Attentional Correlation Filter Network for Adaptive Visual Tracking <Supplementary Material>

Jongwon Choi[1] Hyung Jin Chang[2] Sangdoo Yun[1] Tobias Fischer[2]
Yiannis Demiris[2] Jin Young Choi[1]
[1]ASRI, Dept. of Electrical and Computer Eng., Seoul National University, South Korea
[2]Personal Robotics Laboratory, Department of Electrical and Electronic Engineering
Imperial College London, United Kingdom
jwchoi.pil@gmail.com {hj.chang,t.fischer,y.demiris}@imperial.ac.uk {yunsd101,jychoi}@snu.ac.kr

1. Evaluation Results on CVPR2013 dataset

Fig. 1 represents detailed evaluation results on the CVPR2013 dataset [1]. As shown in the evaluation plots, the proposed framework showed the state-of-the-art performance among the real-time visual trackers.

2. Validations on the Parameters

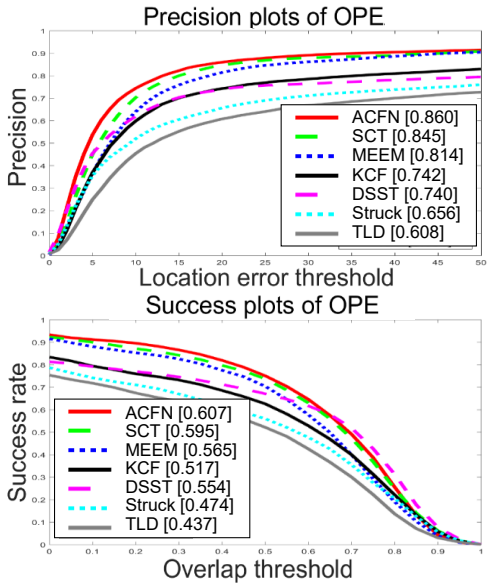
To show the effect of the parameters used in the Attentional Correlation Filter Network (ACFN), two additional experiments were conducted.

In the first experiment, we varied the number of selected tracking modules (N_a) in order to validate the robustness of the attentional mechanism, as shown in Fig. 2 (a). For this experiment, the number of tracking modules with high predicted validation scores (k) was fixed to 13. The result shows that the robustness of the tracker reduces dramatically when N_a was too small, which was due to the insufficient variety of the considered properties in this case. The robustness also decreased with large N_a, which meant that adding the tracking modules without considering the dynamic properties disturbed the robustness of the tracker.

In the second experiment which is depicted in Fig. 2 (b), k was varied, while N_a was fixed to 52. The performance dropped when k was set to small values, which was due to the insufficient number of tracking modules with a high predicted validation score. When k is too big, the performance of the tracker decreased because the prediction errors caused by inactive tracking modules were accumulated over time.

3. Scene-wise Frequency Maps

Fig. 3 shows the frequency maps from the Bolt and Jumping scenes. In the frequency maps of the Bolt scene, the colour-based tracking modules were frequently chosen as the best module. This was due to many shape deformations which occur in this scene, which was hard to track only by



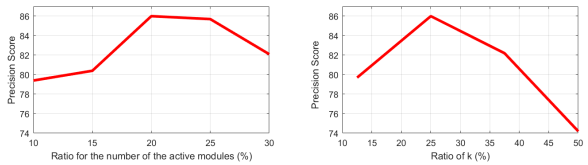
(a) Number of the Active Modules

Figure 1. **Evaluation Results.** In CVPR2013 [1] dataset, ACFN showed the best performance amongst real-time trackers. The numbers within the legend are the average precisions when the centre error threshold equals 20 pixels (top row), or the area under the curve of the success plot (bottom row).

HOG features. In the frequency maps of the Jumping scene, many abrupt movements and blurrinesses happen. As the tracker often suffered from position drift, tracking modules with delayed updates were often chosen as the best module to refine the drifting tracker.

The frequency maps of the active modules and the best module were obtained from each scene of the CVPR2013 dataset [1] and the TPAMI2015 dataset [2]. As shown in Fig. 4 and Fig. 5, the various tracking modules were selected according to the distinct situations of the scenes.

In Fig. 6 and Fig. 7, while the HOG-based tracking mod-



(a) Number of the Active Modules (b) Number of the Modules with the High Predicted Scores

Figure 2. **Parameter variation.** The number of active modules (N_a) and the modules selected by high predicted validation scores (k) are set to various values in order to show the influence on the precision scores obtained in the CVPR2013 dataset [1].

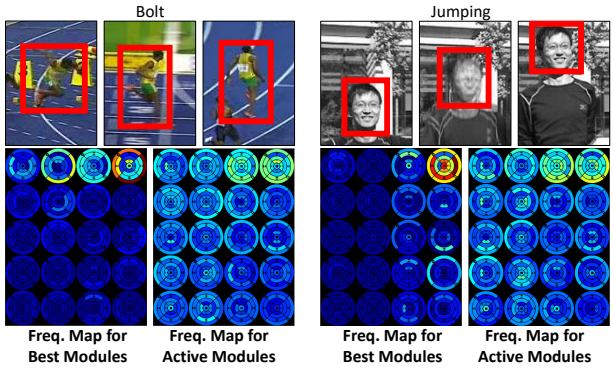


Figure 3. **Frequency map for specific tracking scenes.** According to the various dynamic properties of targets, the modules chosen as the active modules and the best module were distinct.

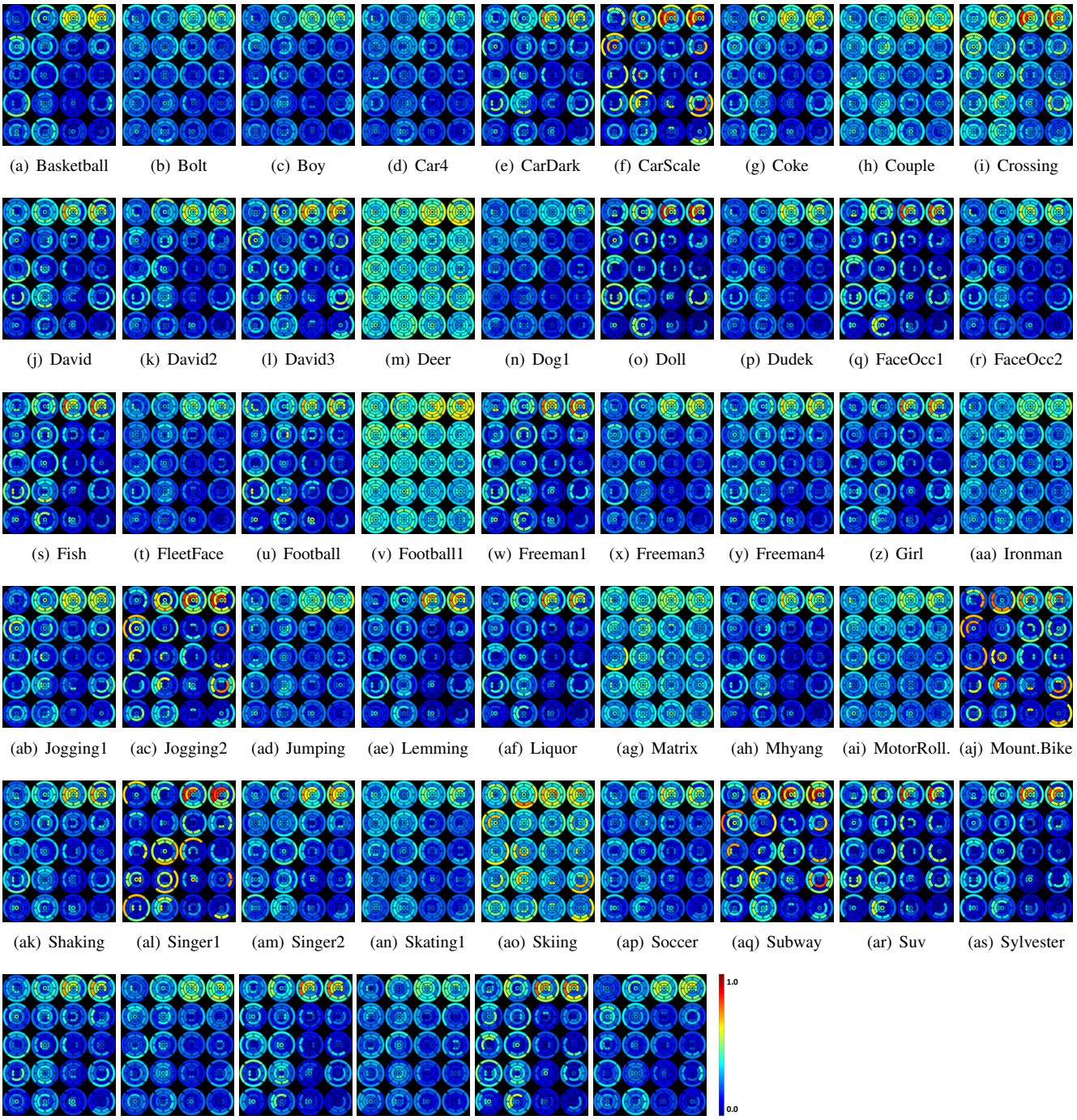
ules were generally chosen as the best module, importantly, other tracking modules also operated to track the target precisely. Because the tracking can fail even with one missing frame, selecting various tracking modules which cover different properties as the best module was critical for the performance of the tracker.

From looking at the frequency of the best module chosen in the scene, one can infer properties of the scene. While most scenes rely mostly on HOG features, the Skating1, Dog1, and Bolt scenes frequently rely on colour features as they contain targets with out-of-the-plane rotation and shape deformations. Interestingly, one could think of using the frequency maps to cluster the scenes. Each cluster would then contain scenes of similar properties. For example, the CarScale and Doll Scene both contain a target with scale changes and partial occlusions, while the Couple, Deer, and Foolball1 scenes contain a target with large motion in a few frames.

References

[1] Y. Wu, J. Lim, and M. H. Yang. Online object tracking: A benchmark. In *CVPR*, pages 2411–2418, 2013. 1, 2, 3, 5

[2] Y. Wu, J. Lim, and M. H. Yang. Object tracking benchmark. *IEEE Trans. on PAMI*, 37(9):1834–1848, 2015. 1, 4, 6



(a) Basketball (b) Bolt (c) Boy (d) Car4 (e) CarDark (f) CarScale (g) Coke (h) Couple (i) Crossing

(j) David (k) David2 (l) David3 (m) Deer (n) Dog1 (o) Doll (p) Dudek (q) FaceOcc1 (r) FaceOcc2

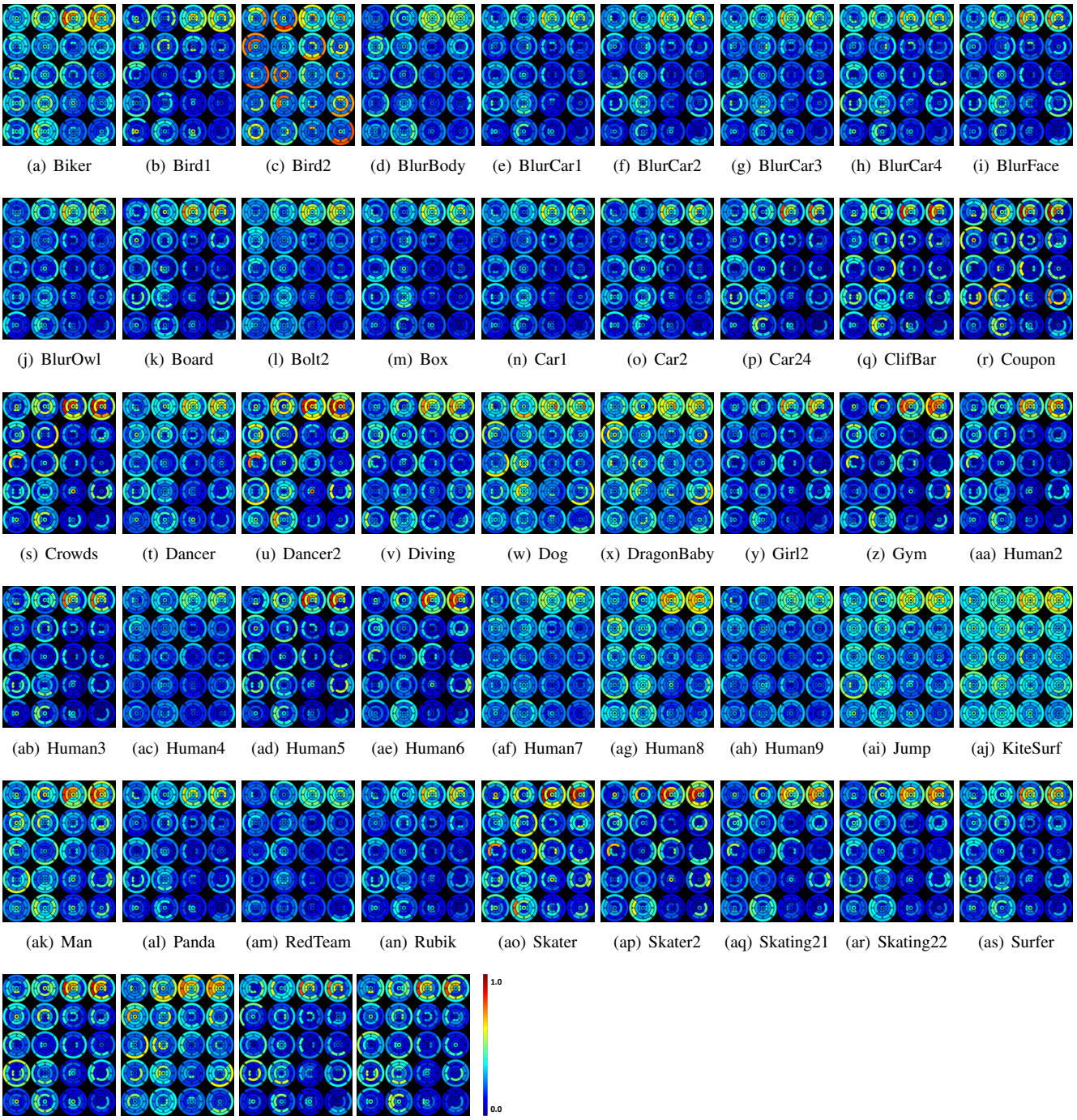
(s) Fish (t) FleetFace (u) Football (v) Football1 (w) Freeman1 (x) Freeman3 (y) Freeman4 (z) Girl (aa) Ironman

(ab) Jogging1 (ac) Jogging2 (ad) Jumping (ae) Lemming (af) Liquor (ag) Matrix (ah) Mhyang (ai) MotorRoll. (aj) Mount.Bike

(ak) Shaking (al) Singer1 (am) Singer2 (an) Skating1 (ao) Skiing (ap) Soccer (aq) Subway (ar) Suv (as) Sylvester

(at) Tiger1 (au) Tiger2 (av) Trellis (aw) Walking (ax) Walking2 (ay) Woman

Figure 4. **Scene-wise frequency maps of the active modules from the CVPR2013 dataset.** From all scenes of the CVPR2013 dataset [1], the frequency maps of the active modules were obtained.



(a) Biker (b) Bird1 (c) Bird2 (d) BlurBody (e) BlurCar1 (f) BlurCar2 (g) BlurCar3 (h) BlurCar4 (i) BlurFace

(j) BlurOwl (k) Board (l) Bolt2 (m) Box (n) Car1 (o) Car2 (p) Car24 (q) ClifBar (r) Coupon

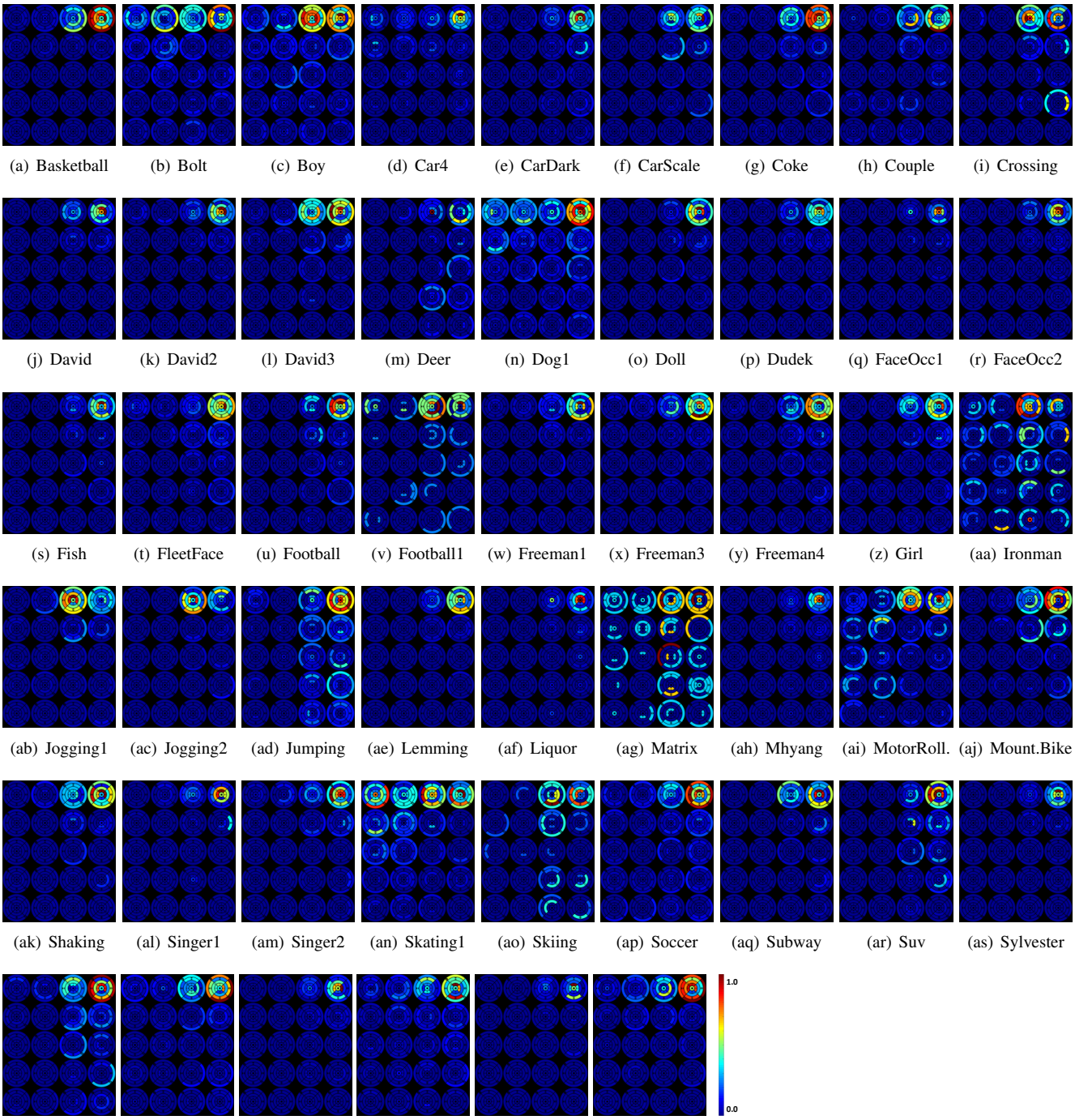
(s) Crowds (t) Dancer (u) Dancer2 (v) Diving (w) Dog (x) DragonBaby (y) Girl2 (z) Gym (aa) Human2

(ab) Human3 (ac) Human4 (ad) Human5 (ae) Human6 (af) Human7 (ag) Human8 (ah) Human9 (ai) Jump (aj) KiteSurf

(ak) Man (al) Panda (am) RedTeam (an) Rubik (ao) Skater (ap) Skater2 (aq) Skating21 (ar) Skating22 (as) Surfer

(at) Toy (au) Trans (av) Twinnings (aw) Vase

Figure 5. **Scene-wise frequency maps of the active modules from the TPAMI2015 dataset.** The frequency maps of the active modules were estimated from the remaining scenes of the TPAMI2015 dataset [2].



(a) Basketball (b) Bolt (c) Boy (d) Car4 (e) CarDark (f) CarScale (g) Coke (h) Couple (i) Crossing

(j) David (k) David2 (l) David3 (m) Deer (n) Dog1 (o) Doll (p) Dudek (q) FaceOcc1 (r) FaceOcc2

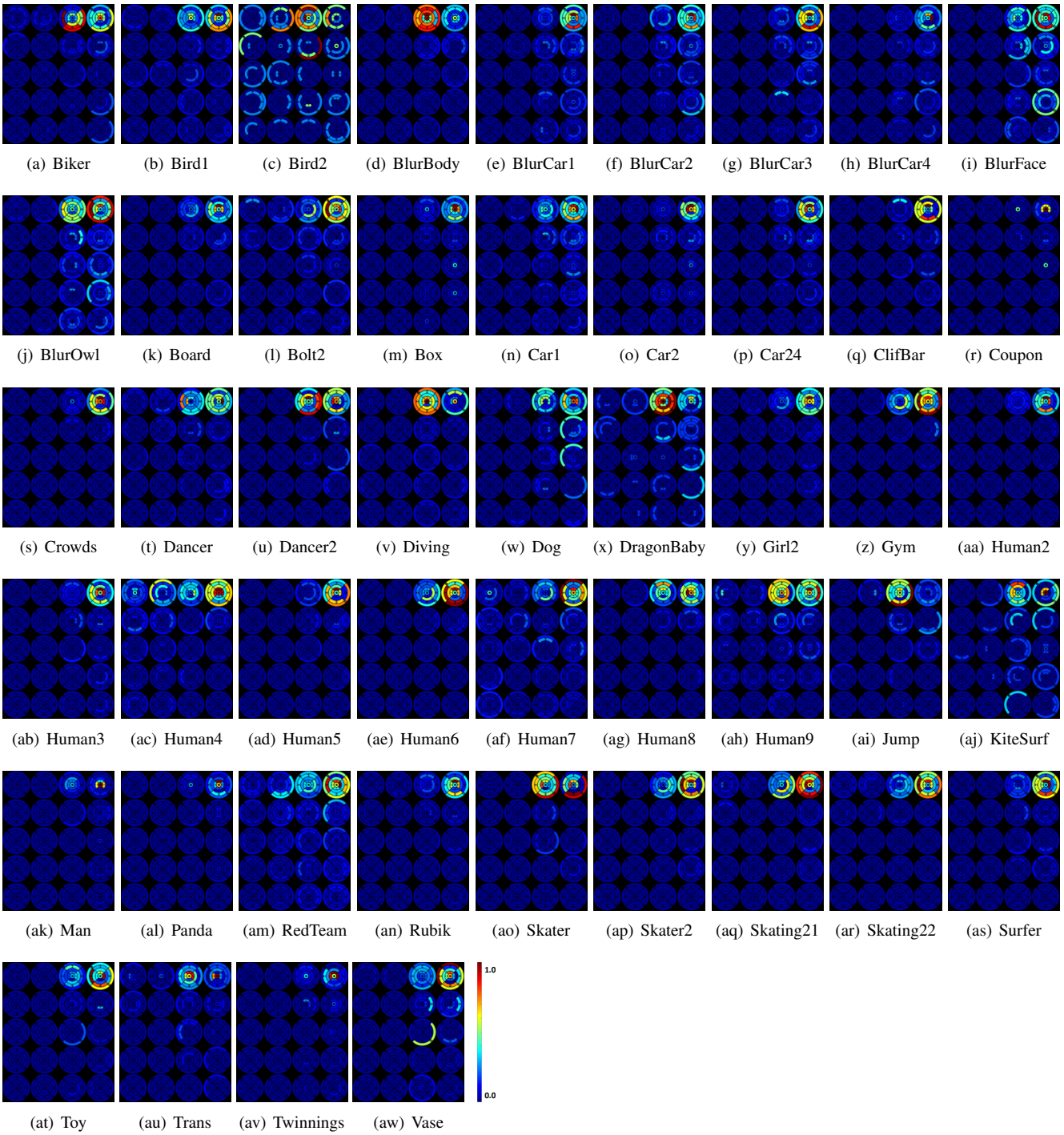
(s) Fish (t) FleetFace (u) Football (v) Football1 (w) Freeman1 (x) Freeman3 (y) Freeman4 (z) Girl (aa) Ironman

(ab) Jogging1 (ac) Jogging2 (ad) Jumping (ae) Lemming (af) Liquor (ag) Matrix (ah) Mhyang (ai) MotorRoll. (aj) Mount.Bike

(ak) Shaking (al) Singer1 (am) Singer2 (an) Skating1 (ao) Skiing (ap) Soccer (aq) Subway (ar) Suv (as) Sylvester

(at) Tiger1 (au) Tiger2 (av) Trellis (aw) Walking (ax) Walking2 (ay) Woman

Figure 6. **Scene-wise frequency maps of the best module from the CVPR2013 dataset.** From all scenes of the CVPR2013 dataset [1], the frequency maps of the best module were obtained.



(a) Biker (b) Bird1 (c) Bird2 (d) BlurBody (e) BlurCar1 (f) BlurCar2 (g) BlurCar3 (h) BlurCar4 (i) BlurFace

(j) BlurOwl (k) Board (l) Bolt2 (m) Box (n) Car1 (o) Car2 (p) Car24 (q) ClifBar (r) Coupon

(s) Crowds (t) Dancer (u) Dancer2 (v) Diving (w) Dog (x) DragonBaby (y) Girl2 (z) Gym (aa) Human2

(ab) Human3 (ac) Human4 (ad) Human5 (ae) Human6 (af) Human7 (ag) Human8 (ah) Human9 (ai) Jump (aj) KiteSurf

(ak) Man (al) Panda (am) RedTeam (an) Rubik (ao) Skater (ap) Skater2 (aq) Skating21 (ar) Skating22 (as) Surfer

(at) Toy (au) Trans (av) Twinnings (aw) Vase

Figure 7. **Scene-wise frequency maps of the best modules from the TPAMI2015 dataset.** The frequency maps of the best modules were estimated from the remaining scenes of the TPAMI2015 dataset [2].