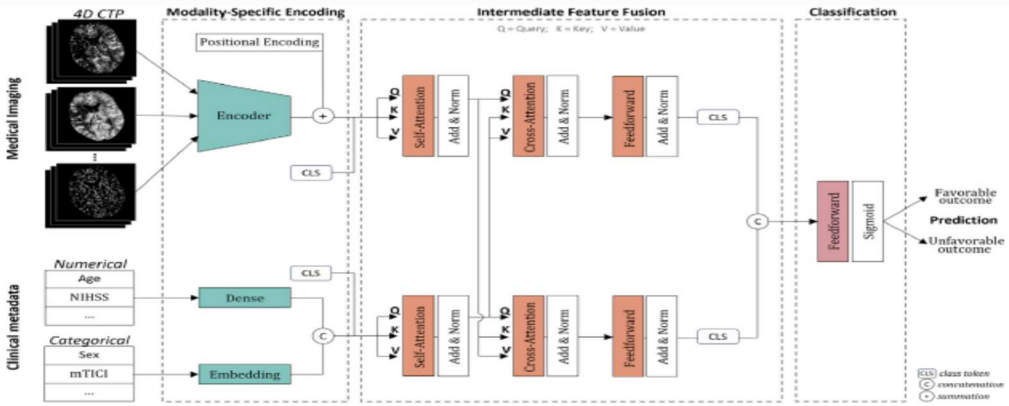


Illustration of model design for abstract goal-based action anticipation.

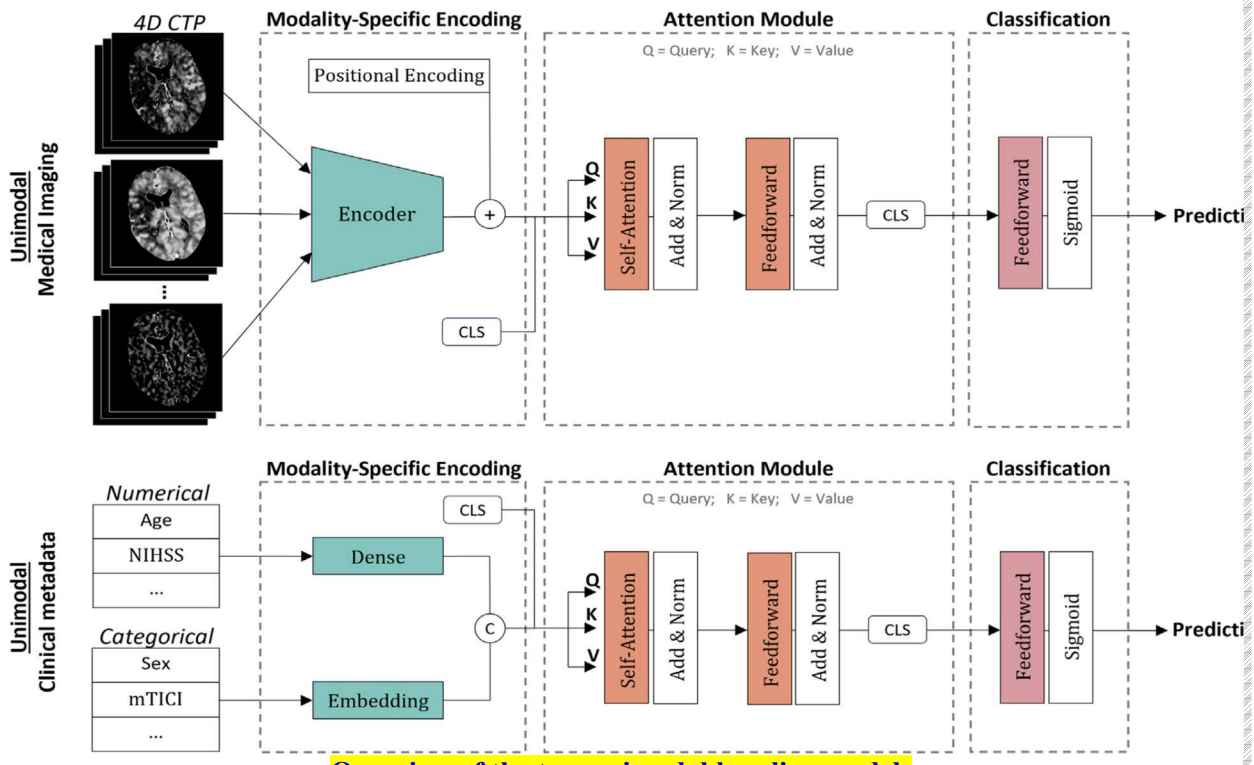
- ✓ Yellow ellipses represent distributions and
- ✓ Pink boxes represent various variables of the model

Transformer Net	2025-118	



Overview of the proposed model architecture

- To predict functional outcomes as favorable (mRS90 of 0-2) or unfavorable (mRS90 of 3-6) using pre-treatment 4D CTP (top) and clinical metadata (bottom)
- It consists of three components:
  - (1) Modality-specific encoders to extract features from each modality
  - (2) An intermediate feature fusion module using self- and cross-attention to integrate and analyze relationships within and between modalities.
  - (3) A classification module generating the final outcome scores



Overview of the two unimodal baseline models

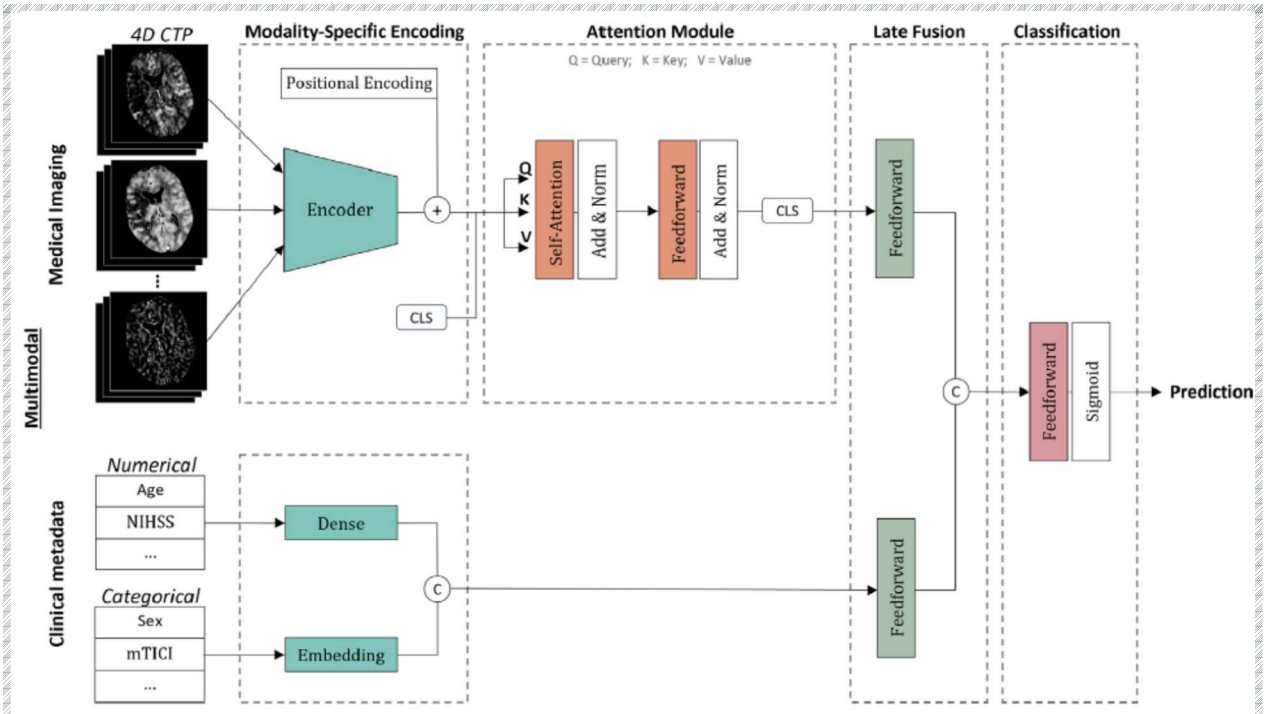


Fig. A.2. Overview of the late fusion (Setup 1) baseline model

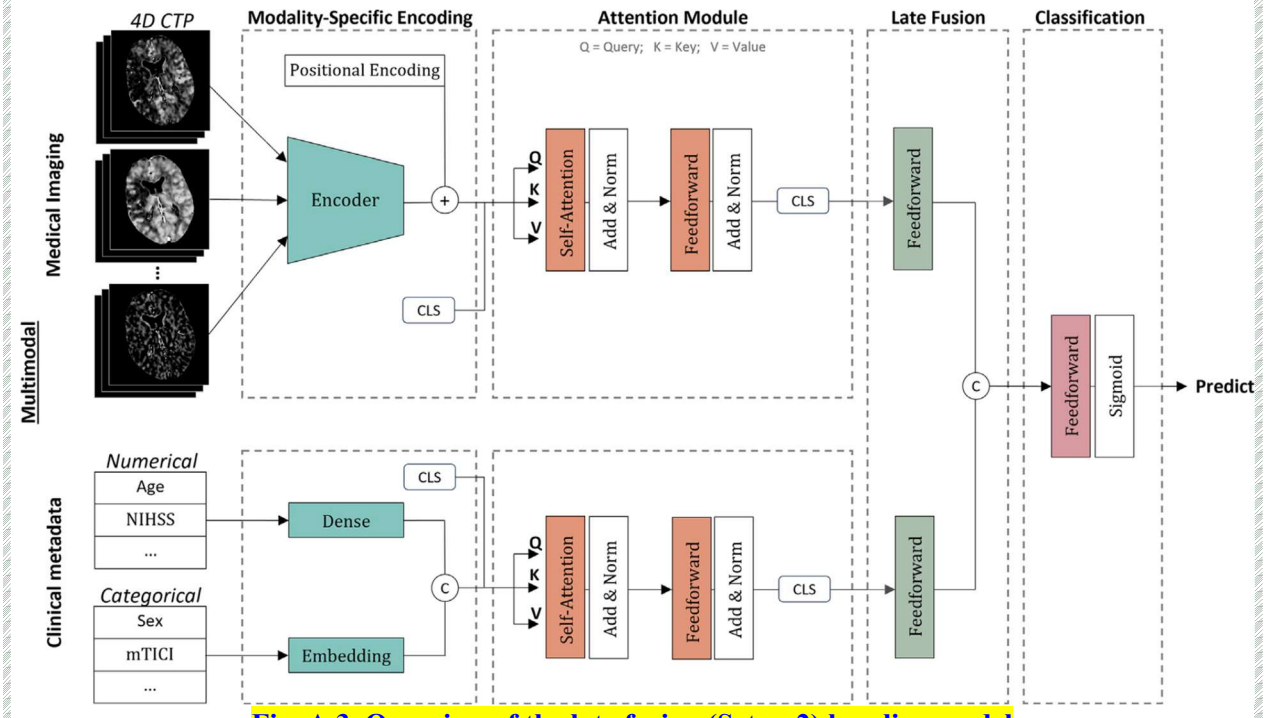


Fig. A.3. Overview of the late fusion (Setup 2) baseline model

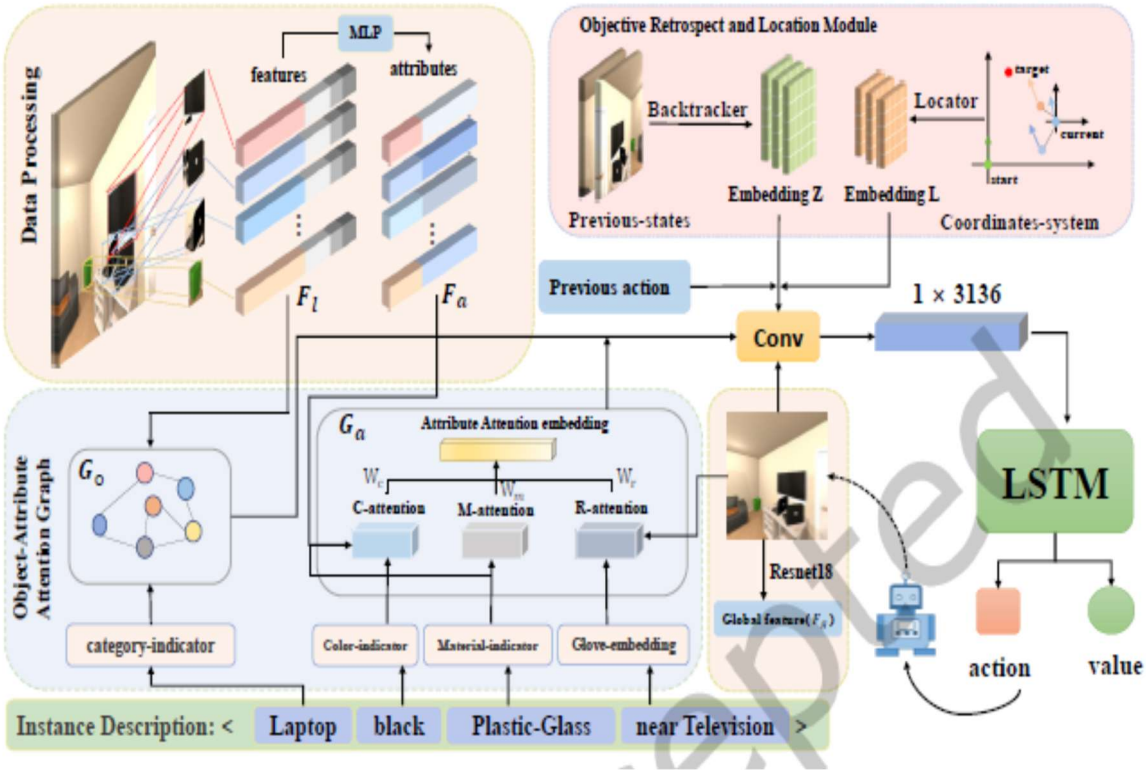
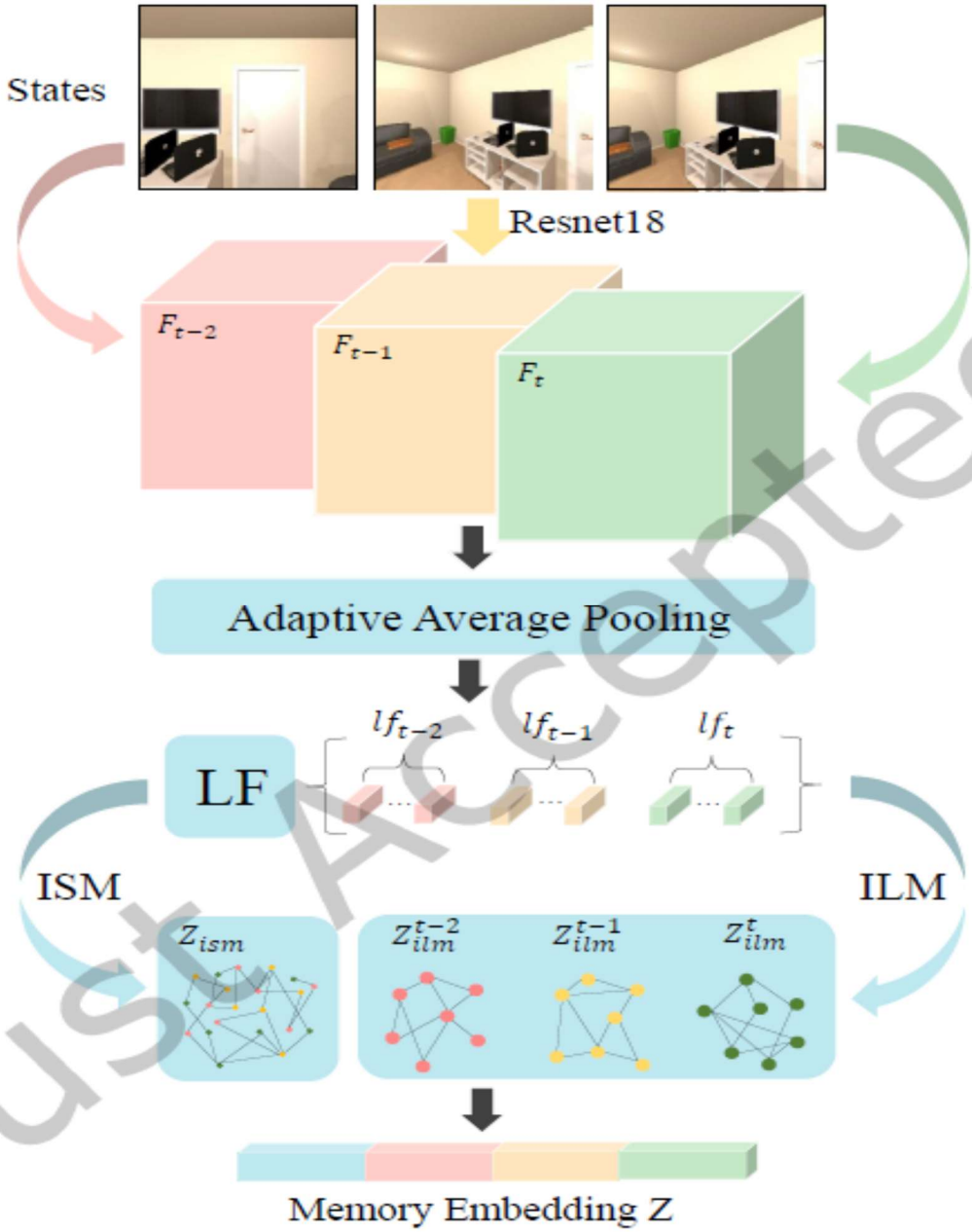


Fig. 2. The framework of the proposed method. The model contains four parts: (i) Data Processing; (ii) Object-Attribute Attention graph (Section 3.3); (iii) Objective Retrospect and Location module (Section 3.4); and (iv) Navigation Strategy (Section 3.5). The global features F_g, the Object-Attribute Attention graph G, the memory embedding Z, and location embedding L are used as inputs to navigation strategy which return an action executable a_t and a value v_t.



Overview. Objective Retrospect Module

- ✓ **ISM represents Inter-frame Sequence Memory retrospect**
- ✓ **ILM represents Intra-frame Layout Memory retrospect**

Fig. 4. Two examples of ablation experimental visualization for Instance-Localization (IN). The descriptions of the targets in the first and second rows are $< Pillow|gray|Fabric|on|Floor >$ and $< CellPhone|red-black|Metal-Glass|on|SideTable >$, respectively. Yellow boxes indicate targets, red lines indicates a navigation failure, and green lines indicates a successful navigation.

Transformer Net	2025-120	

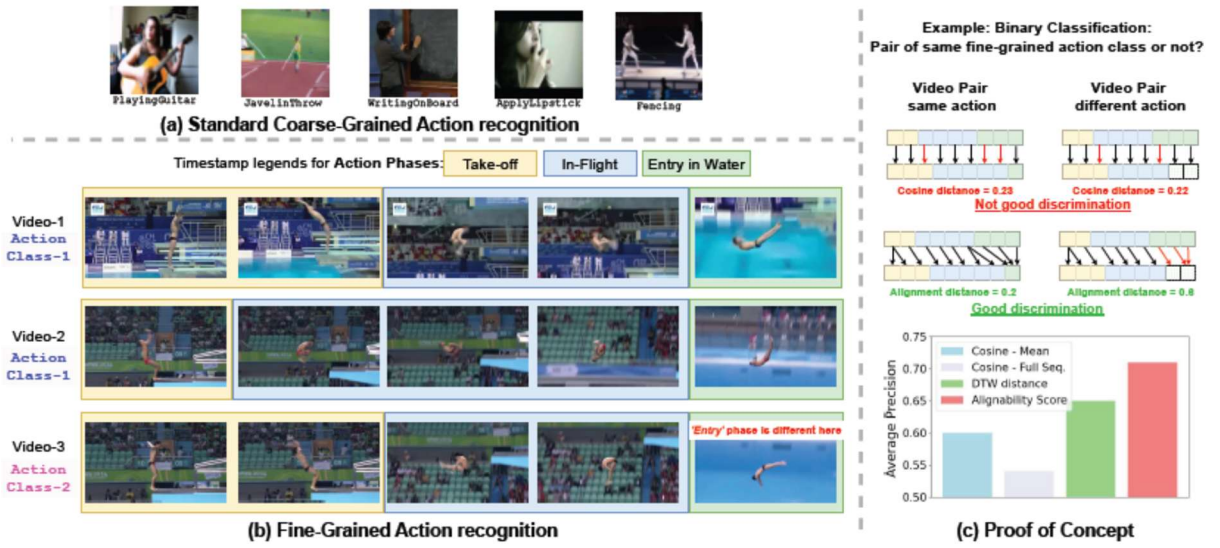


Fig. 1: (a) Sample actions from standard coarse-grained action recognition dataset (UCF101) (b) Sample actions from fine-grained action recognition dataset (Diving48) (c) For proof-of-concept, we choose a binary classification problem of fine-grained actions

✓ **Model has to predict whether the pair of videos belong to the same class or not.**

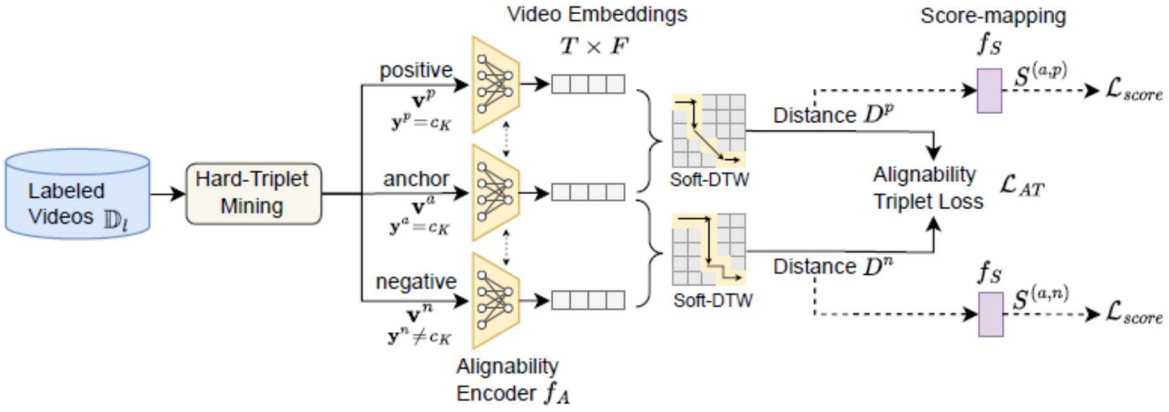


Fig. 2: Alignability-Verification based Metric Learning is proposed to is proposed to decide how well two video instances are alignable and produce an 'alignability score' for effective learning from a limited labeled set $\mathbb{D}_l$. Our approach employs a triplet loss ($\mathcal{L}_{AT}$), considering videos from identical action classes as positive and those from different classes as negative. We selectively mine hard-negatives from the sampled minibatch based on alignment distance, presenting a challenging learning task for the model f_A. Additionally, we incorporate a matching loss $\mathcal{L}_{score}$ to quantify the alignment between videos, serving as a verification task to determine whether a video pair belongs to the same class (i.e. alignable or target label = 1) or different classes (i.e. non-alignable or target label = 0). Further details are provided in Sec. 3.1.

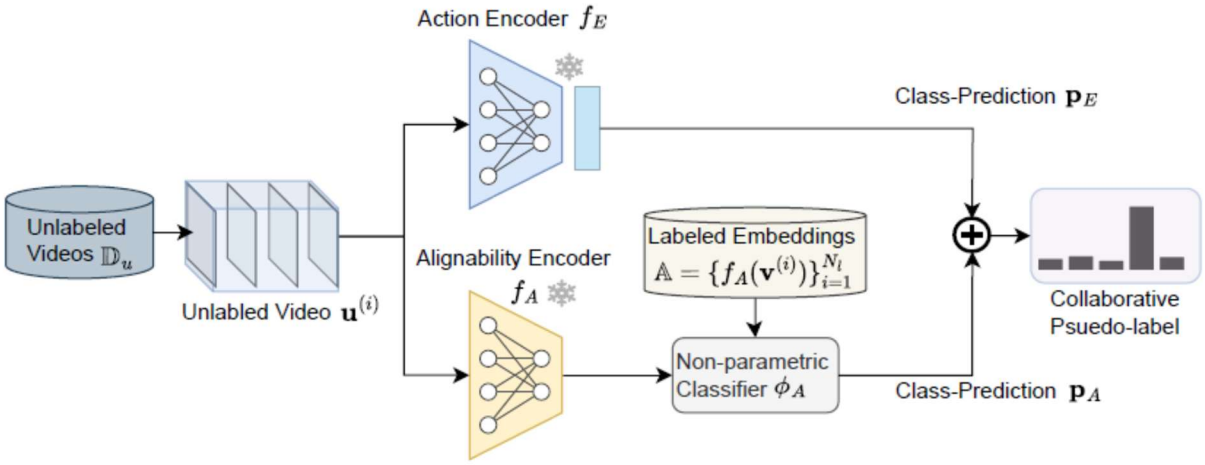
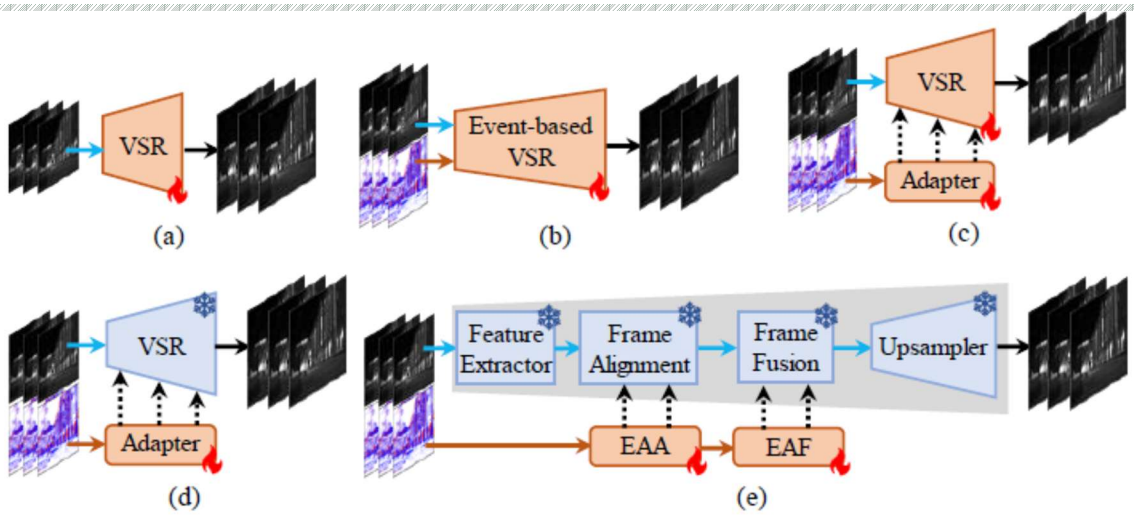


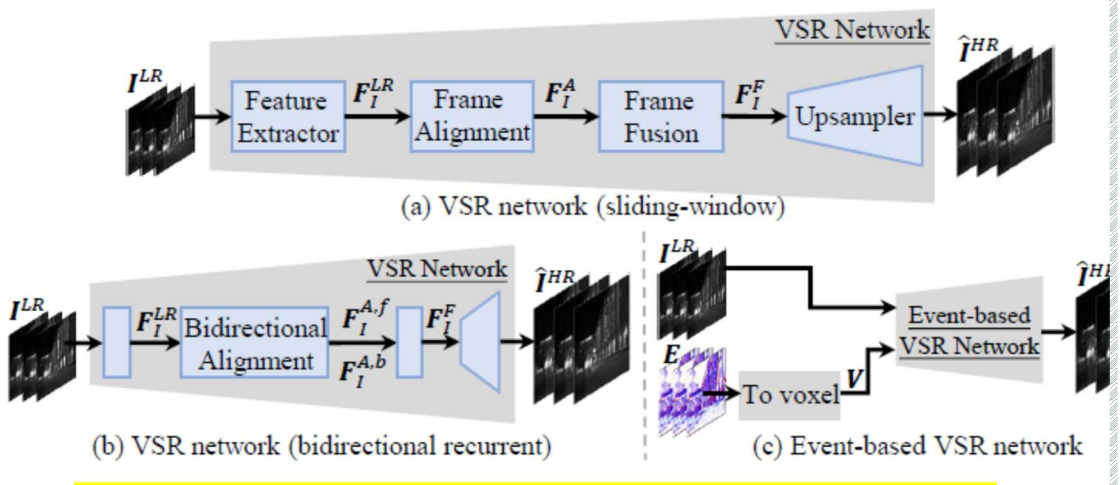
Fig. 3: Collaborative Pseudo-labeling: The unlabeled instance $\mathbf{u}^{(i)}$ undergoes processing by both video encoders (f_E and f_A). For the Action Encoder f_E, its prediction ($\mathbf{p}_E$) is derived via its classification head. For the Alignability Encoder f_A, the embedding of $\mathbf{u}^{(i)}$ computes class-wise alignability scores against a gallery of labeled embeddings $\mathbb{A}$. These scores are then used to generate a class-wise prediction $\mathbf{p}_A$ using the non-parametric classifier ϕ_A. As these predictions stem from distinct supervisory signals—$\mathbf{p}_E$ from video-level and $\mathbf{p}_A$ from alignability-based supervision—they offer complementary insights, resulting in a refined collaborative pseudo-label.



Existing event-based VSR paradigm v.s. proposed Event-Adapted VSR

- Existing VSR methods can generate HR frames from LR observations.
- Existing event-based VSR methods
- Incorporating the event data as a novel modality into an existing VSR method can be approached in two ways: full fine-tuning or retraining from scratch
- (d)-(e) Our proposed event-adapted VSR. By leveraging the advantages of parameter-efficient tuning and insights from well-designed VSR networks, our EATER achieves superior performance with minimal parameter updates.

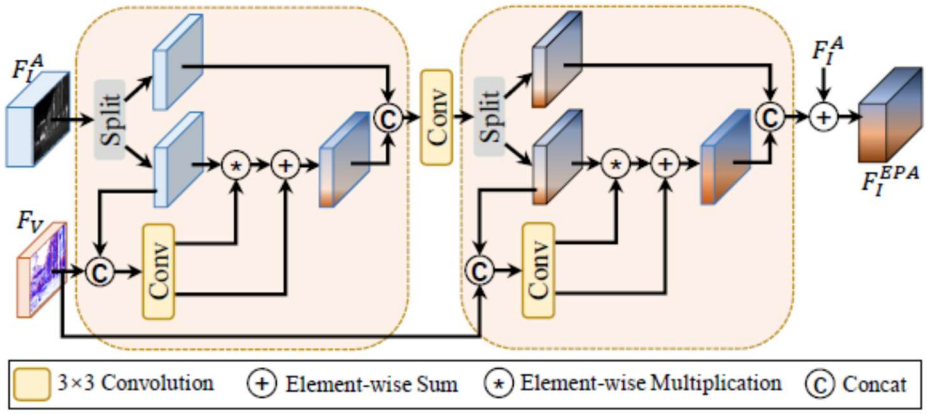
✓



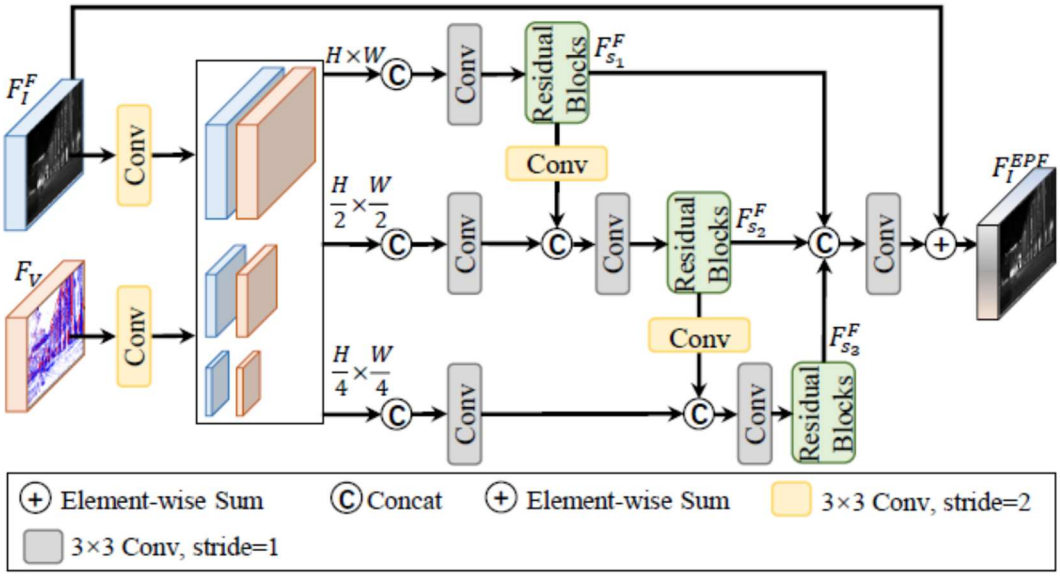
✓

✓ Illustration of typical structure of VSR network and event-based VSR network

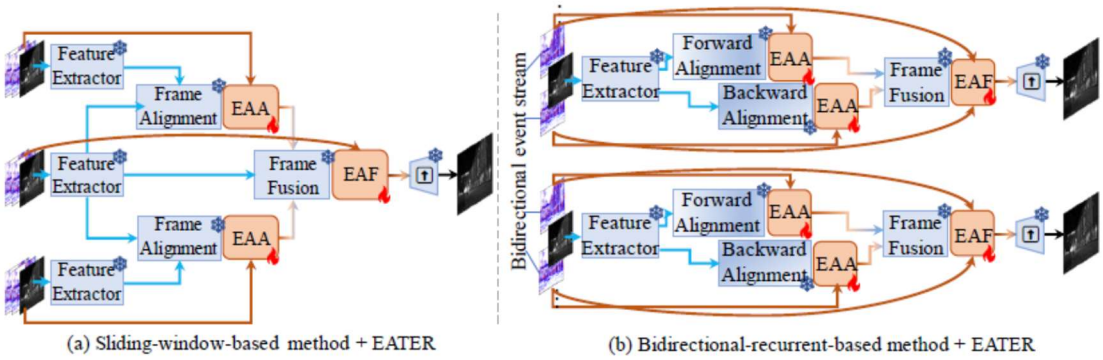
✓ For the sake of simplicity, we remove the bicubic upsampled residual connections and superscripts for the forward and backward event streams



Event-adapted alignment (EAA) unit



Event-adapted fusion (EAF) unit

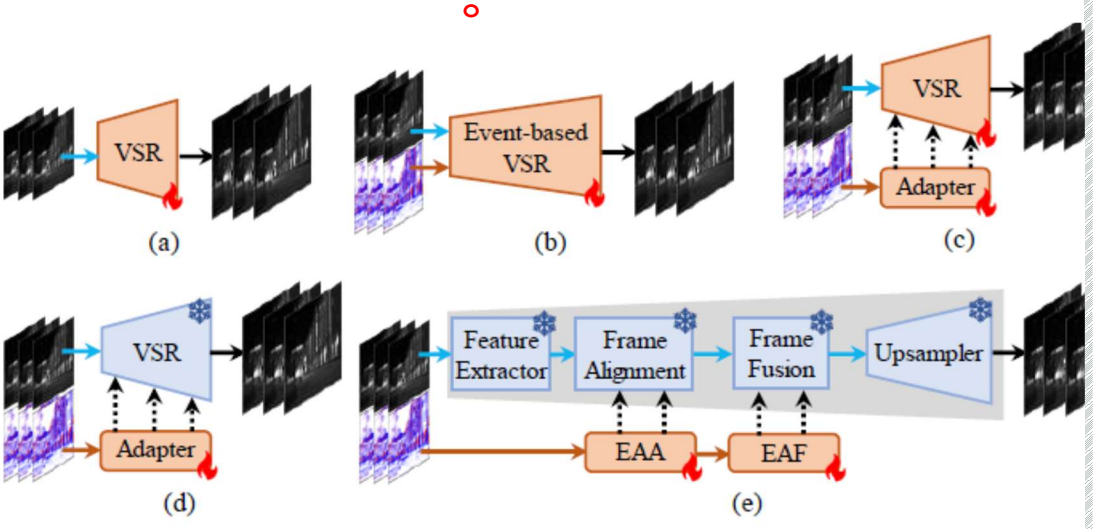


Overview of applying EATER to existing typical EGB-only VSR methods

(a) Apply EAA unit and EAF unit to the sliding-window-based method (e.g., EDVR)

(b) Apply the EAA unit and the EAF unit to the bidirectional recurrent-based method (e.g., BasicVSR

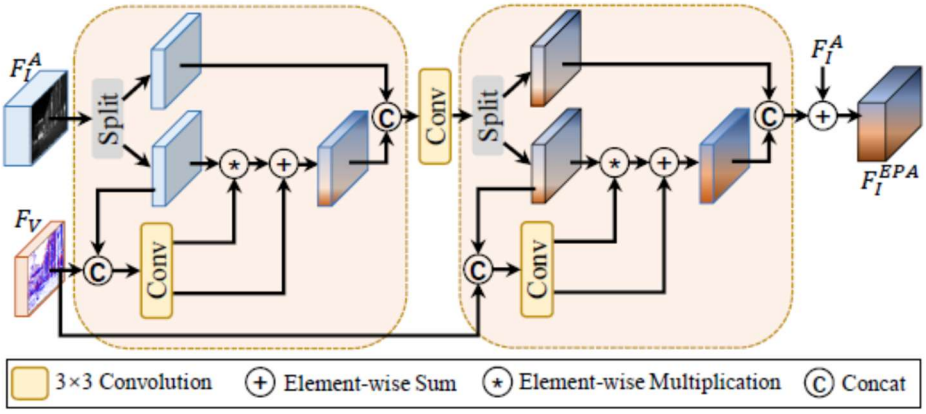
Transformer Net	2025-121	



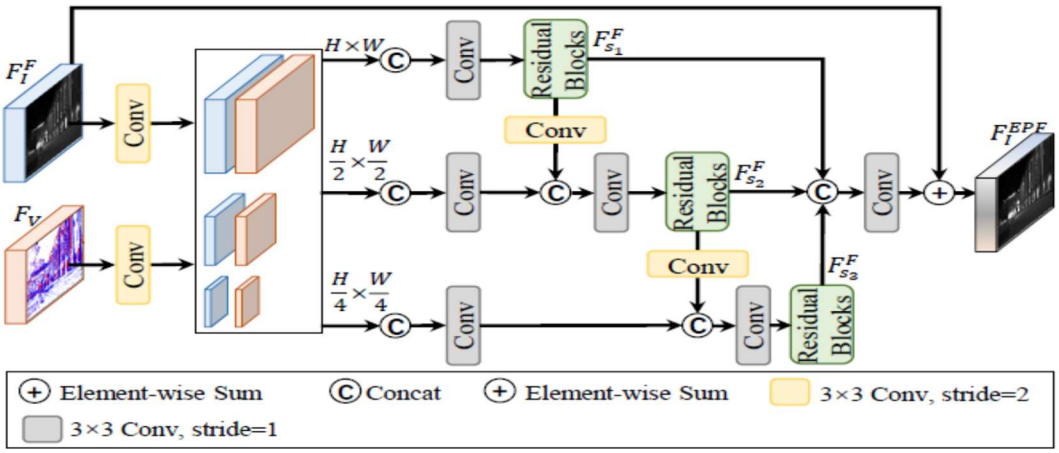
Existing event-based VSR paradigm v.s. proposed Event-Adapted VSR

- (a) Existing VSR methods can generate HR frames from LR observations.
- (b) Existing event-based VSR methods
- (c) Incorporating the event data as a novel modality into an existing VSR method can be approached in two ways: full fine-tuning or retraining from scratch
- (d)-(e) Our proposed event-adapted VSR.

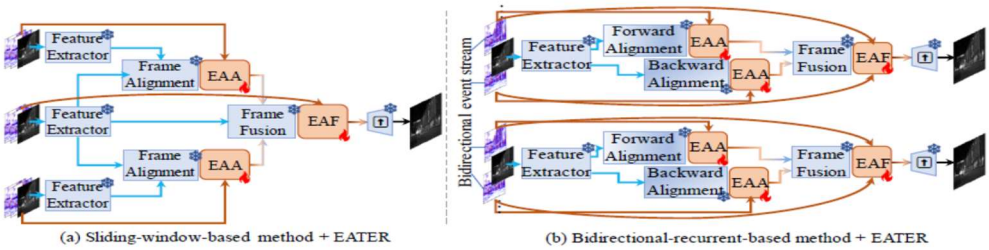
✓ By leveraging the advantages of parameter-efficient tuning and insights from well-designed VSR networks, our EATER achieves superior performance with minimal parameter updates



Event-adapted alignment (EAA) unit



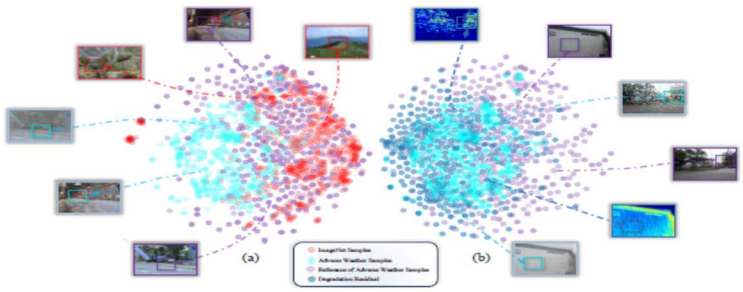
Event-adapted fusion (EAF) unit



Overview of applying proposed EATER to existing typical EGB-only VSR methods

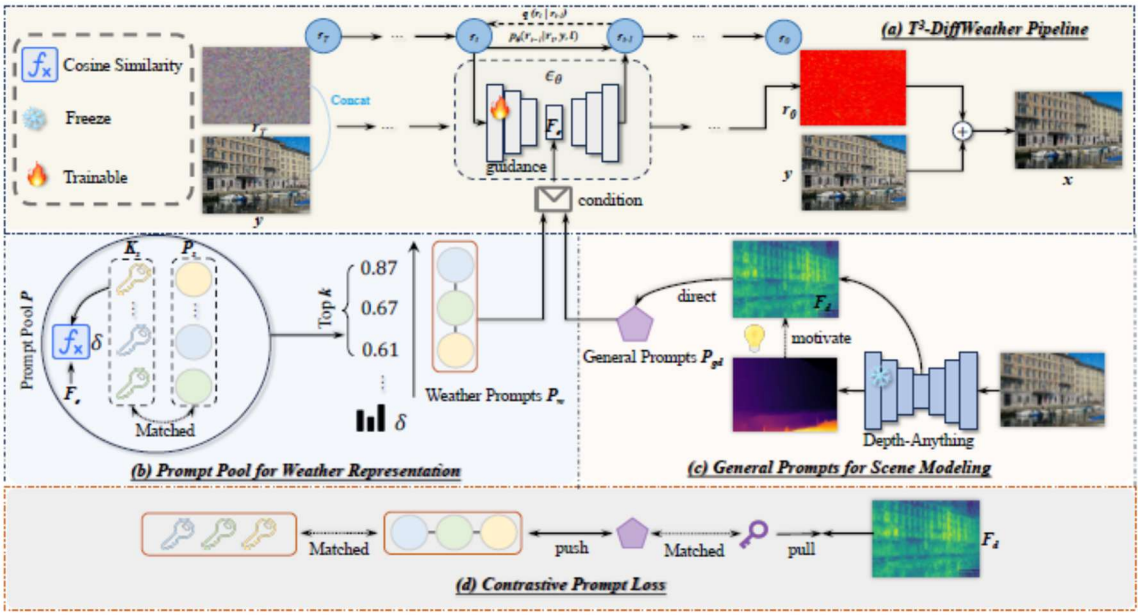
- ✓ (a) We apply the EAA unit and the EAF unit to the sliding-window-based method (e.g., EDVR).
- ✓ (b) We apply the EAA unit and the EAF unit to the bidirectional recurrent-based method (e.g., BasicVSR)

Transformer Net	2025-122	



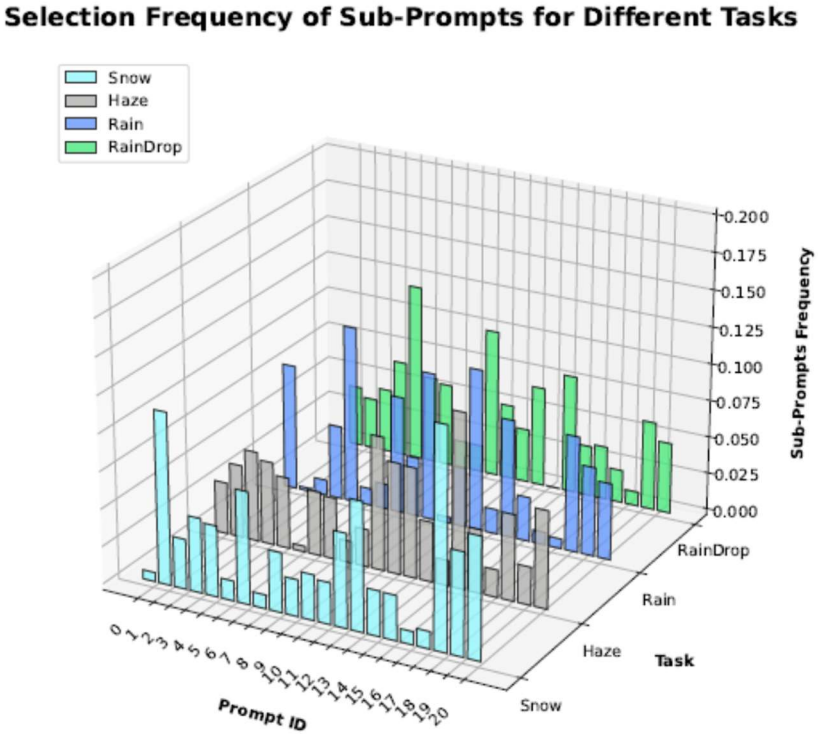
t-SNE visualization of different feature distributions

- (a). Scenes with different contents also have significant commonalities compared to degradations. And there are some differences and commonalities between degradations and degradations
- (b). Degradation residuals can represent degradations to a certain extent and be distinguished from background



Overview of proposed method

(a) showcases our pipeline, which adopts an innovative strategy focused on learning degradation residual and employs the information-rich condition to guide the diffusion process.
(b) illustrates the utilization of our prompt pool, which empowers the network to autonomously select attributes needed to construct adaptive weather-prompts.
(c) depicts the general prompts directed by depth-anything constraint to supply scene information that aids in reconstructing residuals.
(d) shows the contrastive prompt loss, which exerts constraints on prompts driven by two distinct motivations, enhancing their representations



Selection frequency of sub-prompts

- ✓ Some similar selection frequencies reflect the network’ s ability to adaptively exploit common attributes in some similarity between tasks (e.g. rain and raindrop).
- ✓ At the same time, the unique prompt frequencies highlight the flexibility to adapt to the specific characteristics of each weather condition.

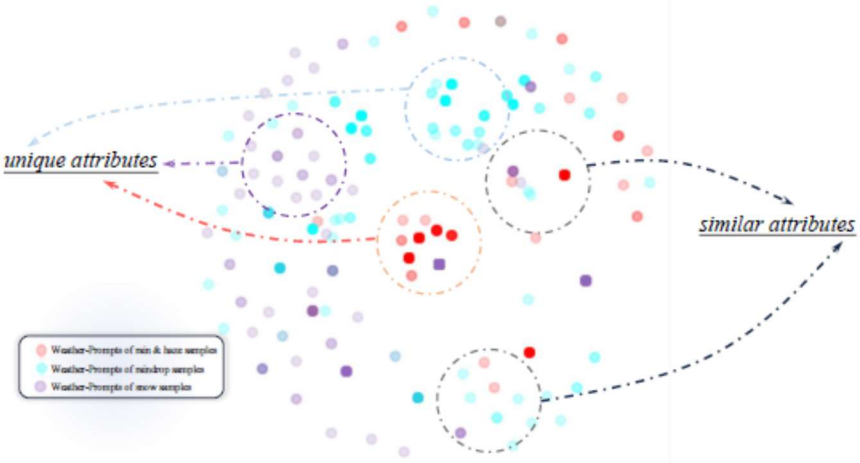
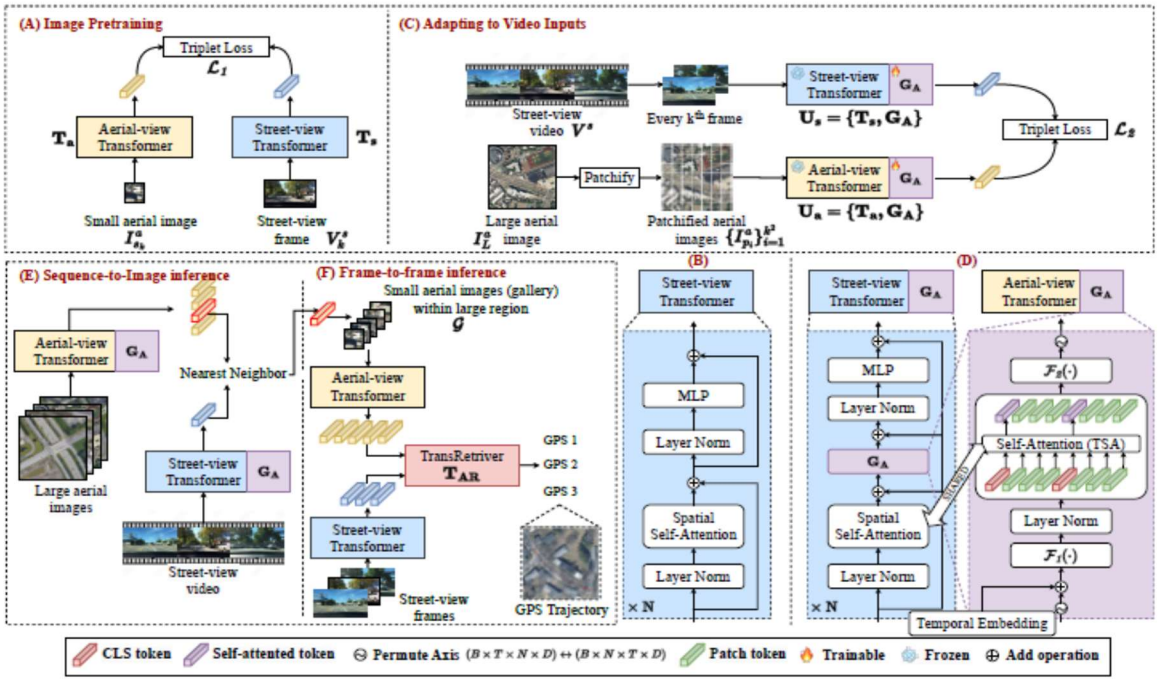


Figure 4. t-SNE visualization of weather-prompts for different weather conditions

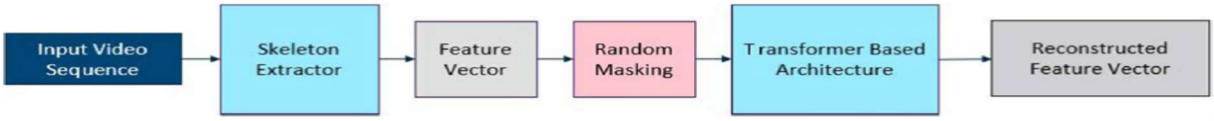
Transformer Net	2025-123	



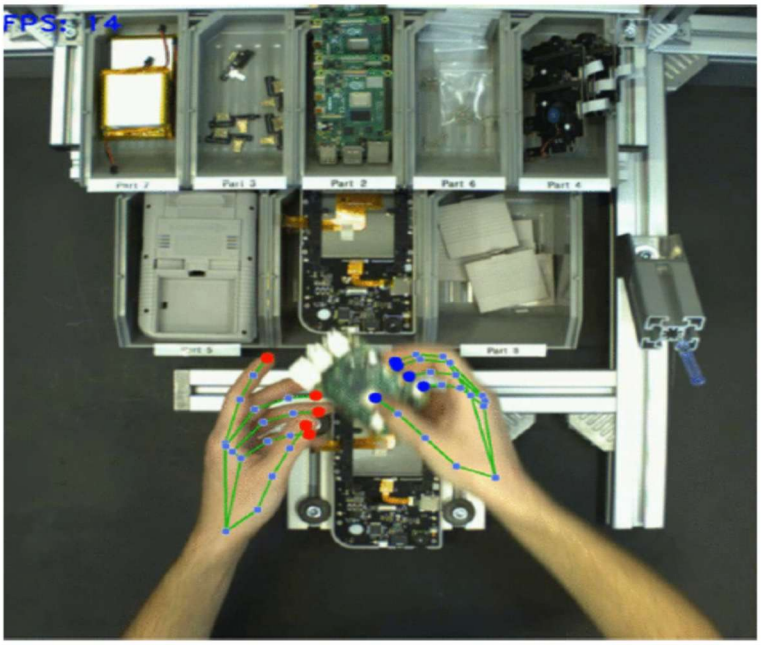
Proposed approach. GAReT
(Cross-view Video Geolocalization with Adapters and Auto-Regressive Transformers)

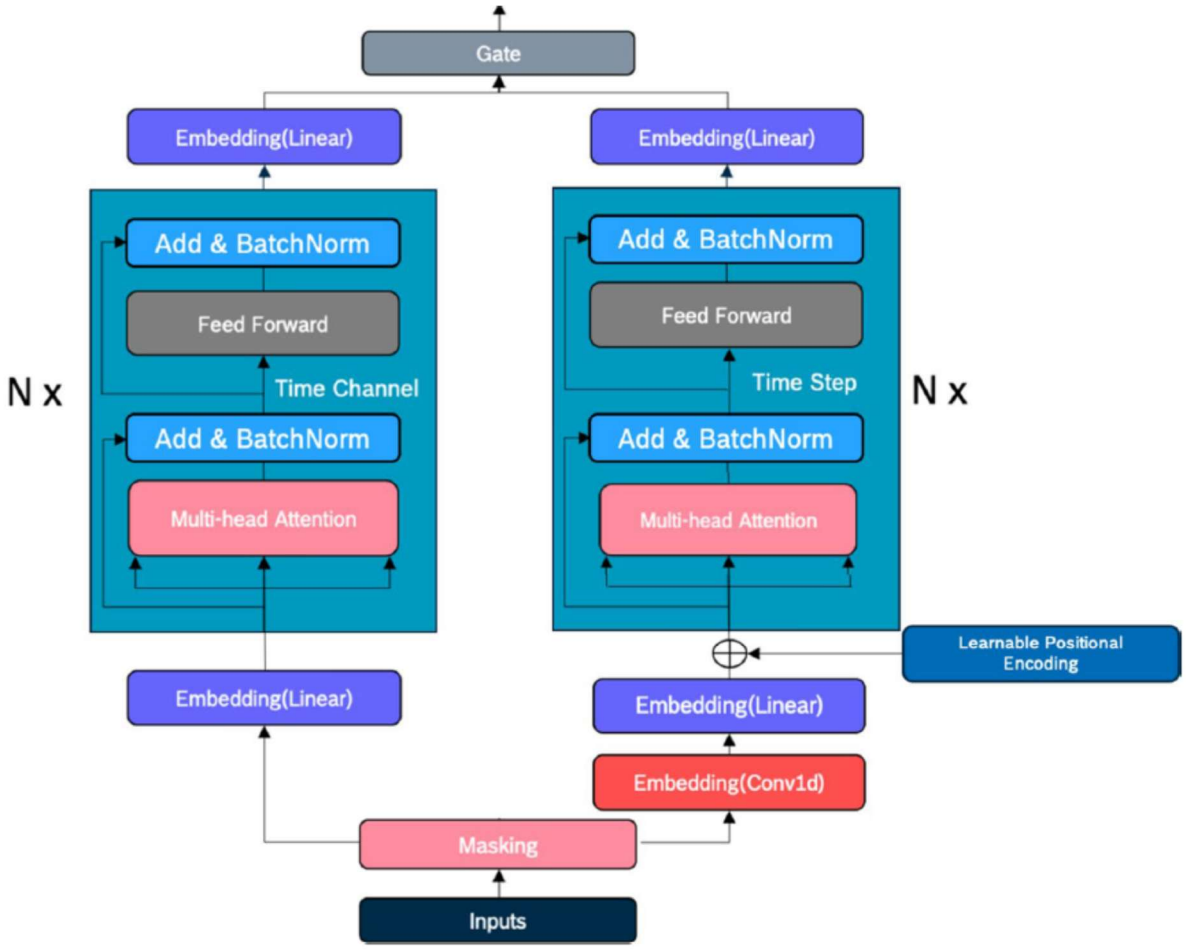
- ✓ (A) We begin by optimizing our image transformer encoders
- ✓ (B) with street-view frame and matching small aerial image pair.
- ✓ (C) Then, for adapting our image encoder to video inputs, we add our GeoAdapter GA module and only optimize the adapter parameters with video pairs as inputs, i.e., a street-view video V s and corresponding large aerial image IaL For training,
- ✓ we sample every kth frame from the street-view video and partition the large aerial image into non-overlapping patches
- ✓ (D) In GA, we apply temporal selfattention (TSA) computation only on the CLS tokens.
- ✓ For TSA computation, we reuse the spatial self-attention weights.
- ✓ (E) During inference, we first perform a Sequence-to-Image inference procedure, where given a query street-view video,
  - The unified module U = {T,GA} produces feature embeddings for both the V s and IaL
  - Then, using embeddings, we retrieve the t nearest neighbor large aerial images (here we show t =
    i. and construct a small aerial image gallery G. (F)
    ii. Finally, GA is removed, and feature embeddings for Ia sk and V s k are obtained.
- ✓ These features are then passed to our TransRetriever TAR model to obtain final frame-by-frame GPS predictions to construct a GPS trajectory.

Transformer Net | **2025-124**

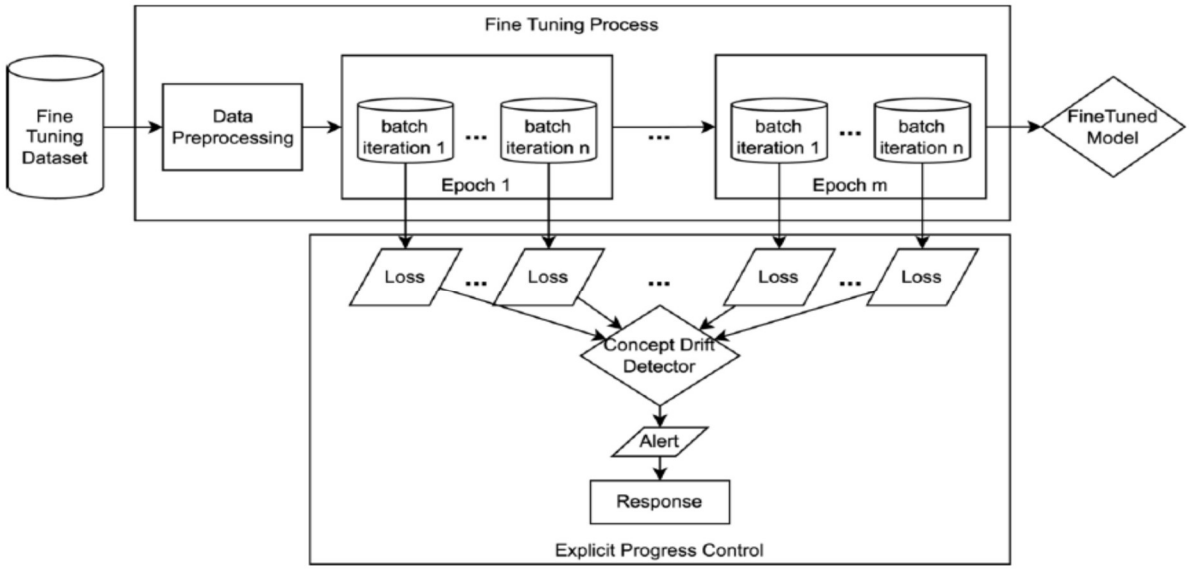


Prototypical description of skeleton-based model architecture in self-supervised learning approach

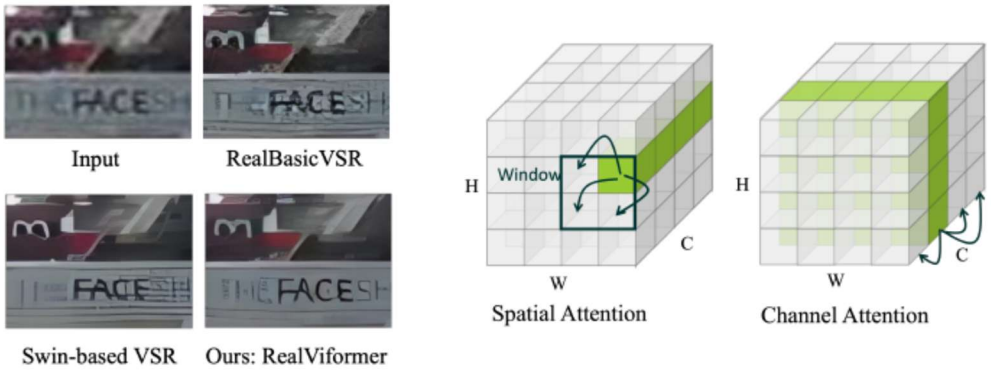




Architecture. ConvGTN model

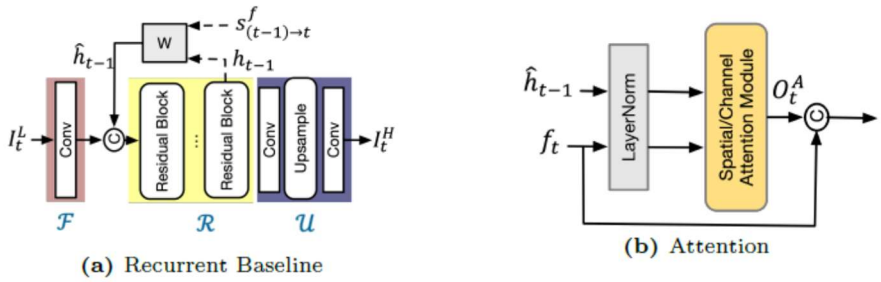


Explicit progress control schema



(a) Visual comparisons. **(b)** Schematics of two attentions.

Fig. 1: (a) Designing a RWVSR transformer is not trivial. A Swin-based transformer suited for standard VSR hallucinates more lines than a RealBasicVSR, a convolutional state-of-the-art. We propose RealViformer based on our investigation of attention under the RWVSR setting. RealViformer generates details with fewer artifacts than RealBasicVSR [3] and the Swin-based VSR model. (b) Schematic for spatial and channel attention. Spatial attention aggregates features based on pixel representations. Channel attention takes $H \times W$ feature map for matching across channels.



(a) Recurrent Baseline **(b)** Attention

Fig. 3: (a) The recurrent baseline in Sec. 3.2 has a shallow mapping module $\mathcal{F}$, reconstruction module $\mathcal{R}$, upsampling module $\mathcal{U}$ and warping function W. W aligns the hidden state h_{t-1} to feature at t based on optical flow $s^f_{(t-1)\to t}$. All residual blocks are convolutional. The concatenation between f_t and $\hat{h}_{t-1}$ are replaced with the spatial or channel attention modules in (b) to compare the effect of attention. (b) The attention module first applies layer normalization to f_t and $\hat{h}_{t-1}$ and then performs channel or spatial attention according to Sec. 3.1. The output feature O^A_t concatenated with f_t is processed by the module $\mathcal{R}$ in (a).

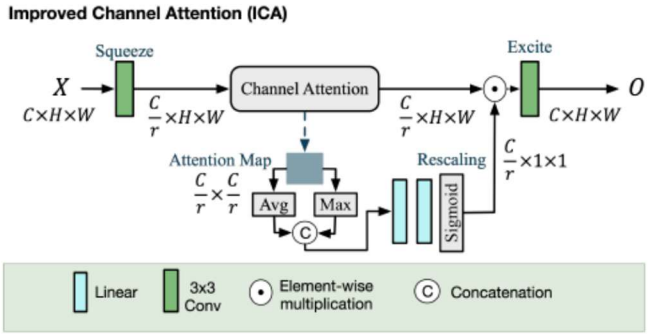


Fig. 5: Improved Channel Attention Module (ICA), showing self-attention for simplicity. The 'squeeze' convolution compresses the number of input feature channels $X \in \mathbb{R}^{C \times H \times W}$ by ratio r. The features are then rescaled by weights predicted from the $\frac{C}{r} \times \frac{C}{r}$ attention map before being expanded by the 'excite' convolution back to the original number of input channels.

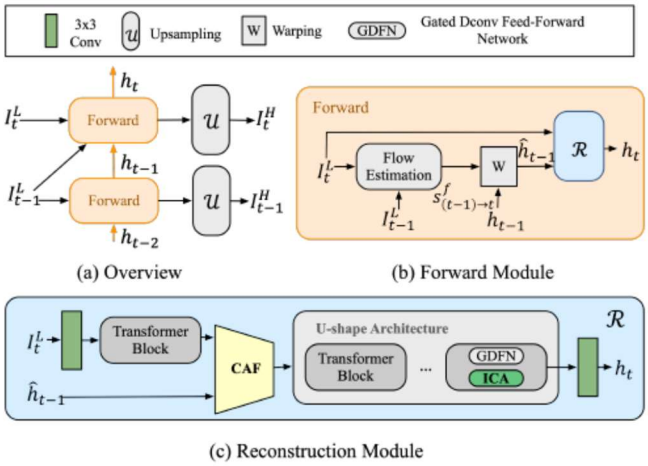


Fig. 6: The framework of RealViformer. (a) Overview of RealViformer, following a unidirectional recurrent framework. The outputs of the Forward module are propagated to the next time step and upsampled by module $\mathcal{U}$ to get HR frames. (b) Explanation of the Forward module in (a), where W denotes the warping function. The reconstruction module $\mathcal{R}$ takes current frame I_t^L and warped hidden state $\hat{h}_{t-1}$ as inputs. (c) Reconstruction module $\mathcal{R}$. The shallow feature of I_t^L and $\hat{h}_{t-1}$ are fused by CAF and then forwarded to Transformer blocks with U-shape connection [39]. Module GDFN follows Restormer [39]. Details of CAF and ICA modules are stated in Fig. 7 and Fig. 5.

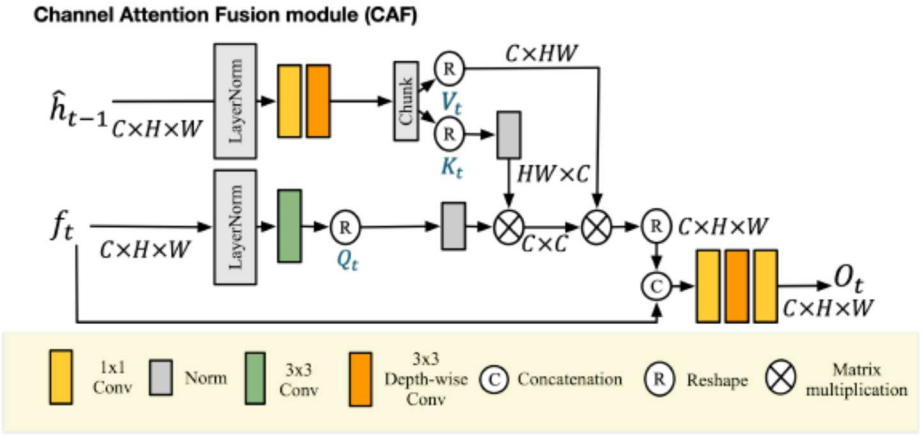
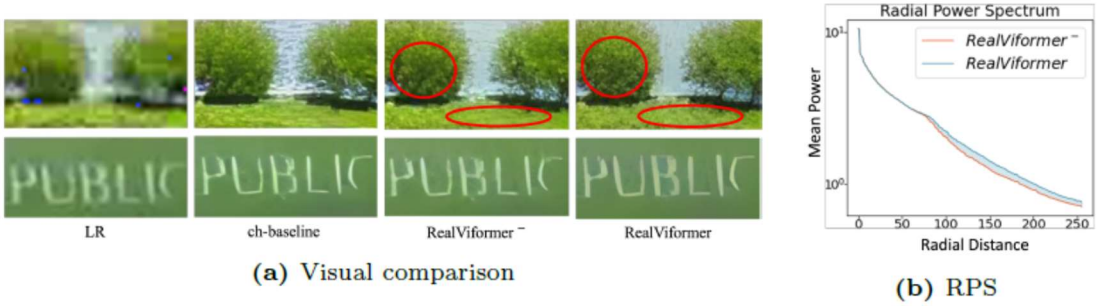


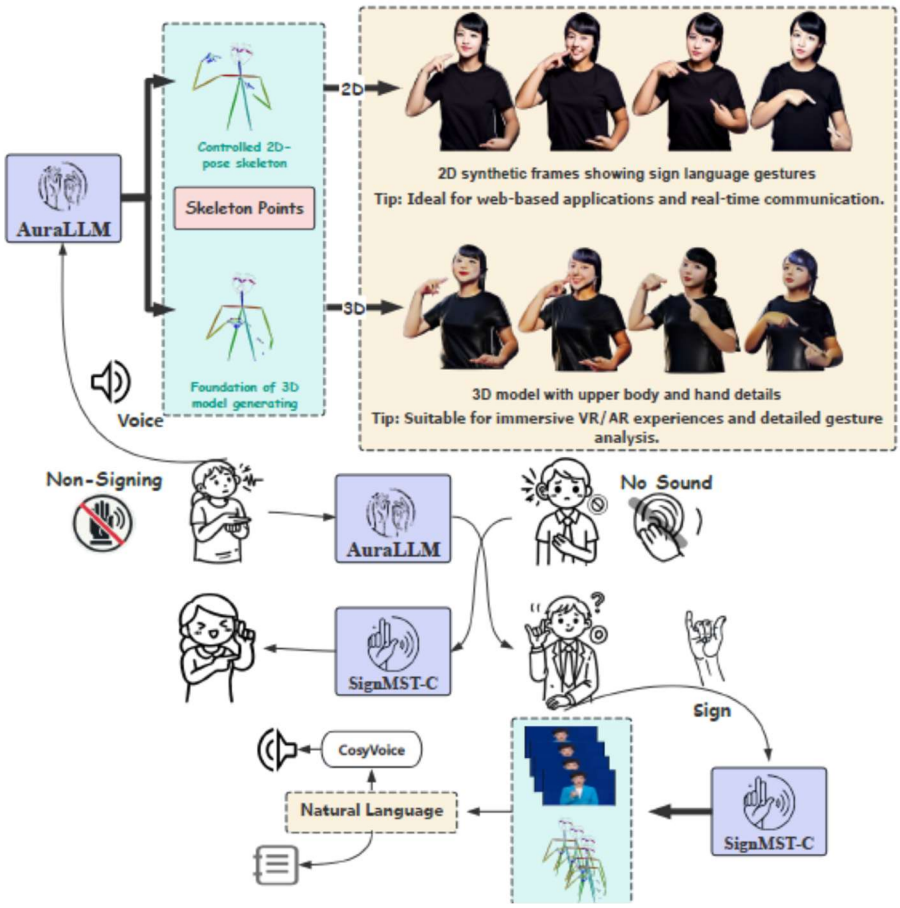
Fig. 7: Details of Channel Attention Fusion (CAF) module. CAF gets the query from current frame feature f_t and {key, value} from hidden state $\hat{h}_{t-1}$. The attention output is concatenated with f_t to process for module output O_t.



(a) Visual comparison

(b) RPS

Fig. 11: (a) Visual comparison between RealViformer and its ablations. Red circles highlight the improved details. (b) Radial Power Spectrum (RPS) of model predictions. Using ICA improves the power of high-frequency components (blue region).



BeyondWords: Enabling Barrier-Free Communication Between Hearing and Hearing-Impaired Individuals

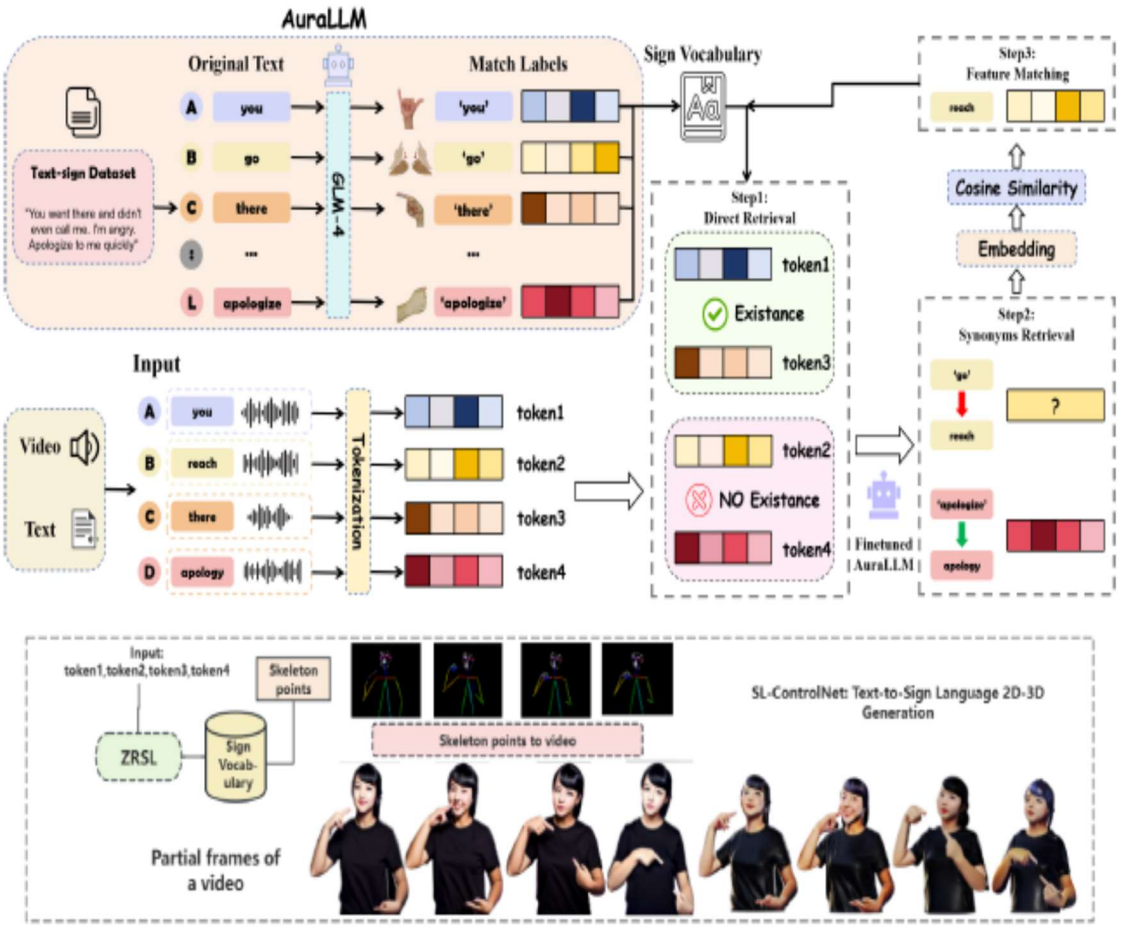
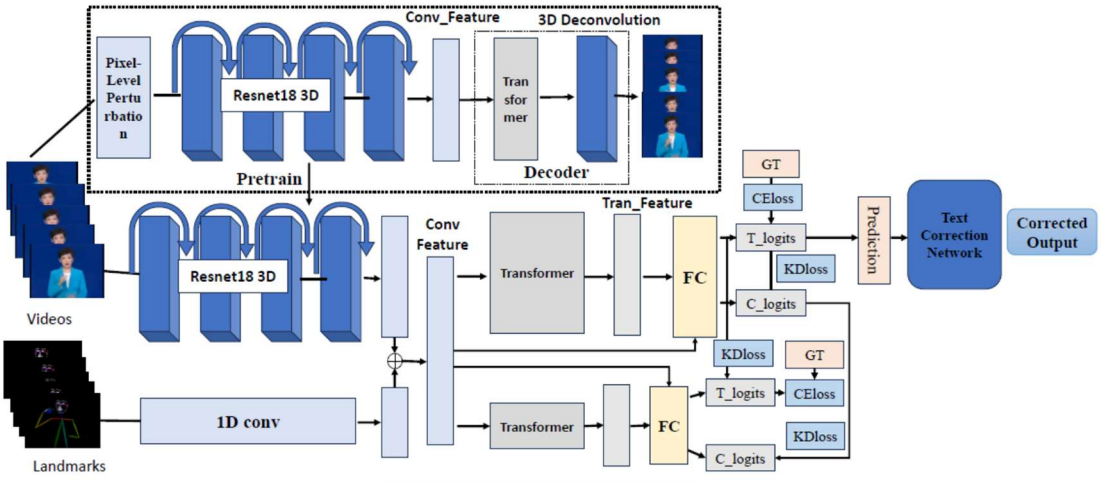
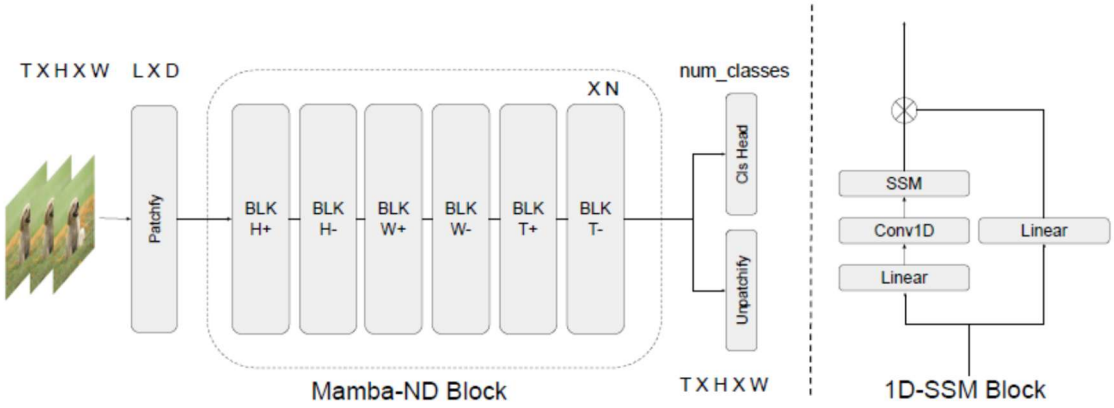


Figure 2. AuraLLM: Combines ZRSL and SL-ControlNet to convert natural language input into high-quality sign language videos.



Overview of SignMST-C

- ✓ Starts with video frames and landmarks processed
  - through 3D ResNet18 and 1D convolution for
    - spatial-temporal and geometric features



Architecture. Mamba-ND

We visualize Mamba-3D as an example.

- Given 3D input, we patchify it into L patches.
- During this process, we maintain the original 3D structure of the input.
- This sequence is then passed through K Mamba-ND blocks, each of which consists of a chain of 1D Mamba layers that process the sequence in alternating orderings.
- In 3D space, we use the order H+H-W+W-T+T-. In 2D space, the sequence would be H+H-W+W-. Finally, the sequence is reshaped back to its original 3D structure and passed to task-specific heads for downstream processing

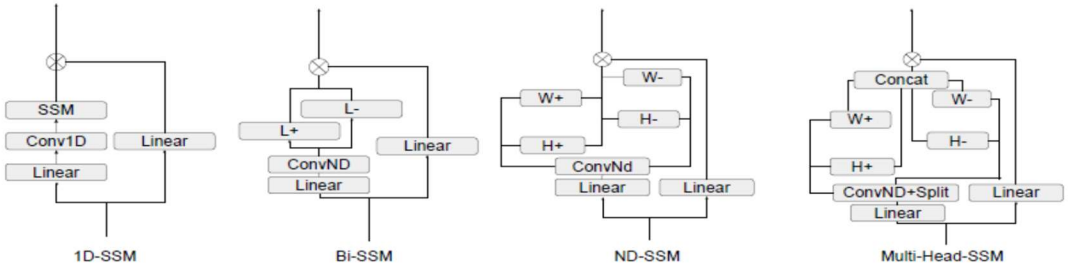
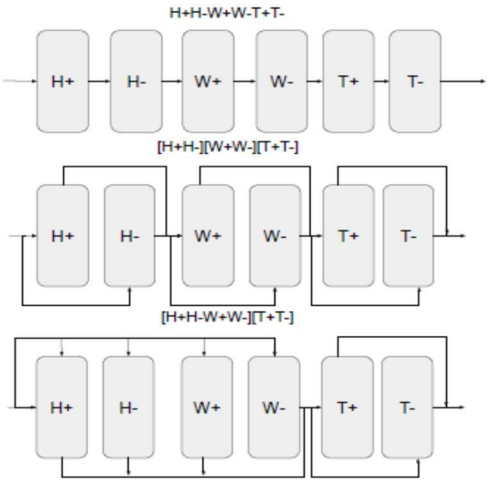
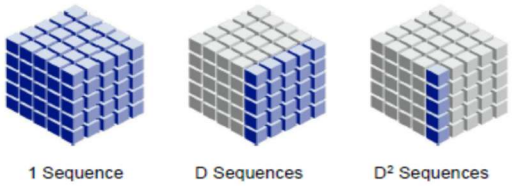


Fig. 3: Variations of SSM Layer Design. Col 1 represents the standard 1D SSM layer. Col 2 represents Bi-SSM, which adds bidirectionality in a similar fashion as LSTM. Col 3 represents ND-SSM block, which extends Bi-SSM to more directions. Col 4 represents multi-head SSM block inspired by multi-head attention in Transformers.



(a) Different ways of arranging Mamba layers. The first row visualizes alternating-directional design. The second row visualizes bidirectional design. The third row visualizes quad-directional design.



(b) Visualization of various Scan-Factorization policies. Col 1: No factorization, there is only 1 sequence. Col 2: Factorizing the 3D sequence into D 2D sequences, where D is the length of a single dimension. Col 3: Factorizing the 3D sequence into D^2 1D sequences.

Visualization of block level design and factorization policies

Table 6: Ablation Study on Layer Designs. We report top-1 accuracy on the ImageNet-1K validation set. The Alt-Directional design is the top-performing one.

		IN1K↑	HMDB-51 ↑
Alt-Directional	Block Level	79.4	59.0
Multi-Head-SSM	Layer-Level	77.6	51.5
ND-SSM	Layer-Level	77.2	46.7
1D-SSM	-	76.4	34.9
Bi-SSM	Layer-Level	74.6	32.1

Transformer Net | **2025-128**

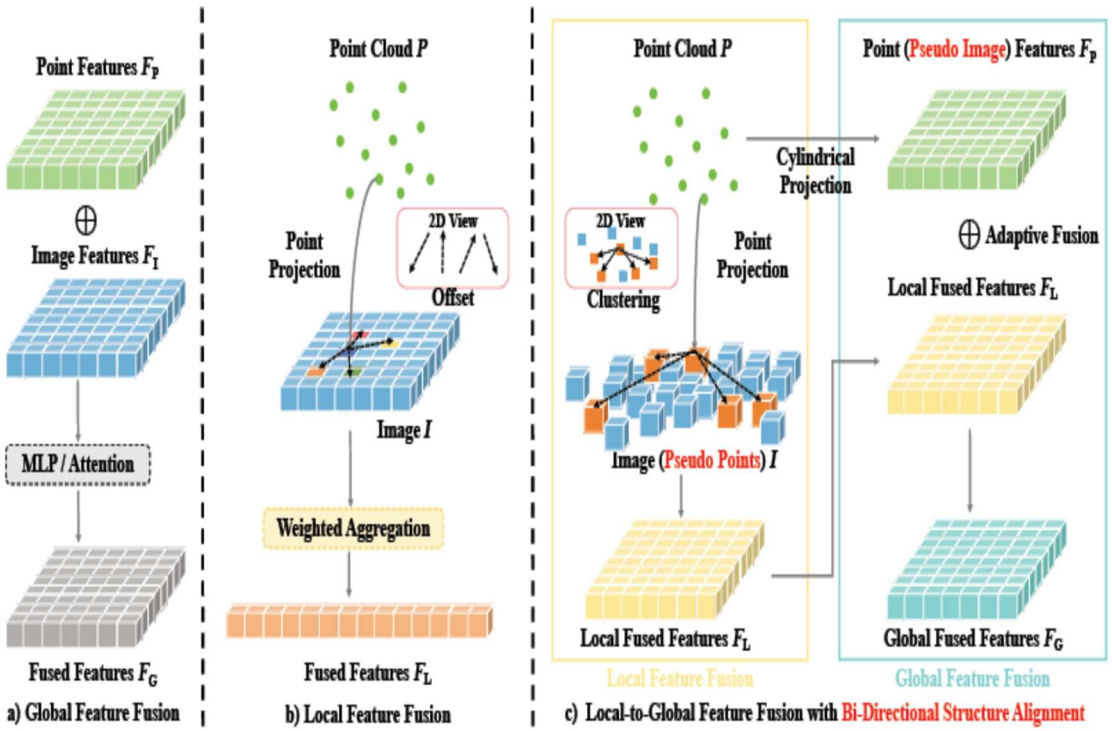
Table 2: Benchmarks with more than one forecasting model using time series

Reference	Methods	Time series	Metrics	Data	Best model
(Masum et al., 2018)	ARIMA and LSMT	Electric load, day	RMSE	Great Britain, Poland and Italy	LSTM
(de Oliveira and Cyrino Oliveira, 2018)	Bagging ARIMA and Exponential Smoothing	Mid-long term electric energy consumption, monthly	ASM, sMAPE, RMSE and TIC	Canada, France, Italy, Japan, Brazil, Mexico and Turkey	Remainder Sieve Bootstrap
(Khan and Osińska, 2021)	Fractional-order Grey Model and ARIMA	2019 statistical review of world energy, yearly	MAPE and MSE	Brazil, Russia, China, India, and the Republic of South Africa (BRICS)	ARIMA
(Banik et al., 2021)	Random Forest, Ensemble Learning, Boosting and XGBoost	Electricity load, hourly, weekly and monthly	R^2 and RMSE	Tripura in India	Ensemble RF-XGBoost
(Farsi et al., 2021)	ARIMA, LSTM, and Parallel LSTM-CNN Network.	Hourly load consumption	MAPE, R^2 and RMSE	Germany and Malaysia	Parallel LSTM-CNN Network
(Zhao et al., 2021)	GRU, LSTM, RNN, TCN, CNN+LSTM and Transformer	Day-ahead load forecasting	MAPE	Australia	Transformer network
(Chaturvedi et al., 2022)	SARIMA, LSTM RNN and Facebook Prophet	Monthly energy demand, monthly	MAPE and RMSE	India	Facebook Prophet
(Panagiotou and Dounis, 2022)	RNN, ANFIS, and LSTM	Energy consumption, hourly	MSE, R, R^2, MAPE and CI	Hospital Building's Energy Consumption	ANFIS LSTM
(Henzel et al., 2022)	Naive Methods, Linear Regression, LSTM, and the Facebook Prophet Method	Energy consumption, hourly	MSE and MAPE	Digital-Twin Model of the Building	Facebook Prophet

(Shohan et al., 2022)	ANN, LSTM, Facebook Prophet and LSTM-Facebook Prophet	Hourly load demand, hourly	MAPE, R^2, SSE and RMSE	Florida	LSTM- Facebook Prophet
(Ribeiro et al., 2022)	ARIMA, SVR, Random Forest, XGBoost, RNN, LSTM, and GRU	Energy consumption, hourly	RMSE, MAPE and MAE	ESCO (Energy Service Company) Ireland	XGBoost
(Shin and Woo, 2022)	Random Forest, XGBoost and LSTM	Energy consumption, monthly	RMSE and MAPE	Korea	LSTM and Random Forest
(Sulandari et al., 2023)	ARIMA, Exponential Smoothing, Facebook Prophet, Neural Networks, and Proposed Ensemble Methods	Generation of electricity, hourly	MAPE and RMSE	US, Ontario, England, Wales and Australia	Proposed Ensemble Methods
(Pierre et al., 2023)	ARIMA, LSTM, GRU, ARIMA-LSTM, and ARIMA-GRU	Electricity demand, hourly	MAPE and RMSE	Benin Electricity Company (CEB)	Hybrid Approach
(Koukaras et al., 2024)	HGBR, LGBMR, ETR, RR, BRR, CBR	Energy consumption, hourly	R^2, RMSE, CVRMSE, NRMSE and MAE	ITI/CERTH Smart House, Thessaloniki (Greece)	HGBR, LGBMR

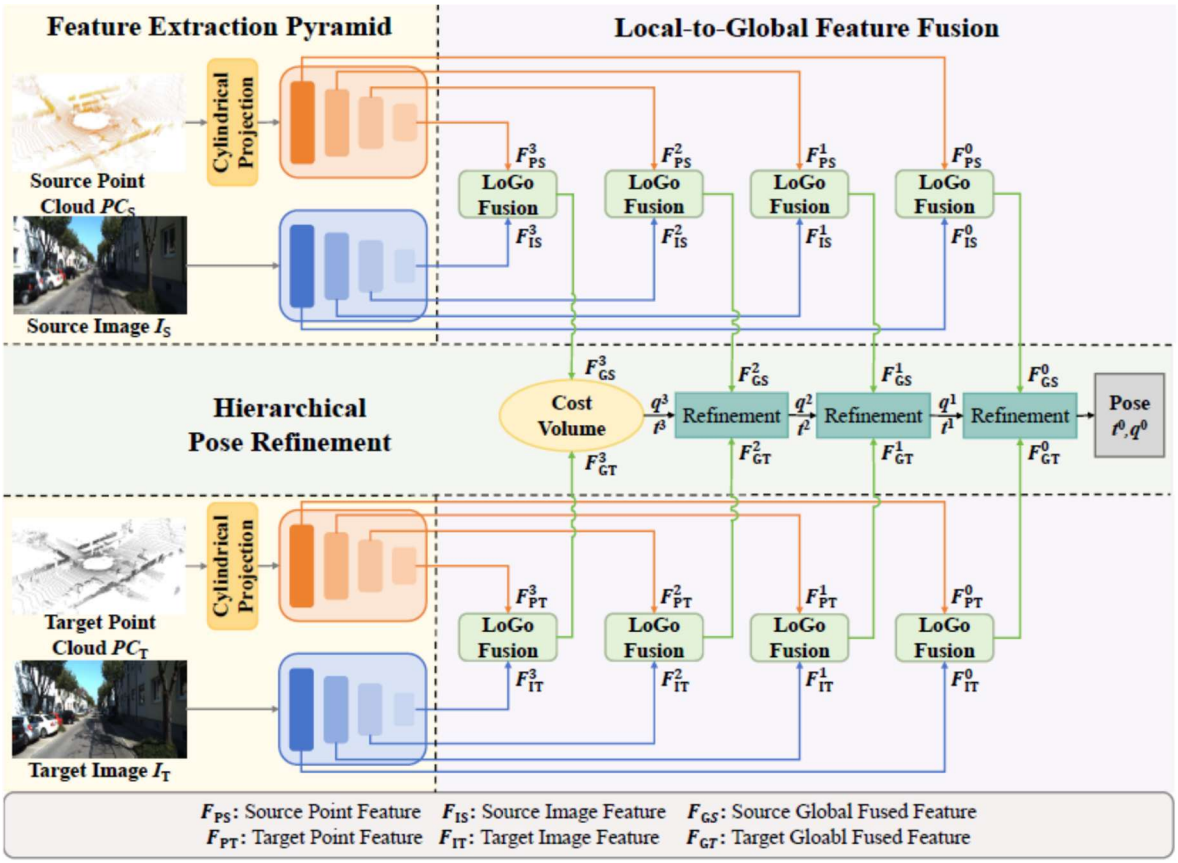
Transformer Net | **2025-129**

DVLO: Deep Visual-LiDAR Odometry with Local-to-Global Feature Fusion and Bi-Directional Structure Alignment



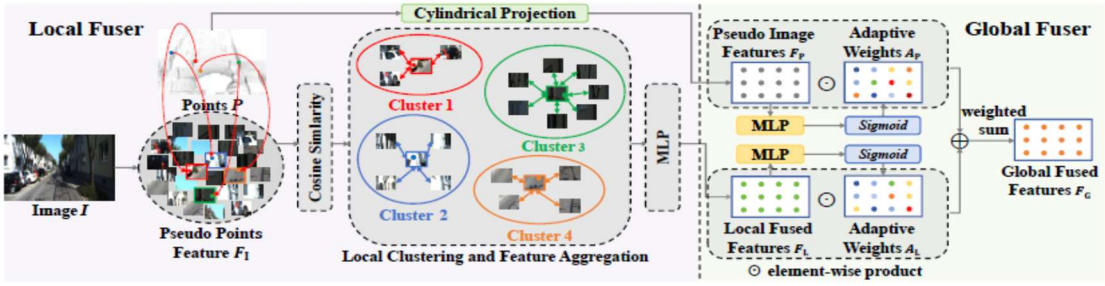
Different fusion strategies for images and points.

- Most previous works only perform the fusion globally or locally [73]
- Present DVLO designs a local-to-global fusion strategy that facilitates the interaction of global information while preserving local fine-grained information
- Furthermore, a bi-directional structure alignment is designed to maximize the inter-modality complementarity



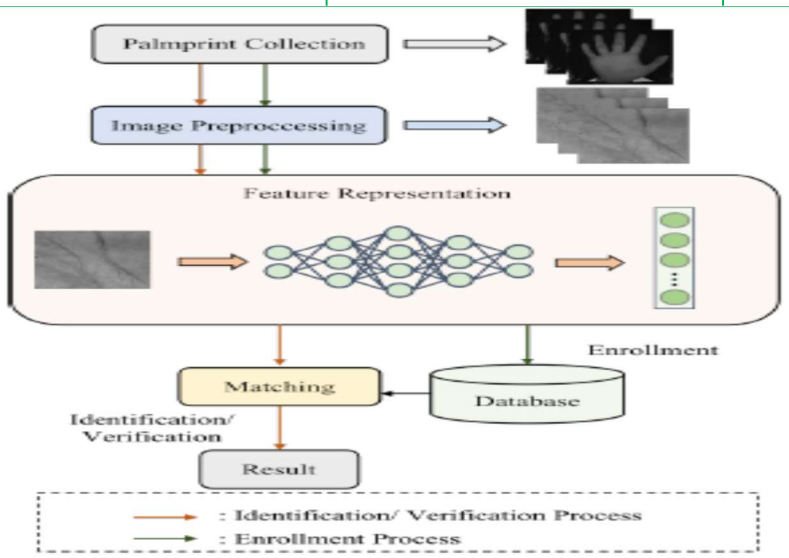
Pipeline. DVLO
Novel Local-to-Global (LoGo) fusion module

- DVLO consists of
  - A clustering-based Local Fuser and
  - an adaptive Global Fuser.
  - The pose is initially regressed from the cost volume of the coarsest fused features and
  - then refined iteratively from fused features in shallower layers

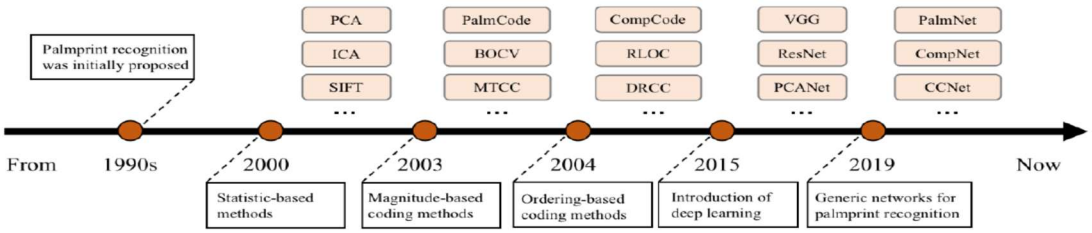


Local-to-Global (LoGo) Fusion module

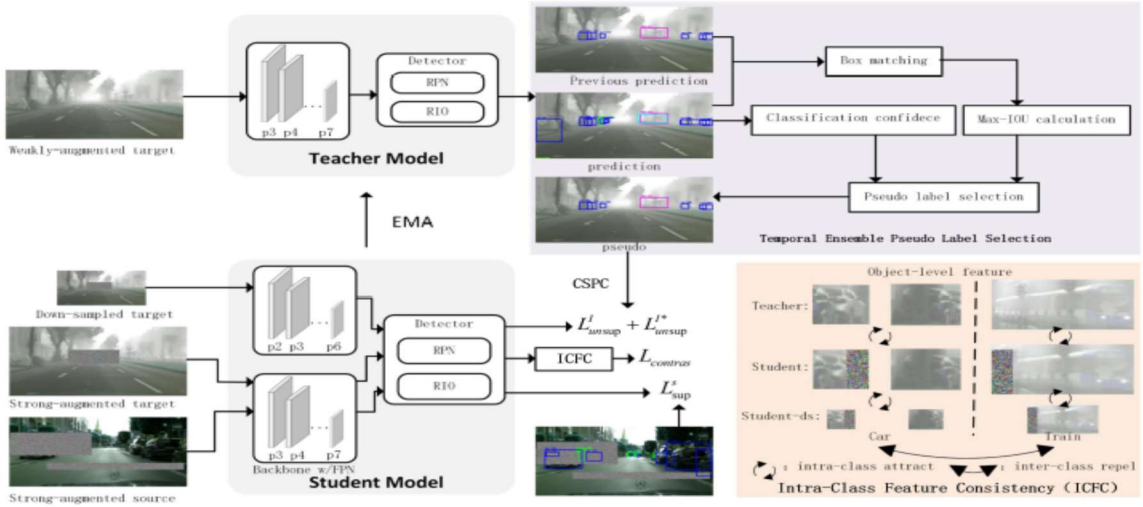
- Project points onto the image plane
  - based on the coordinate system transformation matrix as cluster centers and
- Convert the image into a set of pseudo points
- Then, locally aggregate pseudo point features based on the similarities to each cluster center.



Pipeline. DL-based palmprint recognition system



The evolution of palmprint feature extraction methodologies



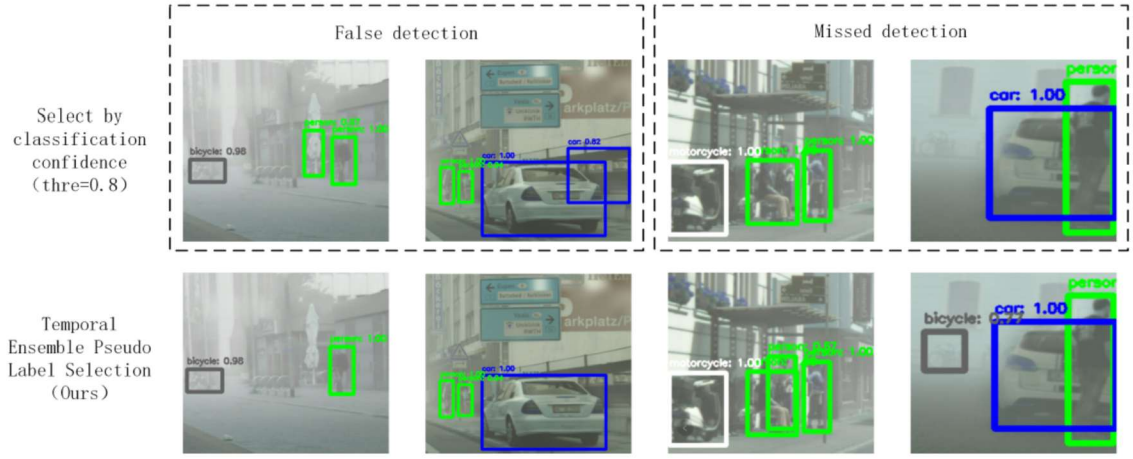
Overview. method for the UDA-OD task.

- ! Architecture. teacher–student Alg.
  - Teacher model generates pseudo-labels based on weakly augmented target domain images
  - Student model is trained using both downsampled and strongly augmented inputs.

- ! CSPC is enforced by supervising the student model with inputs of different resolutions
  - 🔔 improving detection of objects at various scales
- ! Temporal ensemble is employed for
  - Robust pseudo-label selection
    - Combining classification confidence and box matching based on
    - Intersection over Union (IoU) to ensure high-quality pseudo-labels.

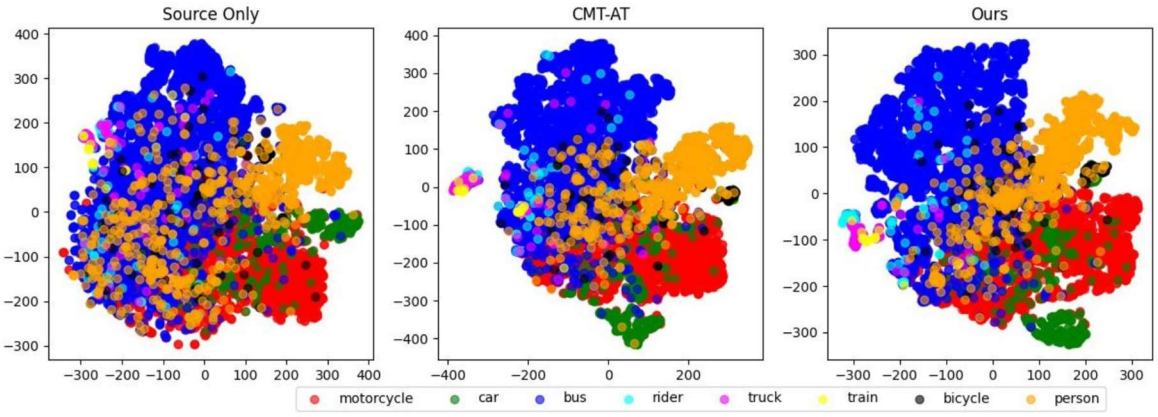
- ! ICFC module
  - Aligns object-level features across scales and augmentations,
    - Utilizing contrastive learning
  - ✓ To ensure intra-class attraction and inter-class repulsion,
    - Enhancing the consistency of object representations

Method	Backbone	Detector	Split	Bicycle	Bus	Car
Source	R50-FPN	Faster R-CNN	0.02	38.4	27.5	44.9
Oracle	R50-FPN	Faster R-CNN	0.02	51.2	52.7	74.1
AT [4]	V16	Faster R-CNN	0.02	51.3	64.9	63.6
CMT [5]	V16	Faster R-CNN	0.02	51.2	66.0	63.7
MOTOR [32]	R50	Faster R-CNN	0.02	35.6	38.6	44.0
SFA [33]	R50	Deform-Detr	0.02	44.0	46.2	62.6
MTTrans [34]	R50	Deform-Detr	0.02	46.5	45.9	65.2
DA-DETR [35]	R50	Deform-Detr	0.02	46.5	45.9	63.1
O^2Net [36]	R50	Deform-Detr	0.02	45.9	47.6	63.6
AQT [37]	R50	Deform-Detr	0.02	46.4	53.7	64.4
MTM [38]	R50	Deform-Detr	0.02	47.7	54.4	67.2
TDD [39]	R50	Faster R-CNN	0.02	49.1	51.1	64.1
MRT [40]	R50	Faster R-CNN	0.02	47.1	58.1	68.7
SA-DA-Faster [7]	R50-FPN	Faster R-CNN	0.02	45.4	50.3	62.1
AT [4]	R50-FPN	Faster R-CNN	0.02	53.3	52.1	66.0
CMT [5]	R50-FPN	Faster R-CNN	0.02	53.1	55.0	66.7
Ours	R50-FPN	Faster R-CNN	0.02	57.7	57.5	69.0
Source	R50-FPN	Faster R-CNN	All	50.8	46.7	62.4
Oracle	R50-FPN	Faster R-CNN	All	55.0	57.9	72.2
PDA [41]	V16	Faster R-CNN	All	35.9	44.1	54.4
ICR-CCR [42]	V16	Faster R-CNN	All	34.6	36.4	49.2
AT [4]	R50-FPN	Faster R-CNN	All	53.1	59.0	69.5
CMT [5]	R50-FPN	Faster R-CNN	All	53.6	58.6	69.9
Ours	R50-FPN	Faster R-CNN	All	53.0	63.0	71.9



Qualitative comparison of pseudo-label selection strategies

- ✓ Top row: shows results using only classification confidence (threshold = 0.8)
  - Leading to false detections (left) and missed detections (right).
- ✓ Bottom row: results using proposed TEPLS strategy



t-SNE visualization of feature distributions for different approaches

- ✓ Source Only (left)
- ✓ CMT-AT (middle),
- ✓ Proposed method (right)
  - Each point represents an object feature with different colors for different object classes

Transformer Net | **2025-134**

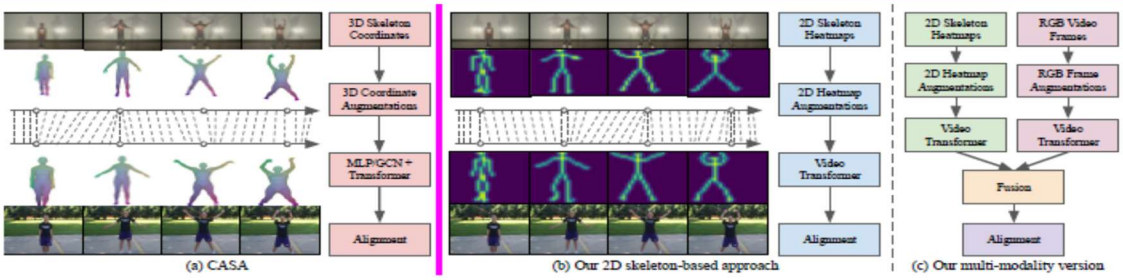
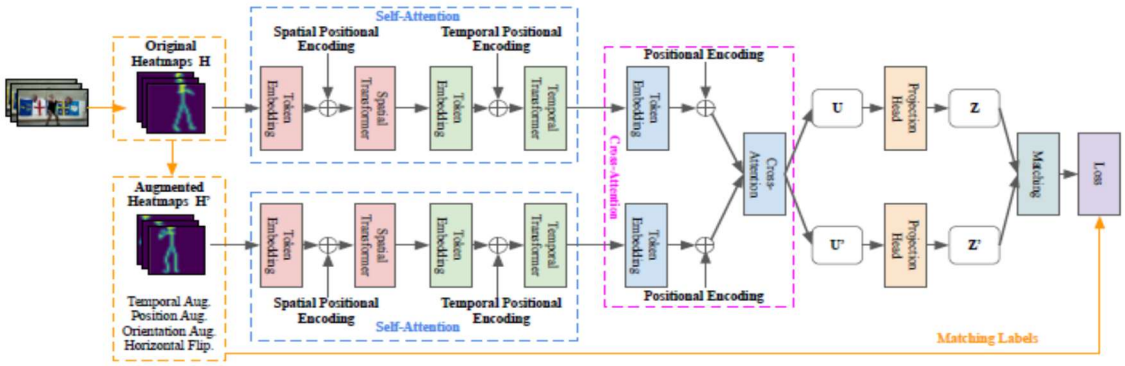


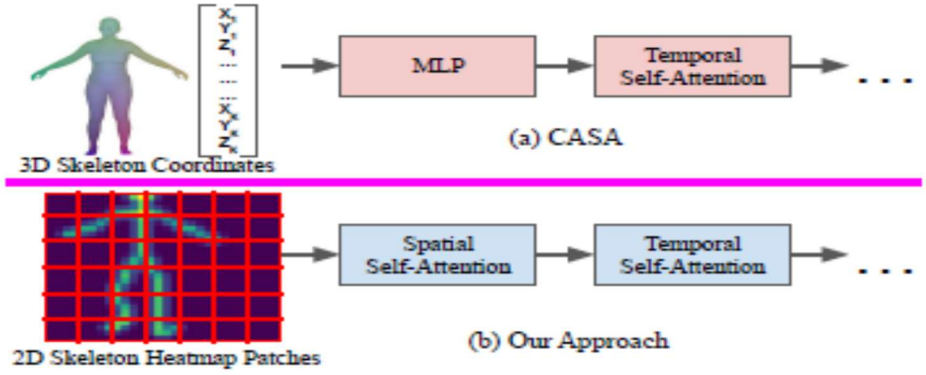
Fig. 1:

- ✓ (a) State-of-the-art method in self-supervised learning, i.e., CASA,
  - uses 3D skeletonbased temporal video alignment as a pretext task and
  - performs 3D skeleton augmentations.
- ✓ (b) This approach relies on 2D skeleton-based temporal video alignment
  - conducts 2D skeleton augmentations. 2D skeleton heatmaps used, which are fed to a video transformer for learning useful spatiotemporal and contextual features.
- ✓ Present method
  - obtains higher accuracy and
  - Better robustness against missing and noisy keypoints,
  - while showing superior performance in various fine-grained human activity understanding tasks.
- ✓ (c) 2D skeleton heatmaps with RGB videos further fused, establishing
  - state-of-the-art across all metrics and datasets

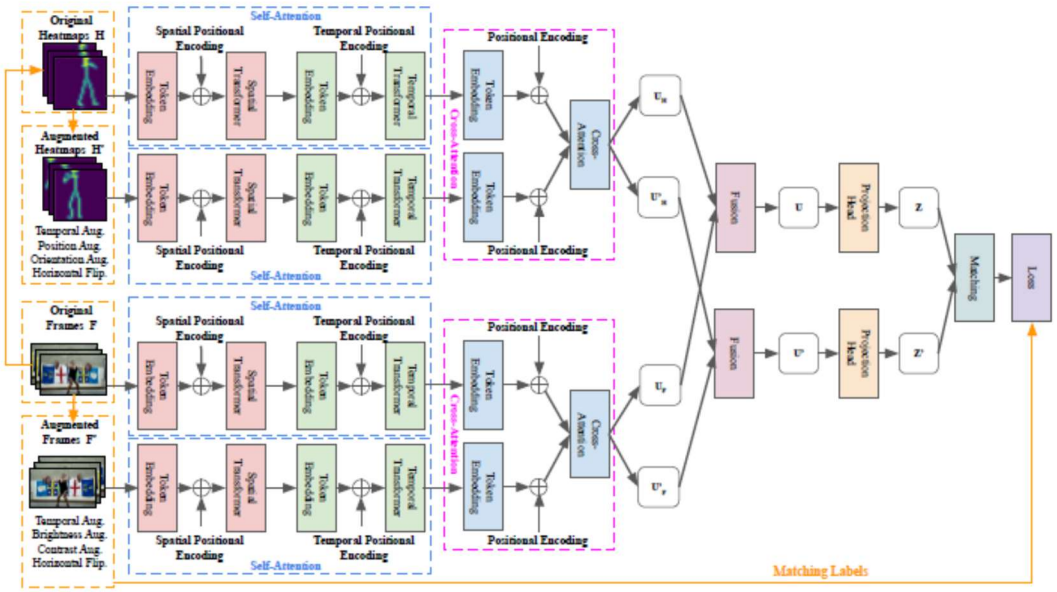


Training

- ! Learning in this method uses input sequences of original heatmaps and augmented heatmaps.
- ! Performs self-attention both in the spatial and temporal domains
  - To extract effective spatiotemporal and contextual cues within each sequence and cross-attention
  - To extract contextual cues across the sequences.
- ! The extracted features after projection heads are passed to a matching module
  - To predict correspondences across the sequences
- ! Matching labels generated by the **augmentation** module are used in the loss function



- 🔔 (a) CASA directly processes 3D skeleton coordinates and
  - conducts self-attention in the temporal domain only
- 🔔 (b) This approach operates on 2D skeleton heatmaps and
  - performs self-attention both in the spatial and temporal domains



Multi-modality model (MMM, MulModModel)

- ! MMM includes various encoders to extract features from
  - Heatmaps (top) and
  - RGB videos (bottom)
- ! Performs late fusion to obtain the combined features

Transformer Net	2025-135	

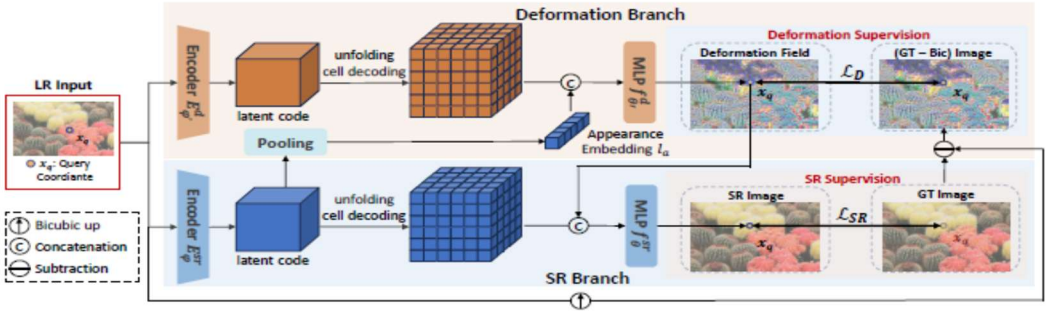


Fig. 3: The training pipeline of our DDIR model. It consists of double branches, which are the deformation branch and the SR branch. Each branch is composed of an encoder and an MLP, taking the LR image and query coordinates as the inputs. The appearance embedding l_a is computed as the spatial average pooling of the 2D feature map from the encoder E_ϕ^{sr} of the SR branch, which is fed into the decoding function $f_{\theta'}^{d}$ of the deformation branch by concatenation. The RGB output of the deformation branch is supervised by the deformation field. Then, the predicted deformation field feeds into the decoding function f_θ^{sr} of the SR branch by concatenation. Finally, the decoding function f_θ^{sr} of the SR branch outputs the target high-resolution RGB values at the query coordinates. Combining the appearance embedding and the deformation field, our DDIR model learns the dual-level deformable implicit representation to address the deformations at the image and pixel levels simultaneously.

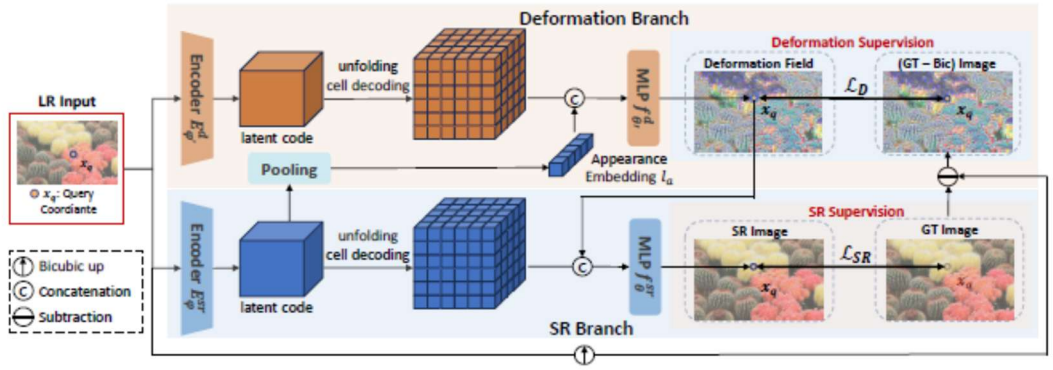
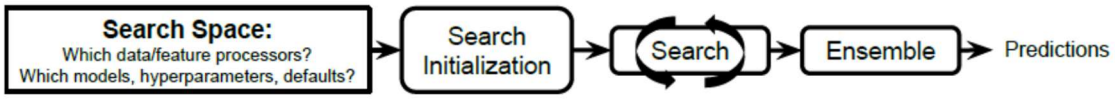


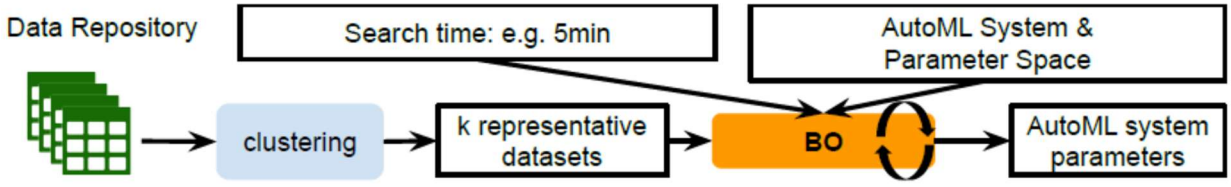
Fig. 3: The training pipeline of our DDIR model. It consists of double branches, which are the deformation branch and the SR branch. Each branch is composed of an encoder and an MLP, taking the LR image and query coordinates as the inputs. The appearance embedding l_a is computed as the spatial average pooling of the 2D feature map from the encoder E_ϕ^{sr} of the SR branch, which is fed into the decoding function $f_{\theta'}^{d}$ of the deformation branch by concatenation. The RGB output of the deformation branch is supervised by the deformation field. Then, the predicted deformation field feeds into the decoding function f_θ^{sr} of the SR branch by concatenation. Finally, the decoding function f_θ^{sr} of the SR branch outputs the target high-resolution RGB values at the query coordinates. Combining the appearance embedding and the deformation field, our DDIR model learns the dual-level deformable implicit representation to address the deformations at the image and pixel levels simultaneously.

Transformer Net | **2025-136**



AutoML Process

- Search Space Setup, Initialization, Search, and Ensembling

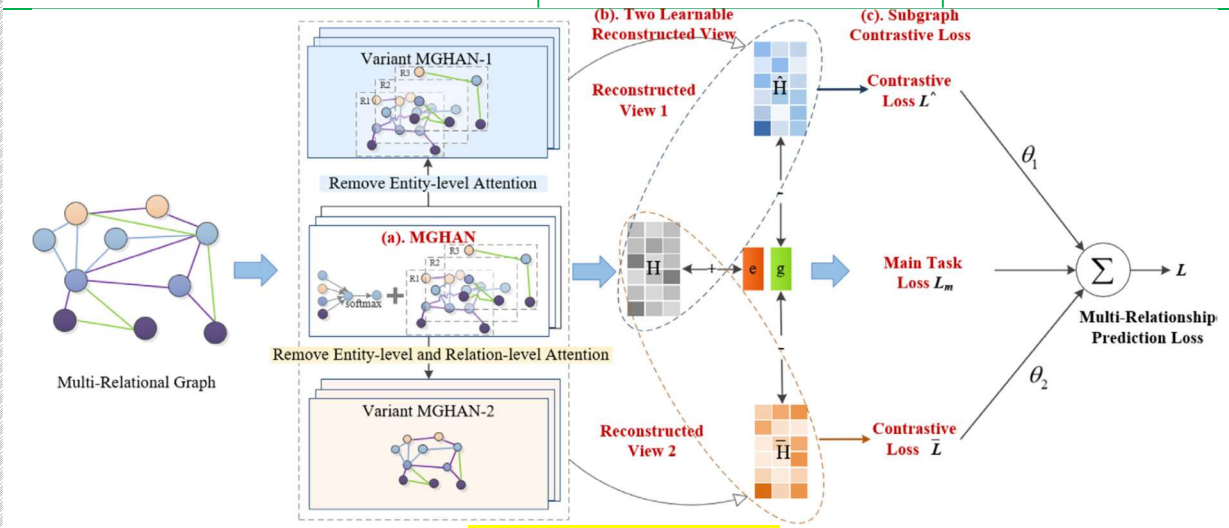


Optimizing AutoML for one search budget

Table: OpenML Test datasets

Name	DatasetID	# instances	# features	# classes
robert	41165	10000	7200	10
riccardo	41161	20000	4296	2
guillermo	41159	20000	4296	2
dilbert	41163	10000	2000	5
christine	41142	5418	1636	2
cnae-9	1468	1080	856	9
fabert	41164	8237	800	7
Fashion-MNIST	40996	70000	784	10
KDDCup09_appetency	1111	50000	230	2
mfeat-factors	12	2000	216	10
volkert	41166	58310	180	10
APSFailure	41138	76000	170	2
jasmine	41143	2984	144	2
nomao	1486	34465	118	2
albert	41147	425240	78	2
dionis	41167	416188	60	355
jannis	41168	83733	54	4
covertype	1596	581012	54	7
MiniBooNE	41150	130064	50	2
connect-4	40668	67557	42	3
kr-vs-kp	3	3196	36	2
higgs	23512	98050	28	2
helena	41169	65196	27	100
kc1	1067	2109	21	2
numerai28.6	23517	96320	21	2
credit-g	31	1000	20	2
sylvine	41146	5124	20	2
segment	40984	2310	16	7
vehicle	54	846	18	4
bank-marketing	1461	45211	16	2
Australian	40981	690	14	2
adult	1590	48842	14	2
Amazon_employee_access	4135	32769	9	2
shuttle	40685	58000	9	7
airlines	1169	539383	7	2
car	40975	1728	6	4
jungle_chess_2pcs_raw_endgame_complete	41027	44819	6	3
phoneme	1489	5404	5	2
blood-transfusion-service-center	1464	748	4	2

Transformer Net **2025-137**



Architecture of MRGCL

- Different colored circles: different types of entities and
- Different colored lines: different types of relations

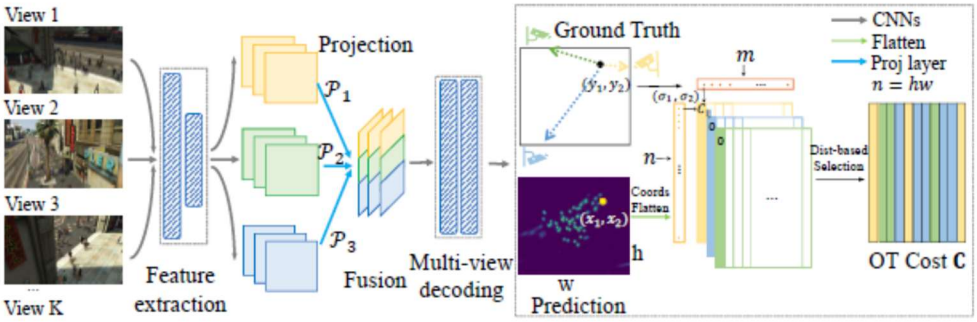
- To describe the multi-relational graph.

! (a): a multi-relational graph hierarchical attention networks consisting of
  - entity-level, relation-level, and layer-level attentions.

! (b): two graph augmented views with adaptive topology are automatically learned by the variant MGHAN-1/2

! (c): a subgraph contrastive loss is designed
  - To generate positives per anchor by
    - calculating strongly connected subgraph embeddings e of the anchor as the supervised signals

Transformer Net	**2025-138**	

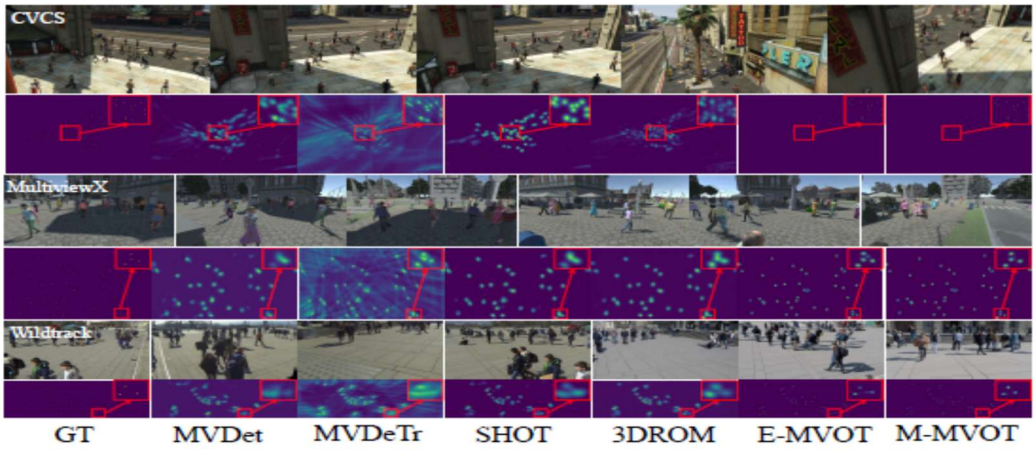


Model architecture M-MVOT

- Proposed Mahalanobis distance-based Multiview optimal transport (M-MVOT) loss for multi-view crowd localization.
  - Modules in model
    - Feature extraction,
    - Projection,
    - multi-view fusion and
    - Decoding.

Proposed M-MVOT,

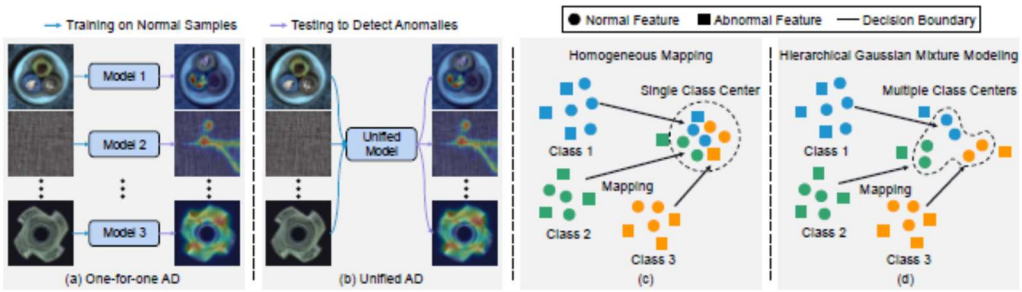
- Each point's transport cost C is calculated via the Mahalanobis distance instead of the common Euclidean distance under the closest camera, which
  - is directed by the view ray and
    - adjusted by the object-to-camera distance



Predicted crowd occupancy maps of different methods
On the 3 datasets

- **CVCS, MultiviewX, and Wildtrack (zoom in for better view)**
  - **The value of crowd occupancy maps indicates the person probability of each location**

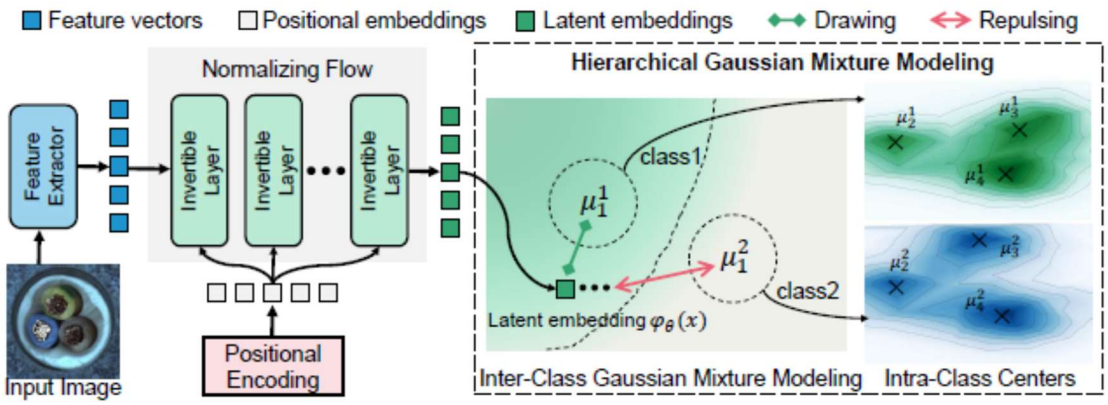
Transformer Net	2025-139	



Anomaly detection task settings

Aim to implement one unified AD model (b).

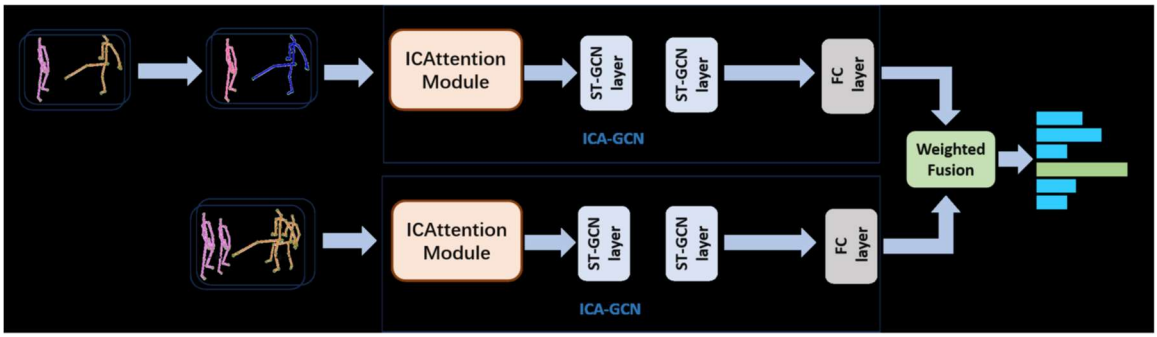
- ✓ (c) Mapping all input features to the same latent class center
  - may induce the “homogeneous mapping” issue.
- ✓ (d) Hierarchical Gaussian mixture modeling method proposed for
  - More effectively capturing the complex multi-class distribution



Model overview

- Extracted feature vectors are sent into
  - normalizing flow model
    - For transforming into latent embeddings.
- Add positional embeddings (i.e., sinusoidal position encoding) to each invertible layer as they are
  - Effective for nf-based ad methods
- Employ hierarchical gaussian mixture modeling approach
  - To fit the latent embeddings,
    - Which can assist the model against learning the “homogeneous mapping”.

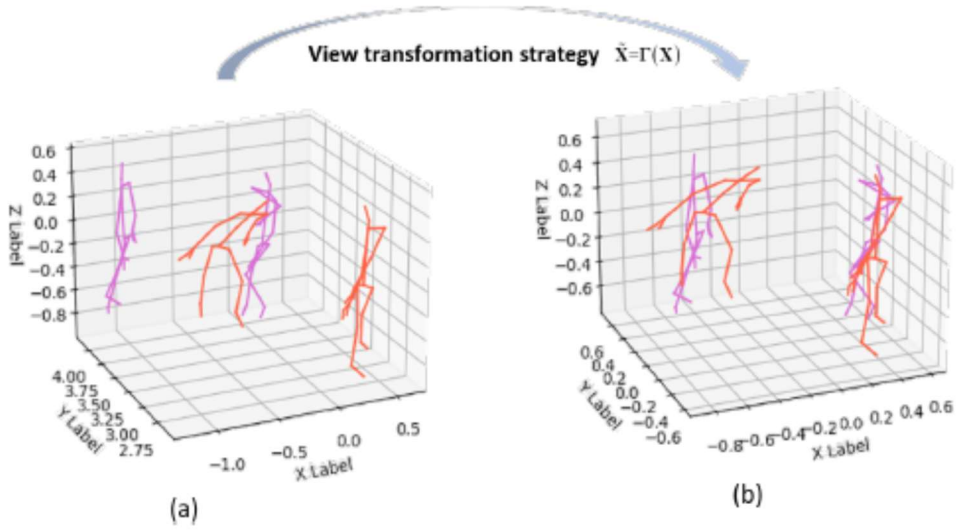
Transformer Net	2025-140	



Architecture

- A two-stream architecture is applied
- Joint coordinates and joint motion data are respectively as the input of each stream.
- Here described the stream with joint coordinates as input.
- Proposed view transformation strategy Γ
  - Transforms the raw skeleton data X to $\tilde{X}$
- Utilize defined individual contribution weights to design the individual contribution based
  - Spatial-temporal attention module(ICAttention)
- Next, muti-layer spatial-temporal graph convolutional network and Proposed attention module are combined
  - To obtain the ica-gcn framework for feature learning
- The output features of ica-gcn are passed through an average global pooling operation
- Then fed to a fully connected layer with Softmax to obtain the classification score

- A weighted fusion strategy is utilized for finally action classification



Comparison of raw skeleton data and transformed skeleton data with proposed view transformation strategy.

- Two different action samples in one frame (in purple color and in orange color)
  - belonging to the same class "kicking other person" are described

- Figure(a): shows the raw skeleton data of the two actions in 3D coordinates
  - which can find the data are significantly mismatched.

- Figure(b) shows the transformed skeleton data of the two actions in 3D coordinates
  - which the proposed interactive center point is the origin.

- It is easy to discover that the transformed skeleton data of the two samples
  - show high similarity.

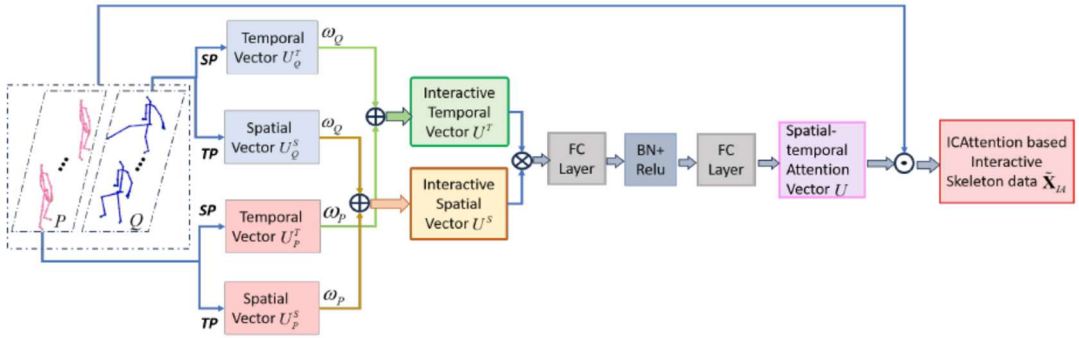


FIGURE 4. The structure of proposed ICAttention module. The transformed skeleton data $\tilde{X}_P$ and $\tilde{X}_Q$ adopt spatial pooling(SP) operation and temporal pooling(TP) operation to respectively obtain the temporal vector and spatial vector. For skeleton data $\tilde{X}_P \in \mathbb{R}^{C\times T\times V}$, we can get temporal vector $U_P^T \in \mathbb{R}^{C\times T\times 1}$ and spatial vector $U_P^S \in \mathbb{R}^{C\times 1\times V}$ by performing global pooling operations along different dimensions on $\tilde{X}_P$. The temporal vector U_Q^T and spatial vector U_Q^S from $\tilde{X}_Q$ can be obtained in similar way. Next, the interactive temporal vector U^T can be calculated by weighted addition operations on U_P^T and U_Q^T where the weights are proposed individual contribution weights ω_P and ω_Q. Similarly, we can calculate interactive spatial vector U^S. U^T through a matrix production operation with U^S can produce a $C \times T \times V$ tensor. After two fully-connected layers, the spatial-temporal attention vector U is obtained, which multiplies by input skeleton data $\tilde{X}$ to finally obtain the ICAttention based interactive skeleton data $\tilde{X}_{IA}$.

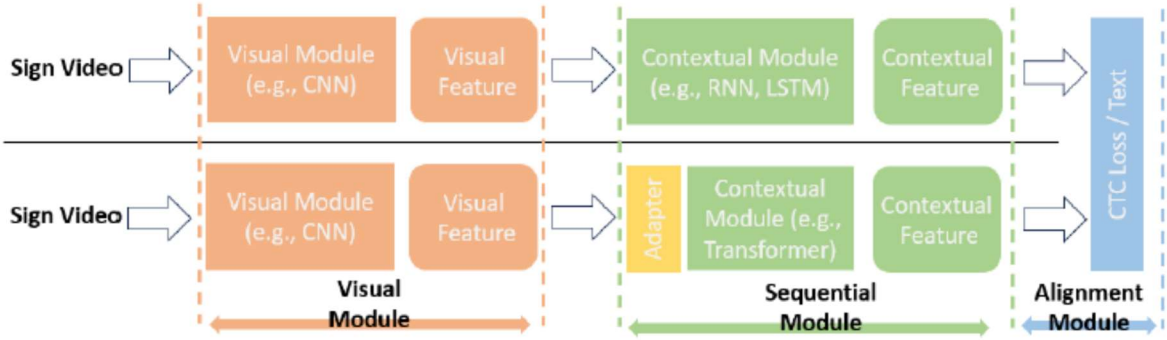
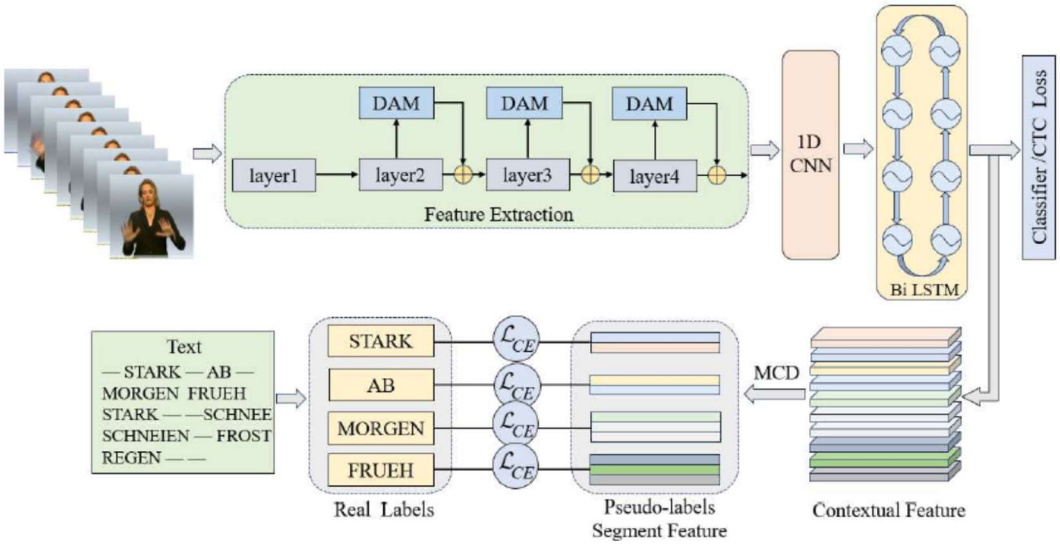


Fig. 1. The main components of mainstream continuous sign language recognition models.



Backbone. DAM-MCD model.

1) CNN is used to capture frame-level features,
2) Followed by 1DCNN+bilstm for temporal modelling
3) Finally, a classifier is used to predict sentences.

✓ Place the proposed **dynamic attention mechanism** module (DAM) that
- Focus on regions with motion changes in the Sequence image feature extraction section, and

✓ Place the dynamic decoding based on
- Maximum cumulative probability (**MCD**) that
  - Segment the sequence features into
    - Single-label features of pseudo-labels in the alignment module section

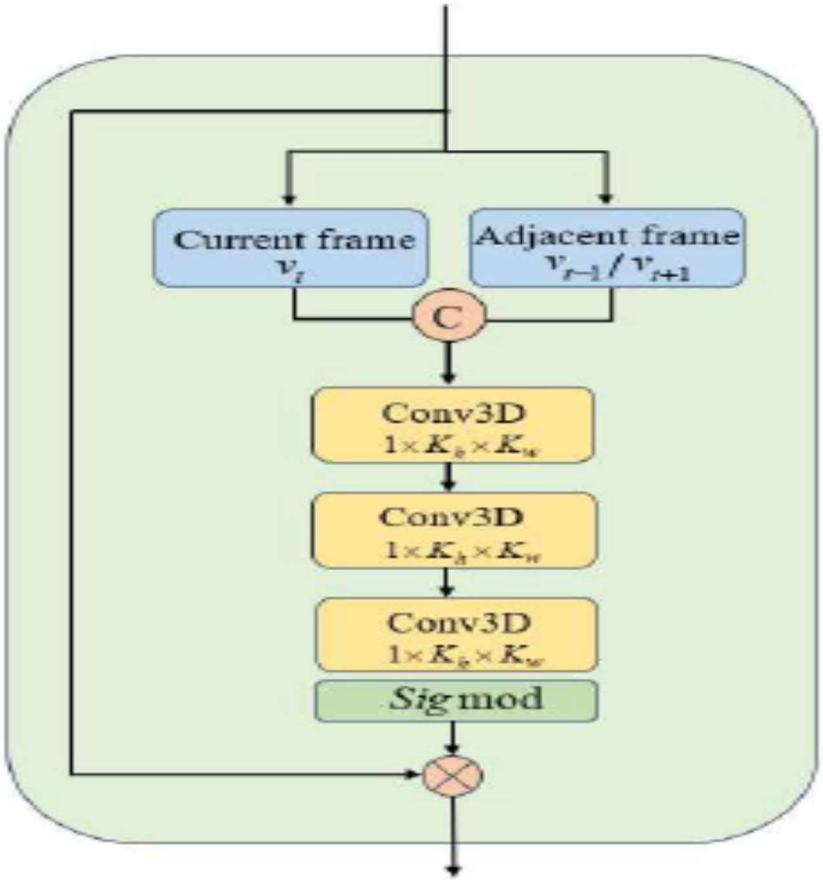
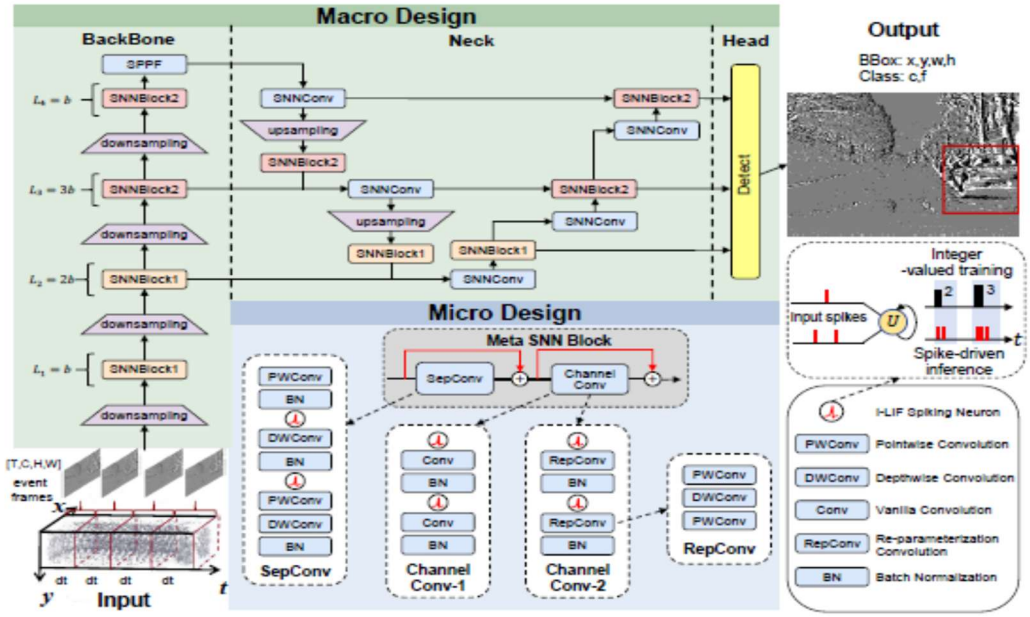


Fig. 3. The structure diagram of dynamic attention mechanism.

Transformer Net	2025-142	



Architecture. SpikeYOLO.

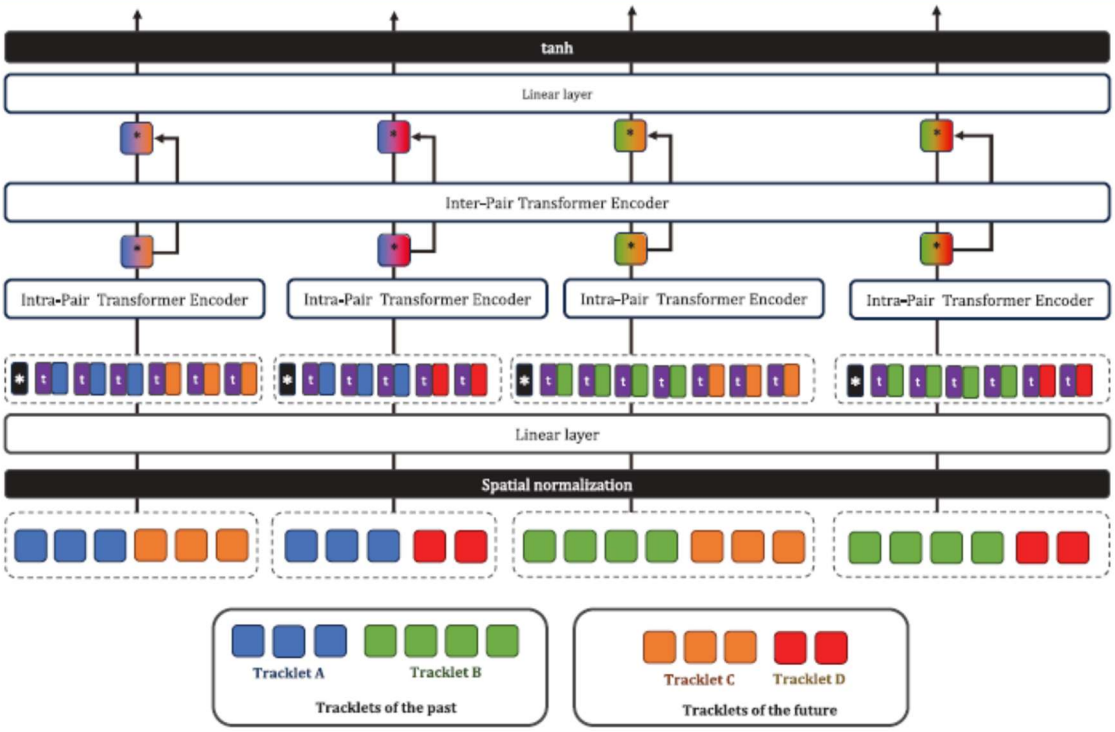
🔔 **Design.SpikeYOLO**

- two SNN blocks,
  - SNNBlock-1 and SNN-Block-2,
- kept other architectures remain as YOLOv8.

✓ SNN-Block-1 employs standard convolution within its ChannelConv (·) component
✓ SNNBlock- 2 utilizes re-parameterization convolution
- The difference between the two is the channel mixer module.

✓ In the low and high stages, SNN-Block-1 and SNN-Block-2, are used respectively.
✓ The spiking neuron is I-LIF, which
- Activates integer values during training while
- Converting them to binary spikes during inference

Transformer Net	**2025-143**	

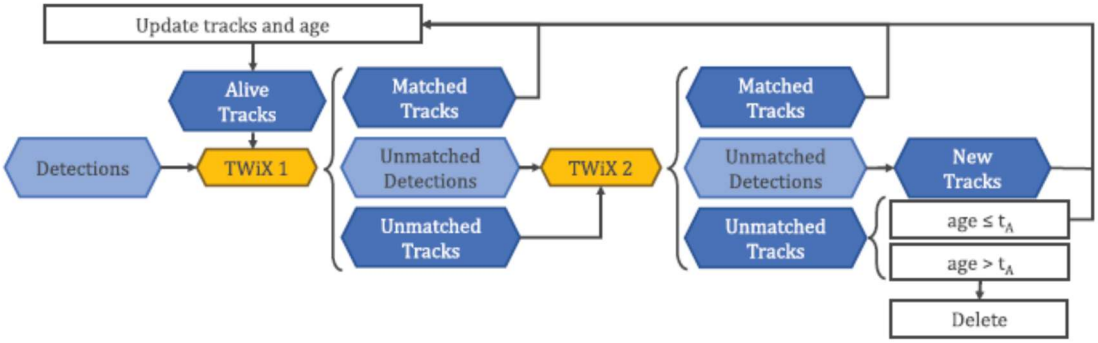


Architecture. TWiX (read from bottom to top)

1) First, pairs of tracklets are normalized and
2) Linearly projected
3) Then encoded with a Transformer

Encoder where

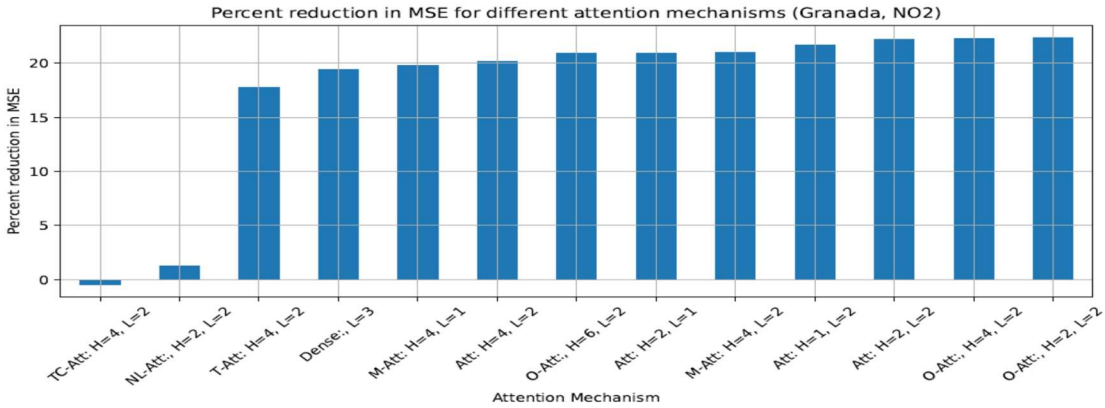
✓ Attention is applied on the temporal dimension.
✓ Refined representations are obtained with a second transformer encoder
✓ Which pays attention to all other pairs.
✓ Linear layer and a hyperbolic tangent function are used to compute an affinity score for each pair.



Tracker C-twix use a cascade matching pipeline for tracking

- The BIoU-computed matrix in C-BIoU is replaced by our TWiX module

Transformer Net	**2025-145**	



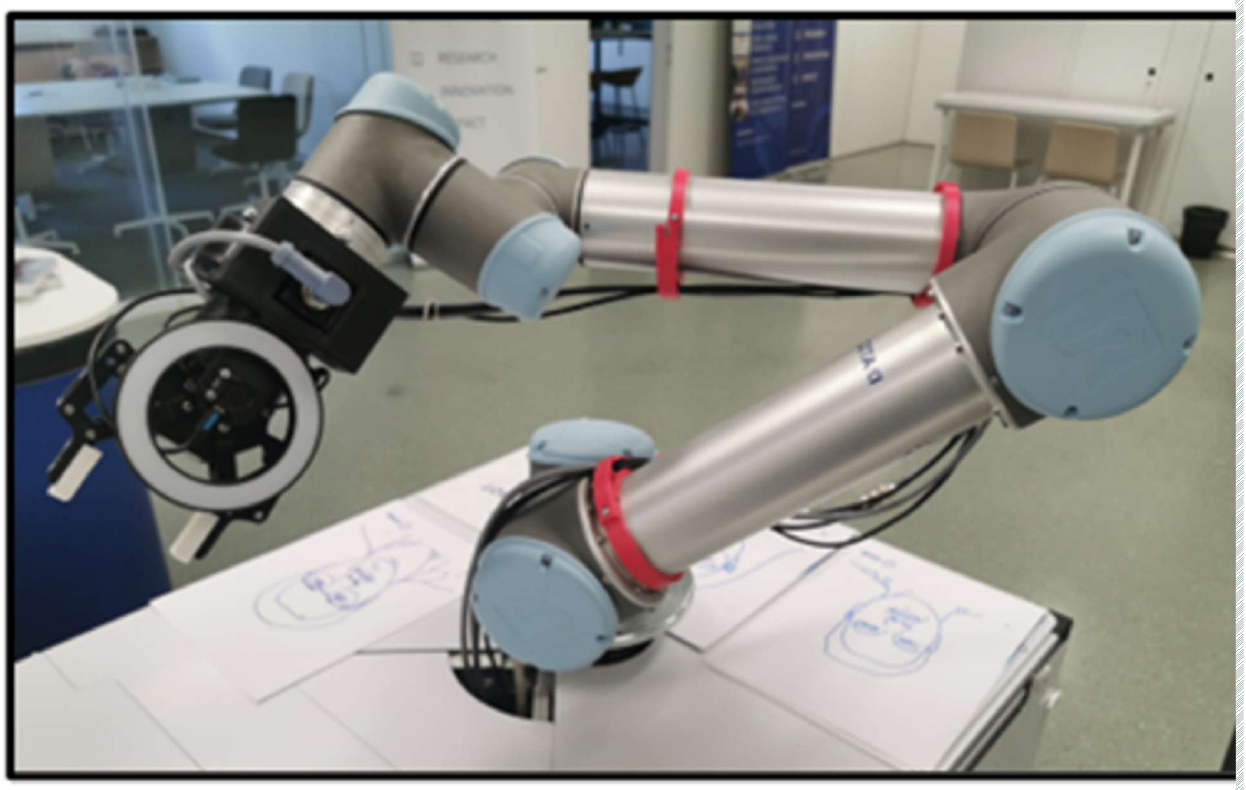
Percentage reduction in MSE as opposed to no attention mechanism for Granada and pollutant NO2

- ✓ **H : number of attention heads,**
- ✓ **L : number of layers in Q,K, V or dense layers**

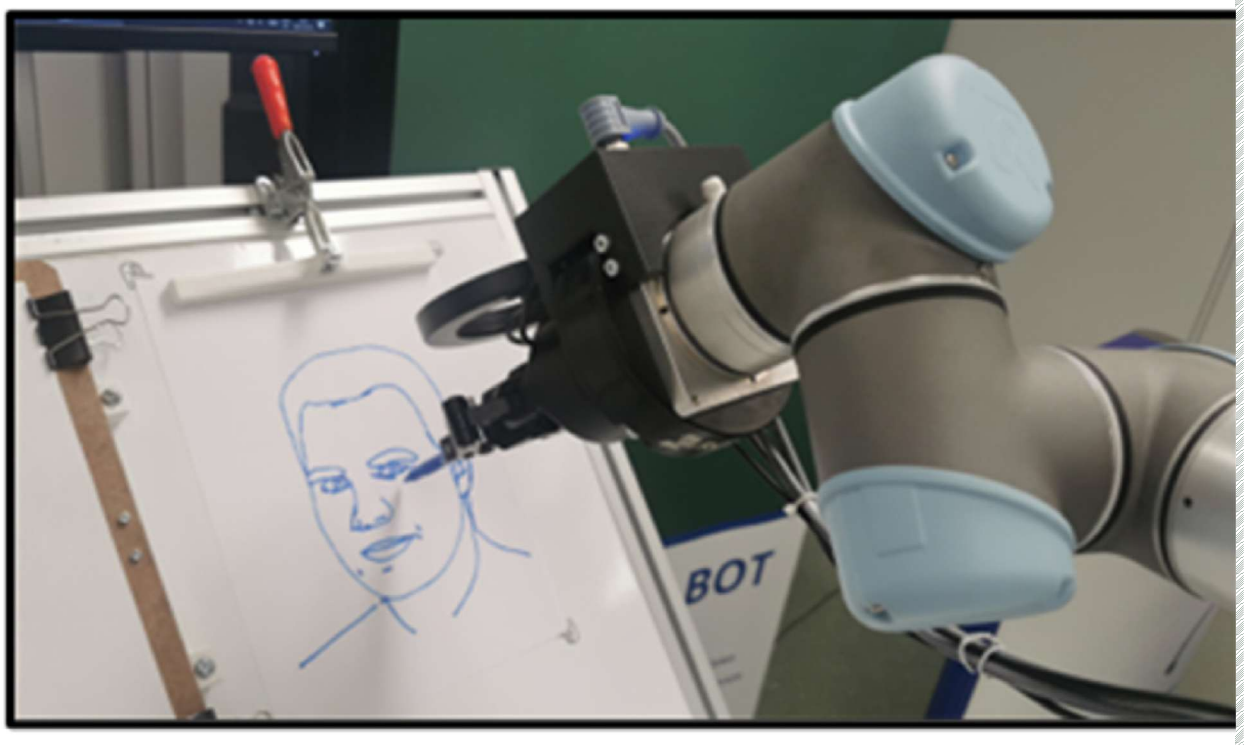
pollutant attention	NO2	O3	PM10	PM25	average
O-Att:, H=2, L=2	**22.4**	12.2	**3.7**	3.3	**10.4**
Att: H=2, L=2	22.2	13.7	1.6	2.6	10.0
O-Att:, H=4, L=2	22.3	13.1	-0.1	4.5	9.9
Att: H=4, L=2	20.2	**14.6**	0.0	1.4	9.1
Att: H=1, L=2	21.6	13.6	-7.0	**4.8**	8.3
M-Att: H=4, L=2	21.0	13.5	-3.3	-0.5	7.7
O-Att:, H=6, L=2	20.9	13.1	-0.0	-3.7	7.6
M-Att: H=4, L=1	19.8	11.9	-2.5	-0.3	7.2
Att: H=2, L=1	20.9	14.0	-6.6	-0.2	7.0

Percent reduction in MSE for different attention mechanisms for Granada

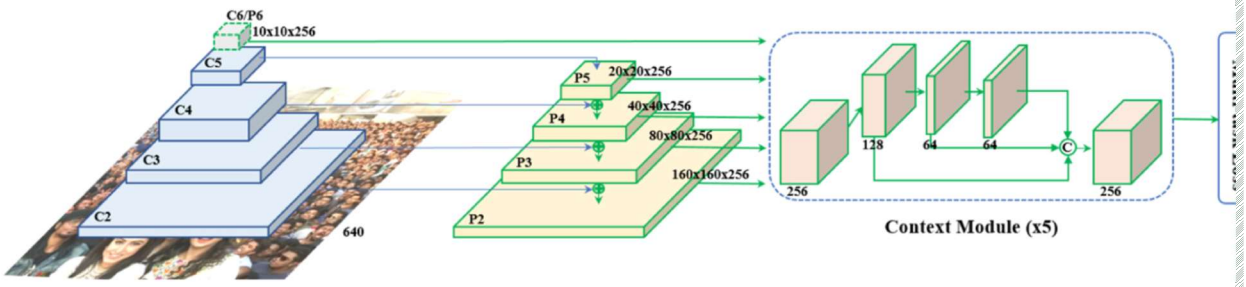
High interest of people for Piktor-o-bot presented in various public fairs and events.



Piktor-o-bot: The portrait drawing robot



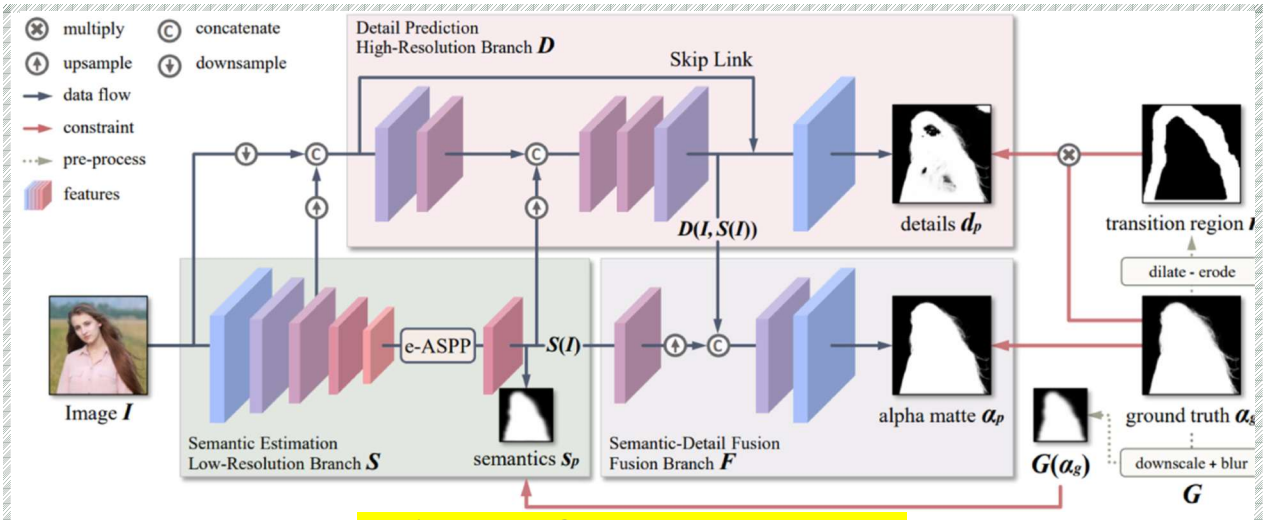
Drawing procedure of the Piktor-o-bot.



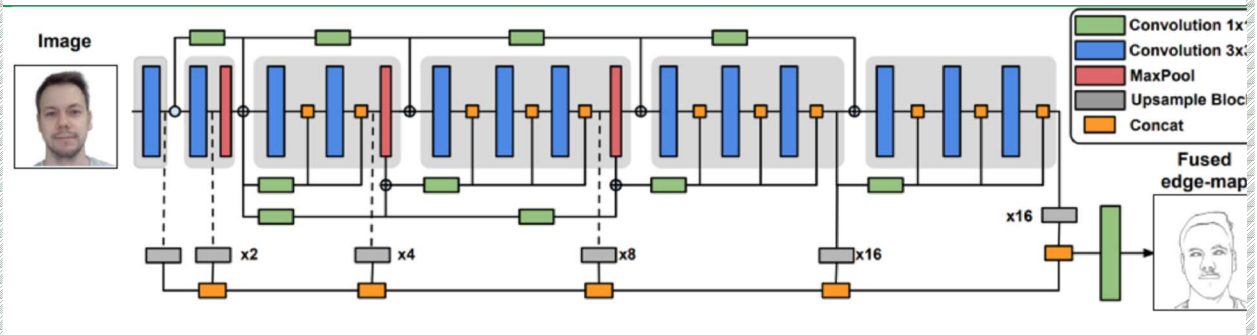
Architecture. RetinaFace deep neural network [14].

Architecture comprises

- ResNet-derived pyramid levels and
- Five context modules with
  - Deformable convolutional layers



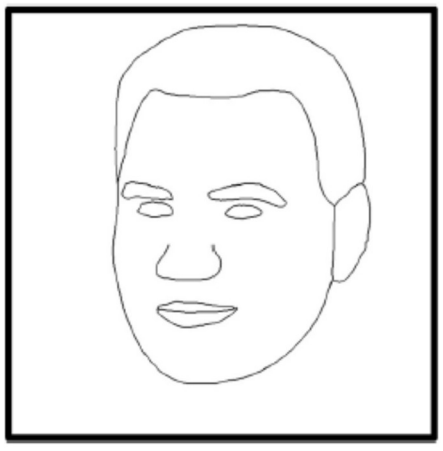
architecture. MODNet deep neural network



Architecture. Dexined deep neural network [16].

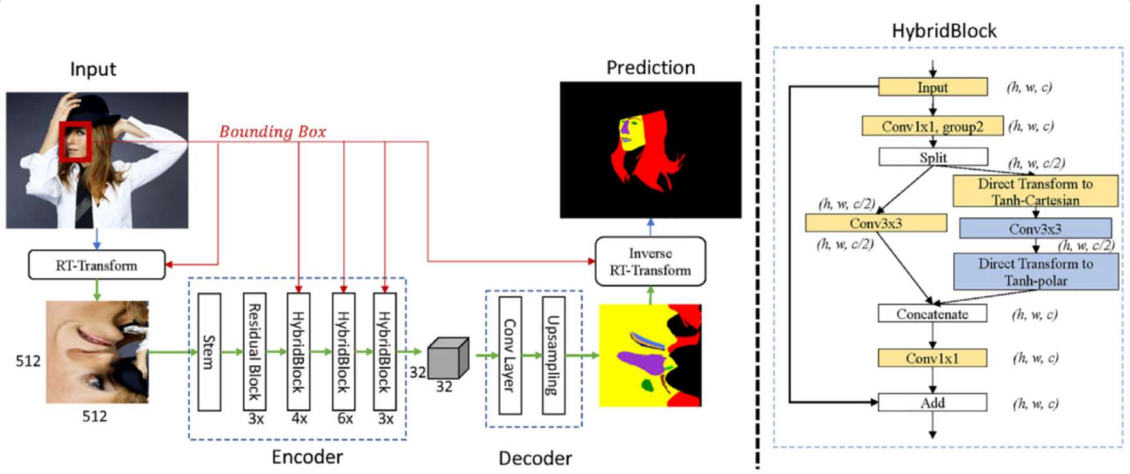
- Architecture consists of
- Convolution, pooling layers, and
- Upsampling blocks for scalability

Result of the edge detection using DexiNed deep NN



Result of face segmentation using RTNet deep neural network

Fig.



Process and architecture of RTNet deep neural network

✓ **Architecture consists of Residual Blocks and Hybrid Blocks**

Visualizing the sequencing algorithm

- Black colour: actual drawing
- Blue colour: paths travelled in the air by the robot

Final result: image drawn by the robot arm

Transformer Net	2025-149	

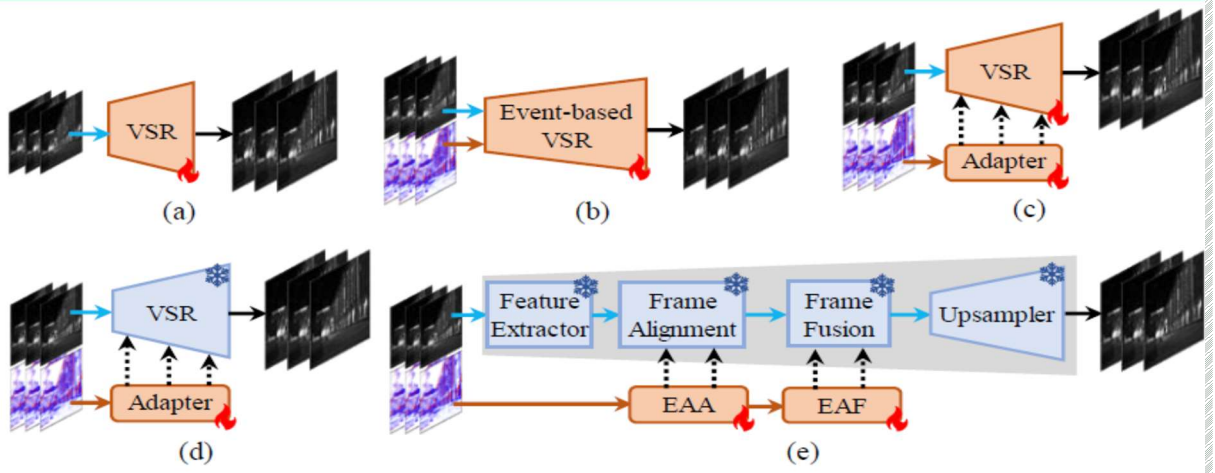
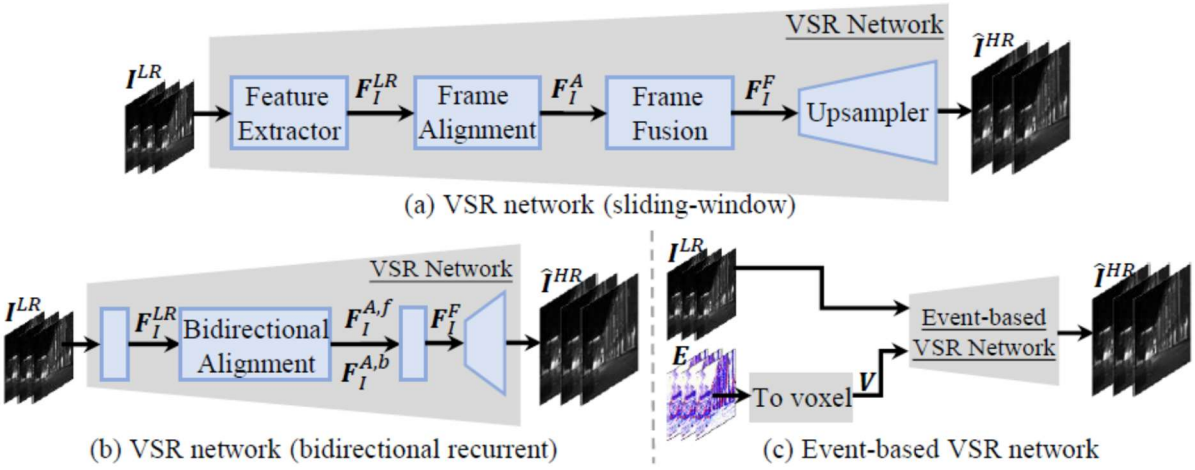


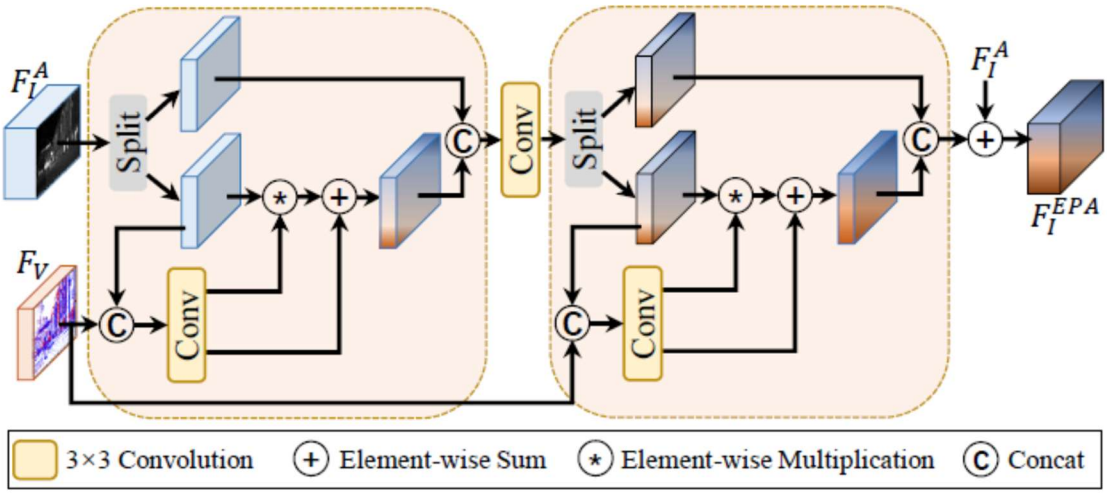
Fig. 1: Existing event-based VSR paradigm vs. our proposed Event-Adapted VSR.

- ✓ (a) Existing VSR methods can generate HR frames from LR observations.
- ✓ (b) Existing event-based VSR methods.
- ✓ (c) Incorporating the event data as a novel modality into an existing VSR method can be approached in two ways: full fine-tuning or retraining from scratch.
- ✓ (d)-(e) Our proposed event-adapted VSR. By leveraging the advantages ofparameter-efficient tuning and insights from well-designed VSR networks, our EATER achieves superior performance with minimal parameter updates

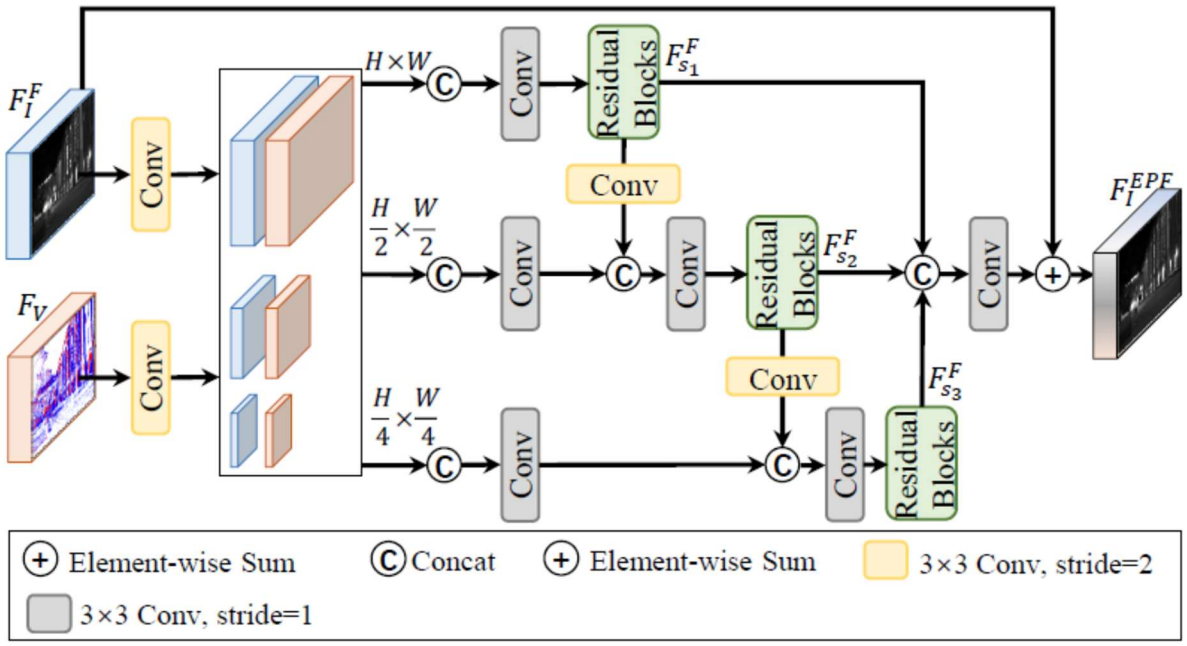


Structure. VSR network and the event-based VSR network

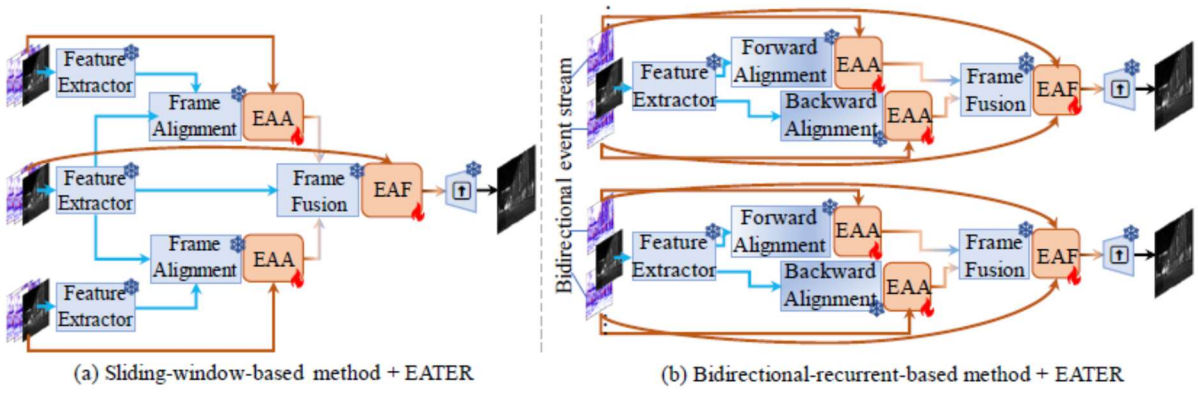
- ! For the sake of simplicity, bicubic upsampled residual connections and
- ! superscripts for the forward and backward event streams removed



Event-adapted _alignment (EAA) unit



Event-Adapted Fusion (EAF) unit



Overview of applying the proposed EATER to existing typical EGB-only VSR methods.

- ✓ (a) We apply the EAA unit and the EAF unit to the sliding-window-based method (e.g., EDVR [66]).
- ✓ (b) We apply the EAA unit and the EAF unit to the bidirectional recurrent-based method (e.g., BasicVSR [4]).

Transformer Net	**2025-150**	

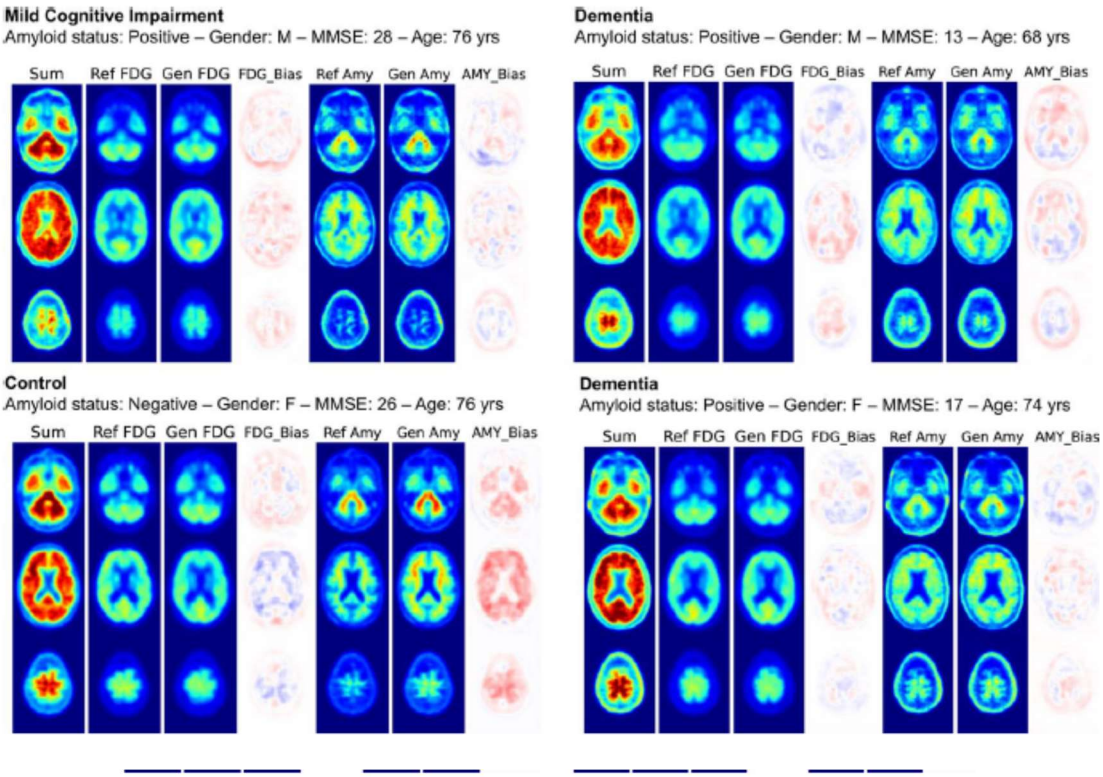
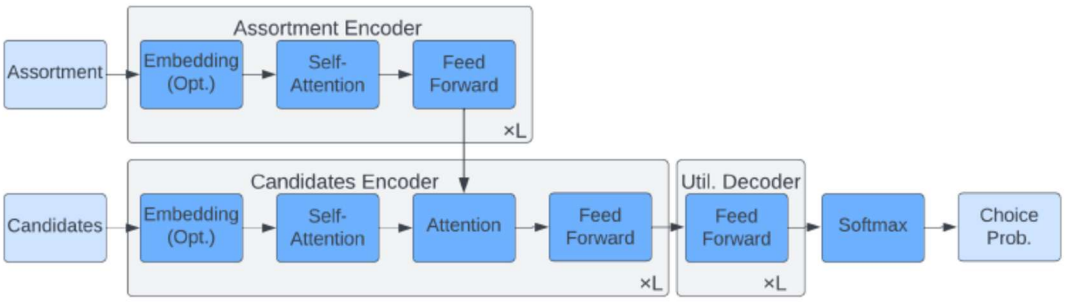
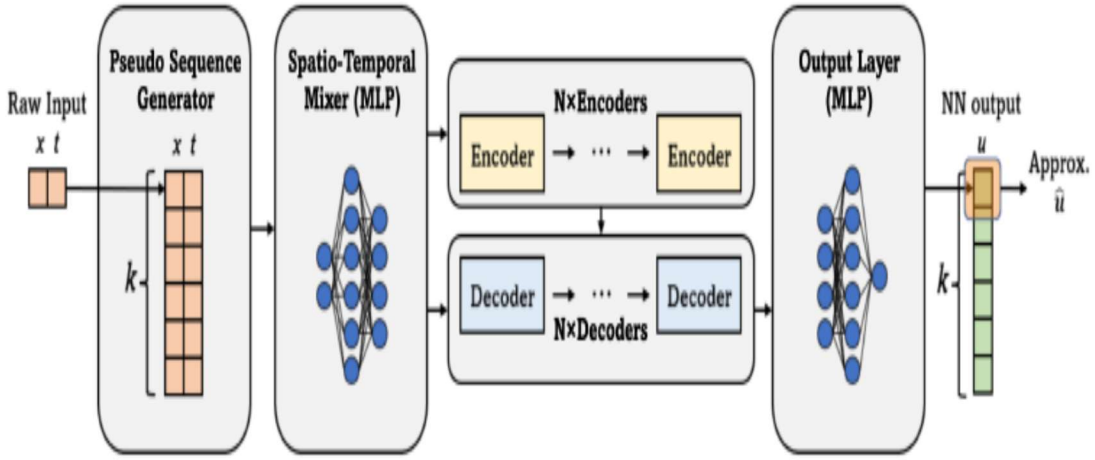


FIGURE 1. Four different subjects depicting various clinical statuses (normal, MCI, DEM) for FDG, FBP, and FMM. The "Sum" column displays dual tracer images (the combined FDG and Amyloid [FBP and FMM]), "Ref FDG" represents the reference FDG, "Gen FDG" denotes the generated FDG, and "FDG_Bias" signifies the difference map between reference and generated FDG. "Ref Amy" represents reference Amyloid, "Gen Amy" refers to generated Amyloid, and "Amy_Bias" indicates the difference map between reference and generated Amyloid. The image range spans from 0 to 3 SUVR, whereas the difference map range is between –0.2 and +0.2 SUVR. Subject-related metrics, including amyloid status, gender, MMSE, and age, are summarized atop each panel.

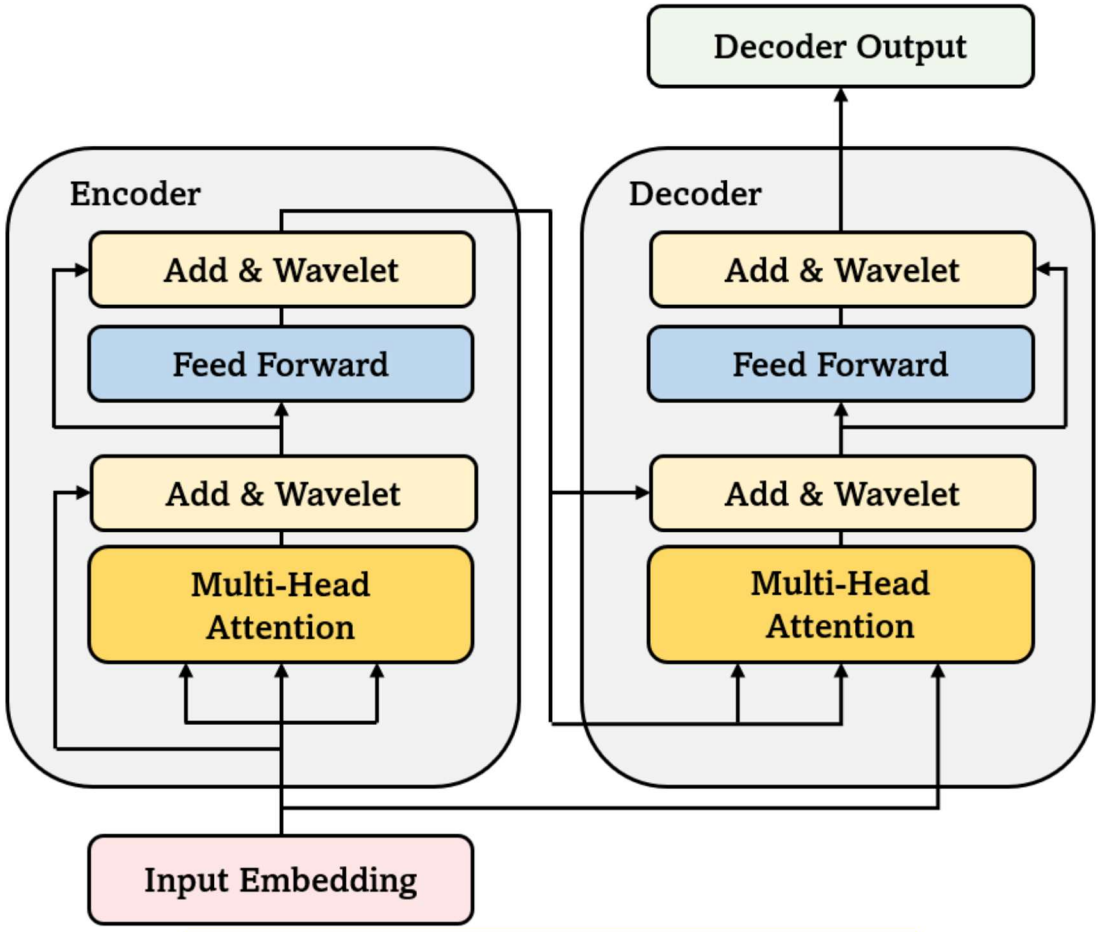
Transformer Net	**2025-151**	



Architecture of Transformer Choice Net (TCNet). omitted the residual connections within each sub-layer.



Architecture of proposed PINNsFormer. PINNsFormer generates a pseudo sequence based on pointwise input features. It outputs the corresponding sequential approximated solution. The first approximation of the sequence is the desired solution ˆu(x, t).



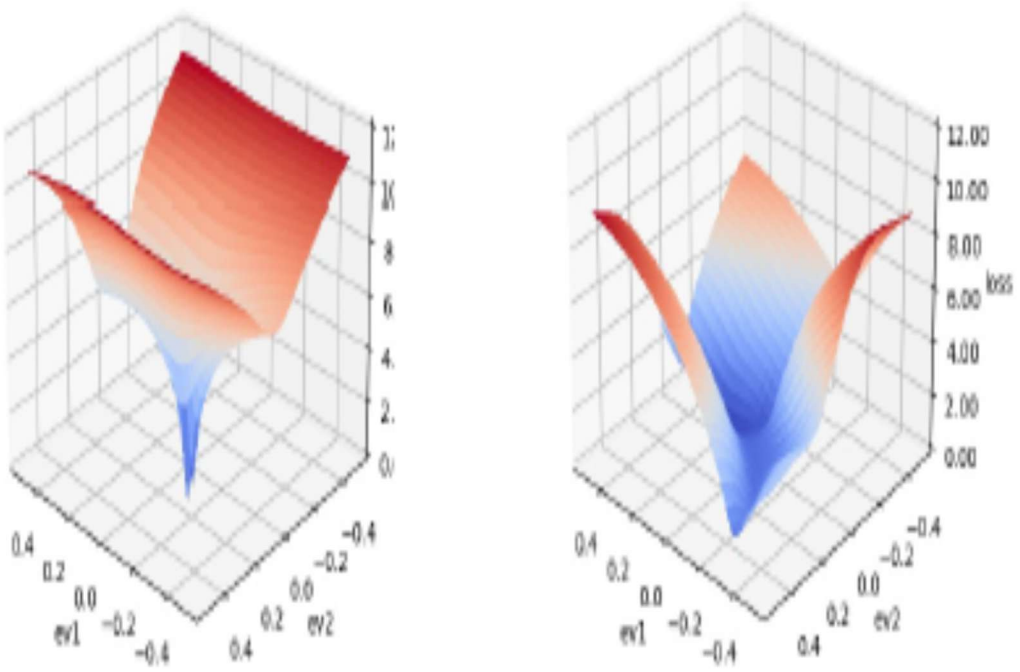
Architecture of PINNs-Former's Encoder-Decoder Layers

decoder is not equipped with self-attentions

Model	1D-Wave		
	Loss	rMAE	rRMSE
PINNs	1.93e-2	0.326	0.335
PINNsFormer	1.38e-2	0.270	0.283
PINNs + *NTK*	6.34e-3	0.140	0.149
PINNsFormer + *NTK*	**4.21e-3**	**0.054**	**0.058**

Results for solving the 1D-wave equation, incorporating NTK method.

- ✓ PINNsFormer combined with NTK outperforms all other methods on all metrics



Visualization of the loss landscape for PINNs (left) and PINNsFormer (right) on a logarithmic scale

The loss landscape of PINNsFormer is significantly smoother than conventional PINNs.

Results for solving Navier-Stokes equation

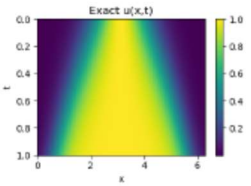
Model	Navier-Stokes		
	Loss	rMAE	rRMSE
PINNs	6.72e-5	13.08	9.08
QRes	2.24e-4	6.41	4.45
FLS	9.54e-6	3.98	2.77
PINNsFormer	**6.66e-6**	**0.384**	**0.280**

✓ **PINNsFormer outperforms all baselines on all metrics.**

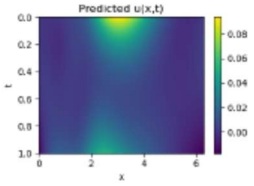
Results for solving convection and 1D-reaction equations using Transformer architecture with different activation functions

Activation	Convection			1D-Reaction		
	Loss	rMAE	rRMSE	Loss	rMAE	rRMSE
ReLU	0.5256	1.001	1.001	0.2083	0.994	0.996
Sigmoid	0.1618	1.112	1.223	0.1998	0.991	0.993
Sin	0.3159	1.074	1.141	4.9e-6	0.017	0.032
ReLU+LN	0.7818	1.001	1.002	0.2028	0.992	0.993
Sigmoid+LN	0.0549	0.941	0.967	0.2063	0.992	0.990
Sin+LN	0.3219	1.083	1.156	4.7e-6	0.016	0.033
Wavelet	**3.7e-5**	**0.023**	**0.027**	**3.0e-6**	**0.015**	**0.030**
Wavelet+LN	NaN	NaN	NaN	3.9e-6	0.018	0.037

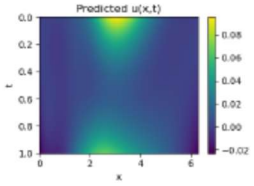
- PINNsFormer (withWavelet activation) consistently outperforms all other activation functions in terms of training loss, rMAE, and rRMSE



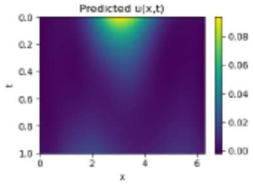
(a) ground truth solution for 1D-reaction equation



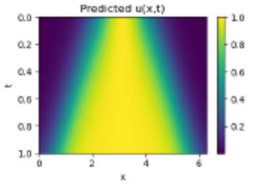
(b) PINNs prediction



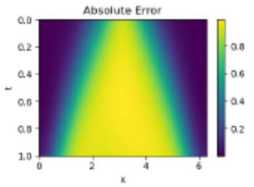
(c) FLS prediction



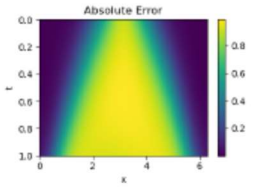
(d) QRes prediction



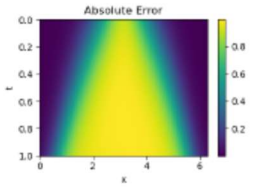
(e) PINNsFormer prediction



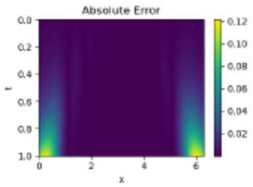
(f) PINNs error



(g) FLS error

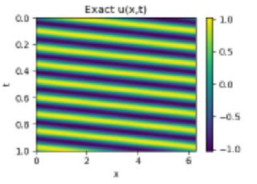


(h) QRes error

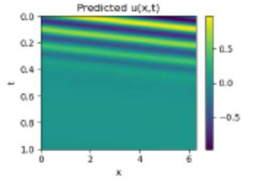


(i) PINNsFormer error

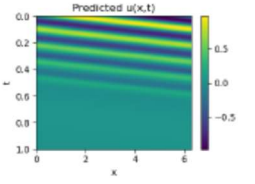
Ground truth solution, predictions, and absolute errors (up to bottom) of PINNs, PINNsFormer, PINNs+NTK, PINNsFormer+NTK (left to right) over 1D-reaction equation.



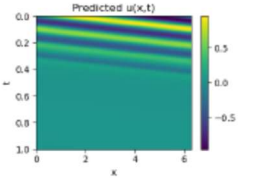
(a) ground truth solution for convection equation



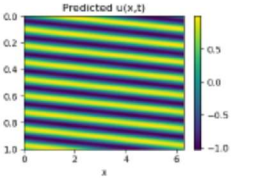
(b) PINNs prediction



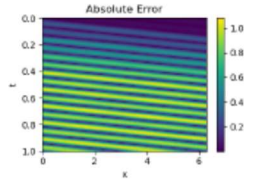
(c) FLS prediction



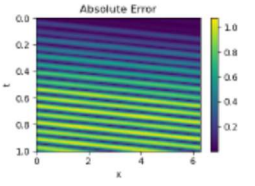
(d) QRes prediction



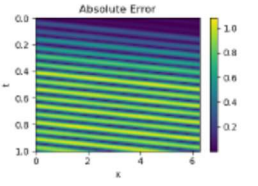
(e) PINNsFormer prediction



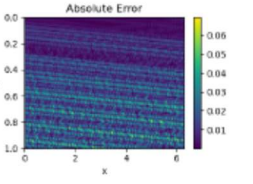
(f) PINNs error



(g) FLS error



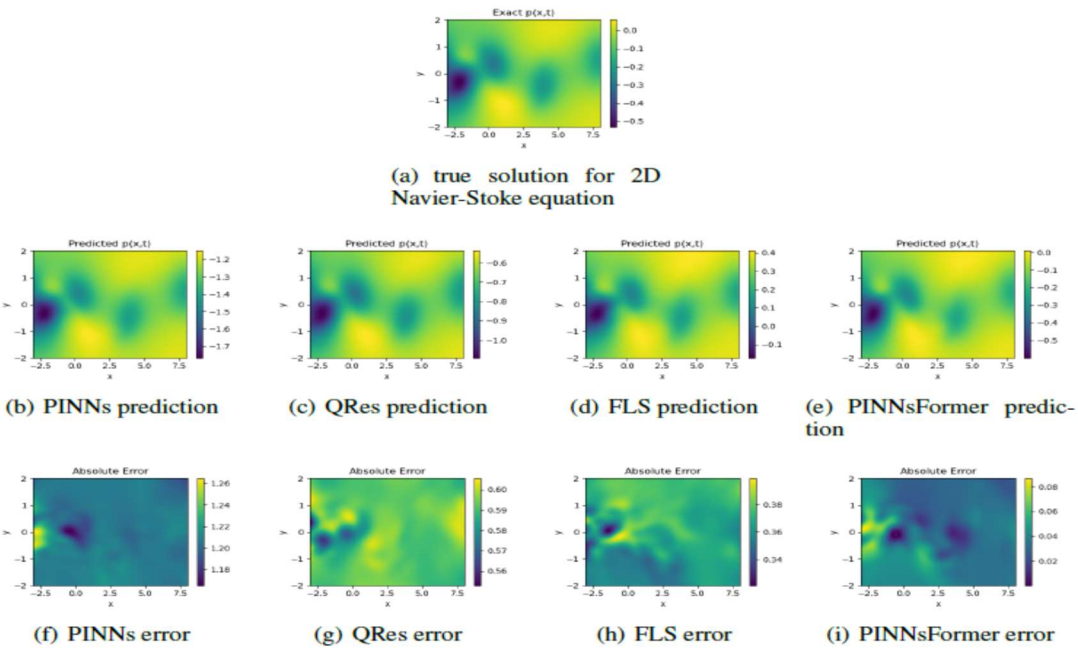
(h) QRes error



(i) PINNsFormer error

Ground truth solution, predictions, and absolute errors (up to bottom) of PINNs, FLS,

QRes, PINNsFormer (left to right) over convection equation



(a) true solution for 2D Navier-Stoke equation

(b) PINNs prediction (c) QRes prediction (d) FLS prediction (e) PINNsFormer prediction

(f) PINNs error (g) QRes error (h) FLS error (i) PINNsFormer error

Ground truth solution, predictions, and absolute errors (up to bottom) of PINNs and PINNsFormer (left to right) over 2D Navier-Stokes equation

Transformer Net	**2025-155**	

Component	$\mu_{\mathbf{x}_{\text{out}}}$	$\sigma^2_{\mathbf{x}_{\text{out}}}$	$\sigma^2_{\mathbf{g}_{\text{in}}}$	$\mathbf{r}^{\mathbf{l}}_{\mathbf{x}_{\text{out}}}$	$\mathbf{r}^{\mathbf{l}}_{\mathbf{g}_{\text{in}}}$
Embeddings	0	$\sum \sigma^2_{w_{\text{embd}}}$	-	$\frac{\pi^2}{18 * \log(\lvert V\rvert)^2} + \frac{2}{9}$	-
Linear ($d_{\text{in}} \to d_{\text{out}}$)	0	$d_{\text{in}}\sigma^2_w(\sigma^2_{x_{\text{in}}} + \mu^2_{x_{\text{in}}})$	$d_{\text{out}}\sigma^2_w\sigma^2_{g_{\text{out}}}$	$\frac{r^l_{x_{\text{in}}} + \mu^2_{x_{\text{in}}}/\sigma^2_{x_{\text{in}}}}{1 + \mu^2_{x_{\text{in}}}/\sigma^2_{x_{\text{in}}}}$	$r^l_{g_{\text{out}}}$
ReLU	$\frac{\sigma_{x_{\text{in}}}}{\sqrt{(2\pi)}}$	$\frac{(\pi-1)}{(2\pi)}\sigma^2_{x_{\text{in}}}$	$\frac{1}{2}\sigma^2_{g_{\text{out}}}$	$0.7r^l_{x_{\text{in}}} + 0.3{r^l_{x_{\text{in}}}}^2$	$(\frac{1}{2} + \frac{\sin^{-1}(r^l_{x_{\text{in}}})}{\pi})\mathrm{r}^{\mathrm{l}}_{g_{\text{out}}}$
LayerNorm (d)	0	1	$\frac{\sigma^2_{g_{\text{out}}}}{\sigma^2_{x_{\text{in}}}}$	$r^l_{x_{\text{in}}}$	$r^l_{g_{\text{out}}}$
Dropout (p)	$\mu_{x_{in}}$	$\frac{\sigma^2_{x_{\text{in}}} + p\mu^2_{x_{\text{in}}}}{1-p}$	$\frac{1}{1-p}\sigma^2_{g_{\text{out}}}$	$\frac{r^l_{x_{\text{in}}}(1-p)}{1 + p\mu^2_{x_{\text{in}}}/\sigma^2_{x_{\text{in}}}}$	$(1-p)r^l_{g_{\text{out}}}$
SHA-without V	0	$r^l_{x_{\text{in}}}\sigma^2_{x_{\text{in}}}$	$r^l_{g_{\text{out}}}\sigma^2_{g_{\text{out}}}$	1	1
Softmax	$\frac{1}{L}$	$\frac{e^{(1-r^d_{x_{\text{in}}})\sigma^2_{x_{\text{in}}}} - 1}{L^2}$	$\frac{e^{(1-r^d_{x_{\text{in}}})\sigma^2_{x_{\text{in}}}}}{L^2}\sigma^2_{g_{\text{out}}}$	-	-

Signal propagation for forward and backward passes through components of a transformer