A. Appendix

A.1. Proof of the error bounds in Theorem 1

In this section, we provide detailed proof for the error bounds in Theorem 1, in particular, regarding the value of $\tilde{d'}(L^p, \tau)$ in Equation (5).

Proof. We first define the concept of the *minimum confidence margin.*

Definition 10 (Minimum Confidence Margin)**.** *Given a network $\mathcal{N}$, an input $\boldsymbol{v}$, and a class c, we define the* minimum confidence margin *as*

$$\mathsf{Mar}(\boldsymbol{v}, c) = \min_{c' \in C, c' \neq c} \{\mathcal{N}(\boldsymbol{v}, c) - \mathcal{N}(\boldsymbol{v}, c')\}. \tag{11}$$

Intuitively, the minimum confidence margin is the discrepancy between the maximum confidence of $\boldsymbol{v}$ being classified as c and the next largest confidence of $\boldsymbol{v}$ being classified as c'. Then, for any input $\boldsymbol{v}'$ whose optical flow sequence is in the subspace of a grid point $\overline{g}$, and the input $\boldsymbol{v}$ corresponding to this optical flow sequence $\overline{g}$, we have

$$\begin{aligned}
&\mathsf{Mar}(\boldsymbol{v}, \mathcal{N}(\boldsymbol{v})) - \mathsf{Mar}(\boldsymbol{v}', \mathcal{N}(\boldsymbol{v})) \\
&= \min_{c \in C, c \neq \mathcal{N}(\boldsymbol{v})} \{\mathcal{N}(\boldsymbol{v}, \mathcal{N}(\boldsymbol{v})) - \mathcal{N}(\boldsymbol{v}, c)\} \\
&\qquad - \min_{c \in C, c \neq \mathcal{N}(\boldsymbol{v})} \{\mathcal{N}(\boldsymbol{v}', \mathcal{N}(\boldsymbol{v})) - \mathcal{N}(\boldsymbol{v}', c)\} \\
&\leq \max_{c \in C, c \neq \mathcal{N}(\boldsymbol{v})} \{\mathcal{N}(\boldsymbol{v}, \mathcal{N}(\boldsymbol{v})) - \mathcal{N}(\boldsymbol{v}, c) \\
&\qquad - \mathcal{N}(\boldsymbol{v}', \mathcal{N}(\boldsymbol{v})) + \mathcal{N}(\boldsymbol{v}', c)\} \\
&\leq \max_{c \in C, c \neq \mathcal{N}(\boldsymbol{v})} \{|\mathcal{N}(\boldsymbol{v}, \mathcal{N}(\boldsymbol{v})) - \mathcal{N}(\boldsymbol{v}', \mathcal{N}(\boldsymbol{v}))| \\
&\qquad + |\mathcal{N}(\boldsymbol{v}', c) - \mathcal{N}(\boldsymbol{v}, c)|\} \\
&\leq \max_{c \in C, c \neq \mathcal{N}(\boldsymbol{v})} \mathsf{Lip}_{\mathcal{N}(\boldsymbol{v})} \cdot \|\boldsymbol{v} - \boldsymbol{v}'\|_p + \mathsf{Lip}_c \cdot \|\boldsymbol{v} - \boldsymbol{v}'\|_p \\
&\leq \max_{c \in C, c \neq \mathcal{N}(\boldsymbol{v})} (\mathsf{Lip}_{\mathcal{N}(\boldsymbol{v})} + \mathsf{Lip}_c) \cdot \|\boldsymbol{v} - \boldsymbol{v}'\|_p \\
&\leq \max_{c \in C, c \neq \mathcal{N}(\boldsymbol{v})} (\mathsf{Lip}_{\mathcal{N}(\boldsymbol{v})} + \mathsf{Lip}_c) \cdot \tilde{d'}(L^p, \tau)
\end{aligned} \tag{12}$$

Now, since the optical flow sequence of $\boldsymbol{v}'$ is in the subspace of $\overline{g}$, we need to ensure that no class change occurs between $\boldsymbol{v}$ and $\boldsymbol{v}'$. That is, $\mathsf{Mar}(\boldsymbol{v}', \mathcal{N}(\boldsymbol{v})) \geq 0$, which means $\mathsf{Mar}(\boldsymbol{v}, \mathcal{N}(\boldsymbol{v})) - \mathsf{Mar}(\boldsymbol{v}', \mathcal{N}(\boldsymbol{v})) \leq \mathsf{Mar}(\boldsymbol{v}, \mathcal{N}(\boldsymbol{v}))$. Therefore, we have

$$\max_{c \in C, c \neq \mathcal{N}(\boldsymbol{v})} (\mathsf{Lip}_{\mathcal{N}(\boldsymbol{v})} + \mathsf{Lip}_c) \cdot \tilde{d'}(L^p, \tau) \leq \mathsf{Mar}(\boldsymbol{v}, \mathcal{N}(\boldsymbol{v})). \tag{13}$$

And as $\overline{g}$ is a grid point, the minimum confidence margin for its corresponding input $\boldsymbol{v}$ can be computed. Finally, we replace $\mathsf{Mar}(\boldsymbol{v}, \mathcal{N}(\boldsymbol{v}))$ with its definition, then we have

$$\tilde{d'}(L^p, \tau) \leq \frac{\min_{c \in C, c \neq \mathcal{N}(\boldsymbol{v})} \{\mathcal{N}(\boldsymbol{v}, \mathcal{N}(\boldsymbol{v})) - \mathcal{N}(\boldsymbol{v}, c)\}}{\max_{c \in C, c \neq \mathcal{N}(\boldsymbol{v})} (\mathsf{Lip}_{\mathcal{N}(\boldsymbol{v})} + \mathsf{Lip}_c)}. \tag{14}$$

□

A.2. Proof of the guarantees in Theorem 2

In this section, we provide a detailed proof for the robustness guarantees in Theorem 2.

Proof. On one hand, we show that $\|\mathsf{P}(\boldsymbol{v}') - \mathsf{P}(\boldsymbol{v})\|_p \geq R(\sigma, s_0)$ for any optical flow sequence $\mathsf{P}(\boldsymbol{v}')$ as a τ-grid point, such that $\mathsf{P}(\boldsymbol{v}') \in \mathsf{B}(\mathsf{P}(\boldsymbol{v}), L^p, d)$ and its corresponding input is an adversarial example. Intuitively, it means that Player `I`'s reward from the game $\mathcal{G}$ in the initial state s_0 is no greater than the L^p distance to any τ-grid manipulated optical flow sequence. That is, the reward value $R(\sigma, s_0)$, once computed, is a lower bound of the optimisation problem $\texttt{FMSR}(\mathcal{N}, \mathsf{P}(\boldsymbol{v}), L^p, d, \tau)$. Note that the reward value can be obtained as every τ-grid point can be reached by some game play, i.e., a sequence of atomic manipulations.

On the other hand, from the termination condition $tc(\rho)$ of the game, we observe that, for some $\mathsf{P}(\boldsymbol{v}')$, if $R(\sigma, s_0) \leq \|\mathsf{P}(\boldsymbol{v}') - \mathsf{P}(\boldsymbol{v})\|_p$ holds, then there must exist some other $\mathsf{P}(\boldsymbol{v}'')$ such that $R(\sigma, s_0) = \|\mathsf{P}(\boldsymbol{v}'' - \mathsf{P}(\boldsymbol{v}))\|_p$. Therefore, we have that $R(\sigma, s_0)$ is the minimum value of $\|\mathsf{P}(\boldsymbol{v}'' - \mathsf{P}(\boldsymbol{v}))\|_p$ among all the τ-grid points $\mathsf{P}(\boldsymbol{v}')$ such that $\mathsf{P}(\boldsymbol{v}') \in \mathsf{B}(\mathsf{P}(\boldsymbol{v}), L^p, d)$ and their corresponding inputs are adversarial examples.

Finally, we observe that the minimum value of $\|\mathsf{P}(\boldsymbol{v}') - \mathsf{P}(\boldsymbol{v})\|_p$ is equivalent to the optical flow value required by Equation (3). □

A.3. Details of the video dataset and the network

As a popular benchmark for human action recognition in videos, *UCF101* [22] consists of 101 annotated action classes, e.g., JugglingBalls (human-object interaction), HandstandPushups (body-motion only), HairCut (human-human interaction), PlayingPiano (playing musical instruments), and FloorGymnastics (sports). It labels 13 320 video clips of 27 hours in total, and each frame has dimension 320 × 240 × 3.

In the experiments, we exploit a VGG16 + LSTM architecture, in the sense of utilising the *VGG16* network to extract the spatial features from the UCF101 video dataset and then passing these features to a separate RNN unit *LSTM*. For each video, we sample a frame every 1000 ms and stitch them together into a sequence of frames. Specifically, we run every frame from every video through VGG16 with input size 224 × 224 × 3, excluding the top classification part of the network, i.e., saving the output from the final MaxPooling layer. Hence, for each video, we retrieve a sequence of extracted spatial features. Subsequently, we pass the features into a single LSTM layer, followed by a Dense layer with some Dropout in between. Eventually, after the final Dense layer with activation function Softmax, we get the classification outcome.

We use the categorical cross-entropy loss function and the accuracy metrics for both the VGG16 and LSTM mod-

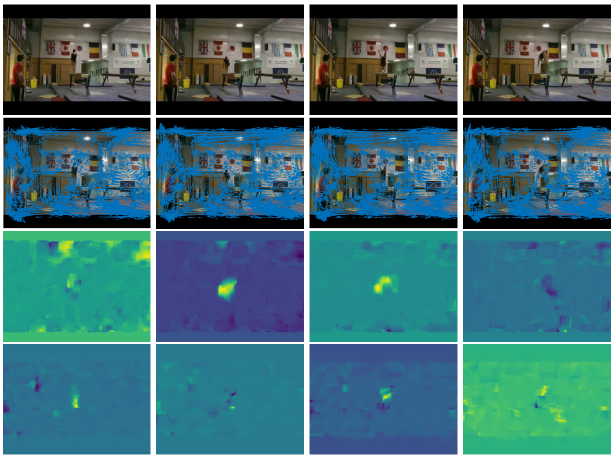

Figure 11. Examples of the optical flows extracted from a BalanceBeam video. Top row: four sampled frames from 0 s to 3 s with original size 320 × 240 × 3. 2nd row: the optical flows (blue arrows) extracted between the frames. 3rd row: one of optical flow's characteristics: magnitude. Bottom row: the other optical flow characteristics: direction.

els. Whilst the former has a SGD optimiser and directly exploits the imagenet weights, we train the latter through a rmsprop optimiser and get 99.15% training accuracy as well as 99.72% testing accuracy. Specifically, when the *loss* difference cannot reflect the subtle perturbation on optical flow during the computation of upper bounds, we use the discrepancy of logit values instead.

A.4. More examples of the optical flows extracted from different videos

Apart from Figure 4 in Section 6, here we include more examples of the optical flows extracted from another two videos with classifications BalanceBeam (Figure 11) and FrontCrawl (Figure 12).

A.5. Another example of the converging upper and lower bounds

Apart from the HammerThrow example (Figures 6 and 7, Section 6), we include another example to illustrate the convergence of the upper and lower bounds. Similarly, Figure 13 exhibits five sampled frames (top row) from a FloorGymnastics video and the optical flows extracted between them (2nd row). The descending upper bounds (red) and the ascending lower bounds (blue) to approximate the value of MSR are presented in Figure 14. Intuitively, after 20 iterations of the gradient-based algorithm, the upper bound, i.e., minimum distance to an adversarial example, is 2100.45 based on the L^2 distance metric. That is, manipulations imposed on the flows exceeding this upper bound may be *unsafe*. Figure 13 (3rd row) shows some of such unsafe perturbations on each optical flow, which result in the misclassification of the video into FrontCrawl with confidence 97.04%. As for the lower bound, we observe that, after 1500 iterations of the admissible A* algorithm, the lower bound reaches 146.61. That is, manipulations within this L^2 norm ball are absolutely *safe*. Some of such safe perturbations can be found in the bottom row of Figure 13.

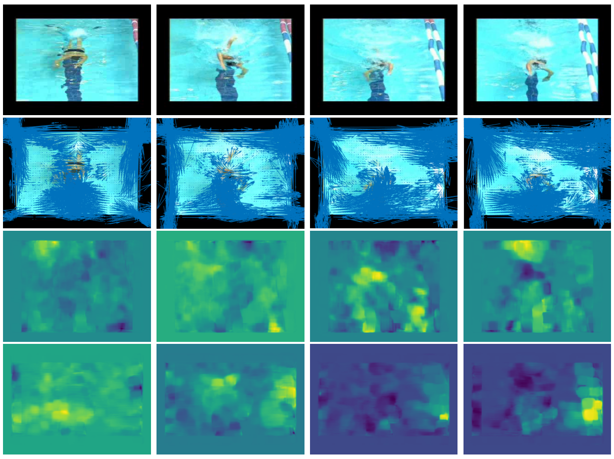

Figure 12. Examples of the optical flows extracted from a FrontCrawl video. Top row: four sampled frames from 0 s to 3 s with original size 320 × 240 × 3. 2nd row: the optical flows (blue arrows) extracted between the frames. 3rd row: one of optical flow's characteristics: magnitude. Bottom row: the other optical flow characteristics: direction.

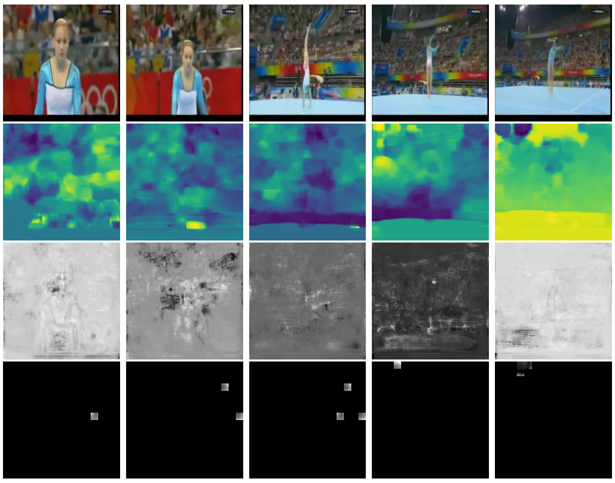

Figure 13. Examples of *unsafe* and *safe* perturbations on the optical flows of a FloorGymnastics video. Top row: five sampled frames from 0 s to 4 s. 2nd row: optical flows of the frames from 0 s to 5 s. 3rd row: *unsafe* perturbations on the flows corresponding to the upper bound. Bottom row: *safe* perturbations on the flows corresponding to the lower bound.

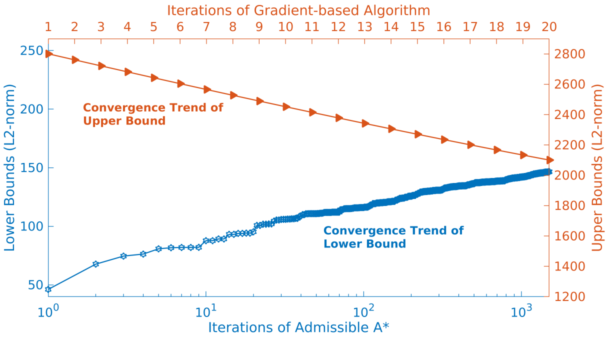

Figure 14. *Converging bounds* of the maximum safe radius of the FloorGymnastics video with respect to manipulations on extracted optical flows. The red line denotes the decreasing *upper* bound from the gradient-based algorithm, and the blue line denotes the increasing *lower* bound from admissible A*.