D2-Net: Weakly-Supervised Action Localization via Discriminative Embeddings and Denoised Activations (Supplementary)

Sanath Narayan[1] Hisham Cholakkal[2] Munawar Hayat[3] Fahad Shahbaz Khan[2,4]
Ming-Hsuan Yang[5,6,7] Ling Shao[1]
[1]Inception Institute of Artificial Intelligence [2]Mohamed Bin Zayed University of AI [3]Monash University
[4]Linköping University [5]University of California, Merced [6]Google Research [7]Yonsei University

In this supplementary material, we present additional qualitative and quantitative analysis of the weakly-supervised action localization performance of our proposed `D2-Net`. The quantitative analysis w.r.t. robustness and impact of design choices are presented in Sec. 1, followed by the qualitative results in Sec. 2.

1. Additional Quantitative Analysis

In this section, we present additional quantitative results w.r.t. model sensitivity, ablations and state-of-the-art comparison on the Charades [8] dataset.

Ablations for penalty term in $\mathcal{L}_{Dis}$: Here, we present an ablation to analyse the impact of the weights in the penalty term of our proposed discriminative loss term (Eq. 4 in main paper). Tab. A1 shows the performance comparison on the THUMOS14 dataset for ablating the penalty term. The penalty term in standard focal loss ($\mathcal{L}_F$ in Tab. A1) comprises only the prediction dependent term (*e.g.*, $(1-\mathbf{p}[c])$ for a positive class). In contrast, our $\mathcal{L}_{Dis}$ without focal penalty comprises only the grouping and clustering weights (*e.g.*, $(w_{fg}+w_{fb})$ for a positive class). Furthermore, our final $\mathcal{L}_{Dis}$ includes both the standard focal penalty along with the grouping and clustering weights. Tab. A1 shows that replacing the standard penalty term with our grouping and clustering weights based penalty term (denoted as $\mathcal{L}_{Dis}$ w/o focal penalty) achieves promising performance over $\mathcal{L}_F$. The performance is further improved in our final L_{Dis}, which combines the standard penalty along with our grouping and clustering weights in the penalty term. This shows the efficacy of integrating our grouping and clustering weights (w_{fg}, w_{bg} and w_{fb}) into the penalty term, for improving the localization.

Table A1. **Performance comparison** by ablating the penalty term in $\mathcal{L}_{Dis}$, on the THUMOS14 dataset. The penalty term in our $\mathcal{L}_{Dis}$ includes the standard focal loss penalty along with the proposed grouping and separating terms (w_{fg}, w_{bg} and w_{fb}). In comparison to the standard focal loss $\mathcal{L}_F$, our $\mathcal{L}_{Dis}$ without the focal loss penalty term achieves promising performance. This is further improved by our final $\mathcal{L}_{Dis}$, indicating the efficacy of integrating w_{fg}, w_{bg} and w_{fb} into the penalty term.

Loss term	**mAP @ IoU**				
	0.1	**0.2**	**0.3**	**0.4**	**0.5**
$\mathcal{L}_F$	58.8	52.4	44.3	35.7	26.7
$\mathcal{L}_{Dis}$ w/o focal penalty	62.9	57.5	47.2	37.9	29.2
$\mathcal{L}_{Dis}$	65.4	59.7	50.1	40.4	32.2

Table A2. **Impact of snippet-level and video-level denoising** on the THUMOS14 dataset. Integrating snippet-level ($\mathcal{L}_{DS}$) and video-level ($\mathcal{L}_{DV}$) denoising terms individually with $\mathcal{L}_{Dis}$ improves the localization performance over $\mathcal{L}_{Dis}$ alone. Moreover, integrating both denoising terms with the discriminative loss term (*i.e.*, $\mathcal{L}_{Dis}+\mathcal{L}_D$) in our `D2-Net` achieves improved localization performance, indicating the importance of both snippet-level and video-level denoising for temporal localization.

Loss term	**mAP @ IoU**				
	0.1	**0.2**	**0.3**	**0.4**	**0.5**
$\mathcal{L}_{Dis}$	65.4	59.7	50.1	40.4	32.2
$\mathcal{L}_{Dis}+\mathcal{L}_{DS}$	63.0	57.1	50.1	41.9	34.3
$\mathcal{L}_{Dis}+\mathcal{L}_{DV}$	65.4	59.8	51.3	42.0	33.2
`D2-Net` ($\mathcal{L}_{Dis}+\mathcal{L}_D$)	65.8	60.1	52.3	43.4	36.0

Impact of snippet-level and video-level denoising: Tab. A2 shows the impact of individually integrating the mutual information (MI) based snippet-level ($\mathcal{L}_{DS}$) and video-level ($\mathcal{L}_{DV}$) denoising terms with $\mathcal{L}_{Dis}$. Integrating both these terms individually improves the localization performance over $\mathcal{L}_{Dis}$ alone. While integrating $\mathcal{L}_{DS}$ achieves 34.3% mAP at IoU=0.5, integrating $\mathcal{L}_{DV}$ suppresses more false positives and results in an mAP of 33.2%. Furthermore, our `D2-Net`, which integrates both snippet-level and video-level denoising terms with the discriminative loss term (*i.e.*, $\mathcal{L}_{Dis}+\mathcal{L}_D$) achieves improved localization performance, indicating the importance of both snippet-level and video-level denoising for temporal localization.

Impact of varying γ: Tab. A3 shows the impact of varying the degree of intra-glass grouping on the THUMOS14 dataset. We observe that when there is no/very high intra-

Table A3. **Impact of varying** γ on the THUMOS14 dataset. Sub-optimal localization performances are observed when there is no/very high intra-class grouping, *i.e.*, γ is 0 or 1. Promising localization performance is achieved when the intra-class embeddings are coarsely grouped, *i.e.*, $\gamma \in [0.01, 0.1]$.

Gamma (γ)	**mAP @ IoU**				
	0.1	**0.2**	**0.3**	**0.4**	**0.5**
0.0	64.8	59.3	51.8	42.5	34.2
0.01	65.8	60.1	52.3	43.4	36.0
0.1	65.5	60.0	52.0	43.1	35.7
1.0	65.2	59.9	51.3	41.9	33.7

Table A4. **State-of-the-art comparison** on the Charades dataset. Our `D2-Net` performs favorably compared to existing weakly-supervised approaches.

	`ActGraph` [6]	`WSGN` [2]	**Ours: `D2-Net`**
mAP	15.8	18.3	**19.2**

class grouping amongst the foreground embeddings (or background embeddings), the temporal localization of actions is hampered. Furthermore, promising localization performance is achieved when the intra-class grouping is performed at a coarse level, *i.e.*, $\gamma \in [0.01, 0.1]$. This shows that grouping the intra-class embeddings coarsely amongst themselves helps in learning discriminative embeddings, leading to improved localization performance.

State-of-the-art Comparison: The **Charades** [8] dataset comprises 9848 indoor videos with 157 everyday activity classes. On an average, there are 6.8 activity instances per video, with complex activities co-occurring. As in [7], we use the standard training and validation split and follow the same localization evaluation. Tab. A4 shows the performance comparison of our approach with existing weakly-supervised methods on the Charades dataset. Note that a strongly-supervised approach of `TGM` [5] achieves an mAP of 22.3. Among the weakly-supervised approaches, the graph convolution networks based `ActGraph` [6] achieves 15.8% mAP, while Gaussian networks-based `WSGN` [2] obtains 18.3. Our `D2-Net` performs favorably against existing weakly-supervised methods, achieving a promising performance of 19.2 mAP.

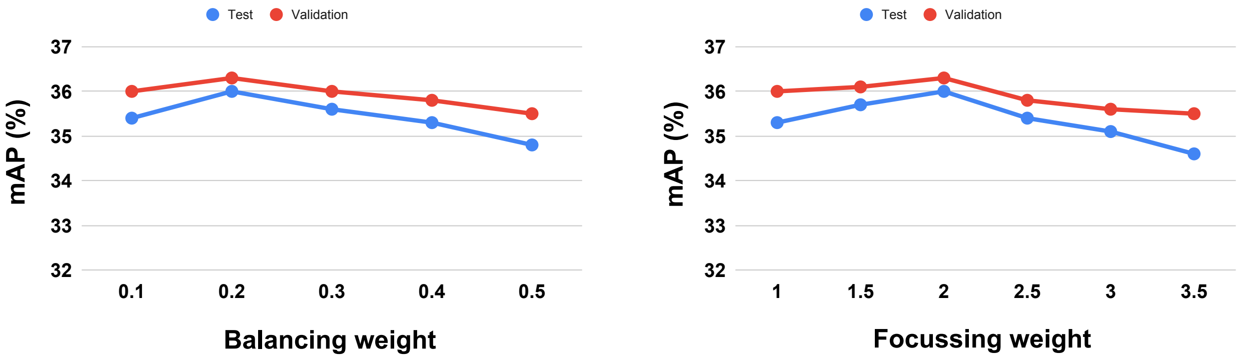
Robustness Analysis: Here, we analyse the robustness of our `D2-Net` w.r.t. variations in the balancing parameter α and focusing parameter β. The performance variations of our approach on both validation and test sets of the THUMOS14 dataset are shown in Fig. 1. The validation accuracy is obtained through cross-validation. The two parameters α and β are varied independently, while keeping the other constant at its respective optimal setting. Varying the balancing weight α results in a performance variation as shown in Fig. 1a. We observe that the performance is optimal when α is around 0.2 and decreases slowly on either side. As α is increased, the denoising loss term ($\mathcal{L}_D$ in Eq. 1 of main paper) overpowers the discriminative loss ($\mathcal{L}_{Dis}$), resulting in a decreased localization performance. In contrast, as α is decreased, the noise in the temporal class activations remains, resulting in reduced localization performance. Hence, we set $\alpha = 0.2$ in our experiments. Similarly, an optimal localization performance of 36.0 mAP is achieved when the focusing parameter β is set to 2 and decreases on either side of it (see Fig. 1b). Note that a similar variation in performance is also observed when using the standard focal loss [4] for generic object detection. Hence, as in [4], we set β as 2 throughout our experiments. These experiments show that our `D2-Net` is reasonably robust to such variations of the balancing and focusing parameters and achieves promising localization performance.

2. Additional Qualitative Results

Here, we present qualitative temporal action localization results of our `D2-Net` framework on example videos from the THUMOS14 [3] and ActivityNet1.2 [1] datasets. In each figure (Fig. 3 to 10), sample frames from a video are shown in the top row followed by the ground-truth segments (green) and predicted detections (blue). The height of a detection is indicative of its score.

THUMOS14: Fig. 3 to 6 illustrate the localization results of our `D2-Net` on example videos, with *Pole Vault*, *Javelin Throw*, *Volleyball Spiking* and *High Jump* actions from the THUMOS14 dataset. Examples show different scenarios: temporally adjacent instances (*Javelin Throw*, *High Jump*), well separated instances (*Pole Vault*) and action pause (*Volleyball Spiking*). Our `D2-Net` detects many of these actions, reasonably well. Generally, well separated actions are detected correctly, as in *Pole Vault* (Fig. 3). Further, an action instance and its slow motion replay are annotated incorrectly as a single action for the fourth instance in *Javelin Throw* (Fig. 4), which is correctly detected as two instances by our approach. Accurately detecting the action instances containing video pauses in between, similar to the first and second instances in *Volleyball Spiking* (Fig. 5), is challenging due to the absence of motion information in the corresponding snippets. The temporally adjacent instances of *High Jump* (Fig. 6) are correctly delineated. These results show that our approach achieves promising localization performance on these variety of actions.

ActivityNet1.2: Fig. 7 to 10 illustrate the localization results of our `D2-Net` on example videos, with *Cricket*, *Washing Hands*, *Playing Harmonica* and *Windsurfing* actions from the ActivityNet1.2 dataset. Examples show different scenarios: well separated instances (*Cricket*), temporally adjacent activities (*Washing Hands*), long and short activity instances (*Playing Harmonica*), and long activity (*Windsurfing*). Well separated activity instances, similar to the



(a) (b)

Figure 1. Action localization performance w.r.t. balancing parameter α in (a) and focusing parameter β in (b) on the THUMOS14 dataset. The performance is shown for both validation and test sets. These experiments show that our `D2-Net` is reasonably robust to such variations of the balancing and focusing parameters and achieves promising localization performance.

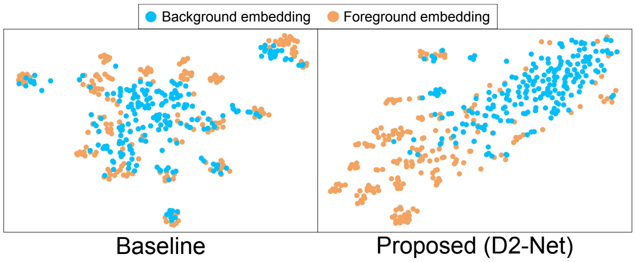


Figure 2. **Illustration of foreground-background separability** obtained in the latent embedding space of (a) the baseline using the standard focal loss and (b) our `D2-Net` via t-SNE scatter plots on the THUMOS14 test set. In both cases, foreground and background embeddings per video are obtained as the mean of latent embeddings at their respective ground-truth locations. Our `D2-Net` better separates the foreground and background, compared to the baseline.

instances of *Cricket* (Fig. 7) are generally detected correctly. The two instances of *Washing Hands* (Fig. 8) are detected as a single instance, since the background that is separating the two instances is indiscriminable from the foreground activity. While the long and short activity instances are both detected correctly for *Playing Harmonica* activity (Fig. 9), an additional false detection is observed due to the visual presence of the performer on stage (but not playing) in the corresponding image frames. Though the annotation for the end of *Windsurfing* activity is inaccurate and includes background regions also as foreground activity, our `D2-Net` correctly detects the end of the temporally long activity (Fig. 10). These qualitative results show that our proposed approach achieves promising action localization performance on a variety of activities.

Foreground-Background Separation: Fig. 2 shows the foreground-background separability comparison, utilizing t-SNE scatter plots, between the baseline and our `D2-Net`. Here, foreground and background embeddings per video are obtained by average pooling (temporally) the latent embeddings at their respective ground-truth snippet locations. Fig. 2 shows that the foreground and background embeddings in the baseline overlap with each other. In contrast, our `D2-Net` better separates the foreground and background, compared to the baseline, leading to improved localization of foreground actions in the videos.

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 2

[2] Basura Fernando, Cheston Tan, and Hakan Bilen. Weakly supervised gaussian networks for action detection. In *WACV*, 2020. 2

[3] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *CVIU*, 2017. 2

[4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2

[5] AJ Piergiovanni and Michael S. Ryoo. Temporal gaussian mixture layer for videos. In *ICML*, 2019. 2

[6] Maheen Rashid, Hedvig Kjellström, and Yong Jae Lee. Action graphs: Weakly-supervised action localization with graph convolution networks. In *WACV*, 2020. 2

[7] Gunnar A. Sigurdsson, Santosh Divvala, Ali Farhadi, and Abhinav Gupta. Asynchronous temporal fields for action recognition. *CVPR*, 2017. 2

[8] Gunnar A. Sigurdsson, Gul Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood

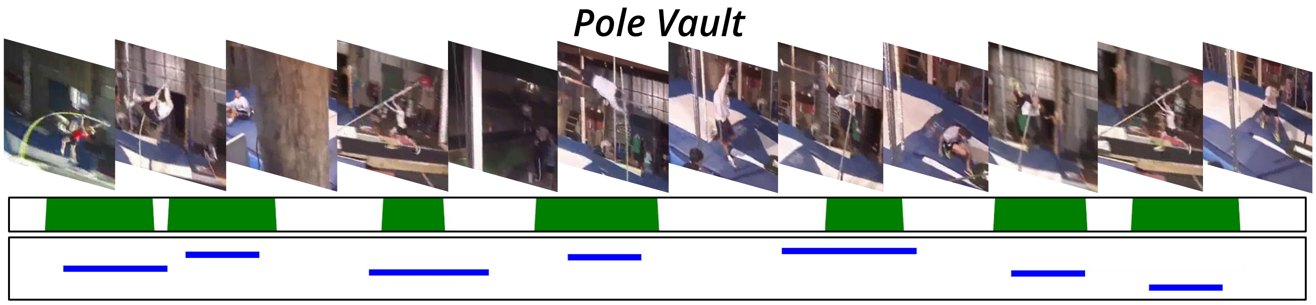


Figure 3. Well separated action instances of *Pole Vault* are generally accurately detected by our `D2-Net`.

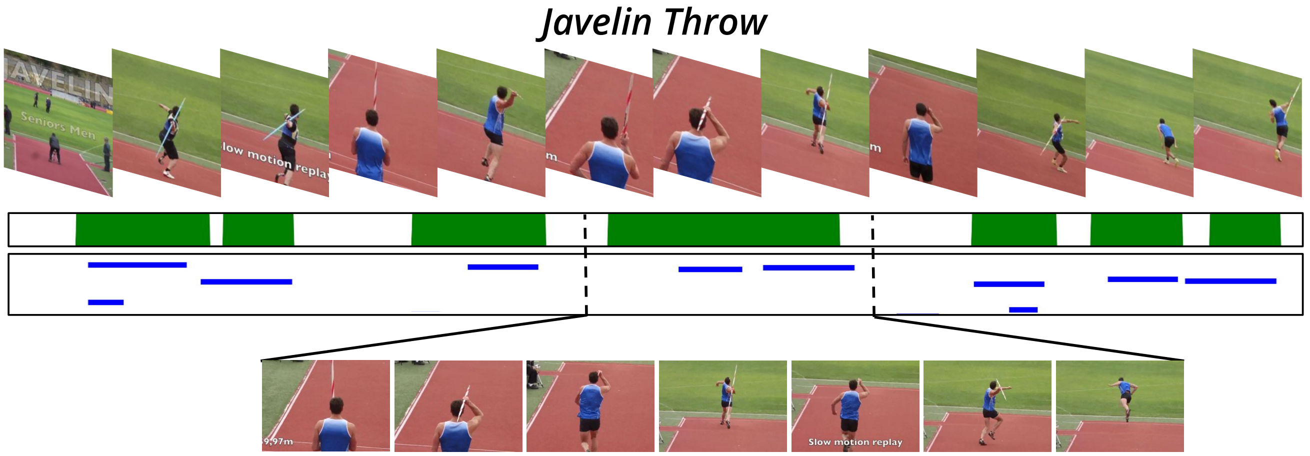


Figure 4. Fourth instance of *Javelin Throw* is incorrectly annotated as a single instance though it has two instances: action and its slow motion replay. Our `D2-Net` correctly detects the two as separate instances.

in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 1, 2

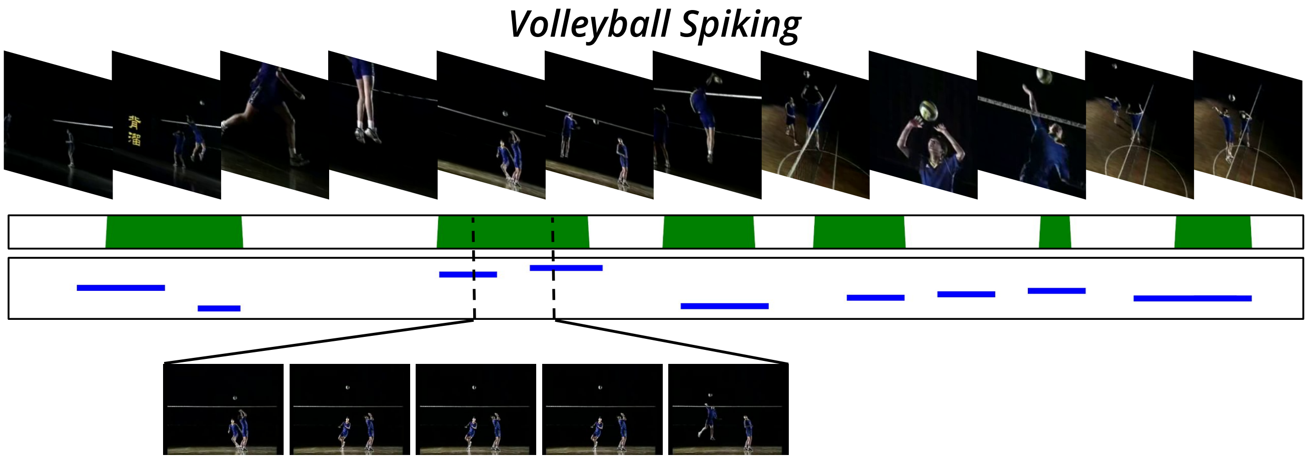


Figure 5. The first two instances of *Volleyball Spiking* have a considerable pause in the video, resulting in the absence of motion for the corresponding frames. *E.g.*, an inset of sample frames in the second instance shows the pause in the video containing zero motion. This absence of discriminative motion information leads to four incorrect detections for these two GT instances.

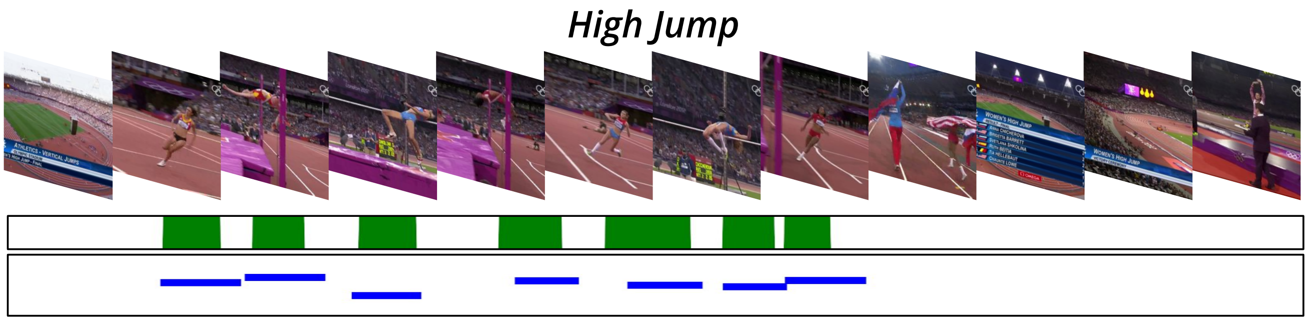


Figure 6. Temporally adjacent action instances of *High Jump* (sixth and seventh instances) are correctly detected as distinct instances by our `D2-Net`.

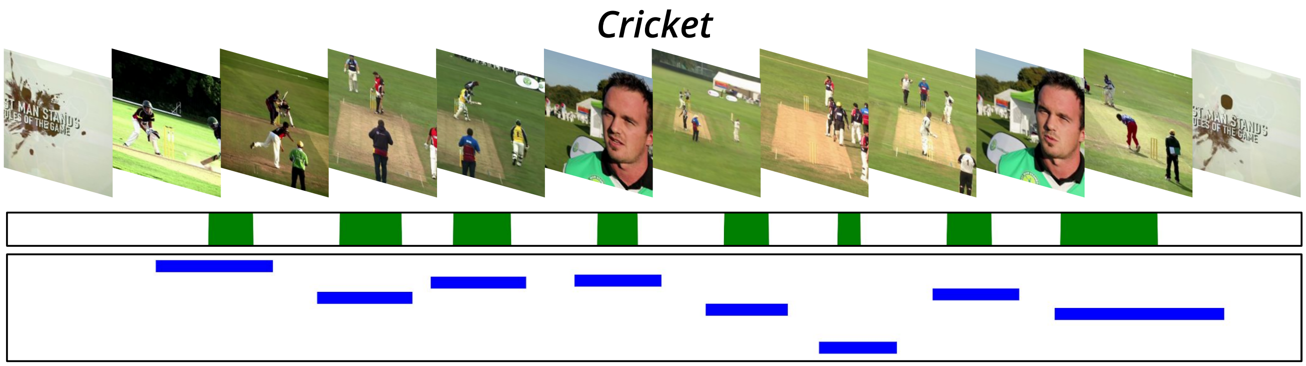


Figure 7. Well separated instances of *Cricket* activity are detected accurately by our `D2-Net`.

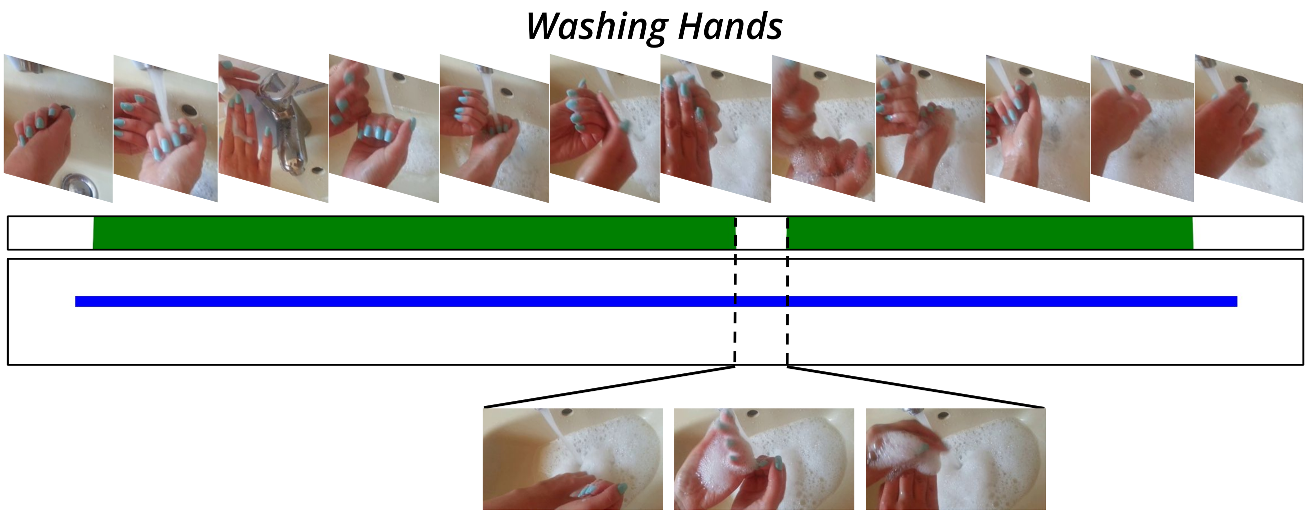


Figure 8. The two adjacent ground-truth *Washing Hands* instances are jointly detected as a single instance by our `D2-Net`, since the separating background is indiscriminable from the foreground activity. Sample background frames, shown inset, contain hands along with soap lather and flowing water and are visually similar to the foreground activity.

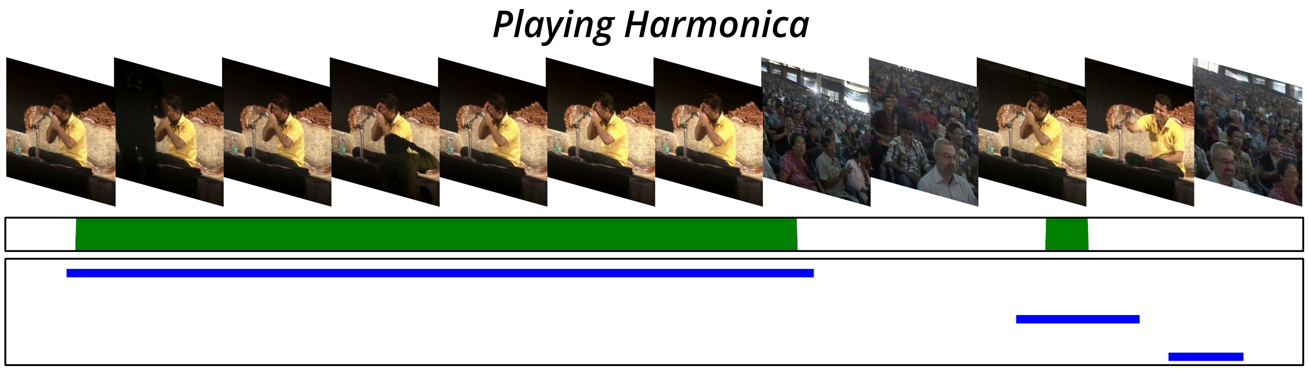


Figure 9. Both the long and short duration instances of *Playing Harmonica* are detected correctly by `D2-Net`. However, a false detection arises due to the presence of the performer on stage (but not playing) in the corresponding image frames.

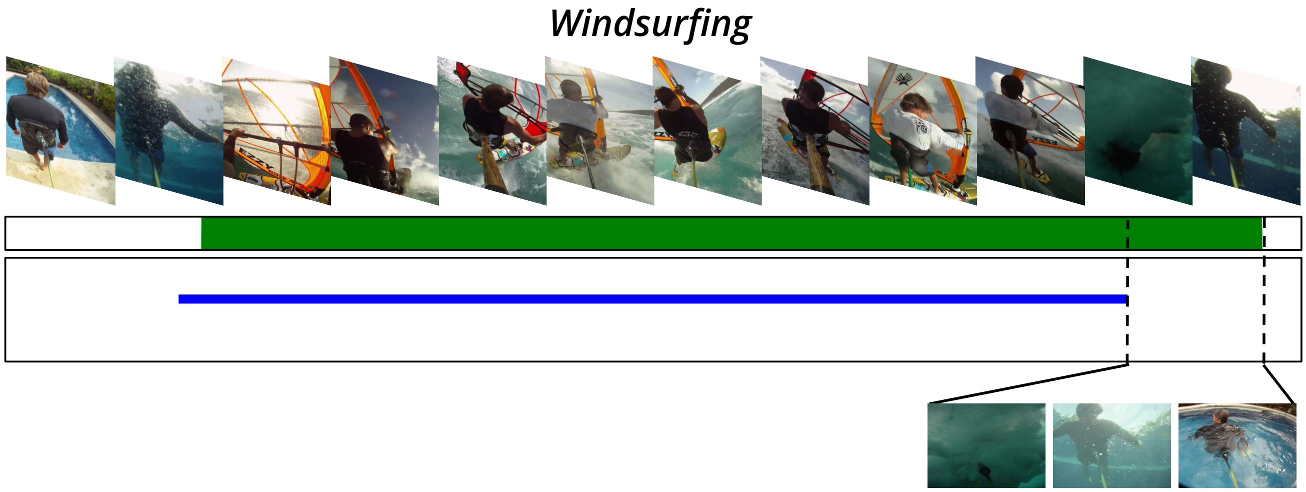


Figure 10. The ground-truth annotation for the end of *Windsurfing* activity is inaccurate since background regions are also included as foreground activity, as shown by the inset frames. Our `D2-Net` accurately detects the temporally long activity.