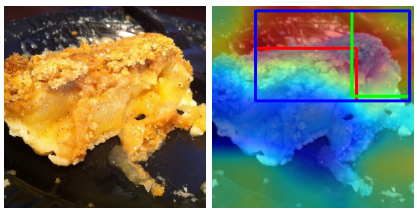
Attend and Guide (AG-Net): A Keypoints-driven Attention-based Deep Network for Image Recognition

Supplementary Document

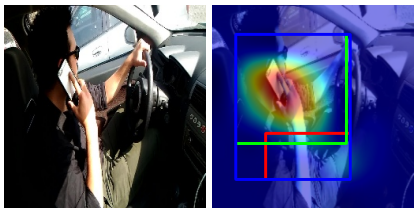
We have included additional visualizations and experimental results (e.g. Confusion matrix) related to our manuscript.



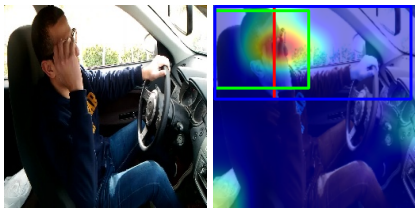
(a) Food-101: Apple-pie

(b) Food-101: Ice-cream

(c) Food-101: Strawberry shortcake

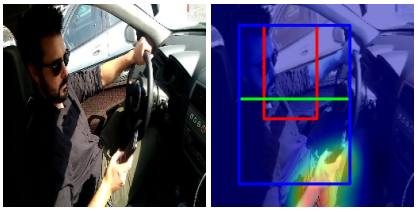


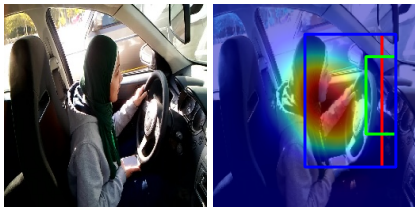
(d) AUC-V1: talking right

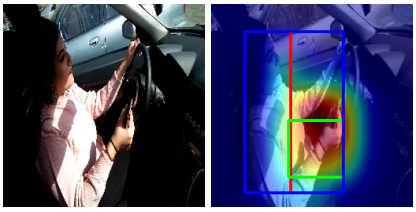


(e) AUC-V1: hair and makeup

(f) AUC-V1: texting right







(g) AUC-V1: texting-right

Figure 12: **Visualization of class activation maps** using the Gradient-weighted Class Activation Mapping (Grad-CAM) [67]. The top and middle rows illustrate **inter-class** variations with three different classes from two datasets: (a-c) Food-101 [31], and (d-f) AUC-V1 [20]. The last row (g) represents the same action from AUC-V1 [20] dataset. Each example contains the original image (left) and corresponding activation map of salient regions (right) on which we have overlaid only three SRs for clarity. It contains two primary SRs (red and green) along with a secondary region (blue) which is derived from those two primary SRs. It is related to Fig.10 (Section V-F) in the manuscript. Best view in color.

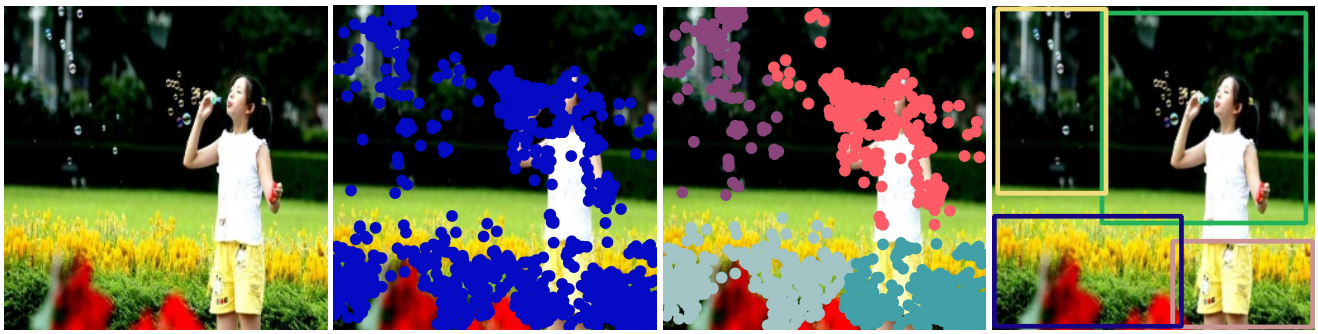


Figure 13: Example images for blowing bubbles from Stanford-40 dataset [24]. Using our keypoints-based clustering for detecting primary SRs. In this example, four primary semantic regions (SRs) are detected: original image ⇒ detected SIFT keypoints ⇒ clustered keypoints ⇒ bounding boxes enclosing SRs (left to right). It is related to Fig.3 (Section III) in the manuscript. Best view in color.

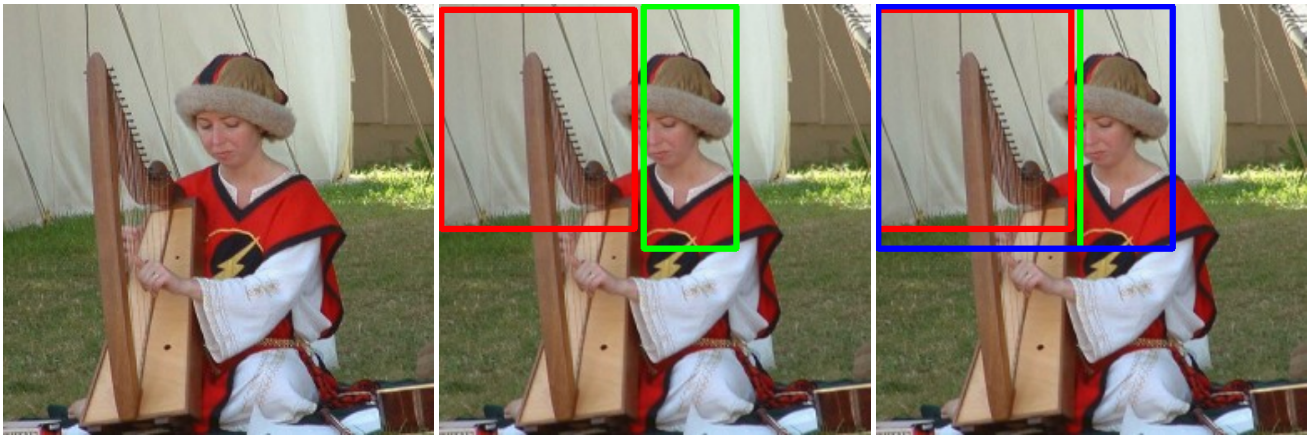


Figure 14: Playing instrument Harp image example from PPMI-24 dataset [27]. Detection of two primary SRs with which a secondary SR is generated. Original image ⇒ detected two primary SRs (red and green) ⇒ Secondary SR (blue) with the primary SRs (left to right). The primary SRs describe a person (green) and Harp instrument (red) separately, which is combined in the secondary region to describe the action that a person is playing Harp. It is related to Fig.4 (Section III) in the manuscript. Best view in color.

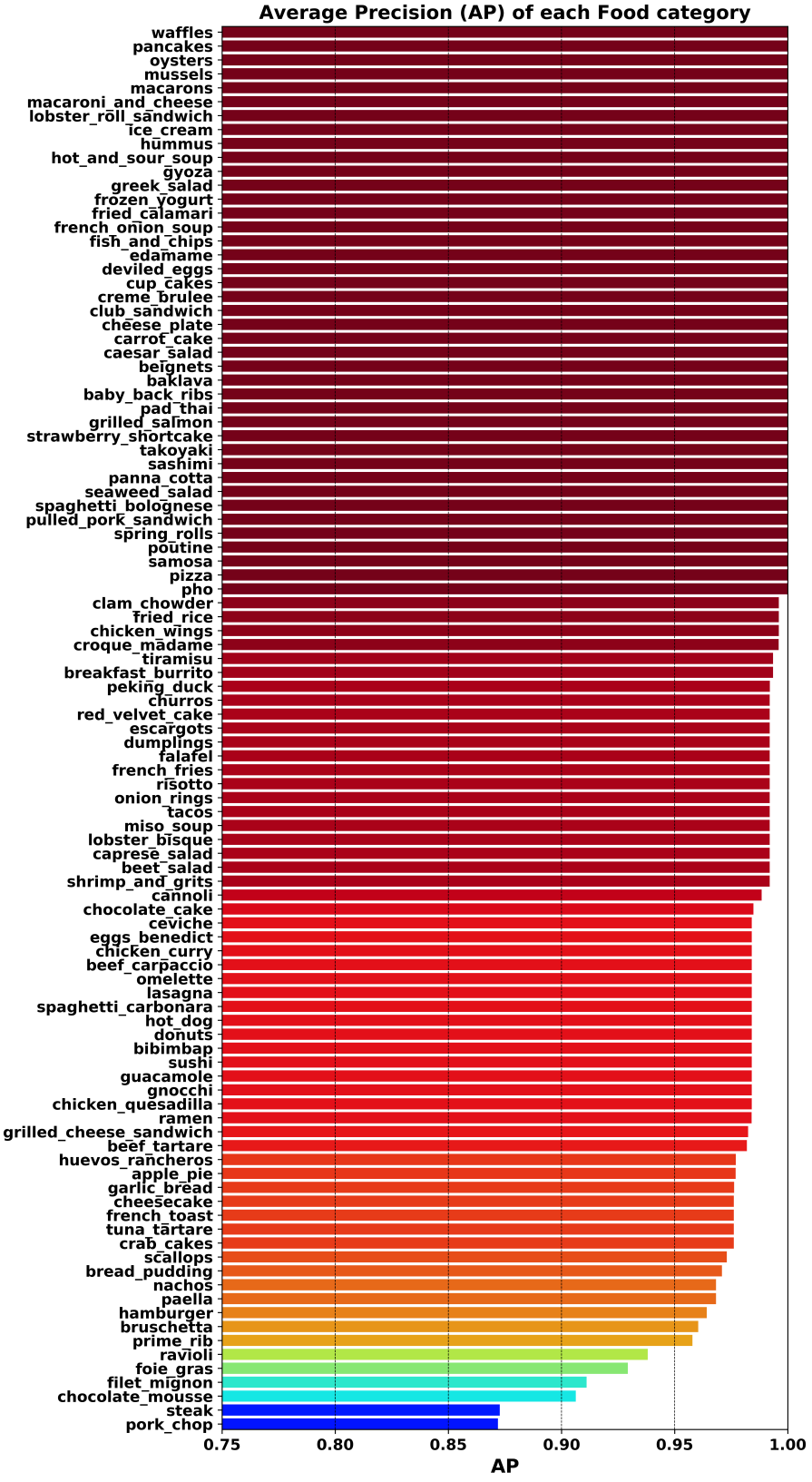


Figure 15: **Average Precision (AP)** of each food category in Food-101 dataset [31] using our AG-Net. It is evident that the top 41 food classes are classified with 100% AP. The least AP is 87.18% for the *pork chop* food class. This has been described in Section V-C in the manuscript.

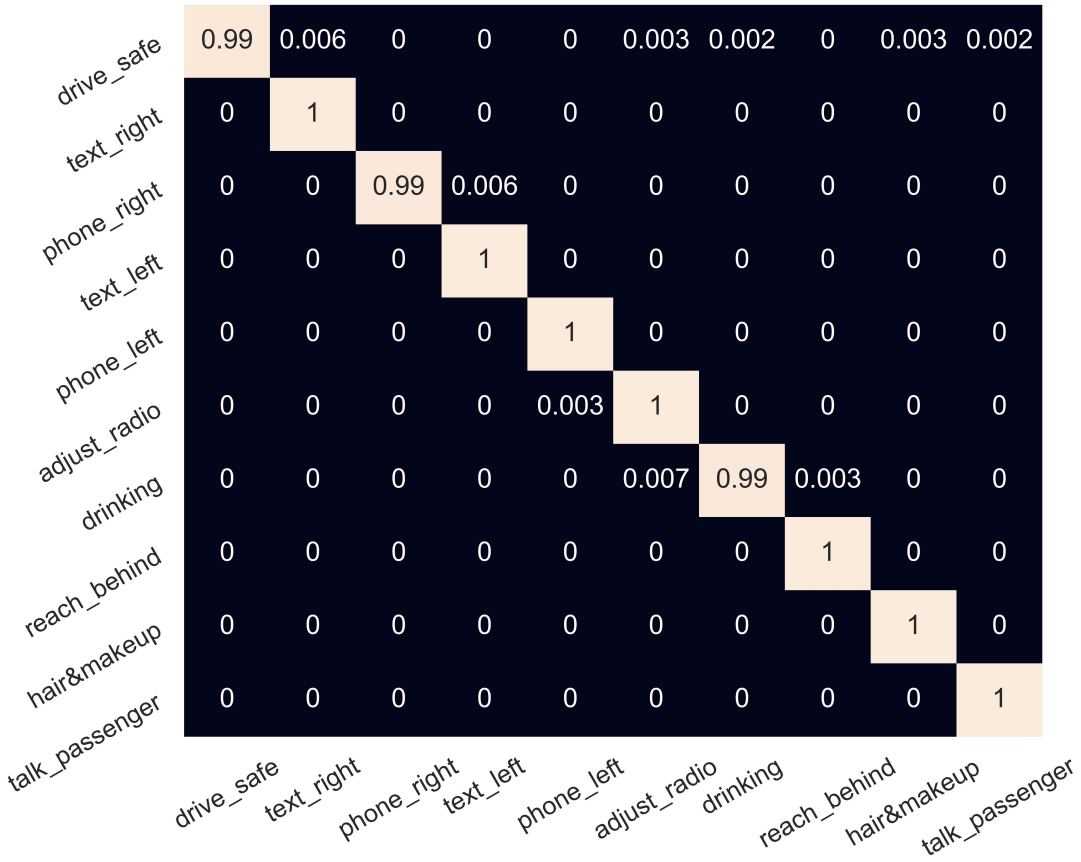


Figure 16: **Confusion matrix** of the proposed AG-Net using AUC-V1 dataset [20]. This is related to Fig. 5a in the manuscript, in which class-wise accuracy has been presented. The x-axis represents predicted labels and y-axis denotes actual class labels of driving activities.

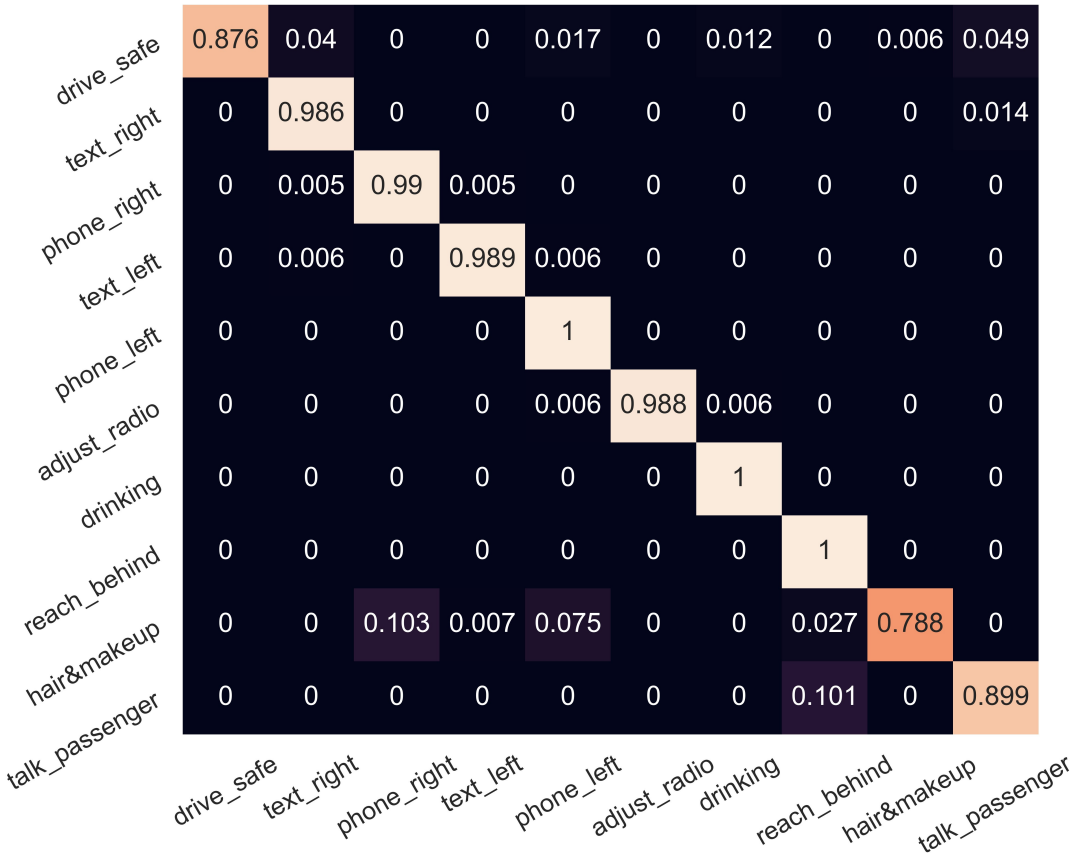


Figure 17: **Confusion matrix** of the proposed AG-Net using AUC-V2 dataset which comprises of unique drivers-wise train-test split [21]. This is related to Fig.5b in the manuscript, in which class-wise accuracy has been depicted. It is clear that *C9: hair and makeup* and *C0: driving safely* fine-grained activities are the low performer and is explained in Section V-A.

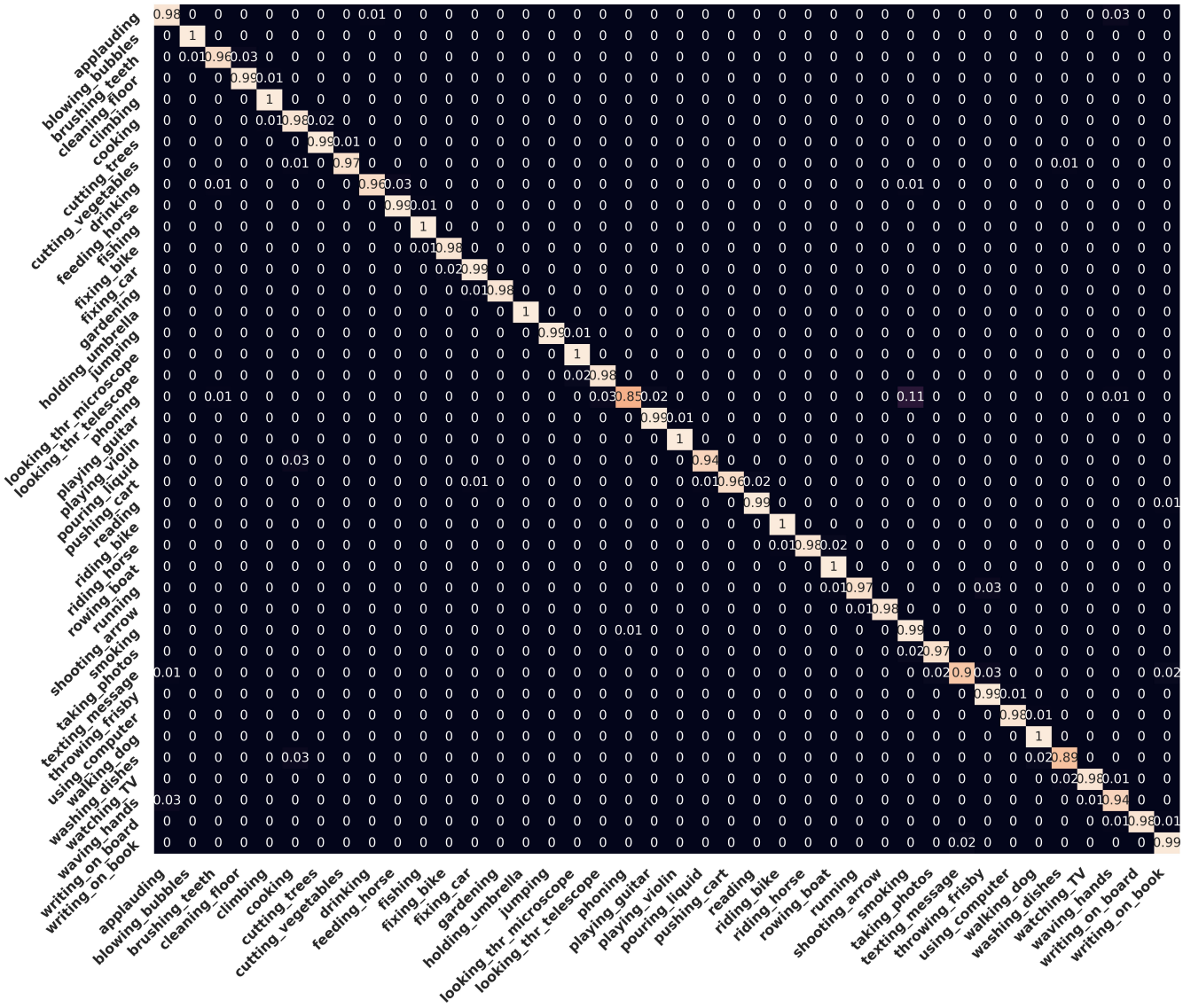


Figure 18: **Confusion matrix** of the proposed AG-Net using Stanford-40 actions dataset [24]. Nine actions offer 100% classification accuracy and it is discussed in Section V-B in the manuscript.

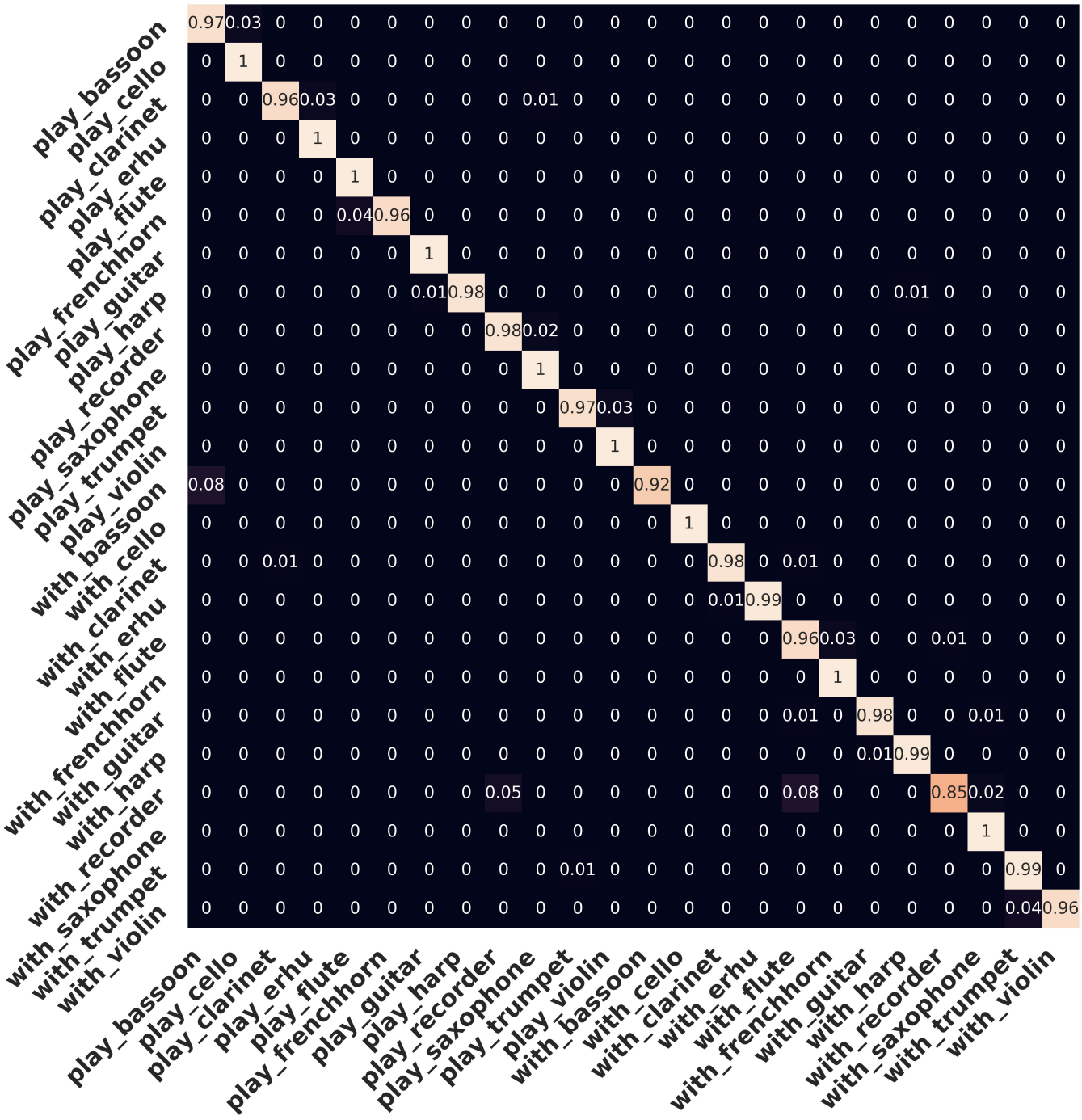


Figure 19: **Confusion matrix** of the proposed AG-Net using PPMI-24 dataset [27]. It is clear that the predicted labels of nine human-instruments interactions are achieved with 100% accuracy. The least accuracy (85%) is achieved by *with recorder* fine-grained interaction. This has been presented in Section V-B in the manuscript.