Temporal Action Localization with Cross Layer Task Decoupling and Refinement

Qiang Li[1,4*], Di Liu[1,2*], Jun Kong[1,3†], Sen Li[1], Hui Xu[4], Jianzhong Wang[1†]

[1] Northeast Normal University [2] Northeast Electric Power University

[3] KLAS of MOE [4] Changchun Humanities and Sciences College

{ liq782, kongjun, lis084, wangjz019 }@nenu.edu.cn, 20102313@neepu.edu.cn, xuhui1@ccrw.edu.cn

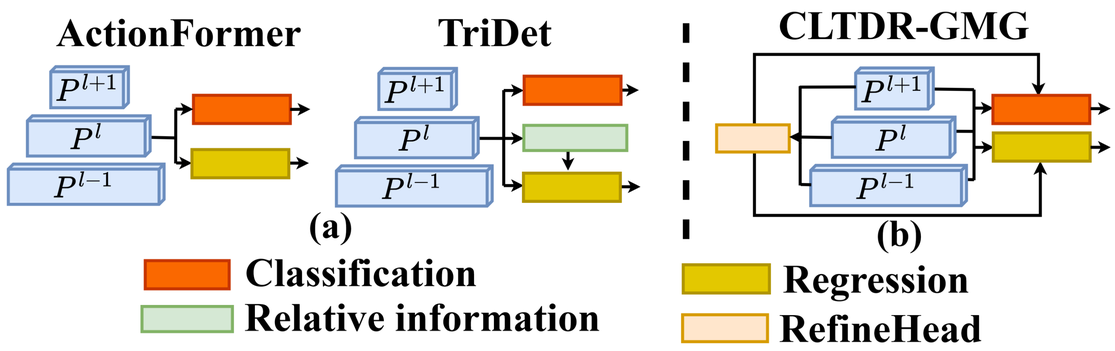
Introduction

Temporal Action Localization

Temporal action localization (TAL) involves dual tasks to classify and localize actions within untrimmed videos.

Challenges

- The trade-off between classification and localization
- Incomprehensive feature encoding and high global computational complexity.

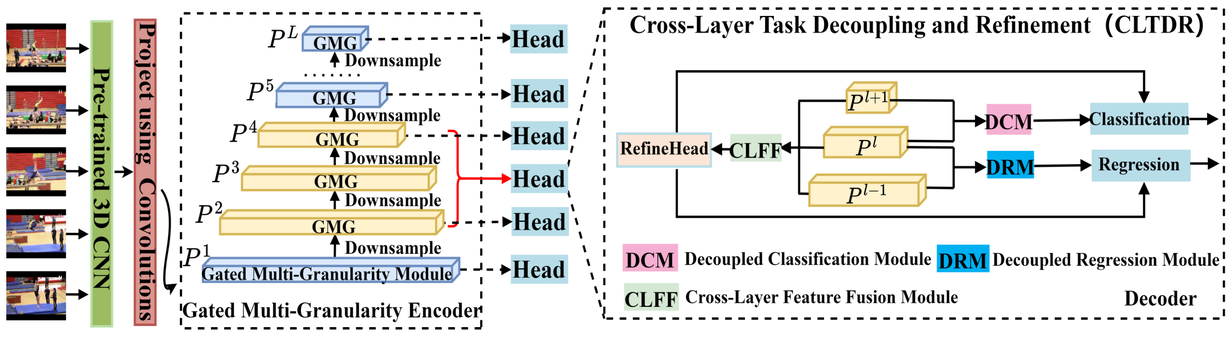


Contribution

- We propose a novel TAL method with Cross Layer Task Decoupling and Refinement (CLTDR). CLTDR strategy integrates semantically strong features from higher pyramid layers and detailed boundary-aware boundary features from lower pyramid layers to effectively disentangle the action classification and localization tasks.
- We propose a lightweight Gated Multi-Granularity (GMG) module to comprehensively extract and aggregate video features at instant, local, and global temporal granularities.
- Our method achieves state-of-the-art performance on five challenging benchmarks: THUMOS14, MultiTHUMOS, EPIC_KITCHENS-100, ActivityNet-1.3, and HACS.

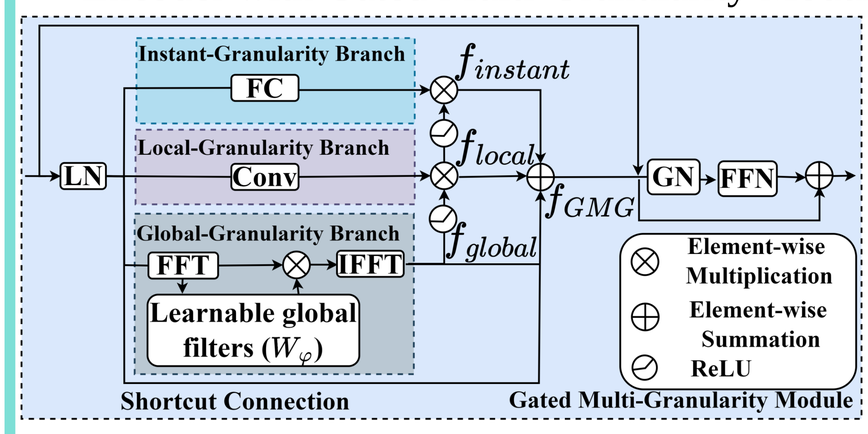
Methods

Framework



We build a feature pyramid with GMG module. The CLTDR decoder at the l-th pyramid layer leverages the features P^{l+1} and P^{l-1} to generate distinct representations for classification and localization tasks, followed by a refinement using RefineHead.

- Encoder with Gated Multi-Granularity Module



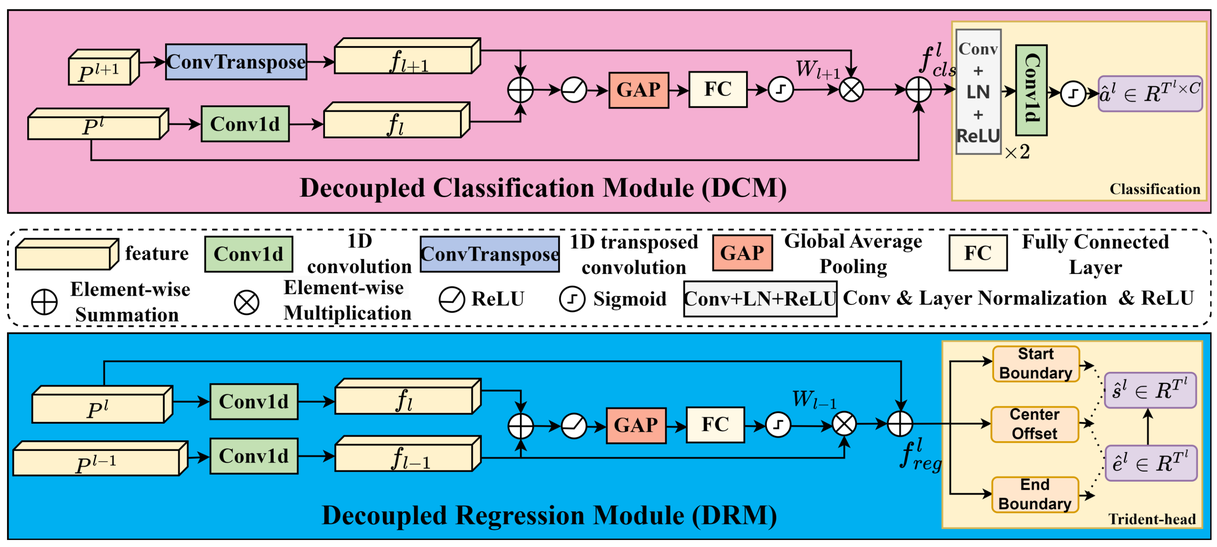
$$X_f[u] = FFT(x) = \sum_{t=0}^{T-1} x_t e^{-j\frac{2\pi}{N}ut}, \quad 0 < u < T-1$$

$$W_\varphi = Conv\left(ReLU\left(Conv(X_f)\right)\right)$$

$$f_{global} = IFFT(X_f \otimes W_\varphi)$$

$$f_{GMG} = x + \underbrace{ReLU(f_{local}) \otimes FC(x)}_{f_{instant}} + \underbrace{ReLU(f_{global}) \otimes Conv(x)}_{f_{local}} + f_{global}$$

- Decoder with Cross Layer Task Decoupling and Refinement



DCM

$$W_{l+1} = Sigmoid\left(FC\left(GAP(ReLU(f_{l+1} + f_l))\right)\right)$$

$$f^l_{cls} = P_l + f_{l+1} \otimes W_{l+1}$$

DRM

$$W_{l-1} = Sigmoid\left(FC\left(GAP(ReLU(f_{l-1} + f_l))\right)\right)$$

$$f^l_{reg} = P_l + f_{l-1} \otimes W_{l-1}$$

RefineHead

$$f^l_c = Conv(P^{l-1}) + P^l + ConvTransposed(P^{l+1})$$

Experiments

Experiments Results

- Performance comparison on five datasets.

Method	0.3	0.4	0.5	0.6	0.7	Avg.
TALLFormer (Cheng and Bertasius 2022) ‡	76.0	—	63.2	—	34.5	59.2
ActionFormer (Zhang, Wu, and Li 2022)	82.1	77.8	71.0	59.4	43.9	66.8
TemporalMaxer (Tang, Kim, and Sohn 2023)	82.8	78.9	71.8	60.5	44.7	67.7
TFFormer (Yang, Wei, and Zheng 2024)	82.1	78.9	72.0	60.8	44.9	67.8
TransGMC (Yang et al. 2024)	82.3	78.8	71.4	60.0	45.1	67.5
TriDet (Shi et al. 2023)	83.6	80.1	72.9	62.4	47.4	69.3
CLTDR-GMG	**84.1**	**80.3**	**73.6**	**62.4**	**48.2**	**69.9**
ActionFormer (Zhang, Wu, and Li 2022) †	84.0	79.6	73.0	63.5	47.7	69.6
TriDet (Shi et al. 2023) †	84.8	80.0	73.3	63.8	48.8	70.1
CLTDR-GMG †	**85.7**	**81.3**	**75.5**	**65.3**	**51.0**	**71.8**
ActionFormer (Zhang, Wu, and Li 2022) §	82.3	81.9	75.1	65.8	50.3	71.9
TemporalMaxer (Tang, Kim, and Sohn 2023) § ∓	87.4	83.1	76.6	65.7	49.6	72.5
ActionMamba (Chen et al. 2024) §	86.9	83.1	76.9	65.9	50.8	72.7
TriDet (Shi et al. 2023) § ∓	86.9	83.4	76.6	66.3	51.7	73.0
CLTDR-GMG §	**87.0**	**84.0**	**78.4**	**67.9**	**54.0**	**74.3**

Table 1: Performance comparison on THUMOS14 dataset. ∗: TSN features. ‡:Swin Transformer features. †: VideoMAEv2 features. §: InterVideo2-6B features. Others: I3D features. ∓:indicates our implementation.

Task	Method	0.1	0.2	0.3	0.4	0.5	Avg.
V.	ActionFormer (Zhang, Wu, and Li 2022)	26.6	25.4	24.2	22.3	19.1	23.5
	TemporalMaxer (Tang, Kim, and Sohn 2023)	27.8	26.6	25.3	23.1	19.9	24.5
	TriDet (Shi et al. 2023)	28.6	27.4	26.1	24.2	20.8	25.4
	TFFormer (Yang, Wei, and Zheng 2024)	28.8	27.7	26.1	24.7	20.5	25.6
	TransGMC (Yang et al. 2024)	27.8	26.7	25.5	23.6	20.7	24.9
	CLTDR-GMG	**29.5**	**28.6**	**27.0**	**24.3**	**20.7**	**26.0**
N.	ActionFormer (Zhang, Wu, and Li 2022)	25.2	24.1	22.7	20.5	17.0	21.9
	TemporalMaxer (Tang, Kim, and Sohn 2023)	26.3	25.2	23.5	21.3	17.6	22.8
	TriDet (Shi et al. 2023)	27.4	26.3	24.6	22.2	18.3	23.8
	TFFormer (Yang, Wei, and Zheng 2024)	27.2	25.9	24.2	21.7	17.9	23.4
	TransGMC (Yang et al. 2024)	26.4	25.2	23.4	21.4	18.1	22.9
	CLTDR-GMG	**28.2**	**26.9**	**25.2**	**22.7**	**19.4**	**24.5**

Table 3: Performance comparison on EPIC-KITCHENS-100 dataset. *V.* and *N.* denote the verb and noun sub-tasks.

Method	tIoU			Avg.
	0.5	0.75	0.95	0.5:0.05:0.95
ReAct (Shi et al. 2022)∗	49.6	33.0	8.6	32.6
TadTR (Liu et al. 2022)∗	51.3	35.0	9.5	34.6
TadTR (Liu et al. 2022)†	53.6	37.5	10.5	36.8
TALLFormer (Cheng and Bertasius 2022)‡	54.1	36.2	7.9	35.6
TFFormer (Yang, Wei, and Zheng 2024)	54.4	36.7	7.5	35.8
ActionFormer (Zhang, Wu, and Li 2022)†	54.7	37.8	8.4	36.6
TransGMC (Yang et al. 2024)†	54.8	37.6	8.5	36.7
TriDet (Shi et al. 2023)†	54.7	38.0	8.4	36.8
CLTDR-GMG†	**55.0**	**38.0**	**8.6**	**37.1**

Table 4: Performance comparison on ActivityNet1.3 dataset. ∗: TSN features. ‡: Swin Transformer features. †: R(2+1)D features. Others: I3D features.

Method	0.5	0.75	0.95	Avg.
TadTR (Liu et al. 2022)	47.1	32.1	10.9	32.1
TALLFormer (Cheng and Bertasius 2022)‡	55.0	36.1	11.8	36.5
TCANet (Qing et al. 2021b)†	54.1	37.2	11.3	36.8
TriDet (Shi et al. 2023)	54.5	36.8	11.5	36.8
CLTDR-GMG	**55.2**	**37.3**	**11.8**	**37.2**
TriDet (Shi et al. 2023)†	56.7	39.3	11.7	38.6
CLTDR-GMG†	**57.6**	**39.9**	**12.0**	**39.3**

Table 5: Performance comparison on HACS dataset. ‡: Swin Transformer features. †: SlowFast features. Others: I3D features.

Project page with code

Method	0.2	0.5	0.7	Avg.
PointTAD (Tan et al. 2022)	39.7	24.9	12.0	23.5
ActionFormer (Zhang, Wu, and Li 2022)	46.4	32.4	15.0	28.6
TemporalMaxer (Tang, Kim, and Sohn 2023)	47.5	33.4	17.4	29.9
TriDet (Shi et al. 2023)	49.1	34.3	17.8	30.7
CLTDR-GMG	**56.7**	**42.2**	**24.1**	**37.1**
TriDet (Shi et al. 2023)∗	55.7	41.0	23.5	36.2
CLTDR-GMG∗	**62.5**	**49.1**	**30.2**	**42.7**
TriDet (Shi et al. 2023)†	57.7	42.7	24.3	37.5
CLTDR-GMG†	**64.9**	**51.4**	**32.9**	**44.9**

Table 2: Performance comparison on MultiTHUMOS dataset. ∗: I3D (RGB+Flow) features. †: VideoMAEv2 features. Others: I3D (only RGB) features.

Ablation Study

We perform various ablation studies on THUMOS14 dataset using InterVideo2-6B features to assess the effectiveness of each component in our CLTDR-GMG.

- Ablation on GMG.

Instant	✓	✗	✗	✓	✓	✗	✓
Local	✗	✓	✗	✓	✗	✓	✓
Global	✗	✗	✓	✗	✓	✓	✓
Avg.	73.4	73.4	73.5	73.6	73.8	73.9	74.3

Table 6: The analysis of different temporal granularities.

Method	0.3	0.5	0.7	Avg.
without Gate	87.0	77.7	53.4	73.9
with Gate	87.3	77.9	53.9	74.3

Table 7: The impact of gated mechanism in GMG.

Global-Level	0.3	0.5	0.7	Avg.	# Params
FFT	87.3	77.9	53.9	74.3	9.5M
Self-Attention	87.5	78.0	53.5	74.2	26.1M

Table 8: The comparison of global feature extractor in GMG.

- Ablation on CLTDR.

classification				regression			
P_{l+1}	P_l	P_{l-1}	mAP	P_{l+1}	P_l	P_{l-1}	mAP
	✓		73.3		✓		73.4
✓	✓		74.3		✓	✓	74.3
	✓	✓	73.8	✓	✓		73.6
✓	✓	✓	74.0	✓	✓	✓	73.2

Table 9: Ablation of pyramid layers in CLTDR.

Method	0.3	0.5	0.7	Avg.
without refinement	87.1	77.3	53.1	73.5
only refine classification	87.4	77.8	53.9	73.9
only refine regression	87.6	77.3	52.4	73.8
refine classification and regression	87.3	77.9	53.9	74.3

Table 10: The effectiveness of refinement in CLTDR.

Method	0.3	0.5	0.7	Avg.
Concatenation	86.7	76.7	53.3	73.4
Addition	86.8	77.3	52.8	73.5
Attention	87.3	77.9	53.9	74.3

Table 11: The comparison of fusion strategies in CLTDR.