大连理工大学
Dalian University of Technology

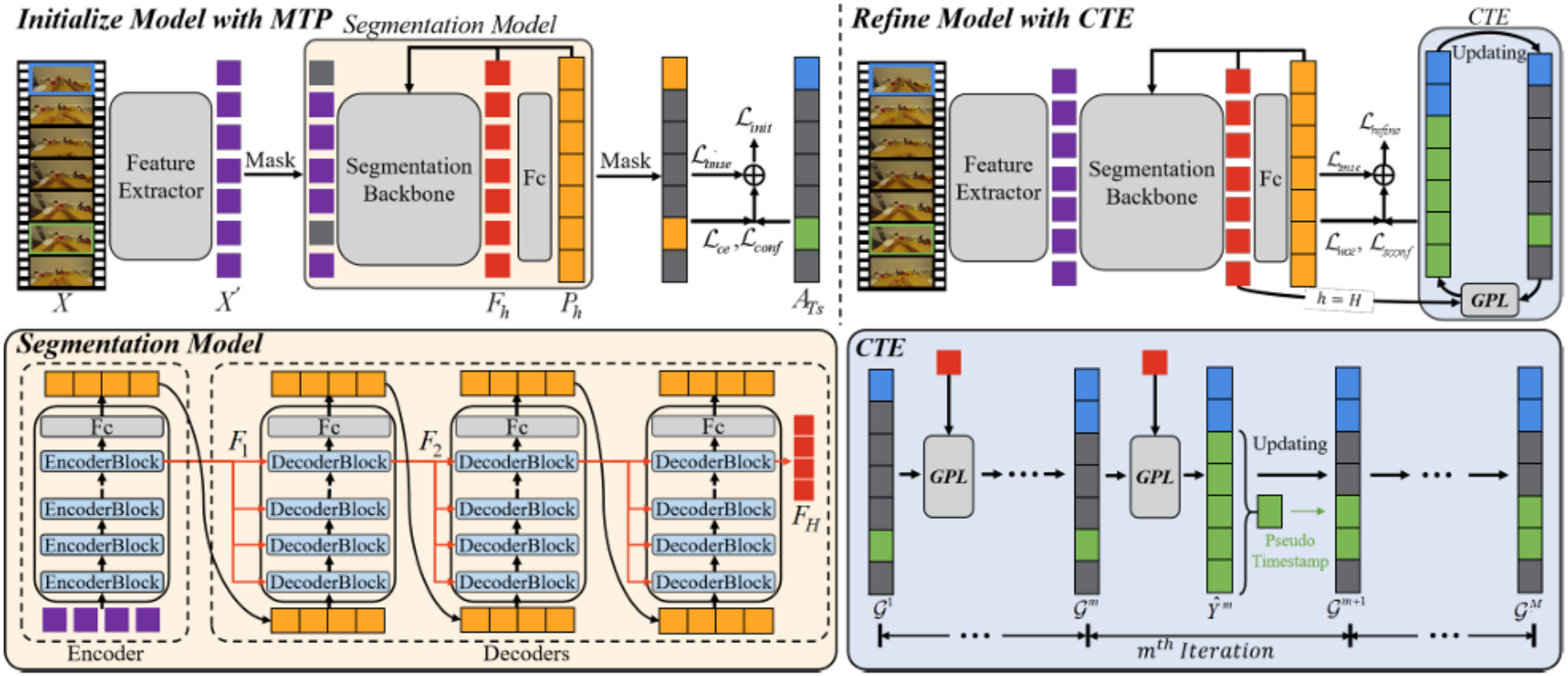
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Reducing the Label Bias for Timestamp Supervised Temporal Action Segmentation

CVPR 2023

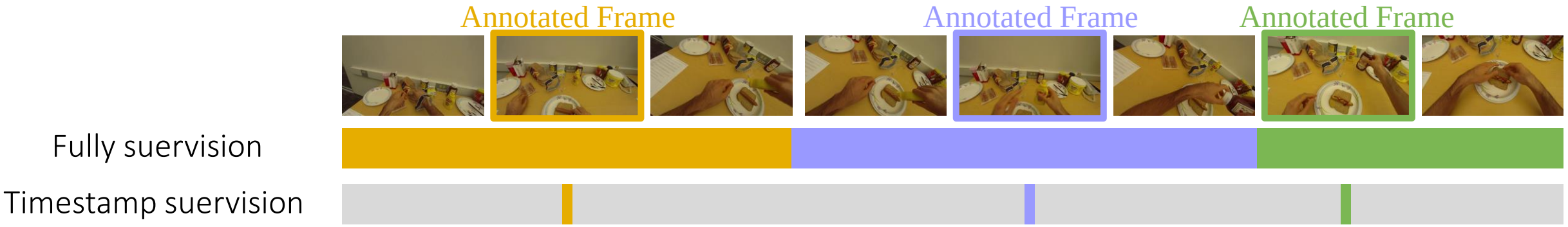
Kaiyuan Liu · Yunheng Li · Shenglan Liu · Chenwei Tan · Zihang Shao

Dalian University of Technology, DaLian, China



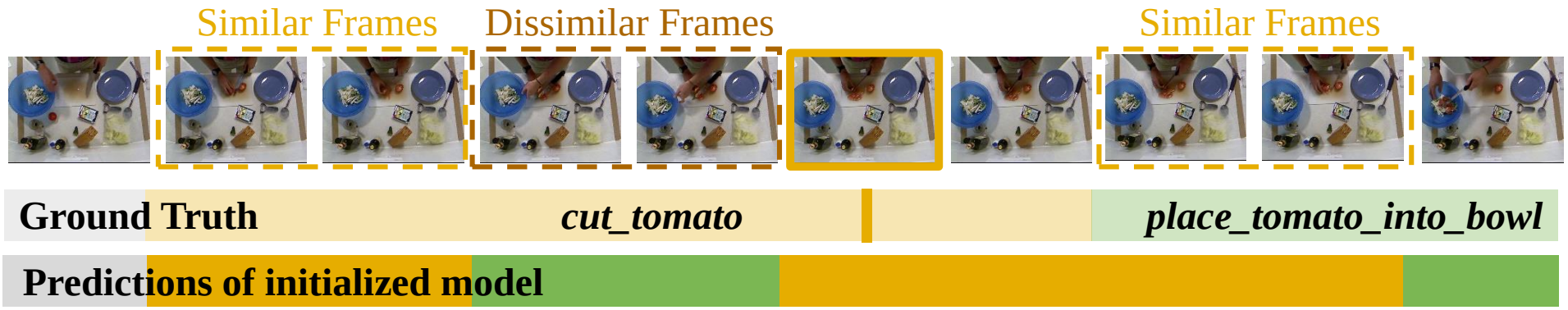
The pipeline of our proposed D-TATAS (Debiasing Timestamp Supervisied Temporal Action Segmentation) consists of two phases: initialize the segmentation model with masked timestamp predictions and refine the model with center-oriented timestamp expansion. GPL is the abbreviation for generating pseudo-labels.

Background and Motivation



(a) An illustrative example of timestamp supervised temporal action segmentation.

Focus bias caused by limited label in the initialization phase.

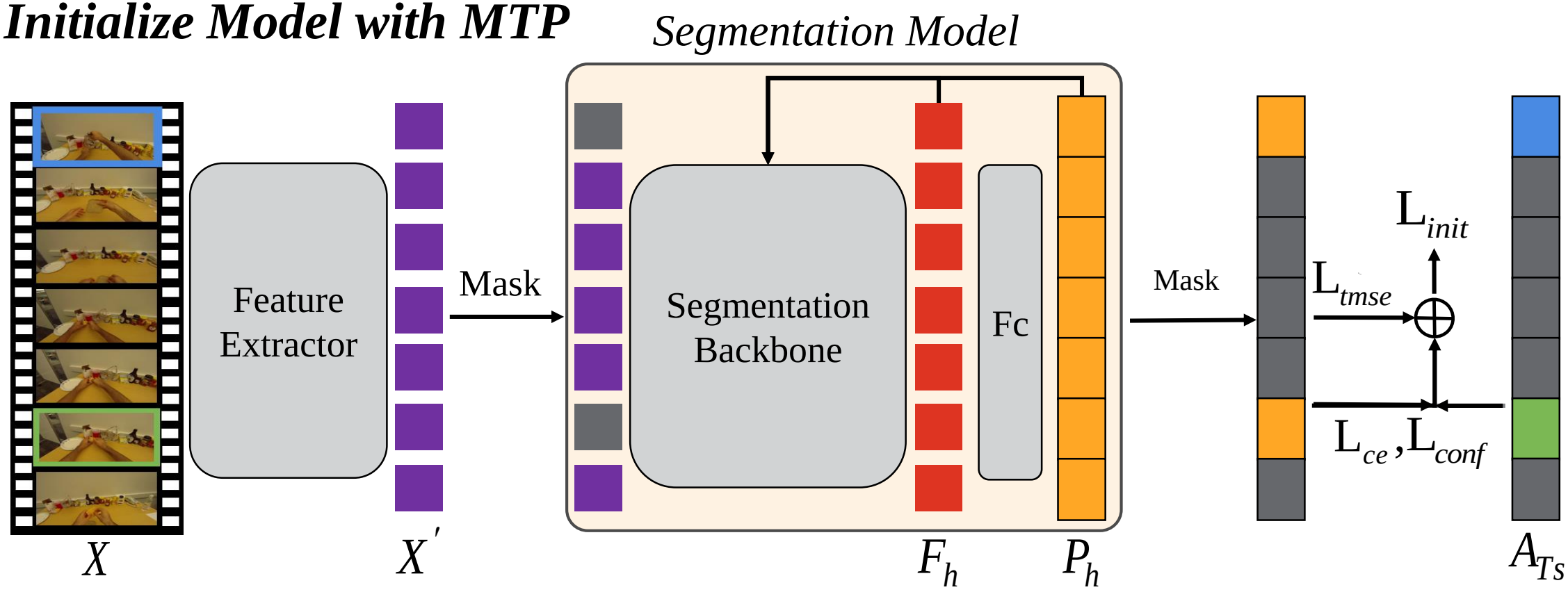


Representation bias caused from limited annatations when refining model.

(b) Label bias in two phases.

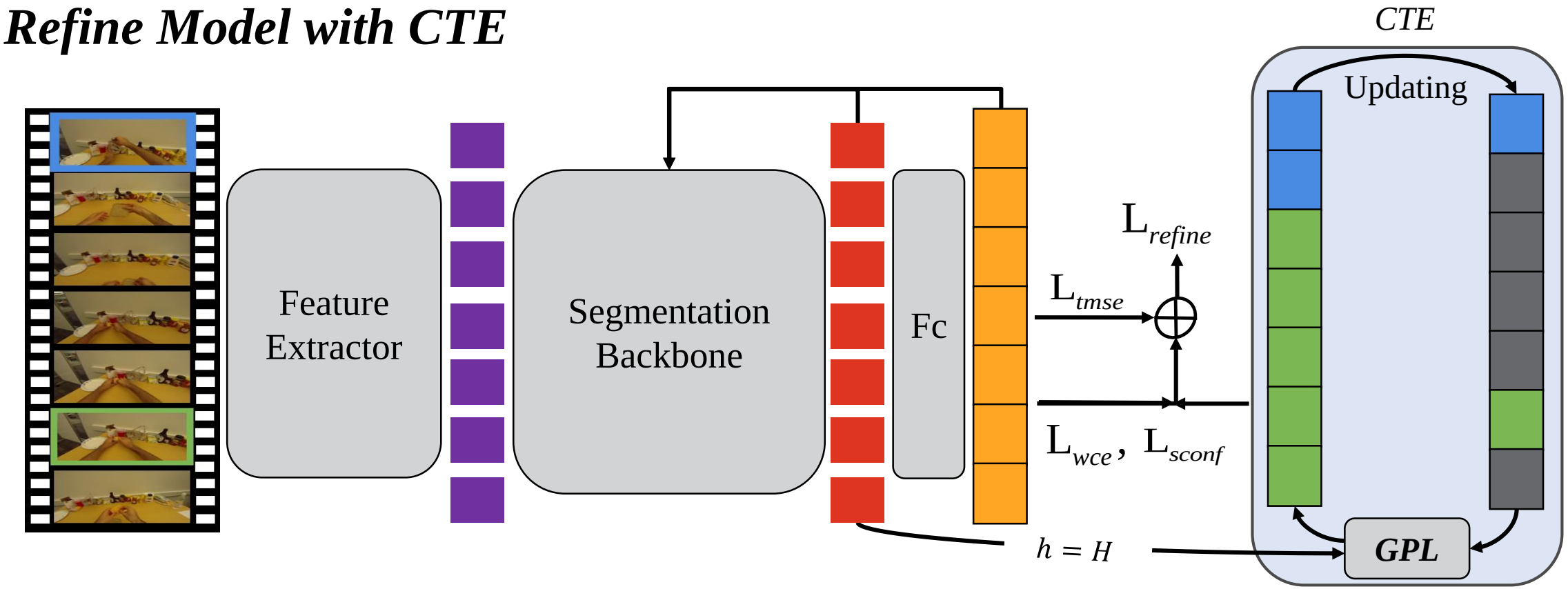
The Pipeline of D-TSTAS

Initialize Model with MTP

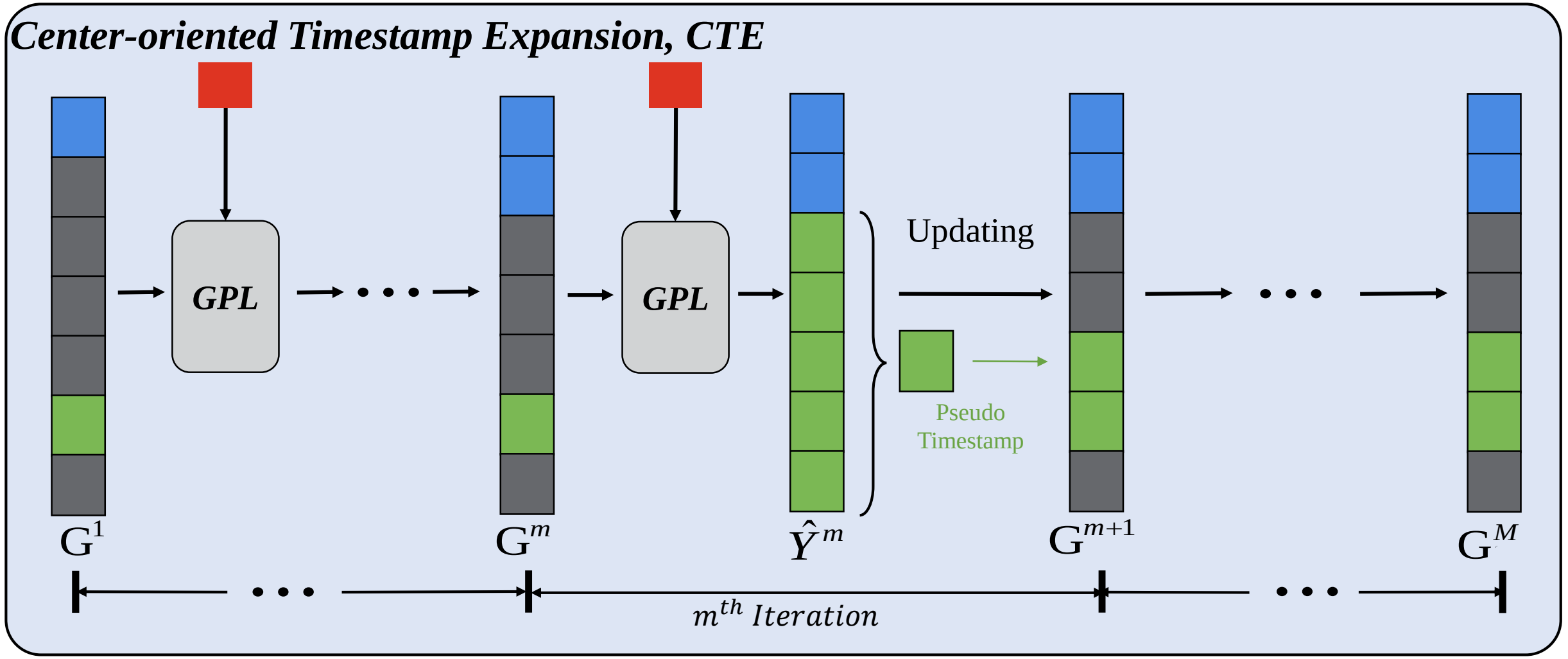


The idea of MTP is to mask the input features of timestamps and use contextual information to predict their action categories. This can force the model to learn more from the unannotated frames and reduce the dependency on the annotated frames.

Refine Model with CTE



The idea of CTE is to expand pseudo-timestamp groups that contain more semantic information than single timestamps. This can help the model to overcome the limitation of the expressiveness of single-frame timestamps and generate better pseudo-labels. And then, we refine the model with pseudo-labels and proposed segmental confidence loss.



CTE can reduce representation bias by capturing more semantic-rich motion representations of action segments. By expanding pseudo-timestamp groups, we can include more frames that have different semantic information within the same segment. This can improve the quality of pseudo-labels and the model predictions.

Quantitative analysis of D-TSTAS

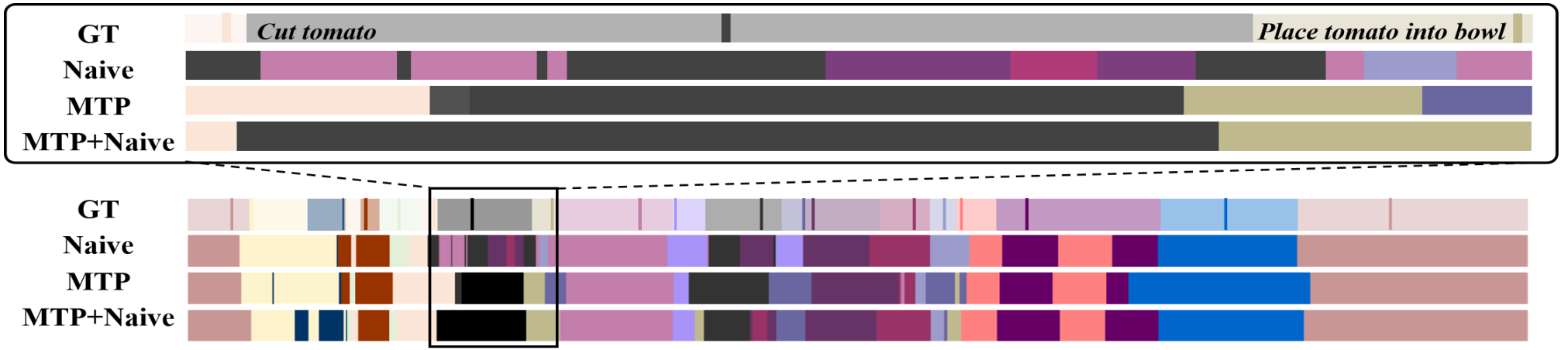
Supervision	Method	GTEA					50Salads					Breakfast				
		F1@{10, 25, 50}			Edit	Acc	F1@{10, 25, 50}			Edit	Acc	F1@{10, 25, 50}			Edit	Acc
Fully	MS-TCN [7]	87.5	85.4	74.6	81.4	79.2	76.3	74.0	64.5	67.9	80.7	52.6	48.1	37.9	61.7	66.3
	MS-TCN++ [25]	88.8	85.7	76.0	83.5	80.1	80.7	78.5	70.1	74.3	83.7	64.1	58.6	45.9	65.6	67.6
	BCN [40]	88.5	87.1	77.3	84.4	79.8	82.3	81.3	74.0	74.3	84.4	68.7	65.5	55.0	66.2	70.4
	ASRF [16]	89.4	87.8	79.8	83.7	77.3	84.9	83.5	77.3	79.3	84.5	74.3	68.9	56.1	72.4	67.6
	ASFormer [49]	90.1	88.8	79.2	84.6	79.7	85.1	83.4	76.0	79.6	85.6	76.0	70.6	57.4	75.0	73.5
	ETSN [26]	91.1	90.0	77.9	86.2	78.2	85.2	83.9	75.4	78.8	82.0	74.0	69.0	56.2	70.3	67.8
	ICC [34]	91.4	89.1	80.5	87.8	82.0	83.8	82.0	74.3	76.1	85.0	72.4	68.5	55.9	68.6	75.2
	UVAST [3]	92.7	91.3	81.0	92.1	80.2	89.1	87.6	81.7	83.9	87.4	76.9	71.5	58.0	77.1	69.7
	DPRN [31]	92.9	92.0	82.9	90.9	82.0	87.8	86.3	79.4	82.0	87.2	75.6	70.5	57.6	75.1	71.7
	Br-Prompt+ ASFormer [23]	94.1	92.0	83.0	91.6	81.2	89.2	87.8	81.3	83.8	88.1	-	-	-	-	-
Semi	ICC(5%) [34]	77.9	71.6	54.6	71.4	68.2	52.9	49.0	36.6	45.6	61.3	60.2	53.5	35.6	56.6	65.3
	ICC(10%) [34]	83.7	81.9	66.6	76.4	73.3	67.3	64.9	49.2	56.9	68.6	64.6	59.0	42.2	61.9	68.8
Timestamp	Li et al. [28]	78.9	73.0	55.4	72.3	66.4	73.9	70.9	60.1	66.8	75.6	70.5	63.6	47.4	69.9	64.1
	Khan et al. [17]	81.5	77.5	60.8	75.6	66.1	75.1	72.3	61.0	67.6	75.1	67.9	61.0	45.3	67.0	61.4
	Zhao et al. [51]	84.3	81.7	64.8	79.8	74.4	78.5	75.5	63.4	71.8	77.7	73.1	66.5	49.4	72.6	64.6
	EM-TSS [32]	-	82.7	66.5	82.3	70.5	-	75.9	64.7	71.6	77.9	-	63.7	49.8	67.2	**67.0**
	UVAST+ alignment decoder [3]	70.8	63.5	49.2	88.2	55.3	75.7	70.6	58.2	78.4	67.8	72.0	64.1	48.6	74.3	60.2
	UVAST+Viterbi [3]	87.2	83.7	66.0	**89.3**	70.5	83.0	79.6	65.9	78.2	77.0	71.3	63.3	48.3	74.1	60.7
	UVAST+FIFA [3]	80.7	75.2	57.4	88.7	66.0	80.2	74.9	61.6	**78.6**	72.5	72.0	64.2	47.6	74.1	60.3
	D-TSTAS	**91.5**	**90.1**	**76.2**	88.5	**75.7**	**84.2**	**82.1**	**71.5**	77.6	**80.0**	**76.7**	**69.3**	**50.7**	**75.8**	65.7

Comparison with different levels of supervision on all three datasets.

Qualitative analysis of D-TSTAS

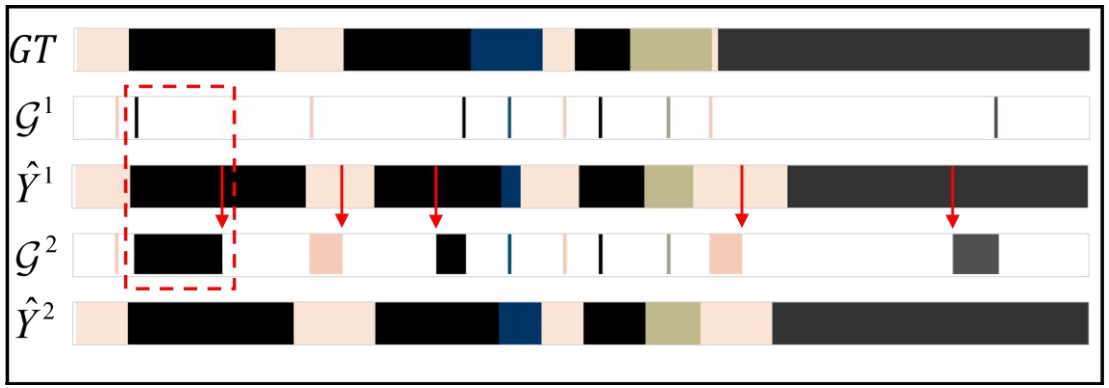
The qualitative result of MTP

- Only MTP: prediction based on context
- MTP+Naive: complement each other to predict the complete segment



The qualitative result of CTE

- more complete semantics by dense timestamps
- more accurate pseudo-labels



大连理工大学
Dalian University of Technology

Reducing the Label Bias for Timestamp Supervised Temporal Action Segmentation

Thanks for your watching