NEURAL INFORMATION PROCESSING SYSTEMS

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Are All Frames Equal? Active Sparse Labeling for Video Action Detection

Aayush J Rana, Yogesh S Rawat

Center For Research In Computer Vision (CRCV)

University Of Central Florida (UCF)

Cost of Video Action Detection

- Spatio-temporal action detection training requires dense training data
  - Dense data $\propto$ large annotation cost
- Dense annotations in videos often have unnecessary cost
  - Repetitive nearby frames
  - Unrelated frames annotated
- Sparse annotation reduces overall annotation cost
  - No standard method to estimate utility of frame for video action detection
- Weakly/semi-supervised approach falls short on performance
  - Large performance gap to fully supervised methods

Motivation

- Build active learning strategy specific to video action detection
  - First work to create AL strategy for frame selection in videos
- Identify most informative frames using active learning
  - Estimate frame utility for video level action detection
  - Only annotate frames that contribute to improving action detection
  - Avoid redundant nearby frame annotation
  - Reduce annotation cost significantly
- Enable sparse learning for video action detection
  - Novel loss formulation that handles sparse annotations
  - Helps in effectively training action detection model from sparse labels

Contributions

- Active Sparse Labeling (ASL)
  - Frame selection strategy using active learning specifically for video action detection task
  - Partial instance annotation by selecting most informative frame for annotation
  - Estimate frame level utility
- Adaptive Proximity-aware Uncertainty (APU)
  - Estimates each frame's utility
  - Uses model uncertainty and proximity to existing annotations
  - Avoids selecting low utility and repetitive frames
- Max-Gaussian Weighted Loss (MGW-Loss)
  - Enables effective action detection learning from sparse labels
  - Uses weighted pseudo-labeling to assign appropriate weight to each frame

Adaptive Proximity-aware Uncertainty (APU)

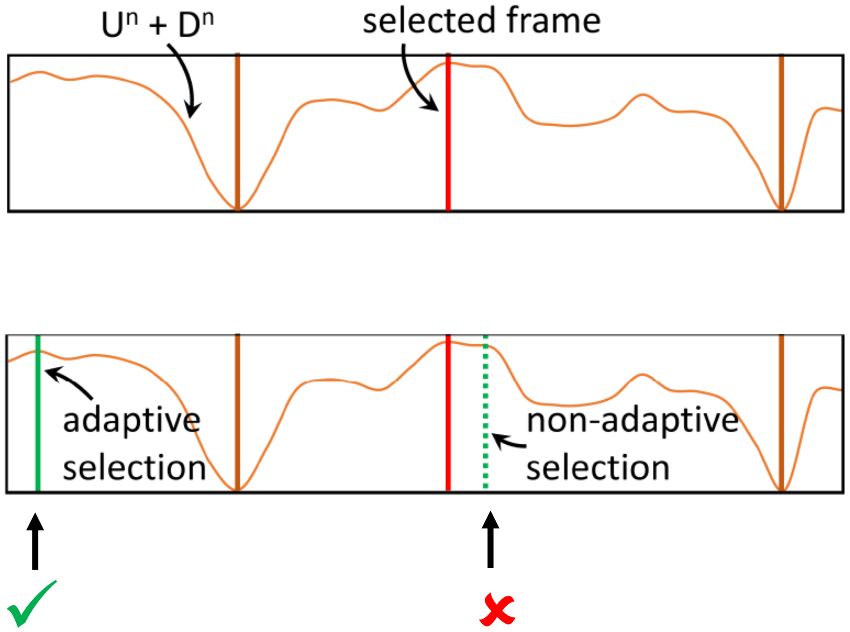
- Uncertainty as frame utility
  - Use MC-dropout as model's uncertainty for each pixel and average them for frame score

$$U^{i\in[1,I]} = \frac{1}{I^p}\sum_{h=1}^{I^p}\frac{1}{T}\sum_{j=1}^{T} -log(P(v_h^i, j))$$

- Adaptive proximity estimation
  - We use a normal distribution centered around annotated frame

$$D^i = 1 - \sum_{j=1}^{K}\varphi_i^j e^{-\frac{1}{2}(\frac{i-\mu_j}{\sigma})^2}$$

- Overall APU is computed as $\mathcal{U}_{APU}^i = \lambda U^i + (1-\lambda)D^i$



Active Sparse Labeling (ASL)

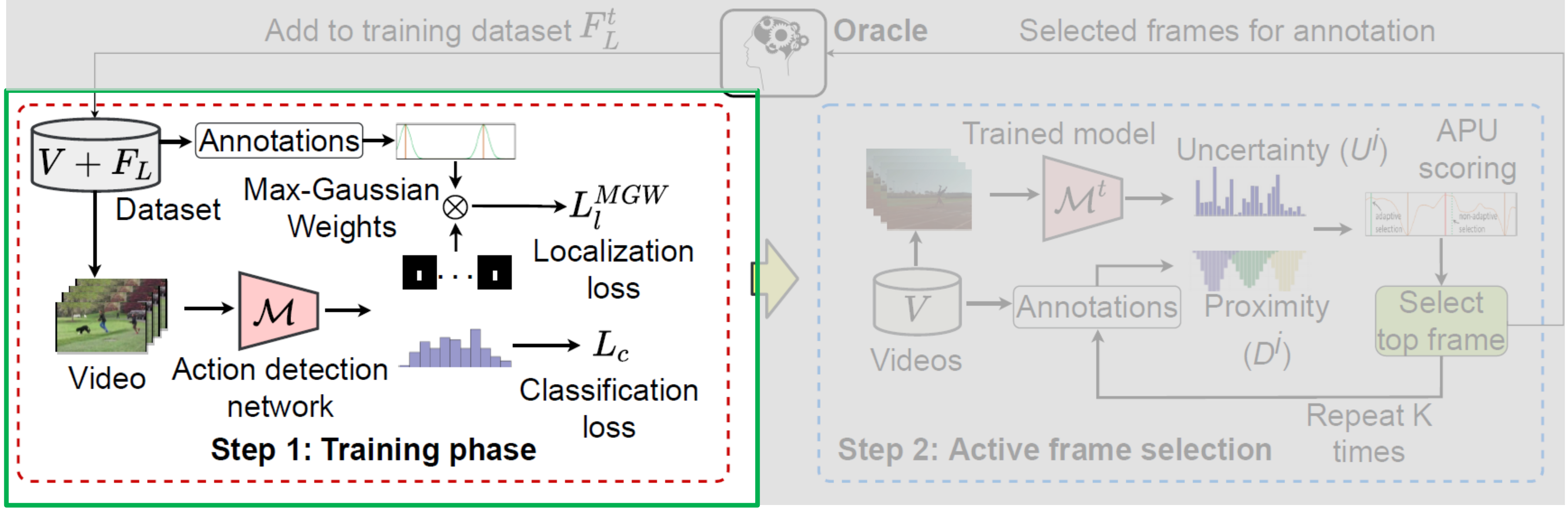
- Estimate each frame's utility using APU
  - APU adjusts for redundancy and diversity of frames
- Informative frame selection
  - Select highest utility frame
  - Re-score remaining frames using APU again
    - Only re-compute distance measure (no model inference required)
  - Select frames based on budget for AL round
- Non-activity suppression
  - Avoid influence of large background regions
  - Ignore highly certain background pixels for APU computation
  - Focus more of possible foregrounds (action region)

Max-Gaussian Weighted Loss (MGW-Loss)

- Handle pseudo-label and actual labels effectively
  - Pseudo-labels closer to ground truth are more reliable
  - Approximated pseudo-labels can still be used but with low weight
  - Use mixture of Gaussian distribution to assign weight

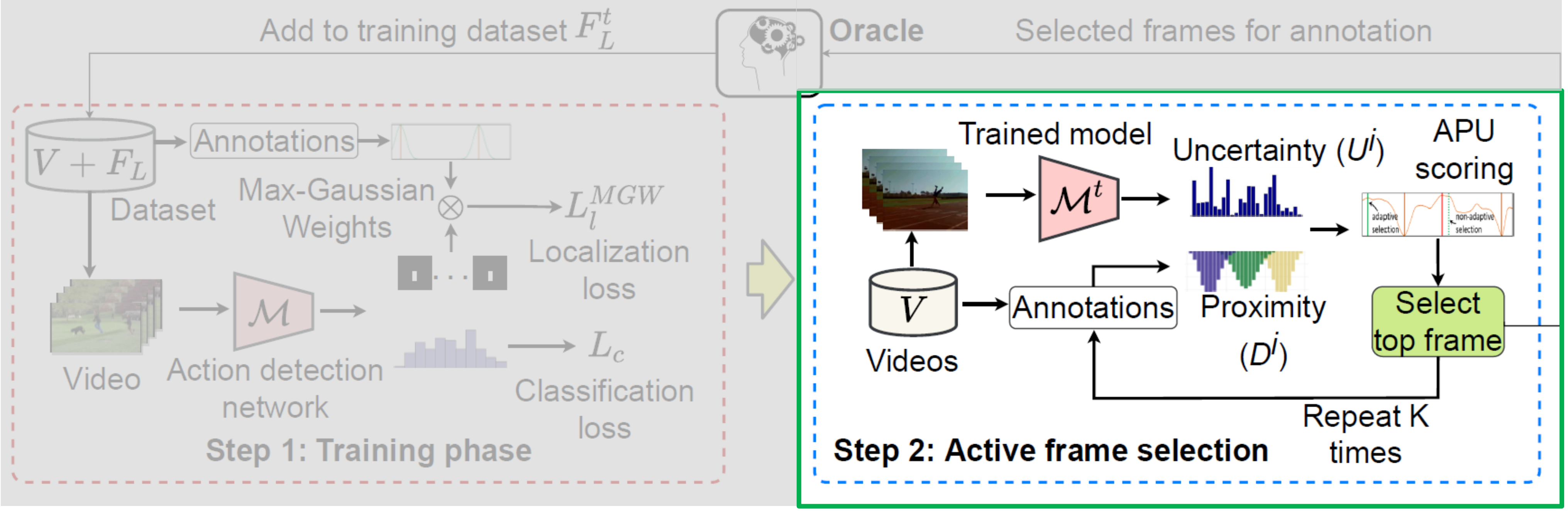
$$L_l^{MGW} = \sum_{i=1}^{N} \left(\sum_{j=1}^{K} \phi_j^i e^{-\frac{1}{2}\left(\frac{i-\mu_j}{\sigma}\right)^2} \right) L_l^i$$

Proposed approach



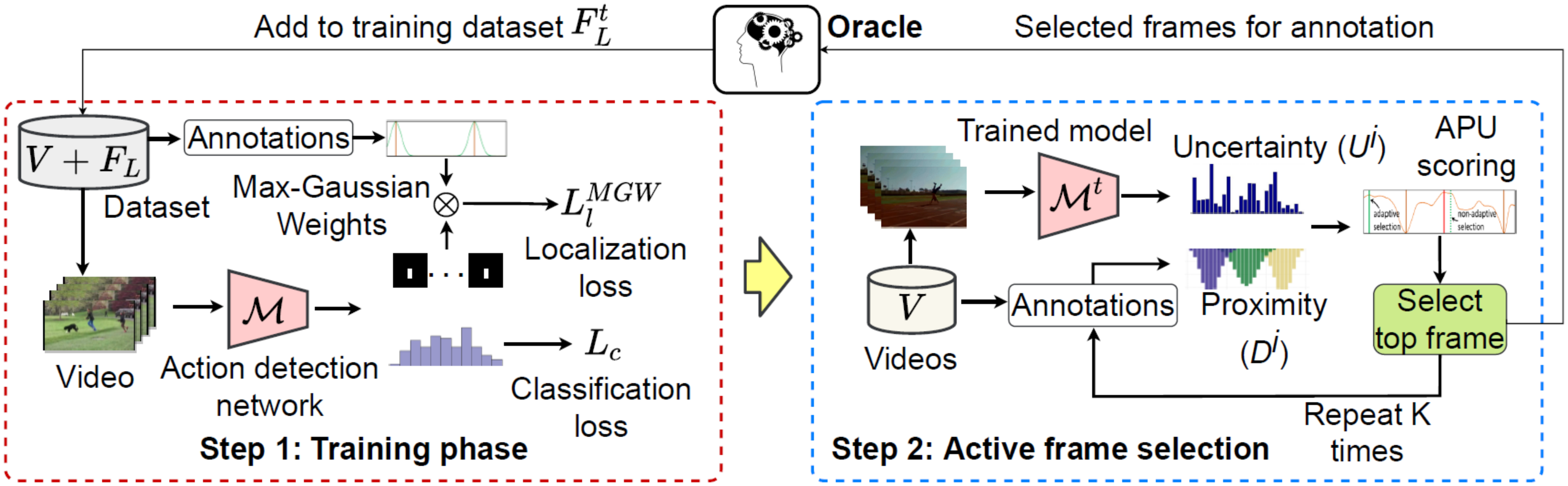
- Train model with pseudo-labels and MGW-loss

Proposed approach



- Select new frames using APU scoring and active sparse labeling strategy

Proposed approach



- Send selected frames to oracle for annotation
- Train new model with increased annotations

Datasets

- UCF-101
  - 3207 videos
  - **24** action classes
  - Spatio-temporal bounding box annotation
- J-HMDB
  - 928 videos
  - **21** action classes
  - Spatio-temporal pixel-wise annotation
- YouTube-VOS
  - 3471 training videos
  - **65** object categories
  - Sparse pixel-level annotation

Results on UCF-101 and J-HMDB

	UCF-101						J-HMDB					
	f-mAP@0.5			v-mAP@0.5			f-mAP@0.5			v-mAP@0.5		
Method	1%	5%	10%	1%	5%	10%	3%	6%	9%	3%	6%	9%
Random	60.7	66.5	69.3	59.2	66.4	69.9	58.3	69.3	71.6	57.4	64.6	70.4
Equidistant	61.8	66.2	68.4	61.7	67.2	69.0	57.4	67.5	71.4	56.9	64.9	66.8
G* [73]	60.9	66.7	68.9	59.4	66.8	69.1	58.2	66.7	67.5	57.4	66.8	67.4
A* [53]	61.4	67.9	69.8	60.1	67.9	70.0	58.8	71.2	71.1	57.7	66.7	71.2
Our	**64.7**	**70.9**	**71.7**	**63.9**	**71.8**	**73.2**	**68.8**	**74.1**	**74.5**	**65.6**	**70.8**	**74.0**

[53] Hamed H Aghdam, Abel Gonzalez-Garcia, Joost van de Weijer, and Antonio M López. Active learning for deep detection neural networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[73] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In International Conference on Machine Learning, 2016.

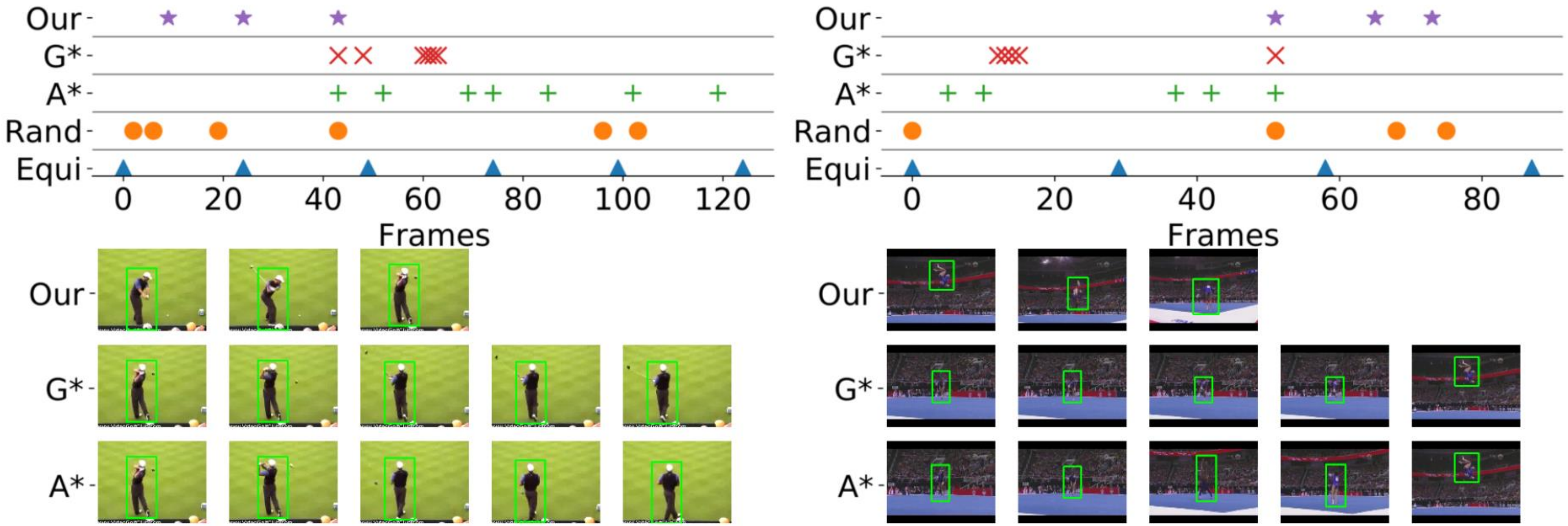
NEURAL INFORMATION PROCESSING SYSTEMS

UCF CENTER FOR RESEARCH IN COMPUTER VISION

Comparison with state-of-the-art

Method	Annot Percent	V	P	B	O	UCF-101 f@0.5	UCF-101 v-mAP@0.1	UCF-101 v-mAP@0.2	UCF-101 v-mAP@0.3	UCF-101 v-mAP@0.5	J-HMDB f@0.5	J-HMDB v-mAP@0.1	J-HMDB v-mAP@0.2	J-HMDB v-mAP@0.3	J-HMDB v-mAP@0.5
Fully supervised															
Peng et al. [7]	100%					65.7	77.3	72.9	65.7	35.9	58.5	-	74.3	-	73.1
TCNN [8]	100%					67.3	77.9	73.1	69.4	-	61.3	-	78.4	-	-
Gu et al. [78]	100%					76.3	-	-	-	59.9	73.3	-	-	-	-
ACT [85]	100%					69.5	-	76.5	-	-	-	-	74.2	-	73.7
STEP [10]	100%					75.0	83.1	76.6	-	-	-	-	-	-	-
VidsCapsNet [9]	100%					78.6	98.6	97.1	93.7	80.3	64.6	98.4	95.1	89.1	61.9
Weakly/Semi-supervised															
Mettes et al. [20]	Video	✓			✓	-	-	37.4	-	-	-	-	-	-	-
Escorcia et al. [24]	Video	✓				-	-	45.5	-	-	-	-	-	-	-
Zhang et al. [25]	Video	✓			✓	30.4	62.1	45.5	-	17.3	65.9	81.5	77.3	-	50.8
Arnab et al. [42]	Video	✓			✓	-	-	61.7	-	35.0	-	-	-	-	-
Weinz. et al. [26]	Partial	✓		✓	✓	63.8	-	57.3	-	46.9	56.5	-	-	-	64.0
Mettes et al. [21]	Partial	✓	✓			-	-	41.8	-	-	-	-	-	-	-
Cheron et al. [23]	Partial	✓			✓	-	-	70.6	-	38.6	-	-	-	-	-
Kumar et al. [86]	20%	✓		✓		69.9	-	95.7	-	72.1	64.4	-	95.4	-	63.5
Ours	10%	✓		✓		71.7	98.1	96.5	91.1	73.2	74.5	99.2	98.4	95.6	74.0
Ours	100%					74.0	98.3	96.9	91.5	75.2	74.9	99.2	99.2	96.4	75.8

Qualitative frame selection analysis

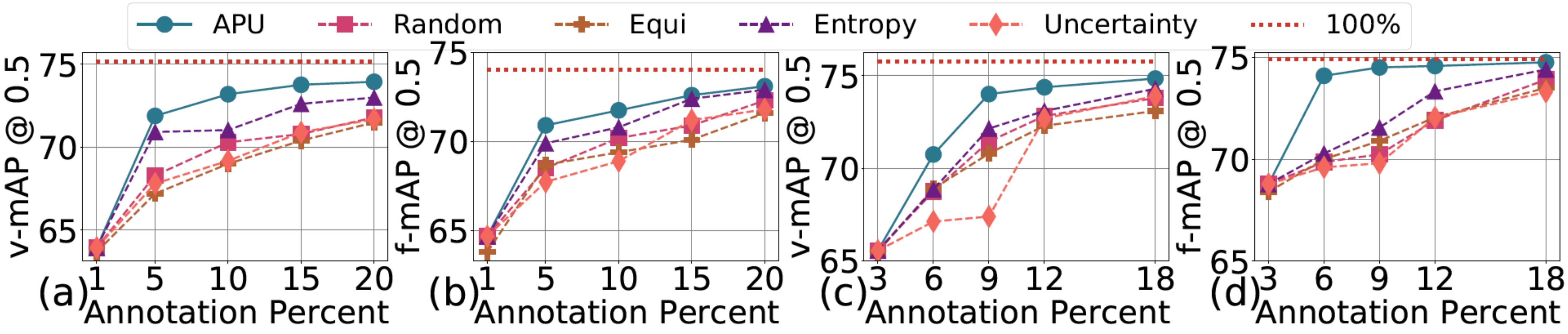


- We select fewer frames with higher diversity and utility
  - Performs better than G* [53], A* [73] (prior methods) and random and equidistant selection

[53] Hamed H Aghdam, Abel Gonzalez-Garcia, Joost van de Weijer, and Antonio M López. Active learning for deep detection neural networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[73] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In ICML, 2016.

Evaluating different frame selection methods

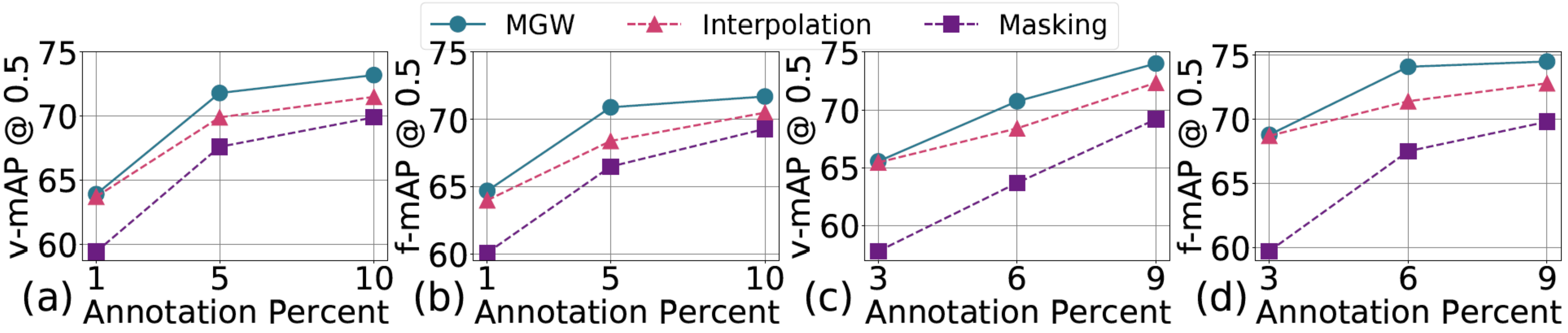


(a-b) UCF-101

(c-d) J-HMDB

- All methods use our MGW-loss to handle sparse labels
- APU performs better at lower annotation cost
  - Handles proximity in videos better than entropy and uncertainty methods

Analyzing various loss formulations



(a-b) UCF-101

(c-d) J-HMDB

- Masking doesn't utilize pseudo-labels and performs lower
- Interpolation improves overall detection
- MGW uses weight based on proximity to ground truth
  - Directs network on how much to trust pseudo-labels based on reliability

Performance on YouTube-VOS

Method	Overall			$\mathcal{J}_S$			$\mathcal{J}_U$			$\mathcal{F}_S$			$\mathcal{F}_U$		
	10%	20%	30%	10%	20%	30%	10%	20%	30%	10%	20%	30%	10%	20%	30%
Random	28.4	42.3	42.5	29.1	42.9	43.8	25.8	38.5	38.6	30.2	44.3	45.0	28.4	43.5	42.7
A * [53]	30.1	45.6	47.2	31.5	45.4	47.6	26.7	43.4	47.9	22.8	46.7	48.8	17.6	46.8	44.6
G * [73]	27.9	45.1	48.8	28.5	50.8	48.5	24.8	42.0	46.6	29.7	42.1	49.8	28.7	45.5	50.4
Our	**31.7**	**58.6**	**66.7**	**33.6**	**58.2**	**66.7**	**27.8**	**54.3**	**61.5**	**35.2**	**60.6**	**69.1**	**30.1**	**60.9**	**69.7**

- Generalization of proposed method on video object segmentation task

Findings

- APU helps select diverse and useful frames for annotation
- MGW-loss is effective in using pseudo-labels to train action detection
- Lower increment step selects fewer frames with higher utility
  - Each step only selects most useful frames
  - Improves overall selection but takes more AL rounds
- Global frame selection outperforms local selection
  - Enables difficult videos to get more frames
- Selecting sparse frames more valuable than annotating entire videos

Conclusion

- ASL is first active learning strategy specific for video action detection
- APU scoring identifies frames with higher diversity and utility
- MGW-loss is simple and effective at handling sparse labels
- ASL saves annotation cost by 90% and performs close to fully supervised
- ASL can generalize to video object segmentation task

Thank You

Project Page

https://www.crcv.ucf.edu/research/projects/active-sparse-labeling-for-video-action-detection/