WEEK-1

COMPUTER VISION RESEARCH PAPERS OF THE WEEK 2022

Computer Vision Group(CVG)

Follow Share

/ashishpatel2604

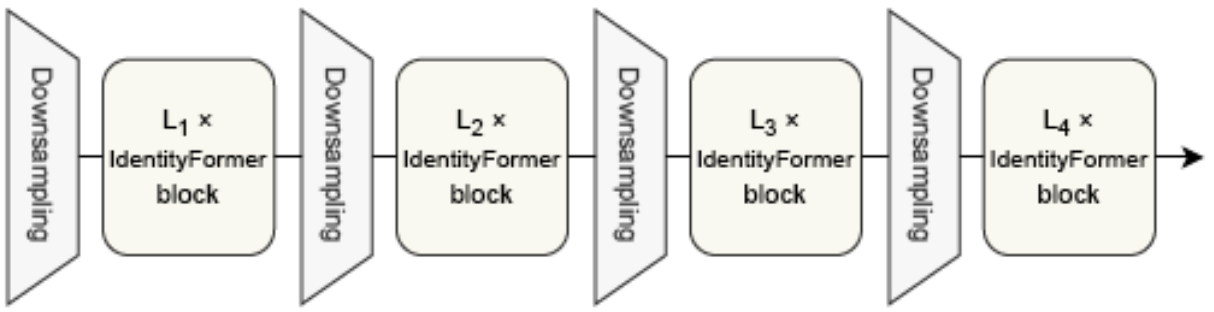
METAFORMER

COMPUTER VISION RESEARCH PAPERS OF THE WEEK 2022

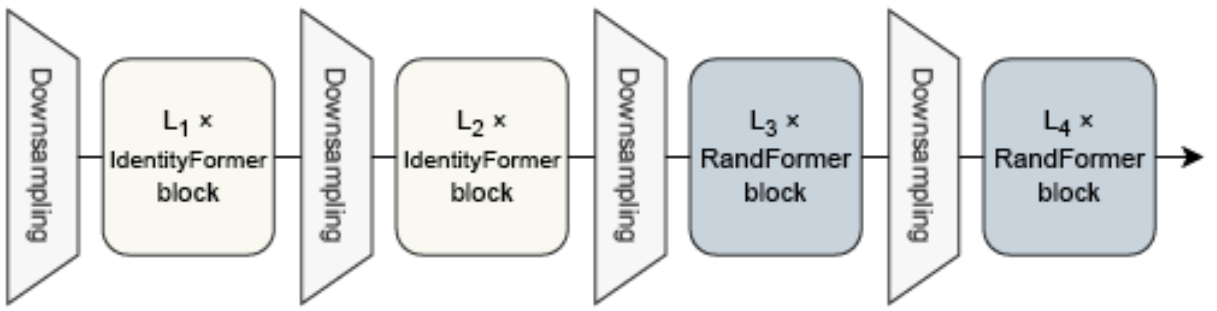
Intput

$D_0 : H \times W \times 3$

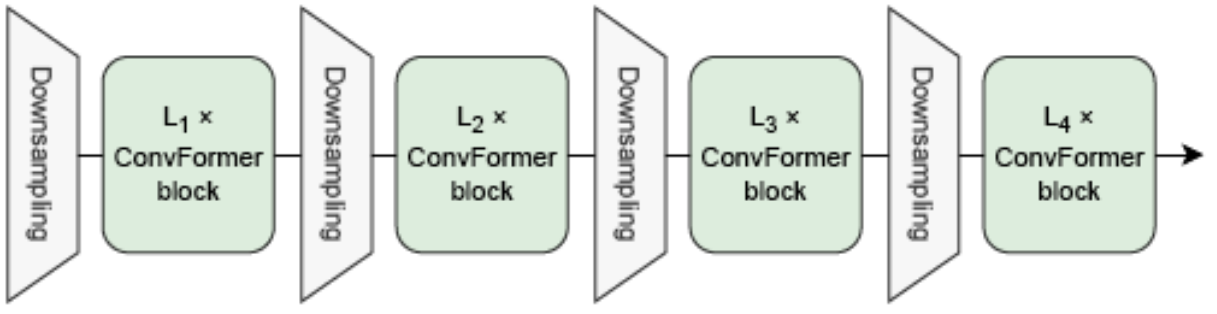
Stage 1 $D_1 : \frac{H}{4} \times \frac{W}{4} \times C_1$ Stage 2 $D_2 : \frac{H}{8} \times \frac{W}{8} \times C_2$ Stage 3 $D_3 : \frac{H}{16} \times \frac{W}{16} \times C_3$ Stage 4 $D_4 : \frac{H}{32} \times \frac{W}{32} \times C_4$



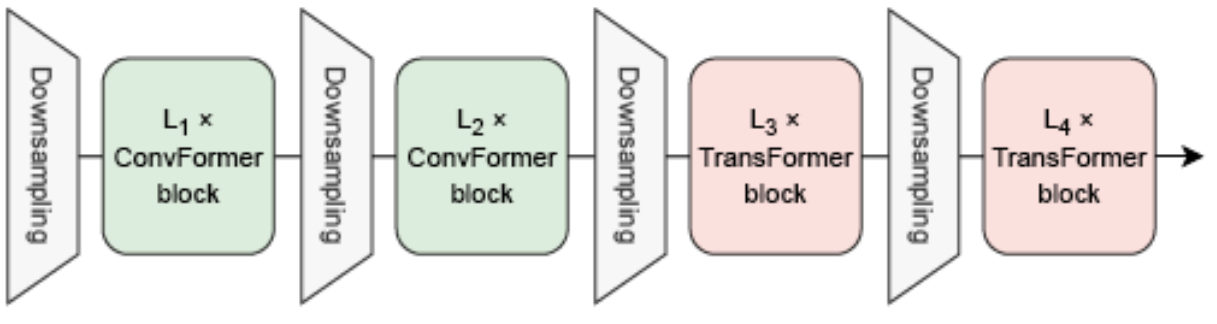
(a) Overall IdentityFormer framework



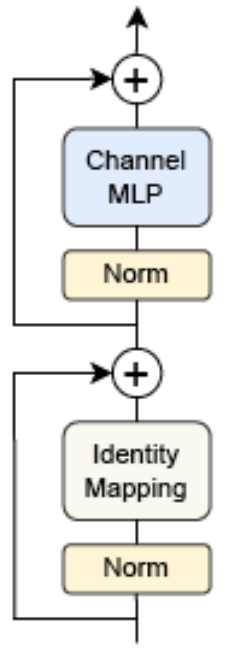
(b) Overal RandFormer framework



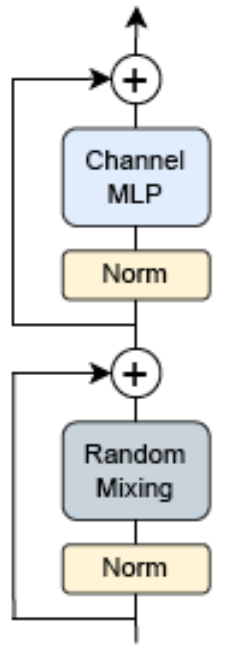
(c) Overal ConvFormer framework



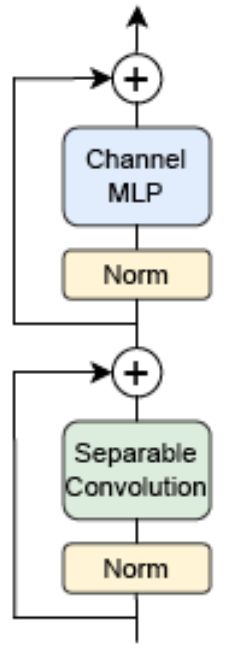
(d) Overal CAFormer framework



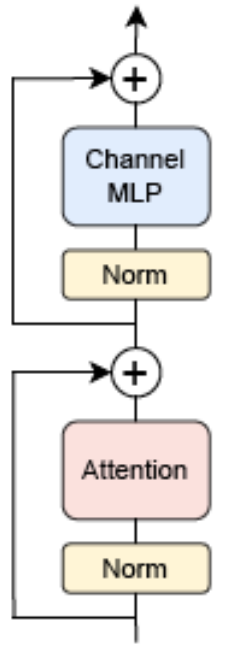
(e) IdentityFormer block



(f) RandFormer block



(g) ConvFormer block



(h) Transformer block

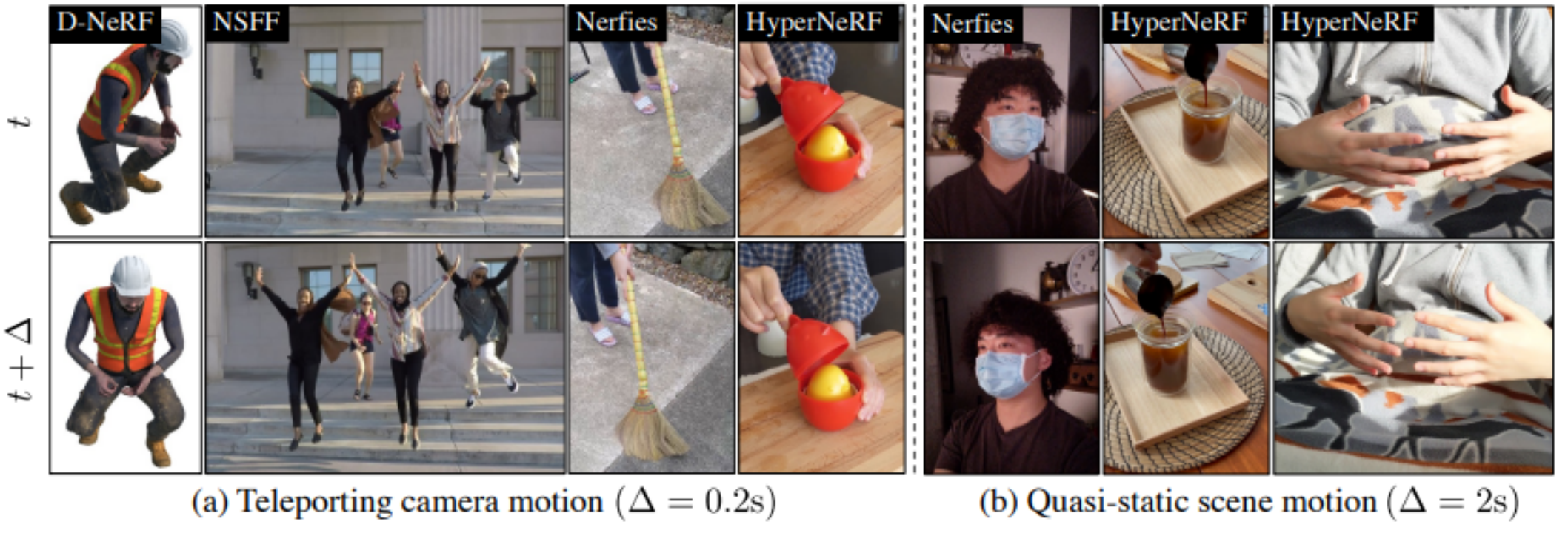
Computer Vision Group(CVG)

Follow Share

/ashishpatel2604

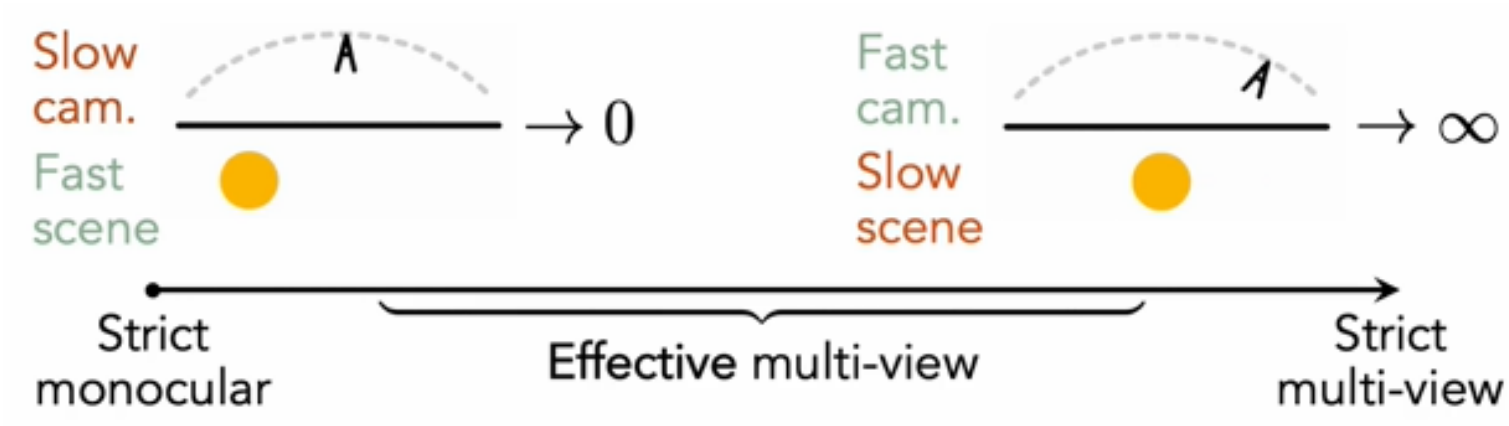
MONOCULAR DYNAMIC VIEW SYNTHESIS

COMPUTER VISION RESEARCH PAPERS OF THE WEEK 2022



(a) Teleporting camera motion ($\Delta = 0.2$s) (b) Quasi-static scene motion ($\Delta = 2$s)

Effective Multi-view



A monocular video contains *effective* multi-view cues when the camera moves much faster than the scene, even though the underlying scene is observed only once at each time step.

GFN : GALLERY FILTER NETWORK

COMPUTER VISION RESEARCH PAPERS OF THE WEEK 2022

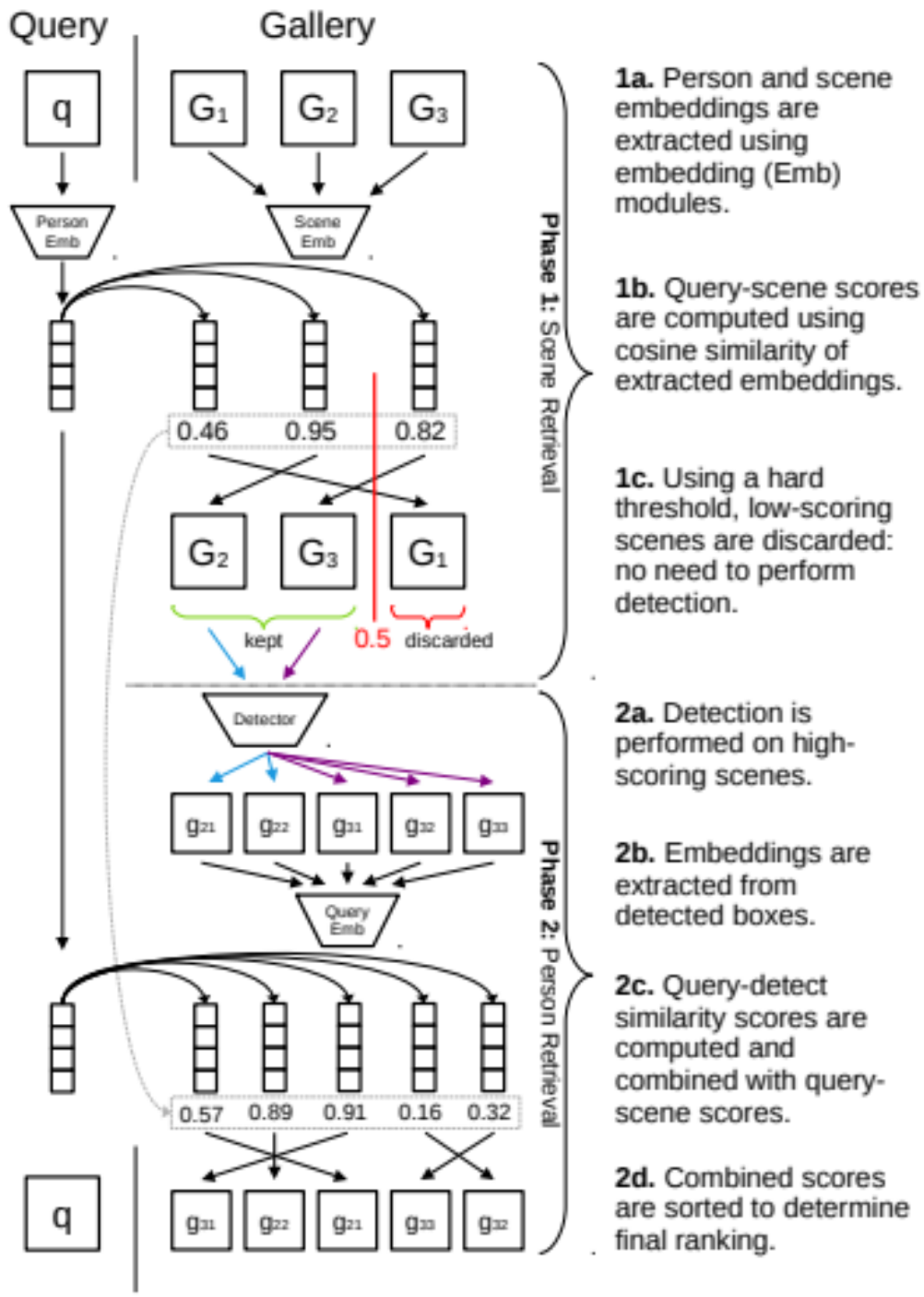


Figure 1: An illustration of our proposed two-phase retrieval inference pipeline. In the first phase, the Gallery Filter Network discards scenes unlikely to contain the query person. The second phase is the standard person retrieval process, in which persons are detected, corresponding embeddings extracted, and these embeddings are compared to the query to produce a ranking.

Method	Backbone	CUHK-SYSU		PRW	
		mAP	top-1	mAP	top-1
Two-step					
IDE [44]	ResNet50	-	-	20.5	48.3
MGTS [6]	VGG16	83.0	83.7	32.6	72.1
CLSA [20]	ResNet50	87.2	88.5	38.7	65.0
IGPN [10]	ResNet50	90.3	91.4	47.2	87.0
RDLR [14]	ResNet50	93.0	94.2	42.9	70.2
TCTS [36]	ResNet50	93.9	95.1	46.8	87.5
End-to-end					
OIM [39]	ResNet50	75.5	78.7	21.3	49.4
IAN [38]	ResNet50	76.3	80.1	23.0	61.9
NPSM [26]	ResNet50	77.9	81.2	24.2	53.1
RCAA [4]	ResNet50	79.3	81.3	-	-
CTXG [41]	ResNet50	84.1	86.5	33.4	73.6
QEEPS [29]	ResNet50	88.9	89.1	37.1	76.7
APNet [45]	ResNet50	88.9	89.3	41.9	81.4
HOIM [5]	ResNet50	89.7	90.8	39.8	80.4
BINet [9]	ResNet50	90.0	90.7	45.3	81.7
NAE+ [7]	ResNet50	92.1	92.9	44.0	81.1
PGSFL [19]	ResNet50	92.3	94.7	44.2	85.2
DKD [43]	ResNet50	93.1	94.2	50.5	87.1
DMRN [15]	ResNet50	93.2	94.2	46.9	83.3
AGWF [13]	ResNet50	93.3	94.2	53.3	87.7
AlignPS [40]	ResNet50	94.0	94.5	46.1	82.1
SeqNet [22]	ResNet50	93.8	94.6	46.7	83.4
SeqNet+CBGM [22]	ResNet50	94.8	95.7	47.6	87.6
COAT [42]	ResNet50	94.2	94.7	53.3	87.4
COAT+CBGM [42]	ResNet50	94.8	95.2	54.0	89.1
MHGAM [21]	ResNet50	94.9	95.9	47.9	88.0
PSTR [3]	ResNet50	94.2	95.2	50.1	87.9
PSTR [3]	PVTv2-B2	95.2	96.2	56.5	89.7
SeqNeXt (ours)	ResNet50	94.1	94.7	50.8	86.0
SeqNeXt+GFN (ours)	ResNet50	94.7	95.3	51.3	90.6
SeqNeXt (ours)	ConvNeXt	96.1	96.5	57.6	89.5
SeqNeXt+GFN (ours)	ConvNeXt	**96.4**	**97.0**	**58.3**	**92.4**

Table 2: Standard performance metrics mAP and top-1 accuracy on the benchmark CUHK-SYSU and PRW datasets are compared for state-of-the-art *two-step* and *end-to-end* models. ConvNeXt backbone = ConvNeXt Base.

DUAL-MIL

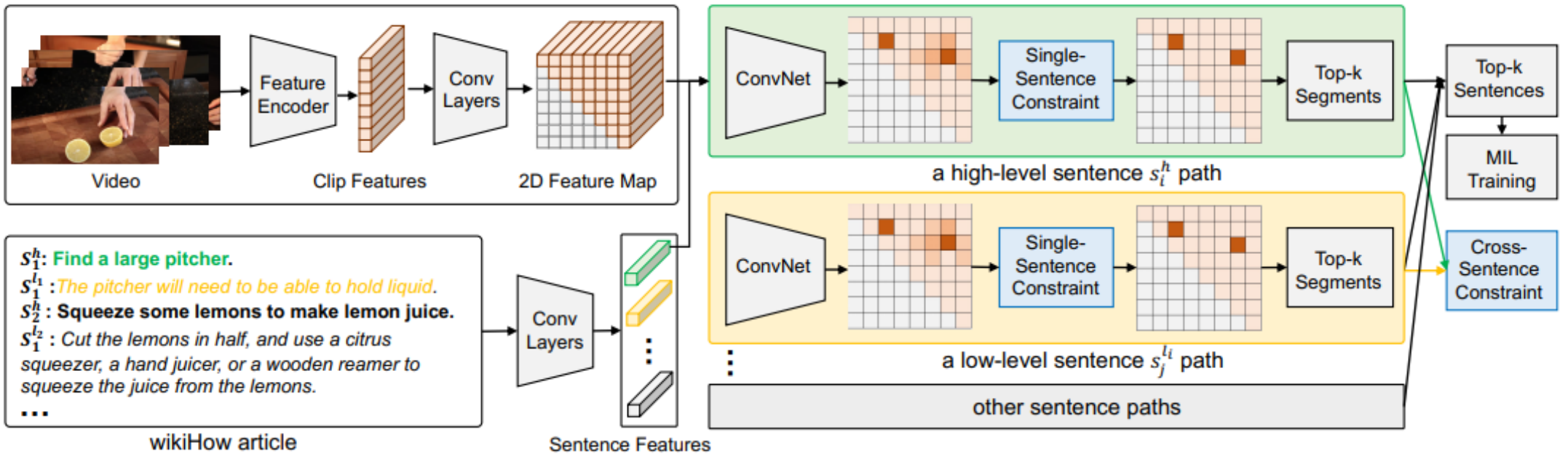


Figure 4: The overview of the article grounding architecture with the proposed DualMIL.

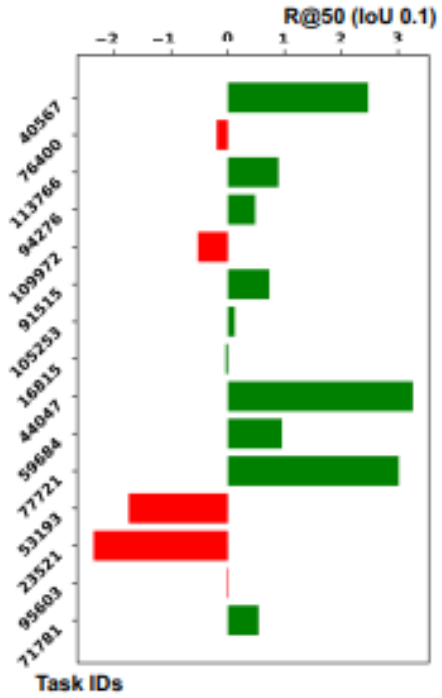


Figure 5: Performance gains (%) between models w/ & w/o structure-NMS. Task ids are ranked by the agreement between the order of groundable sentences and GT segments (cf. Appendix).

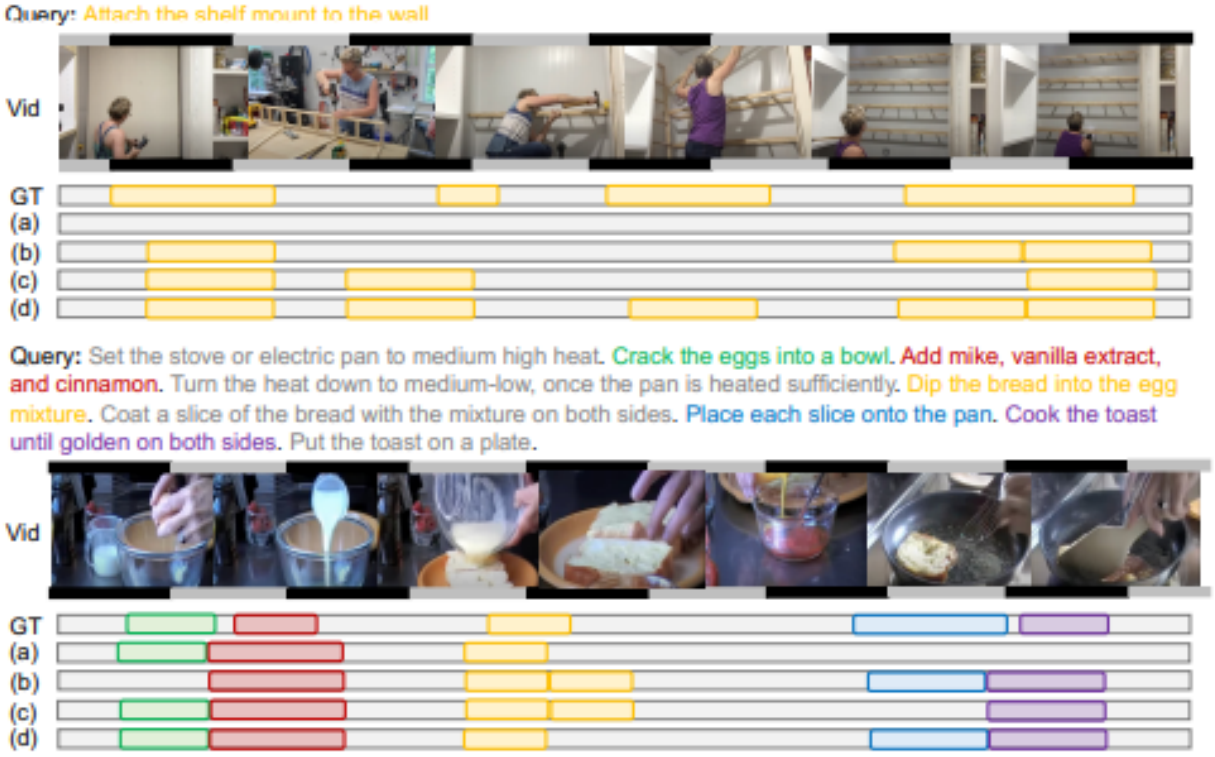


Figure 6: **Upper**: An example of a query with multiple GT segments. **Below**: Given a video and an article (only high-level sentences), GT segments of all groundable sentences are shown (with corresponding colors). (a) - (d) denotes baseline, baseline w/ single-sentence const., baseline w/ cross-sentence const., and full model, respectively. Top-50 predictions overlapped with GT are shown.

TDS-NET

COMPUTER VISION RESEARCH PAPERS OF THE WEEK 2022

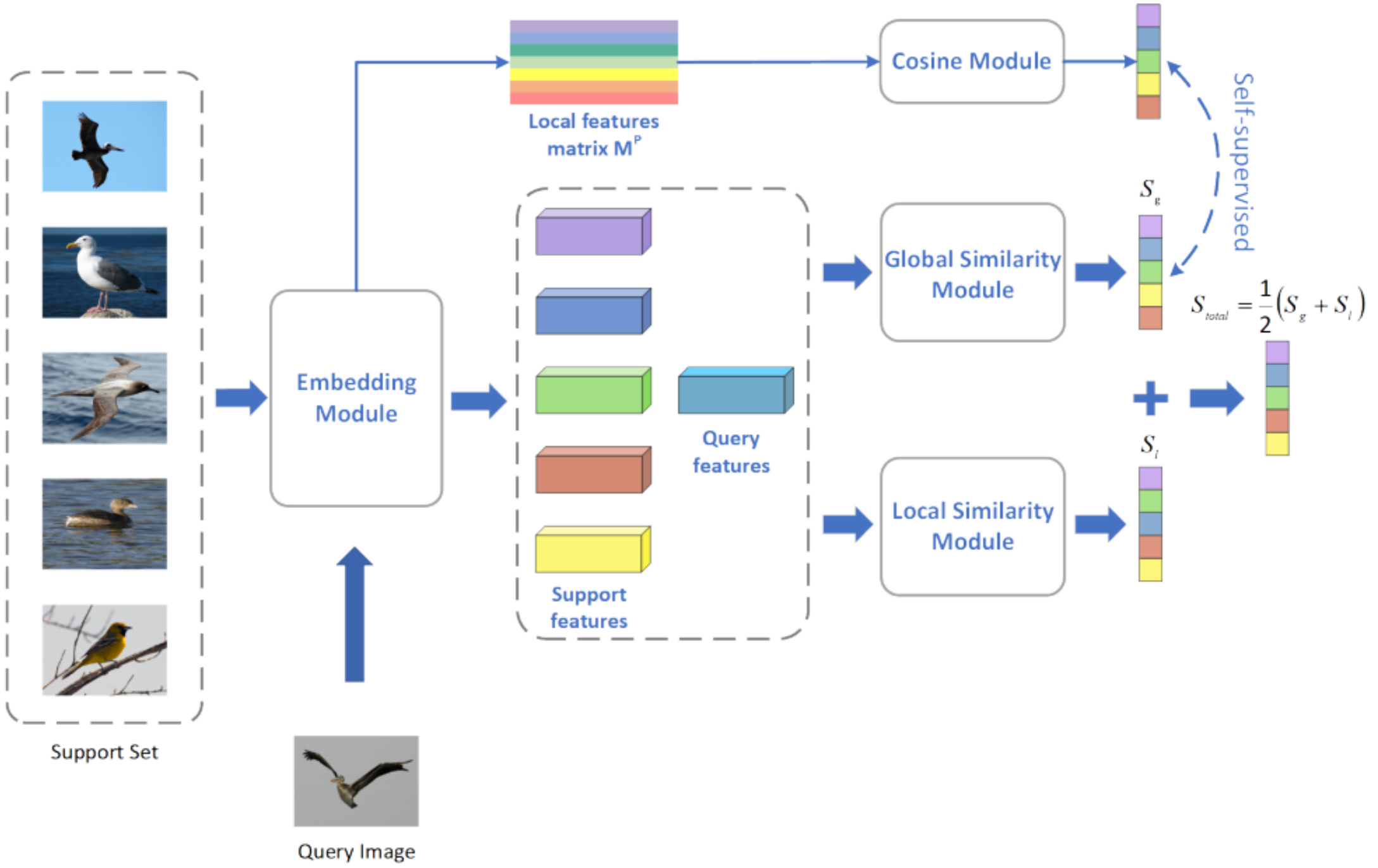
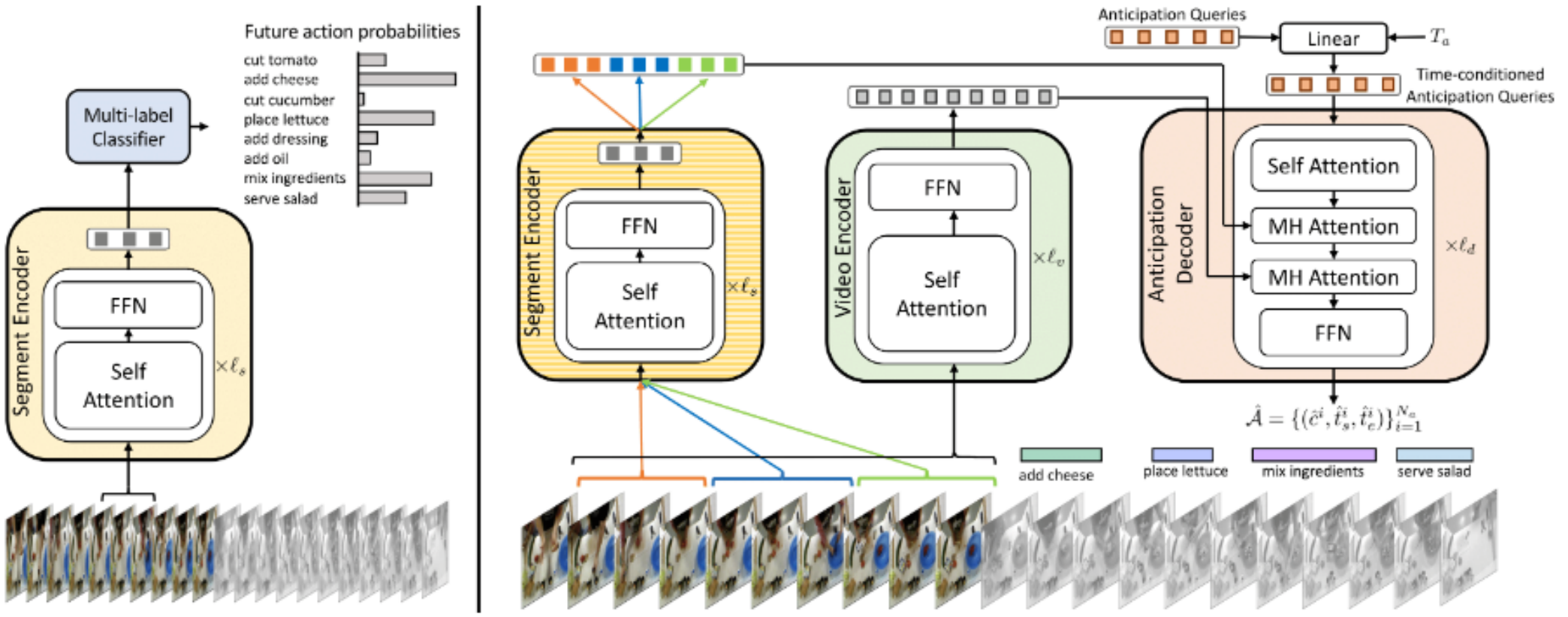


Fig. 1. Framework of the proposed TDSNet.

LONG-TERM ACTION ANTICIPATION

COMPUTER VISION RESEARCH PAPERS OF THE WEEK 2022



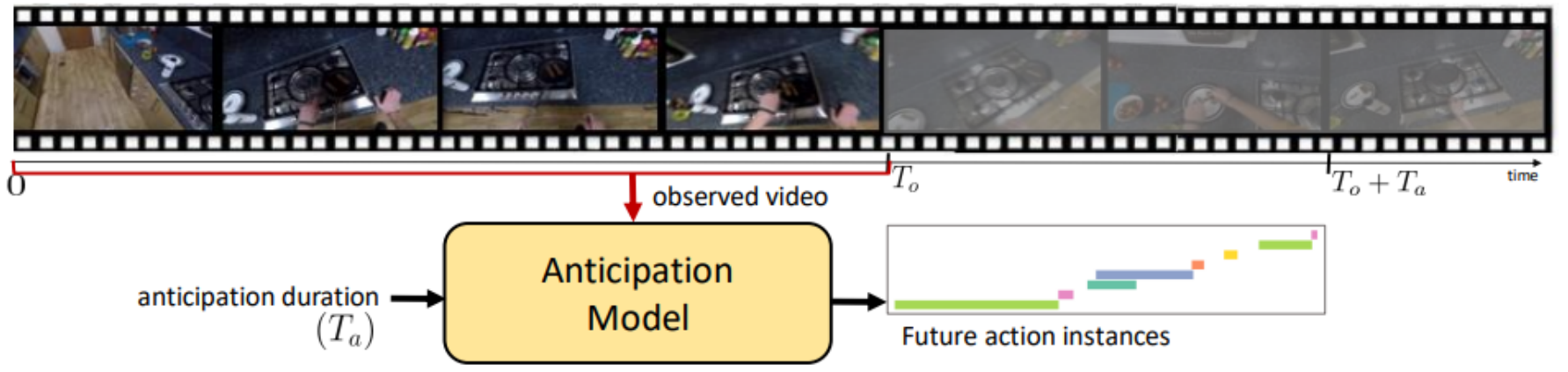


Fig. 1. Long-Term Action Anticipation. Given the initial portion of an activity video $(0, \ldots, T_o)$ and anticipation duration T_a, the task is to predict the actions that would occur from time $T_o + 1$ to $T_o + T_a$. Our proposed anticipation model receives the observed video and the anticipation duration as inputs and directly predicts a set of future action instances. Here, the action anticipation is *long-term* – both the observed duration T_o and the anticipation duration T_a are in the order of minutes.

HUMAN BEHAVIOR ANIMATION

COMPUTER VISION RESEARCH PAPERS OF THE WEEK 2022

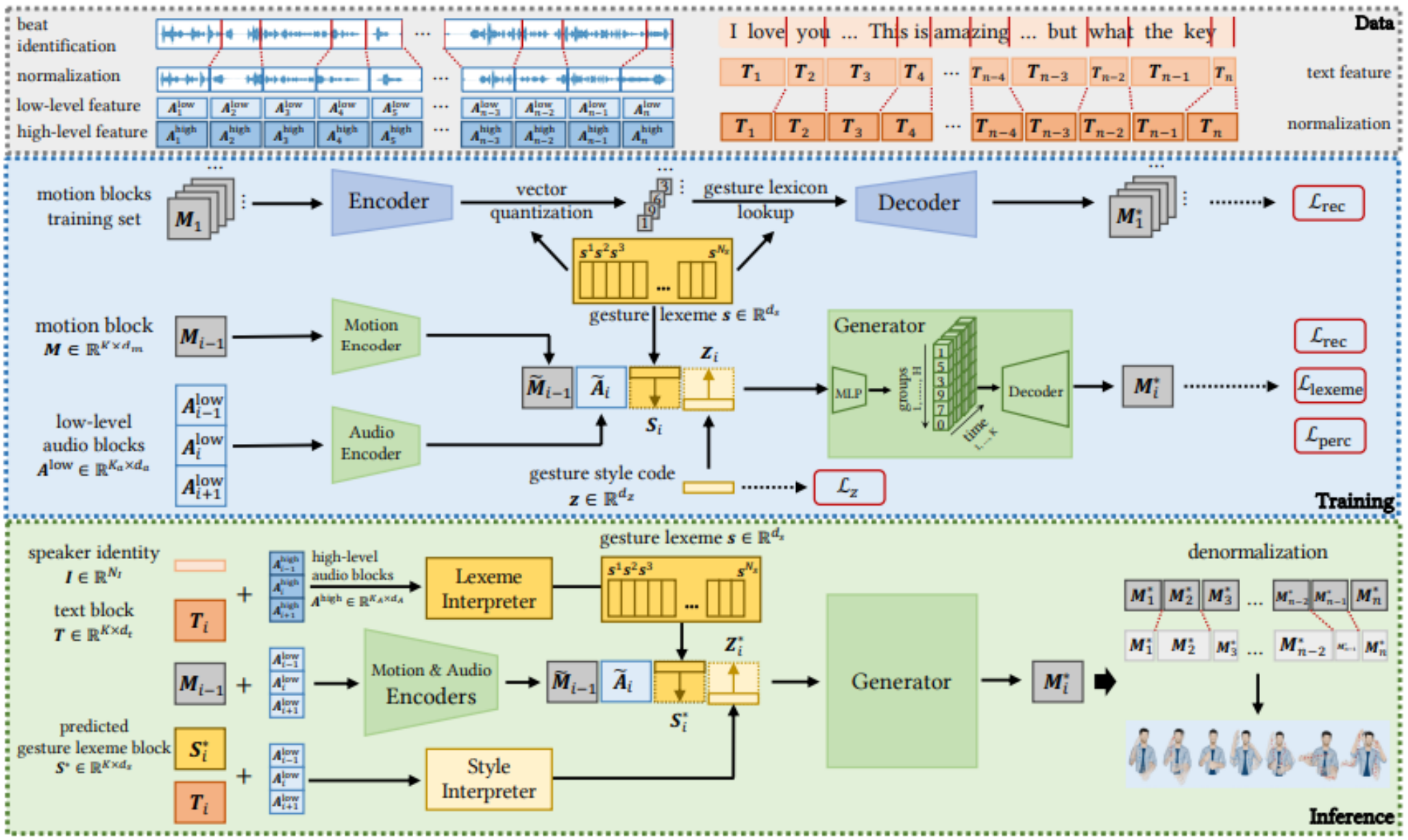


Fig. 2. Our system is composed of three core components: (a) the *data* module preprocesses a speech, segments it into normalized blocks based on the beats, and extracts speech features from these blocks; (b) the *training* module learns a gesture lexicon from the normalized motion blocks and trains the generator to synthesize gesture sequences, conditioned on the gesture lexemes, the style codes, as well as the features of previous motion blocks and adjacent speech blocks; and (c) the *inference* module employs interpreters to transfer the speech features to gesture lexemes and style codes, which are then used by the learned generator to predict future gestures.