Video Action Transformer Network

Giridhar *et al.*, 2018

Presented by – Shashanka Venkataramanan

Outline

- Objective
- Challenges
- Proposed Approach
- Implementation & Training
- Experimental Results
- Conclusion

Objective

- To detect all persons and classify their actions at a given point of time.

Objective

- To detect all persons and classify their actions at a given point of time.
- To extract contextual information for human action recognition without manual supervision.

Objective

- To detect all persons and classify their actions at a given point of time.
- To extract contextual information for human action recognition without manual supervision.
- To build a model capable of representing information using self attention without using RNNs

Challenges

- Human actions are difficult to recognize because of dependence of contextual information.

Challenges

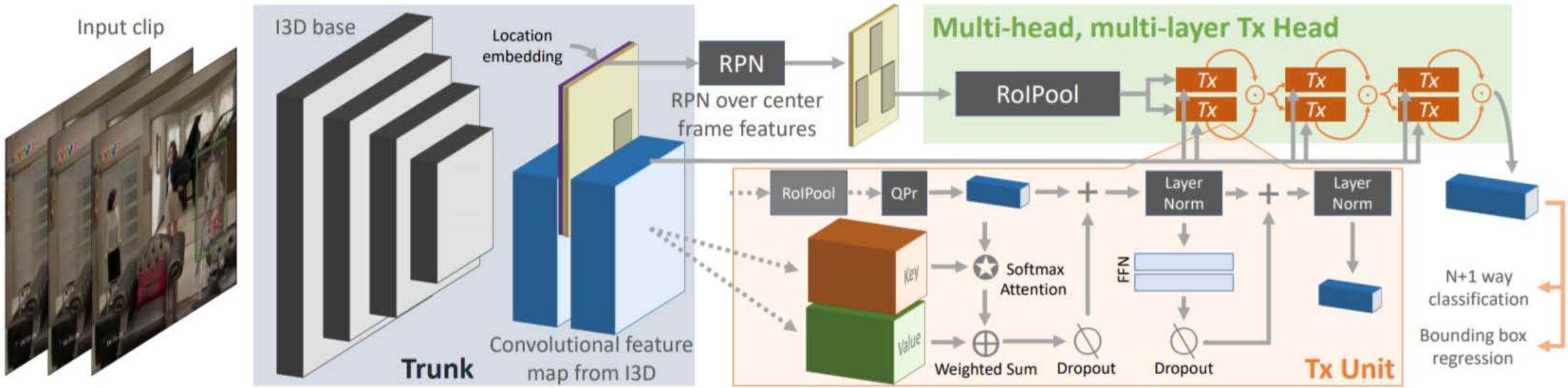
- Human actions are difficult to recognize because of dependence of contextual information.

- Do not have explicit supervision except bounding box information and class labels.

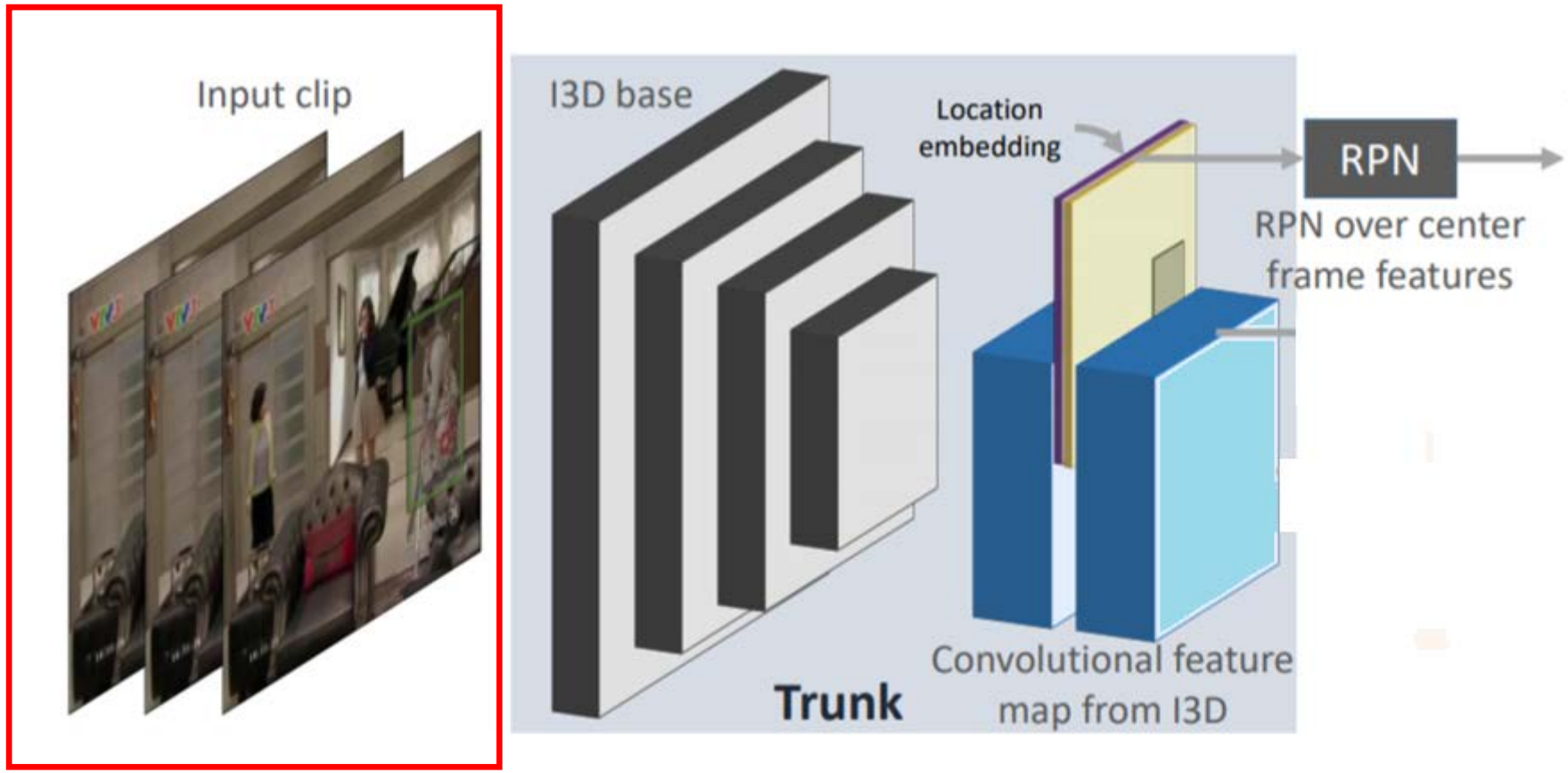
Challenges

- Human actions are difficult to recognize because of dependence of contextual information.
- Do not have explicit supervision except bounding box information and class labels.
- Many classes even with large training sets are still hard to recognize.

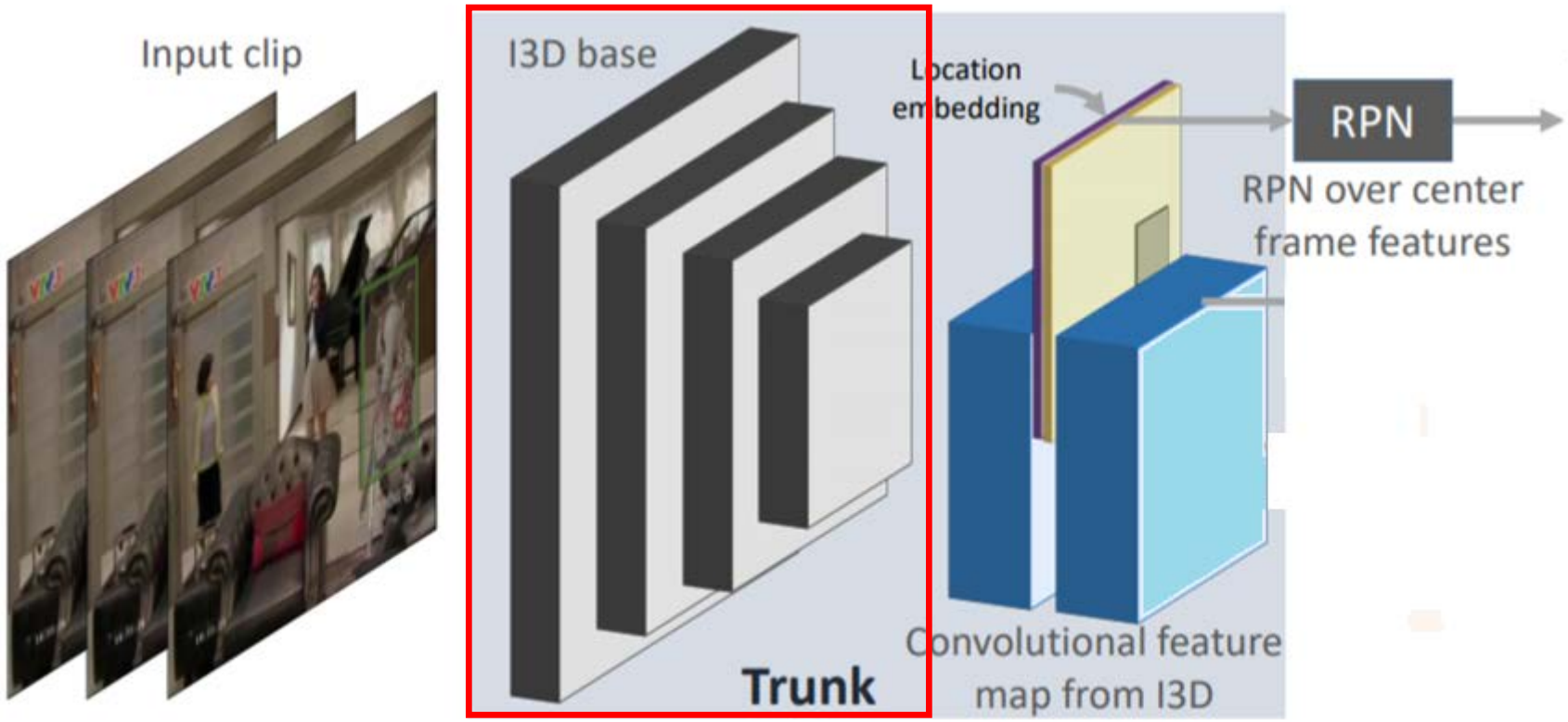
Proposed Approach



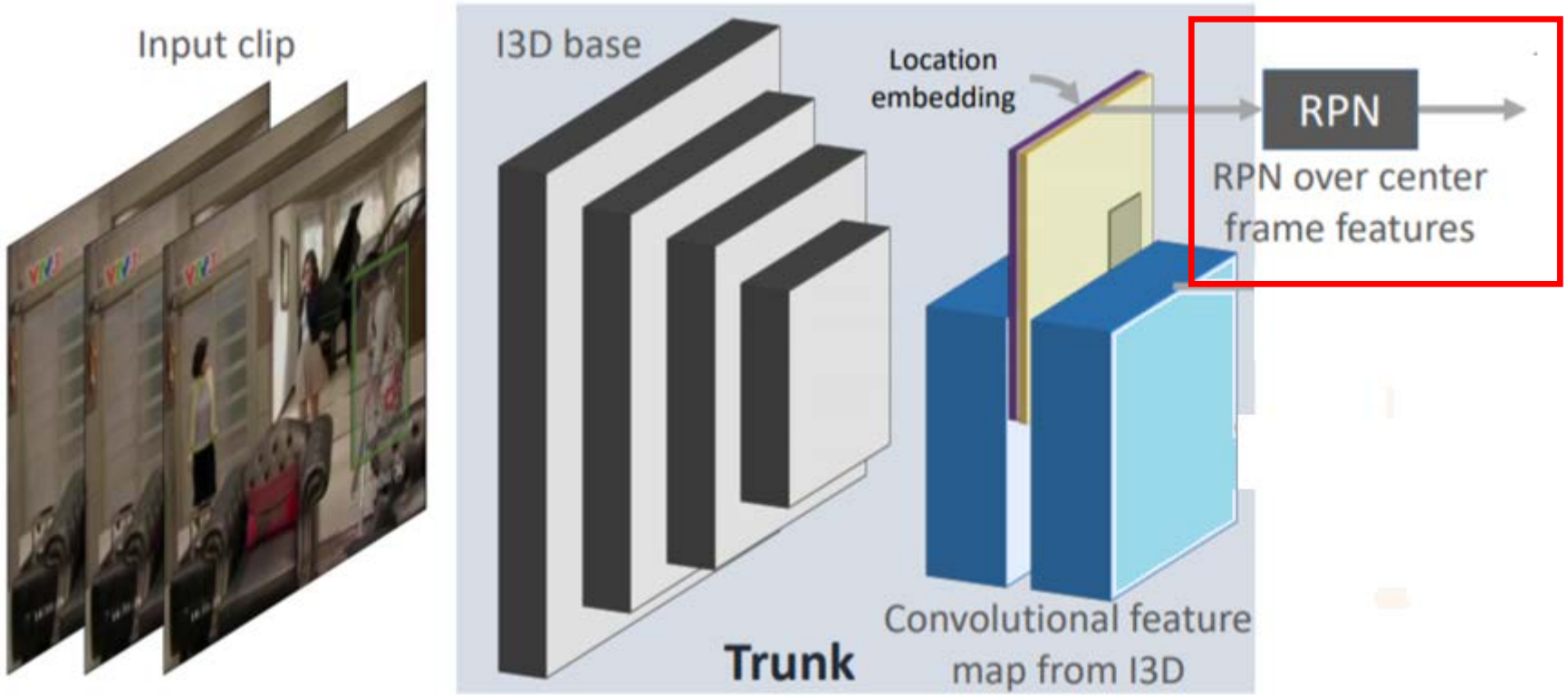
Trunk Region



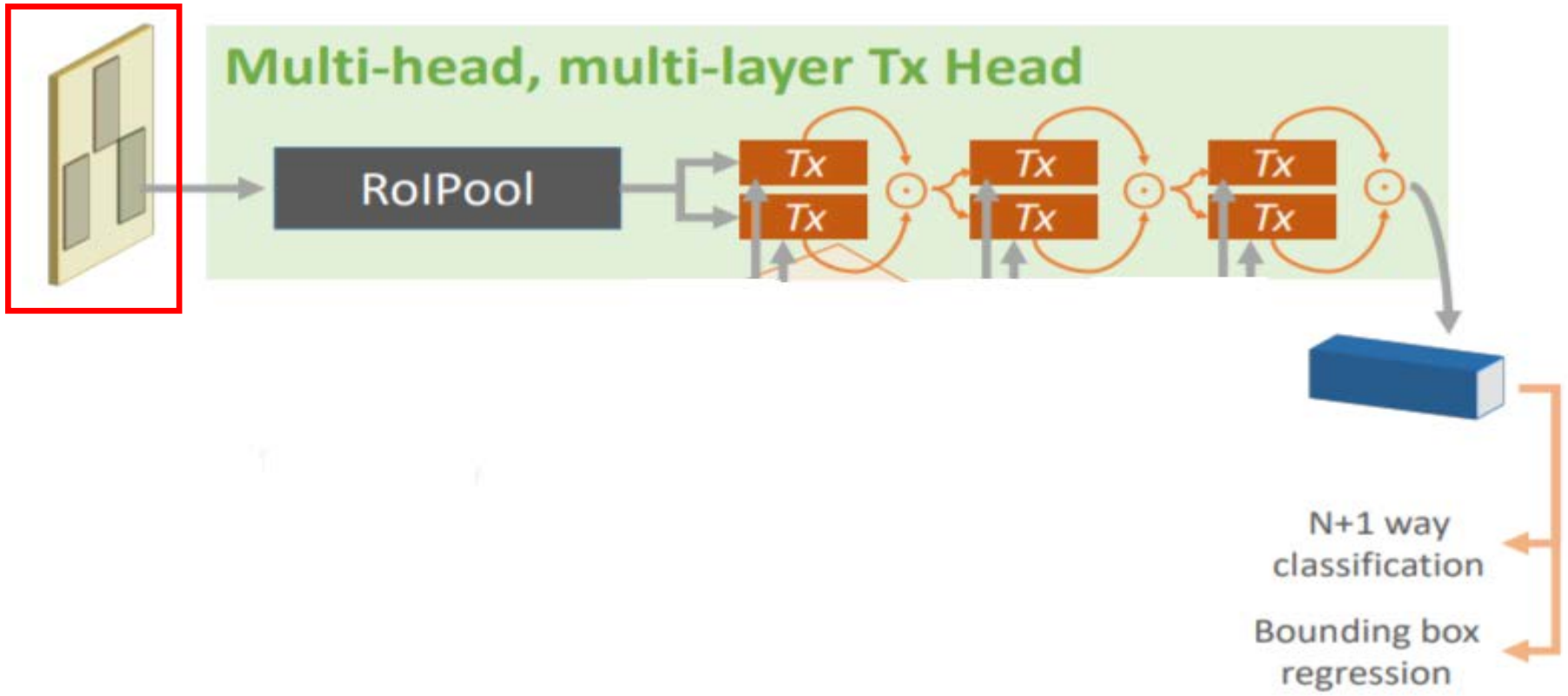
Trunk Region



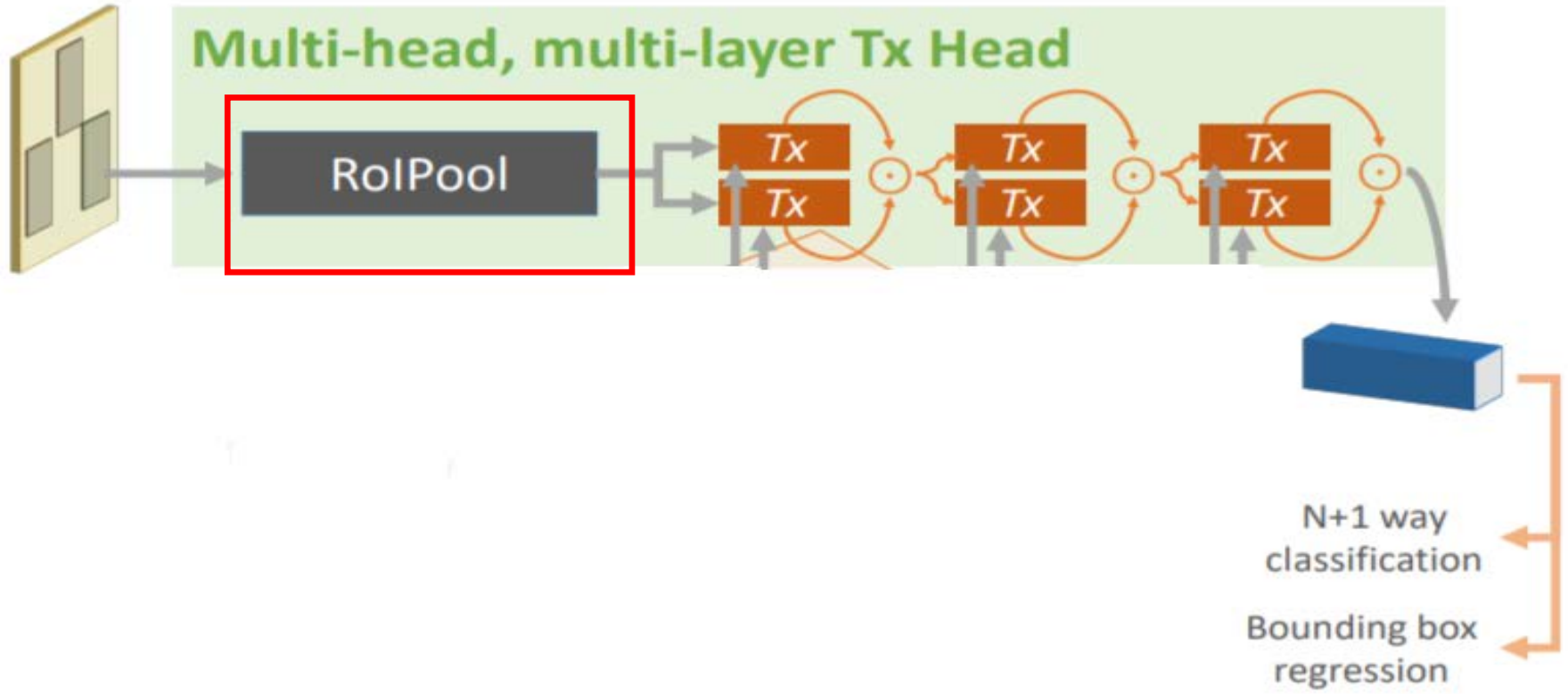
Trunk Region



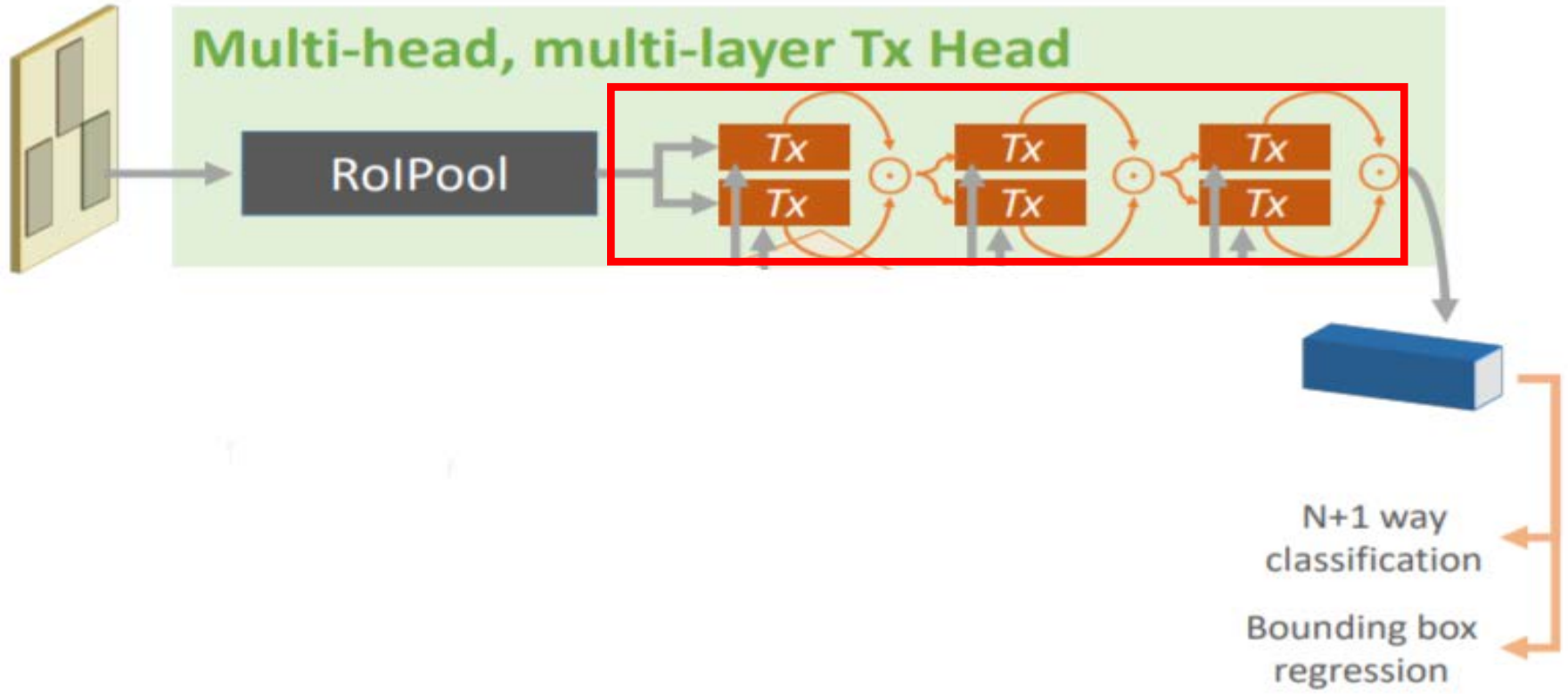
Head Region


Multi-head, multi-layer Tx Head
RoIPool
Tx
Tx
Tx
Tx
Tx
Tx
N+1 way classification
Bounding box regression

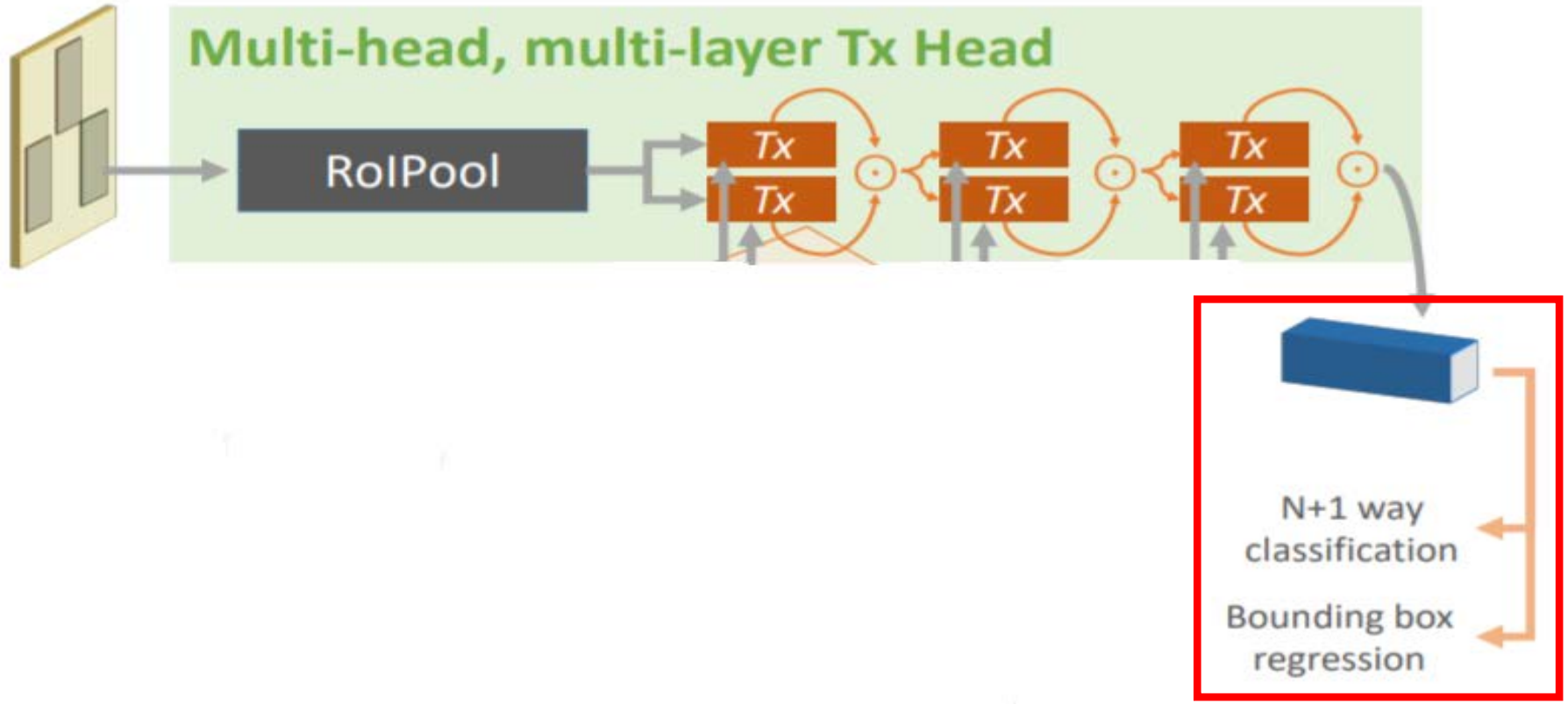
Head Region


Multi-head, multi-layer Tx Head
RoIPool
Tx
Tx
Tx
Tx
Tx
Tx
N+1 way classification
Bounding box regression

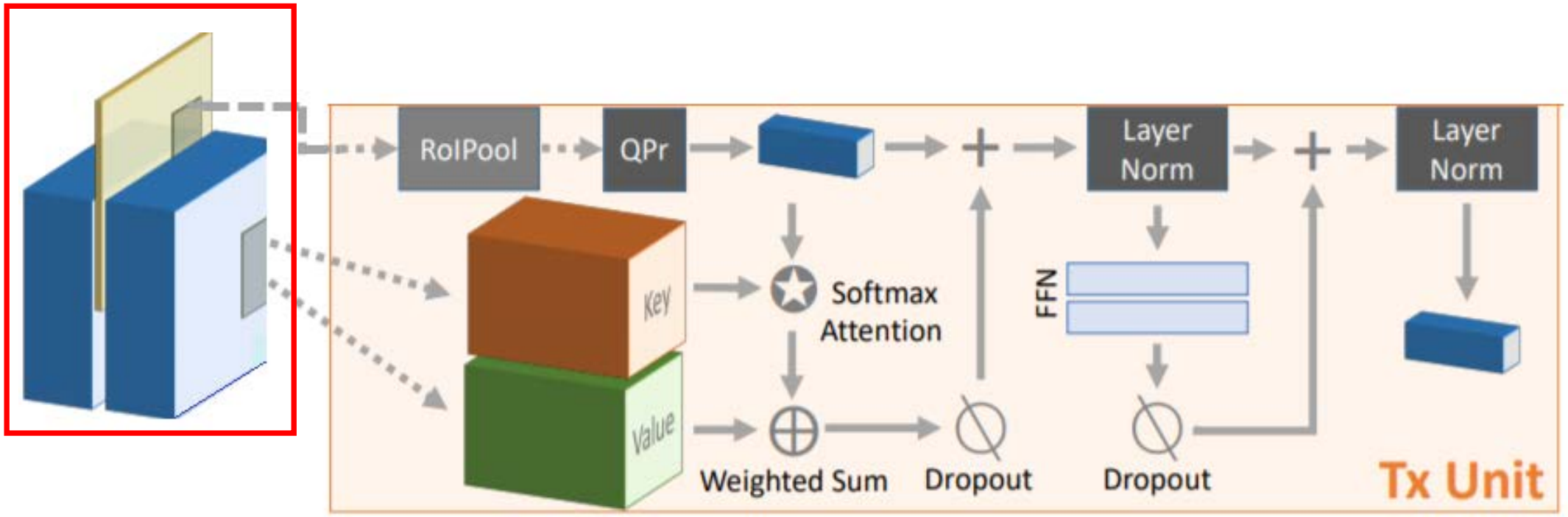
Head Region


Multi-head, multi-layer Tx Head
RoIPool
Tx
Tx
Tx
Tx
Tx
Tx
N+1 way classification
Bounding box regression

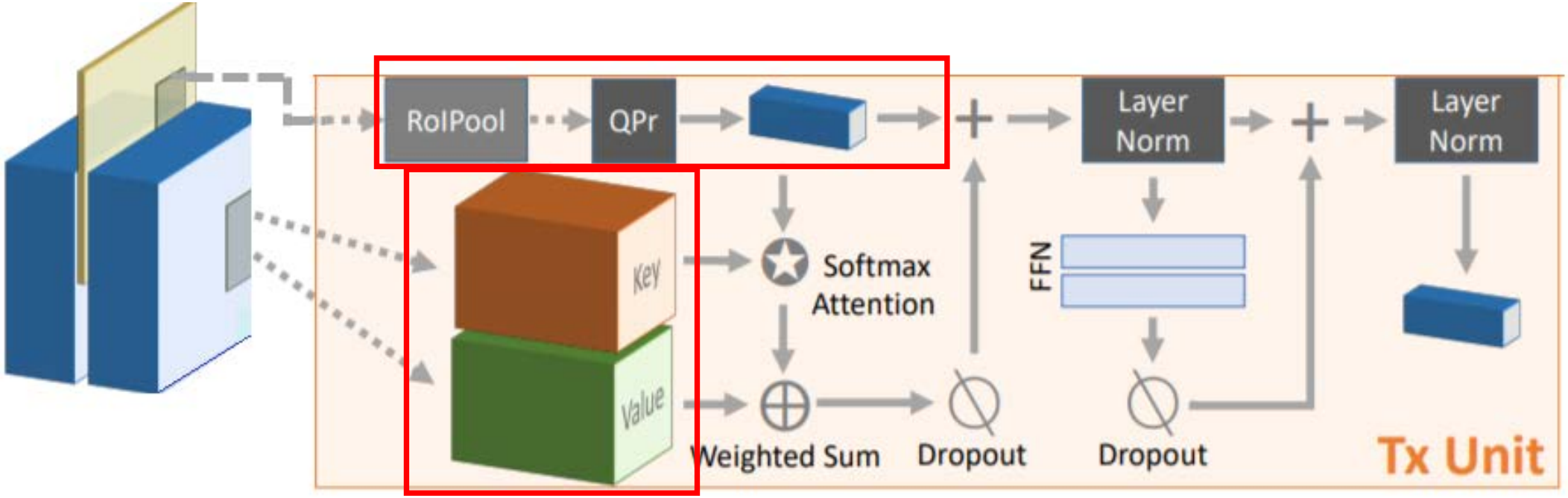
Head Region


Multi-head, multi-layer Tx Head
RoIPool
Tx
Tx
Tx
Tx
Tx
Tx
N+1 way classification
Bounding box regression

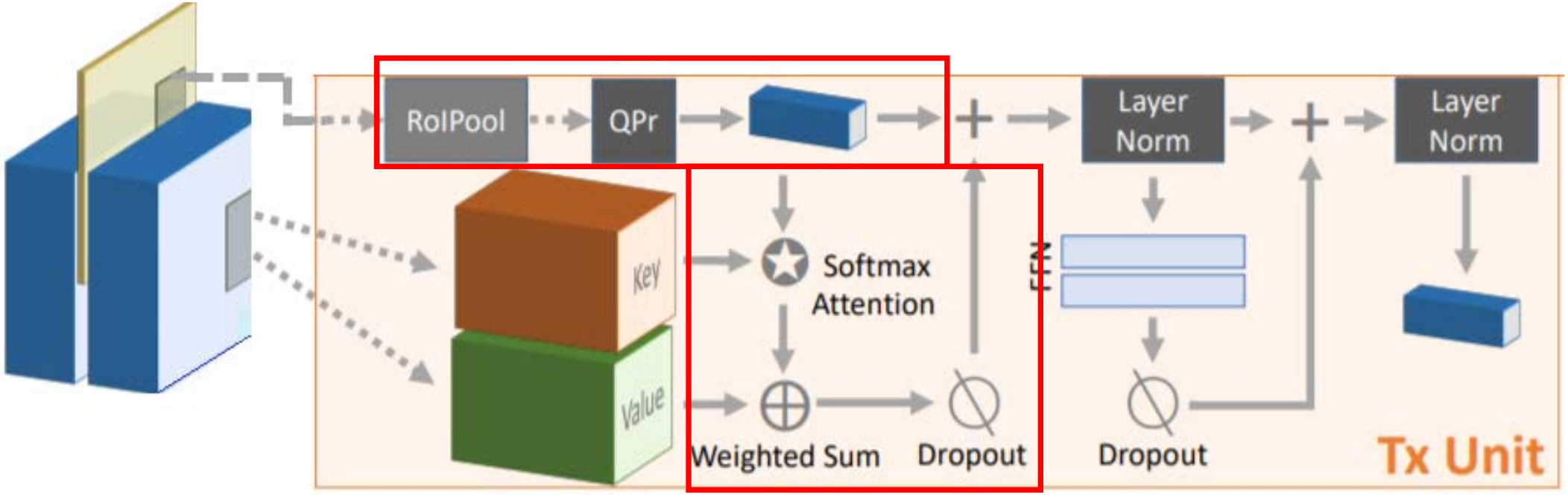
Action Transformer Network



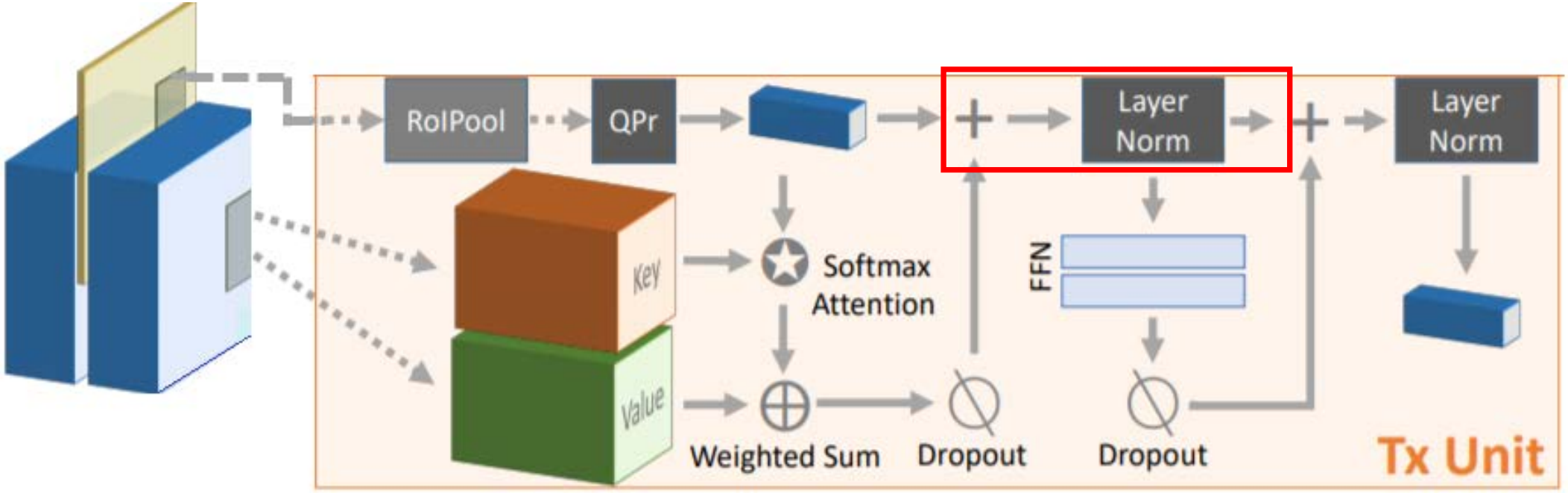
Action Transformer Network



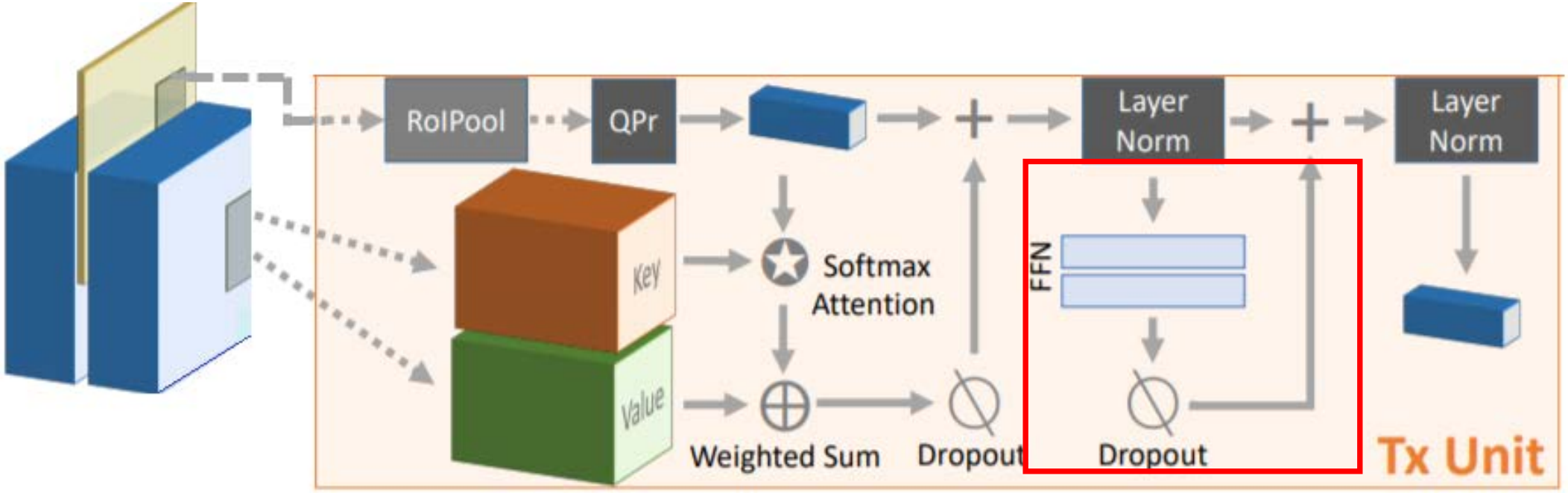
Action Transformer Network



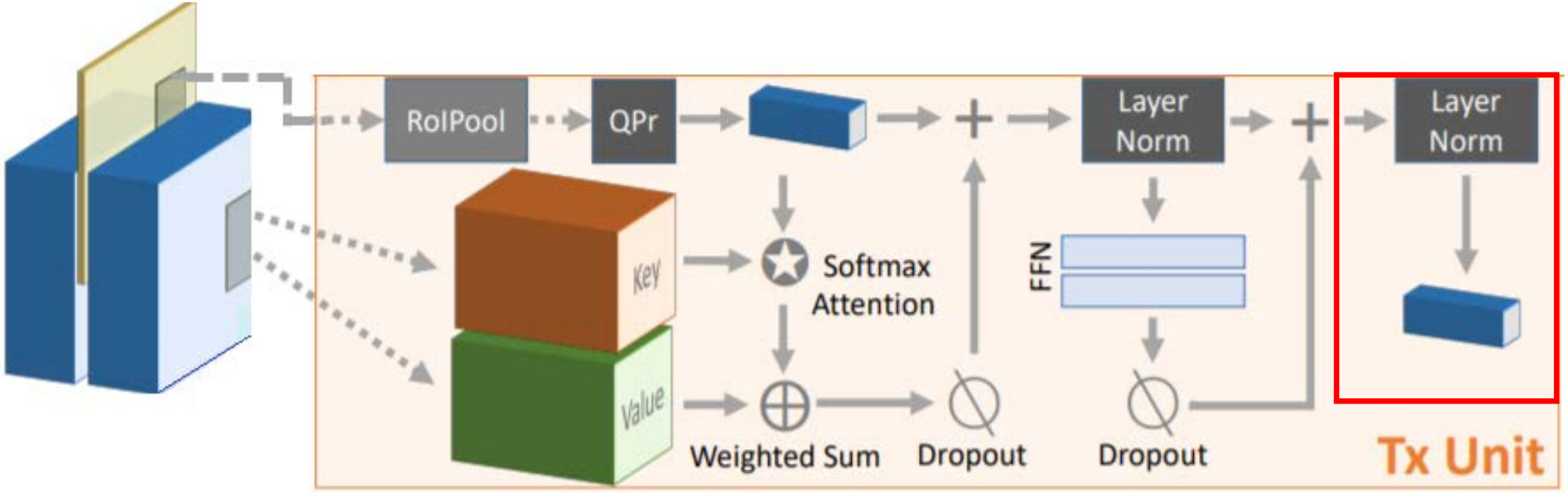
Action Transformer Network



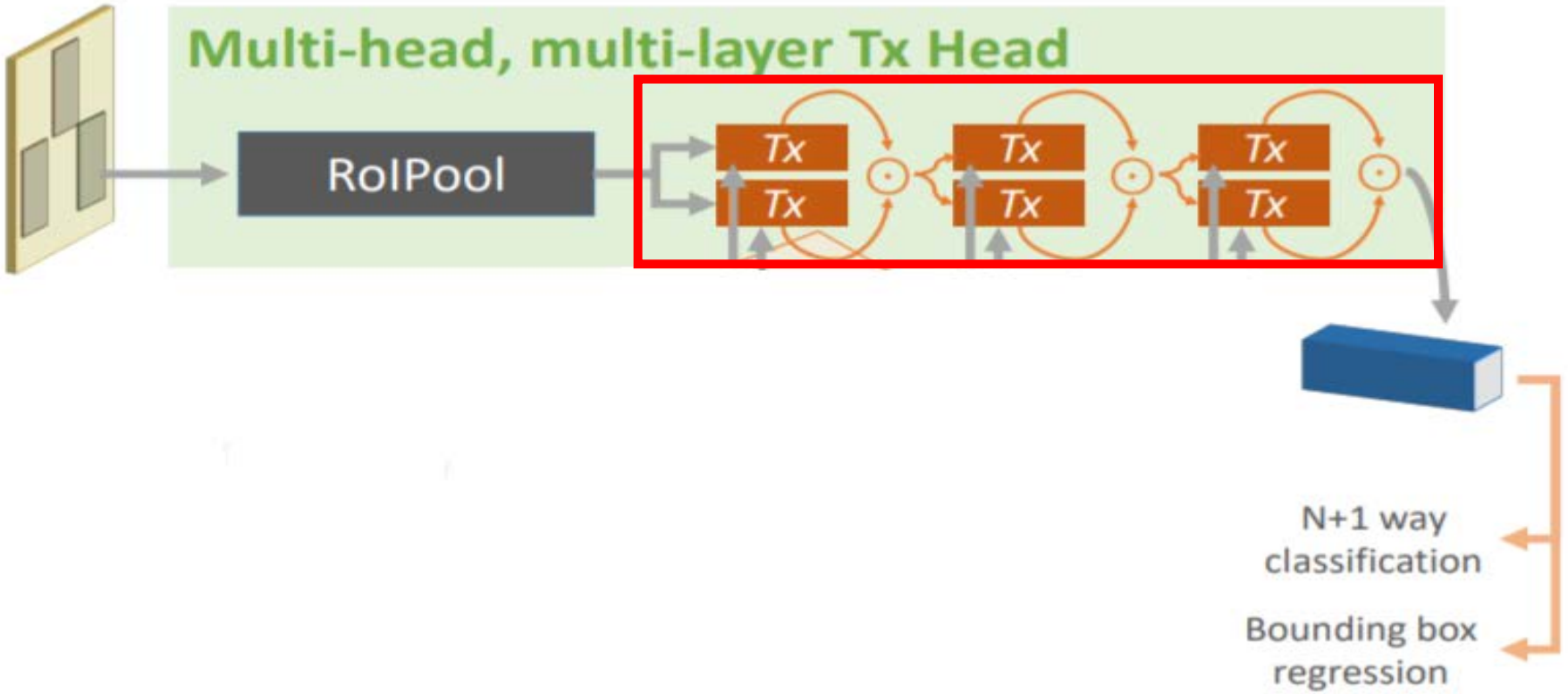
Action Transformer Network



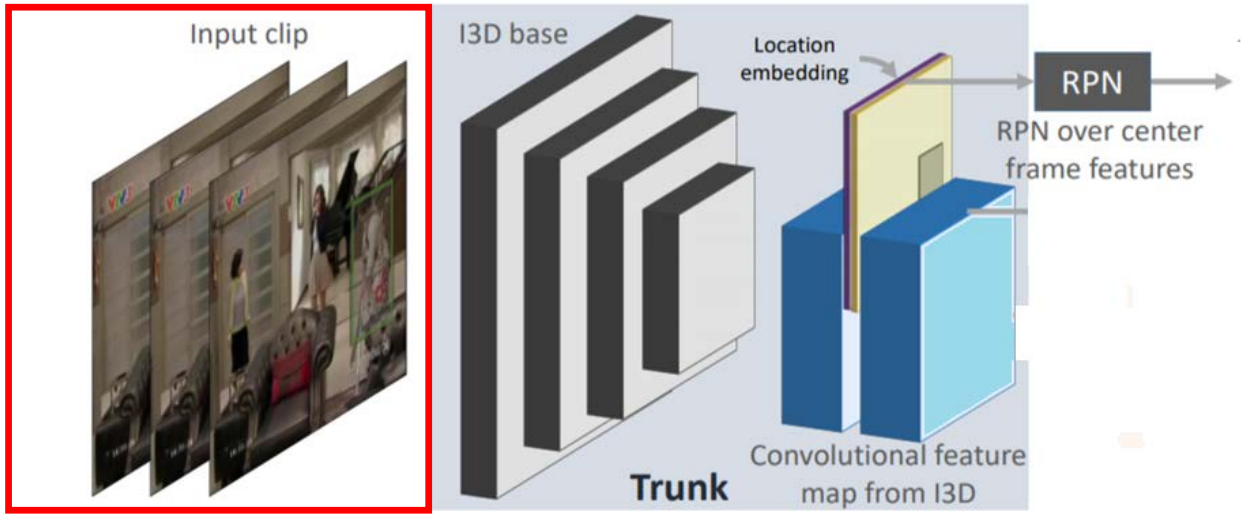
Action Transformer Network



Head Region


Multi-head, multi-layer Tx Head
RoIPool
Tx
Tx
Tx
Tx
Tx
Tx
N+1 way classification
Bounding box regression

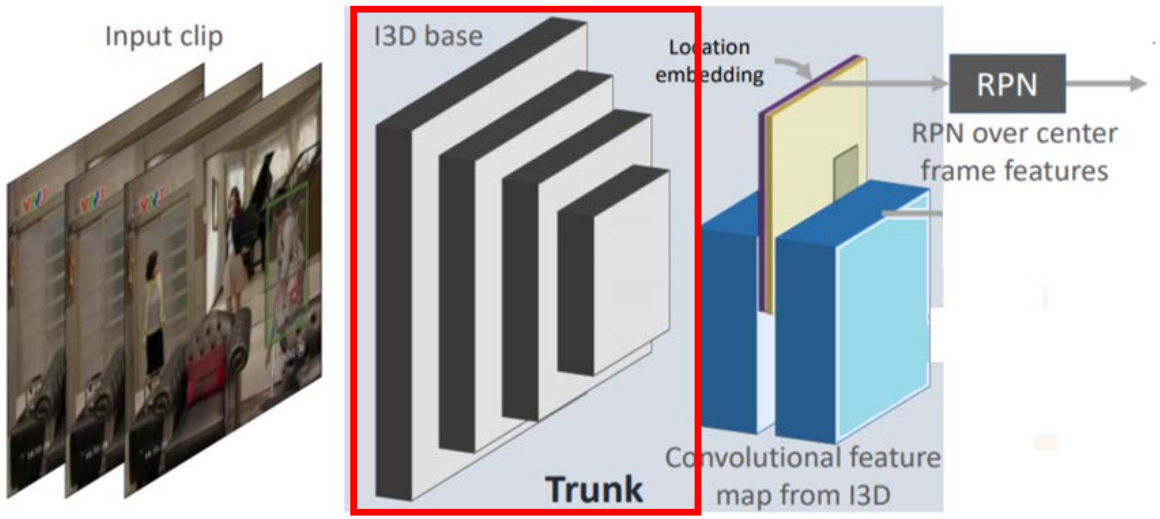
Implementation Details



Input dimension:- [64 x 400 x 400]
[T] x [W] x [H]

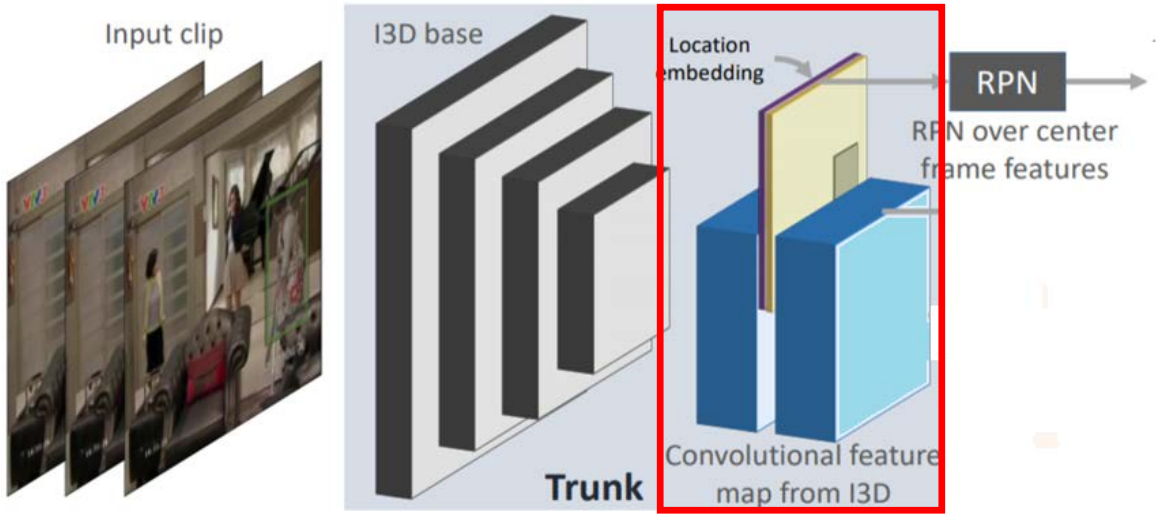
Performed Data Augmentation – Flip, Crop to avoid overfitting.

Implementation Details



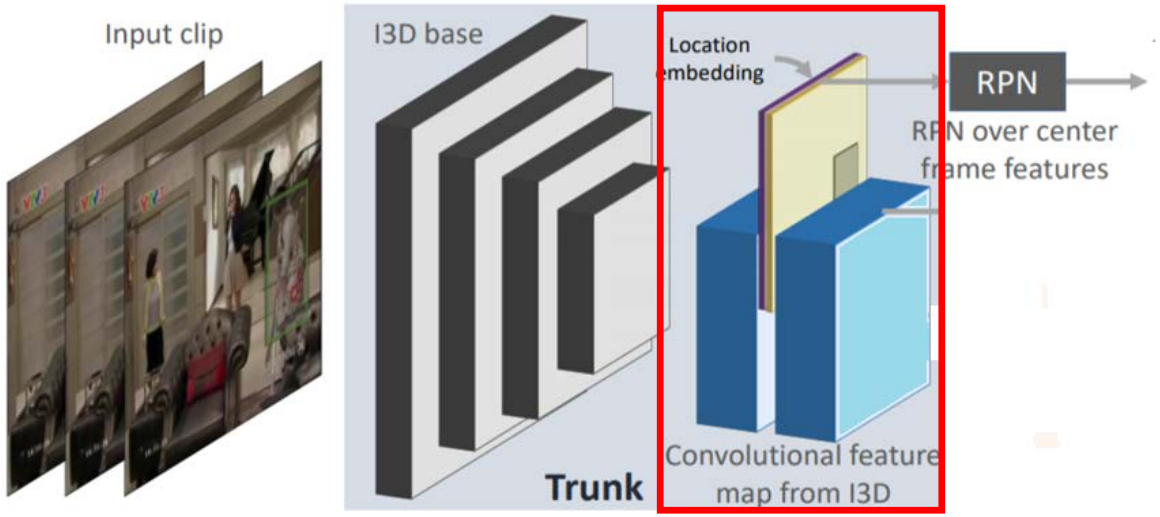
Pre-Trained I3D network (kinetics-400)
Features extracted from `mixed_4f` layer.

Implementation Details



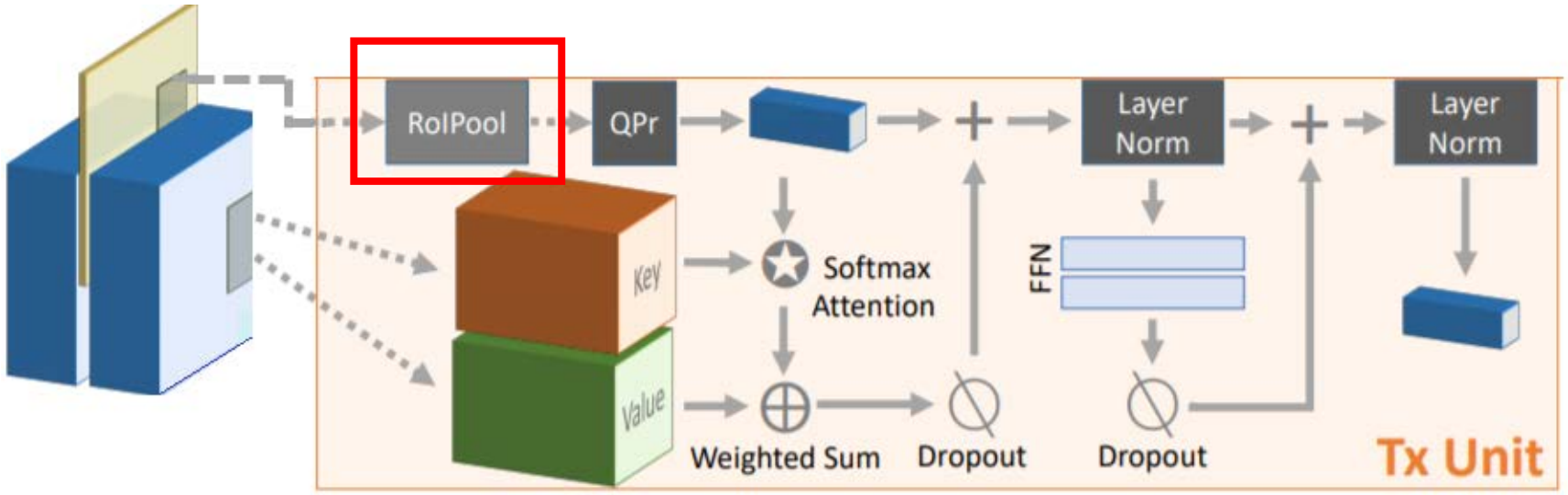
Context feature dimension [16 x 25 x 25].

Implementation Details



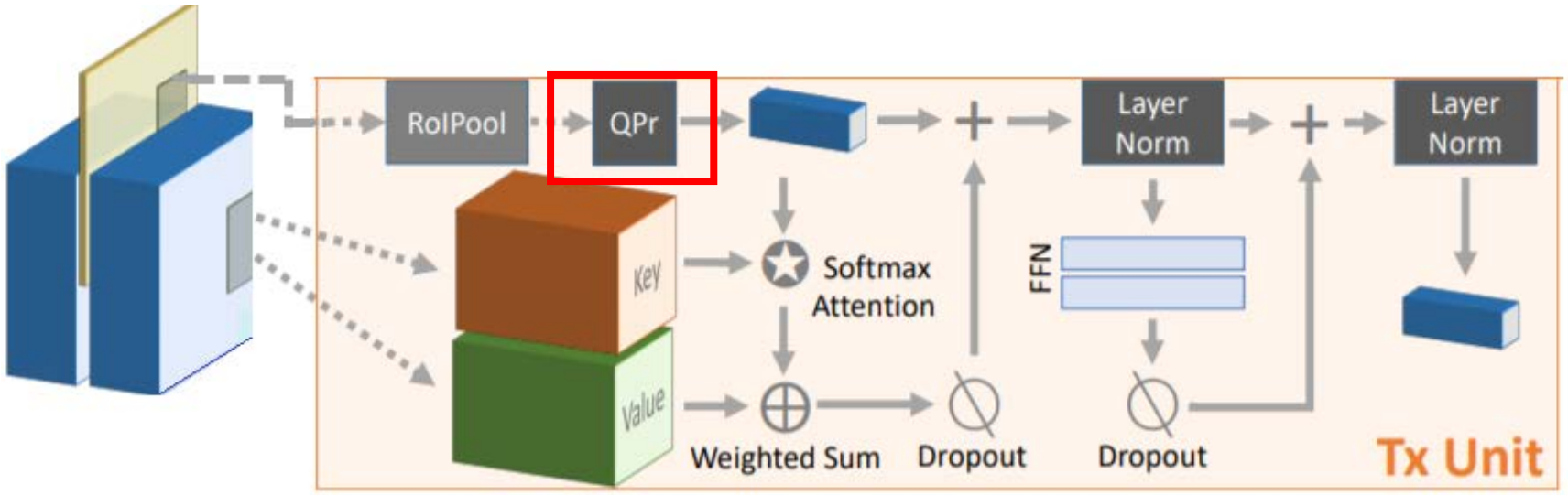
Query feature dimension [16 x 7 x 7].
(Output of RoiPool = 16 x 14 x 14 followed by MaxPool)

Implementation Details



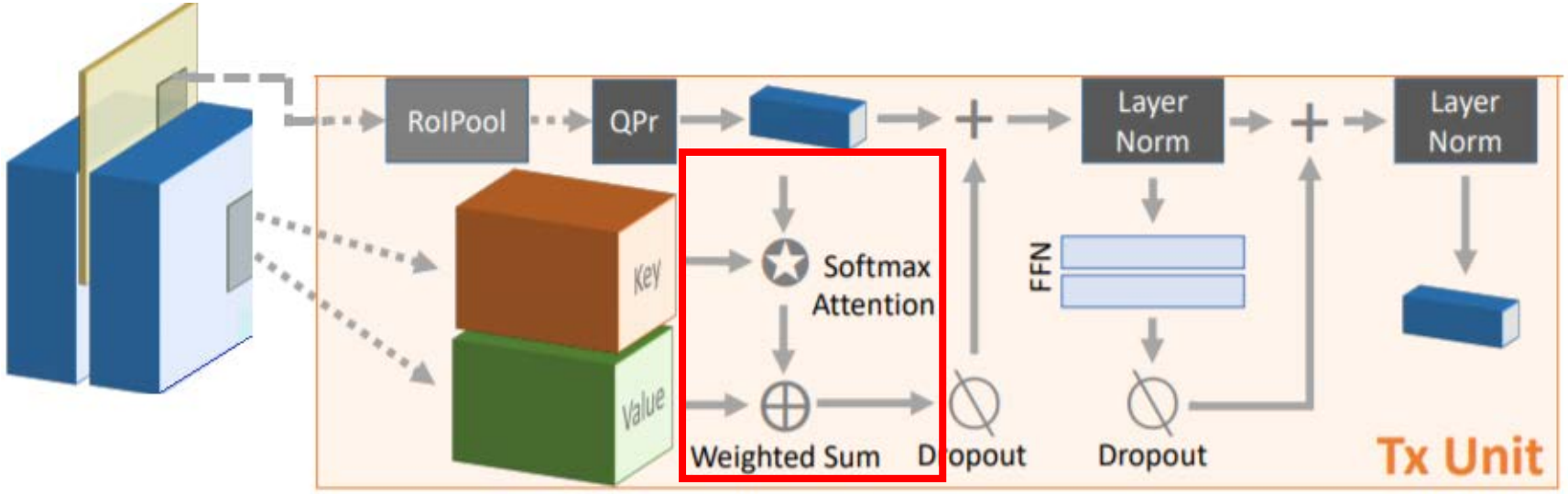
Query -- RoiPool

Implementation Details



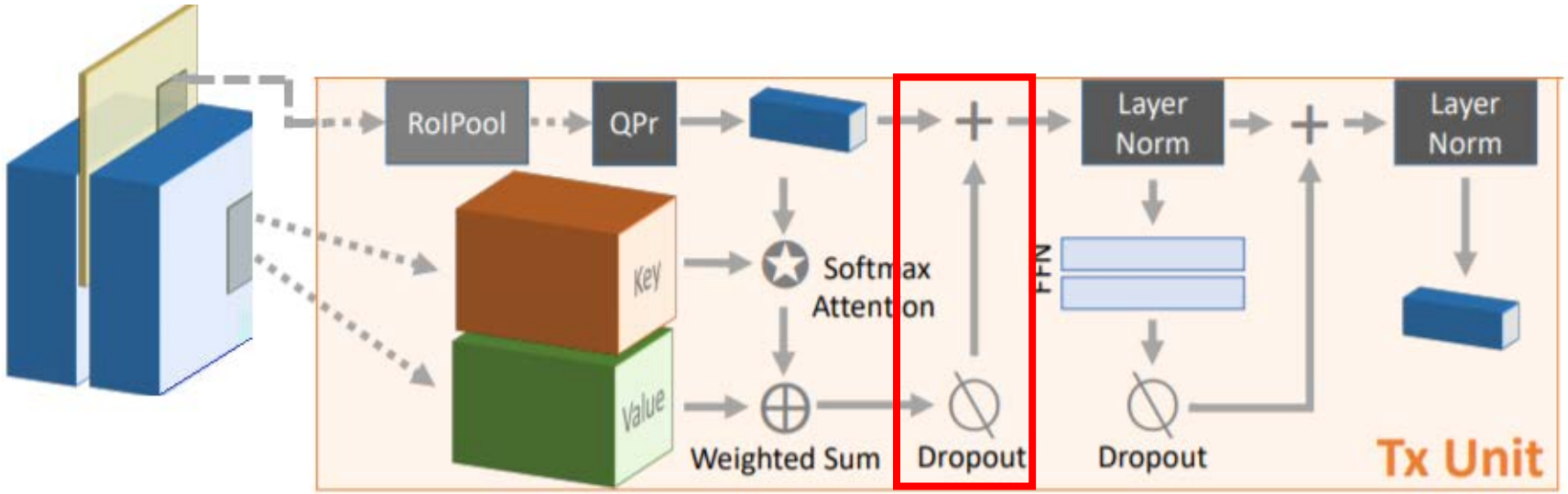
Query – RoiPool – Query Preprocess Q^r

Implementation Details



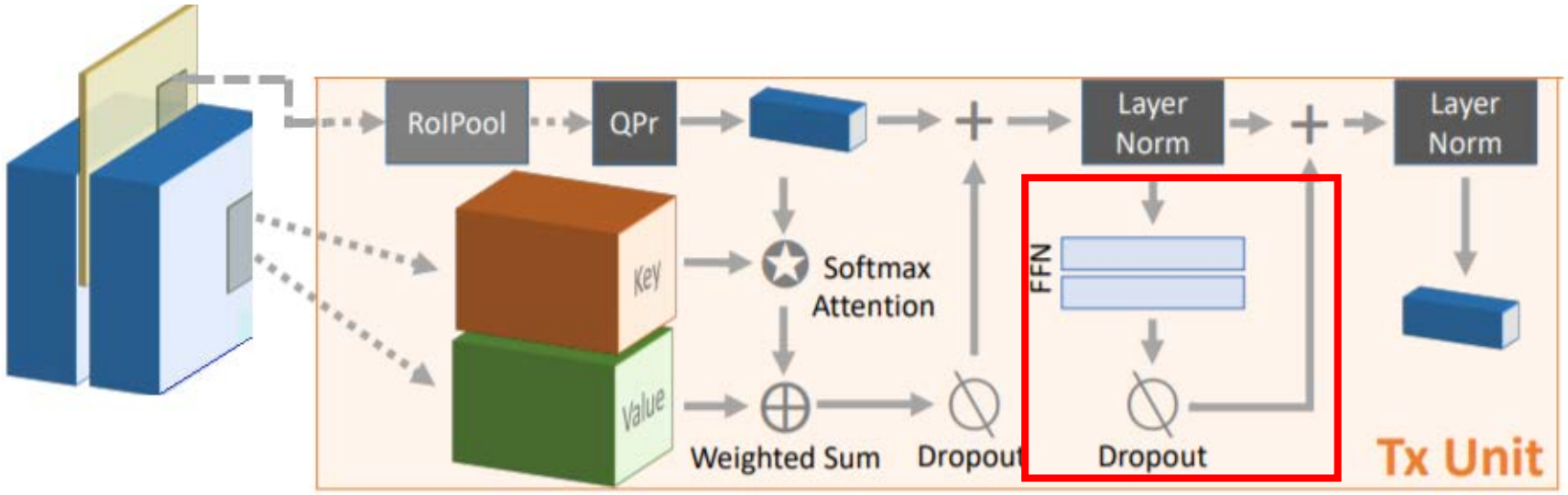
Query – RoiPool – Query Preprocess Q^r – Calculate A^r –

Implementation Details



Query – RoiPool – Query Preprocess Q^r – Calculate A^r – Calculate $Q^{r'}$

Implementation Details



Query – RoiPool – Query Preprocess Q^r – Calculate A^r – Calculate $Q^{r'}$ – Calculate $Q^{r''}$

Experiments – Dataset Description

- Experiments are carried on the AVA dataset.
- 211K training, 57K validation and 117K testing clips.
- Videos captured at 1 fps from 430, 15-minute video clips
- Center frame in clip is labelled with bounding box of all persons available.
- Evaluation metric used – Frame level mAP with IoU of 0.5 threshold.

Experiments

- Case 1: Action classification when GT boxes are used

Trunk	Head	QPr	GT Boxes	Params (M)	Val mAP
I3D	I3D	-		16.2	21.3
I3D	I3D	-	✓	16.2	23.4
I3D	Tx	LowRes		13.9	17.8
I3D	Tx	HighRes		19.3	18.9
I3D	Tx	LowRes	✓	13.9	28.5
I3D	Tx	HighRes	✓	19.3	27.6

Experiments

- Case 2: Localization performance

RoI source	QPr	Head	Val mAP	
			IOU@0.5	IOU@0.75
RPN	-	I3D	92.9	77.5
RPN	LowRes	Tx	77.5	43.5
RPN	HighRes	Tx	87.7	63.3

Experiments

- Case 3: Overall Performance: Action Tx with HighRes preprocessing

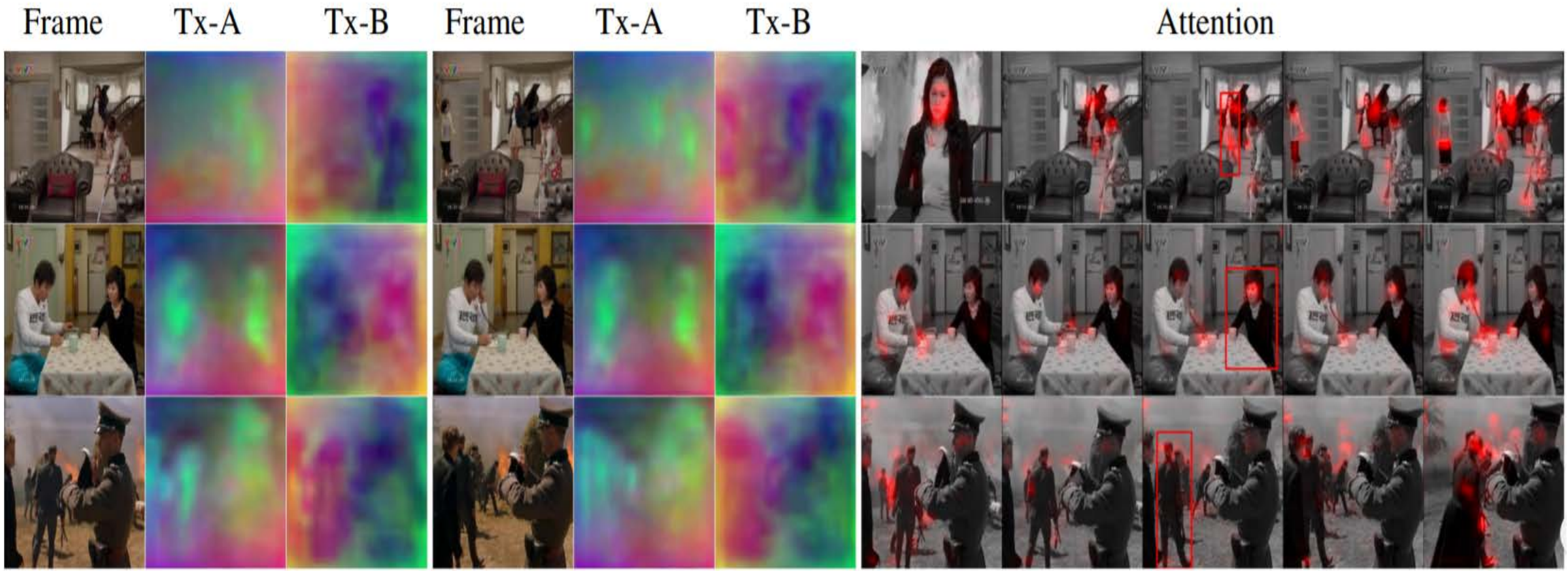
Head	QPr	#proposals	Val mAP
I3D	-	64	21.3
I3D	-	300	20.5
Tx	HighRes	64	18.9
Tx	HighRes	300	24.4
Tx+I3D	HighRes	300	24.9

Experiments

- Case 4: Comparison with State-of-the-Art

Method	Modalities	Architecture	Val mAP	Test mAP
Single frame [14]	RGB, Flow	R-50, FRCNN	14.7	-
AVA baseline [14]	RGB, Flow	I3D, FRCNN, R-50	15.6	-
ARCN [39]	RGB, Flow	S3D-G, RN	17.4	-
Fudan University	-	-	-	17.16
YH Technologies [48]	RGB, Flow	P3D, FRCNN	-	19.60
Tsinghua/Megvii [20]	RGB, Flow	I3D, FRCNN, NL, TSN, C2D, P3D, C3D, FPN	-	21.08
Ours (Tx-only head)	RGB	I3D, Tx	24.4	24.30
Ours (Tx+I3D head)	RGB	I3D, Tx	24.9	24.60
Ours (Tx+I3D+96f)	RGB	I3D, Tx	**25.0**	**24.93**

Qualitative Results



Qualitative Results – Correct Predictions

Qualitative Results – Incorrect Predictions

Smoking Class Incorrectly predicted

Conclusion

- The Action Transformer Network is able to spatio-temporal context from human actions and objects in the clip.
- The embeddings and attention maps have semantic meanings.
- Do not use motion/flow stream but claim that motion information is likely to boost performance.

Thank you