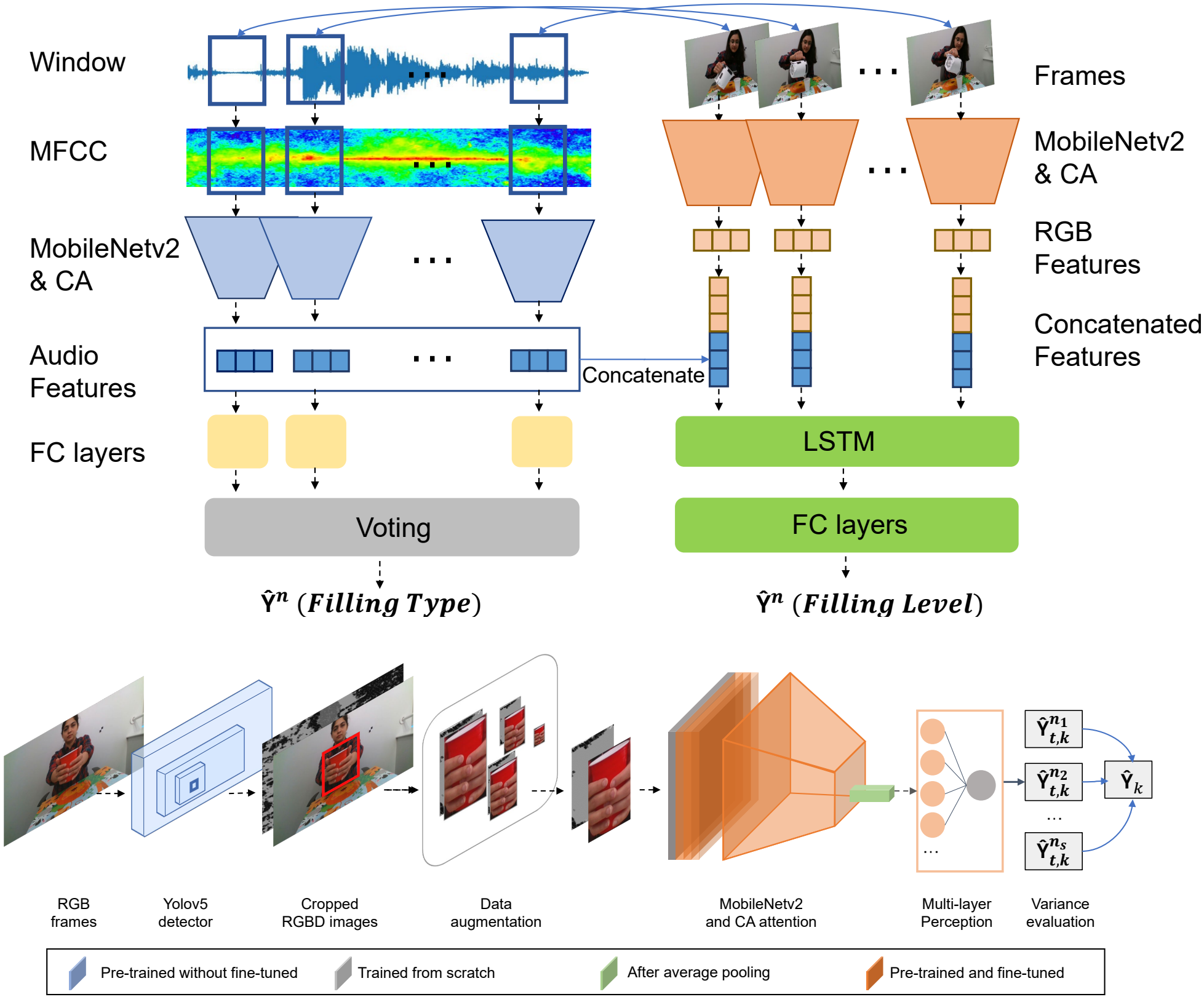
Window
MFCC
MobileNetv2 & CA
Audio Features
FC layers
Voting
$\hat{Y}^n$ (Filling Type)
Concatenate
Frames
MobileNetv2 & CA
RGB Features
Concatenated Features
LSTM
FC layers
$\hat{Y}^n$ (Filling Level)
RGB frames
Yolov5 detector
Cropped RGBD images
Data augmentation
MobileNetv2 and CA attention
Multi-layer Perception
Variance evaluation
$\hat{Y}_{t,k}^{n_1}$
$\hat{Y}_{t,k}^{n_2}$
$\hat{Y}_{t,k}^{n_s}$
$\hat{Y}_k$
Pre-trained without fine-tuned
Trained from scratch
After average pooling
Pre-trained and fine-tuned