Video Threat Detection via Image and Audio Ensemble

TASK FORMULATION

Video Data → Image → YOLOv5 → Cropped bounding box
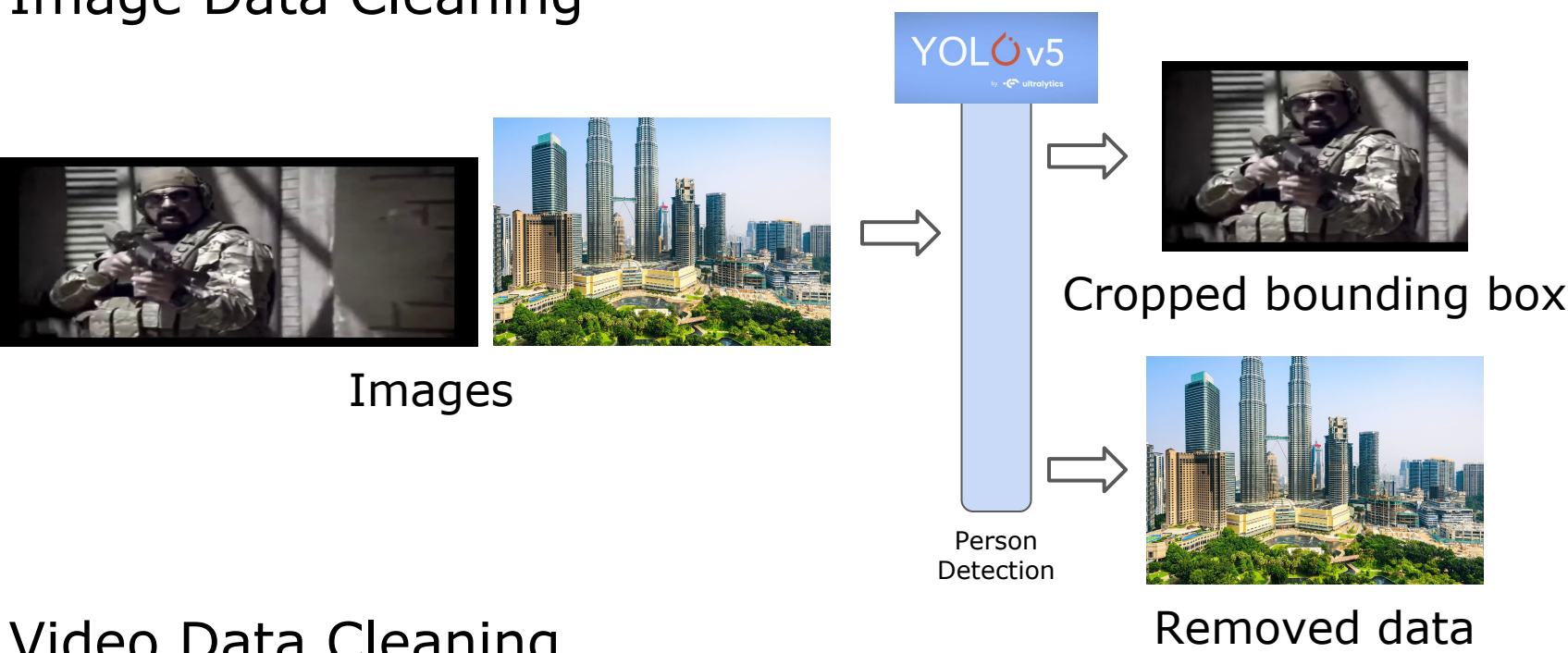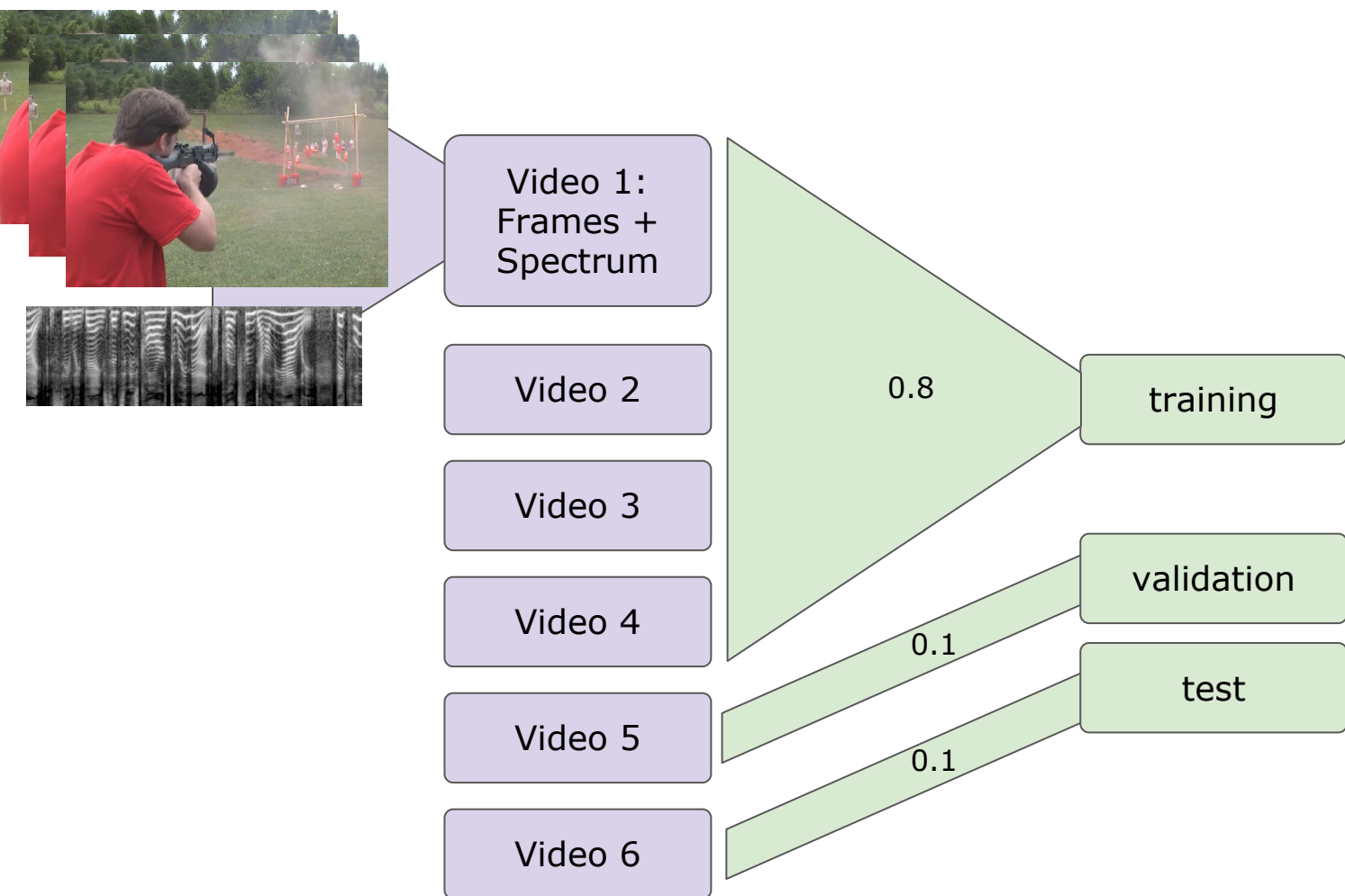Video Data → Audio → librosa → Mel Spectrum
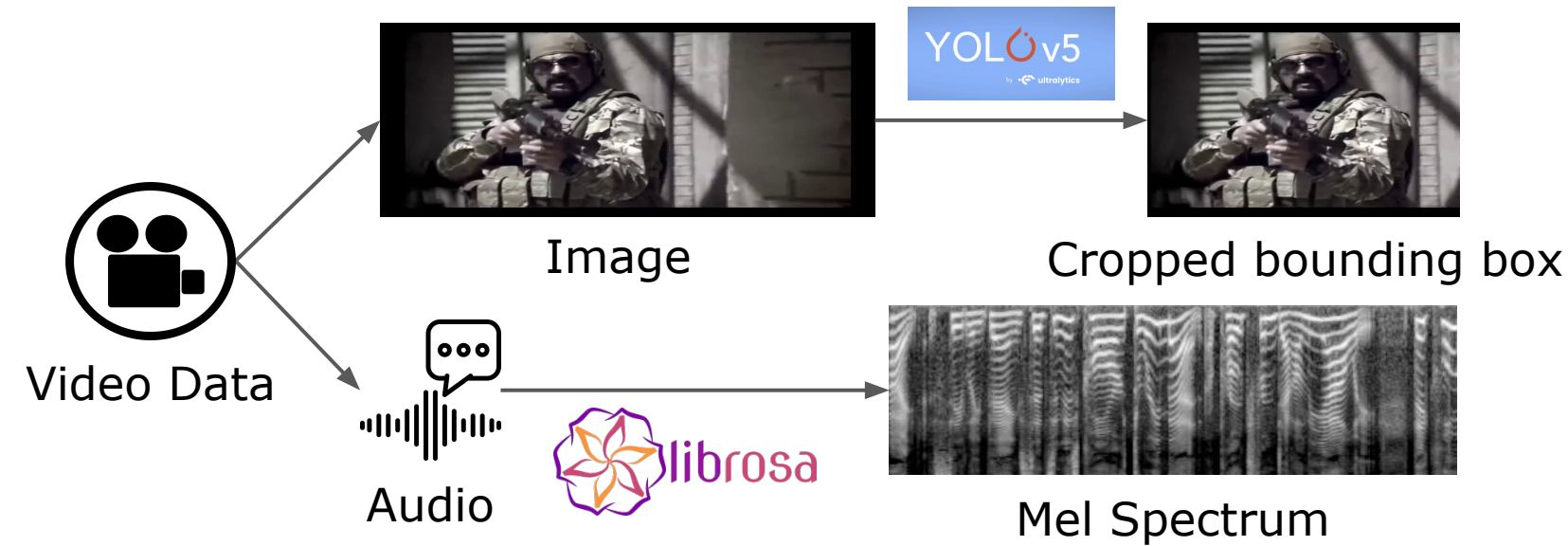
METHODOLOGY

Tools: PyTorch, colab   Models: YOLOv5, INCEPTION

DATA CLEANING

Image Data Cleaning

Images → YOLOv5 (Person Detection) → Cropped bounding box / Removed data

Video Data Cleaning

Video 1: Frames + Spectrum, Video 2, Video 3, Video 4, Video 5, Video 6
0.8 → training
0.1 → validation
0.1 → test

CONTRIBUTIONS
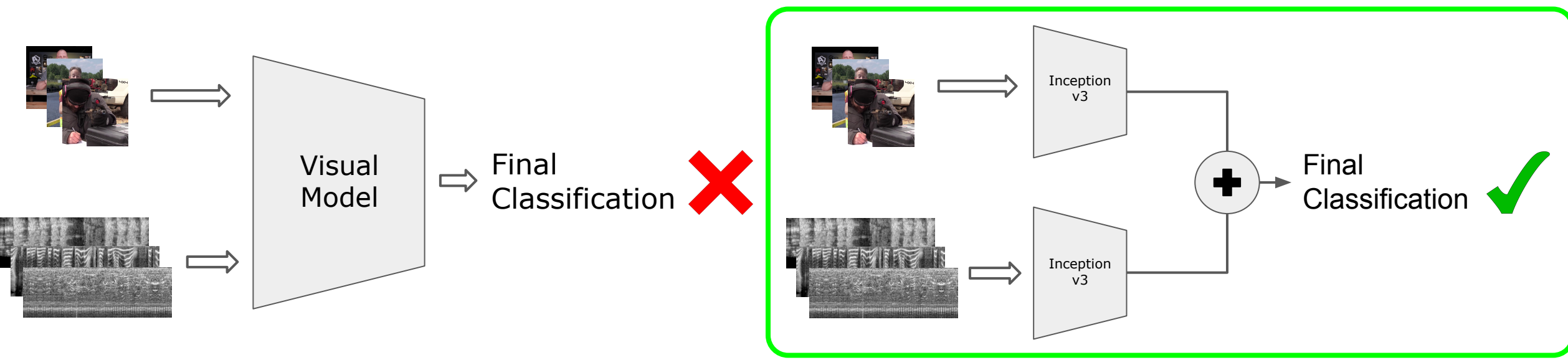
Visual Model → Final Classification ❌

Inception v3 + Inception v3 → Final Classification ✔
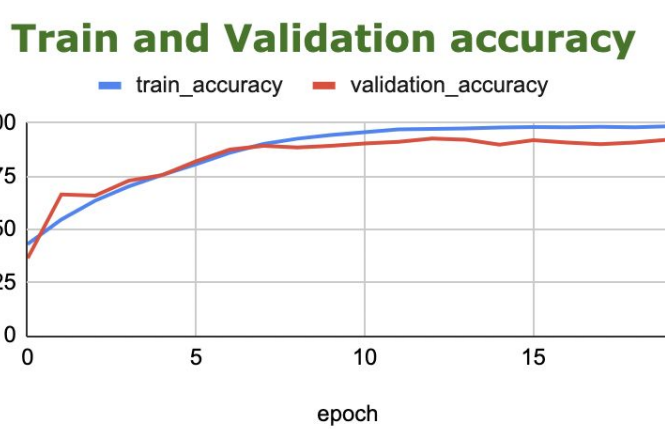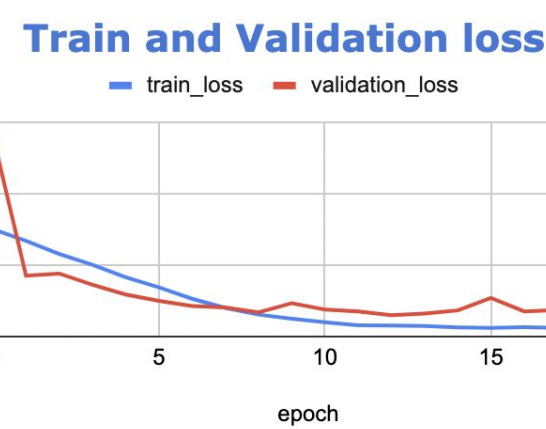
IMAGE CLASSIFICATION

Train data
32 x (299 x 299 x 3) → Inception v3
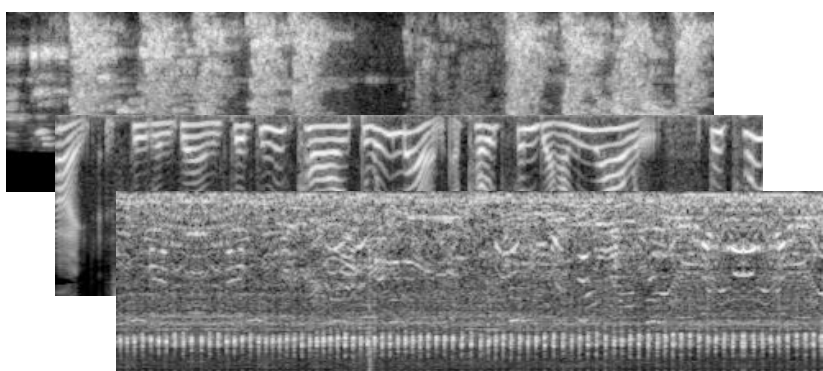
Parameter	Value
optimizer	adam
epochs	20
learning_rate	1e-3
scheduler	Exp decay, gamma = 0.7

Train and Validation loss — train_loss, validation_loss

Train and Validation accuracy — train_accuracy, validation_accuracy

predict			
true	normal	carrying	threat
normal	170	5	4
carrying	19	228	14
threat	0	0	142

Accuracy = 0.9375

AUDIO CLASSIFICATION

Train data
32 x (299 x 299 x 3) → Inception v3

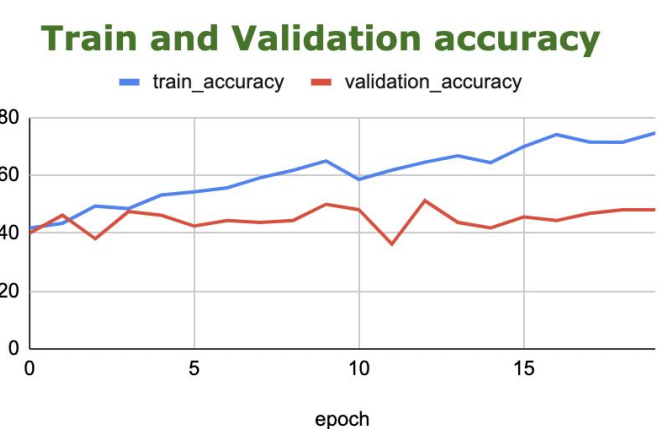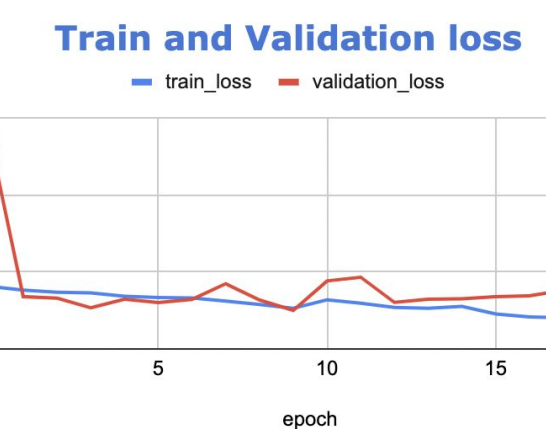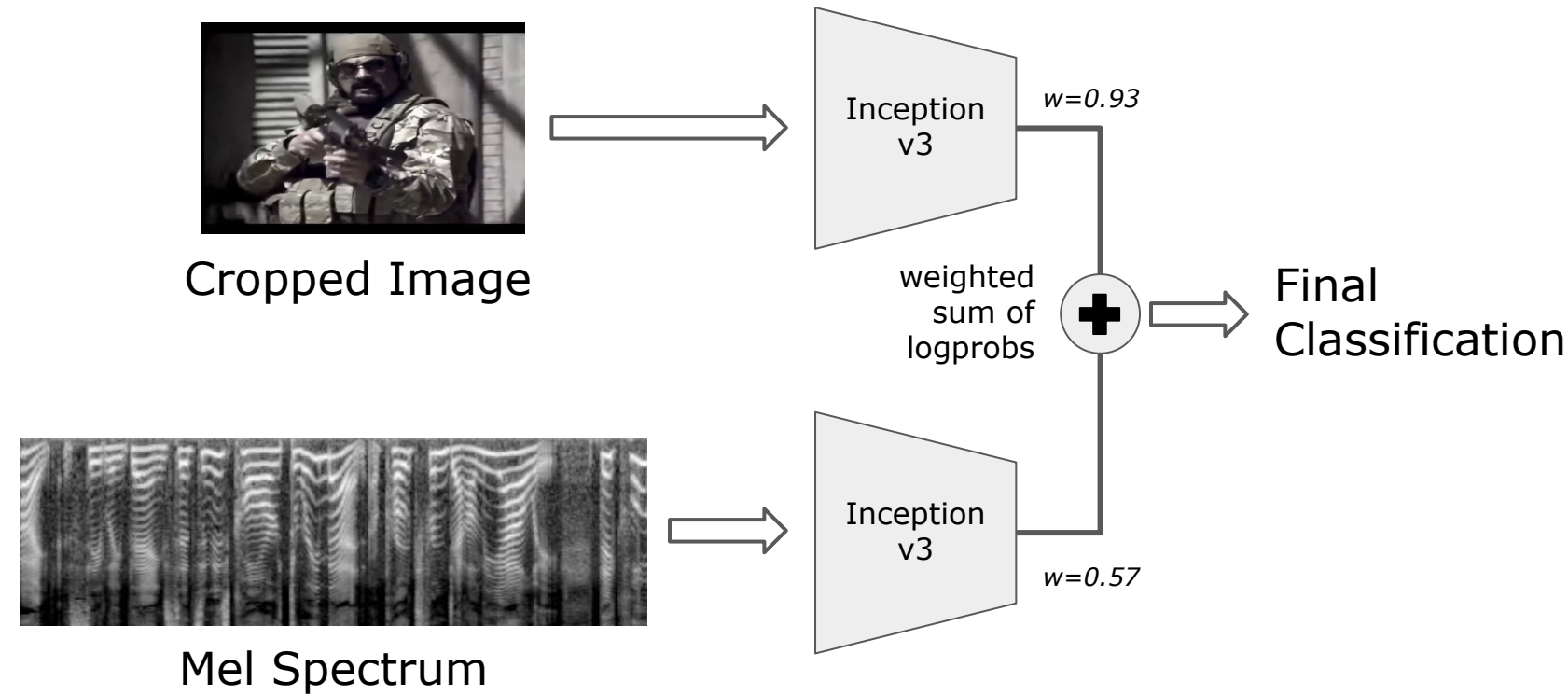| Parameter	Value
optimizer	adam
epochs	20
learning_rate	1e-3
scheduler	Exp decay, gamma = 0.9

Train and Validation loss — train_loss, validation_loss

Train and Validation accuracy — train_accuracy, validation_accuracy

predict			
true	normal	carrying	threat
normal	24	16	10
carrying	2	42	10
threat	5	27	25

Accuracy = 0.5688

ENSEMBLE

Ensemble Schema

Cropped Image → Inception v3 (w=0.93)
Mel Spectrum → Inception v3 (w=0.57)
weighted sum of logprobs → Final Classification

predict			
true	normal	carrying	threat
normal	113	37	29
carrying	7	230	24
threat	1	8	133

Accuracy = 0.8264

DEMO - LINK

Video Data → Sampled Image & Audio → Cropped Image & Audio → Ensemble → Classification → Bounding Box Detection (threat)

FUTURE WORK

Train on Video Data   Cleaner Data   Investigate combined model