PERSIST: Persistent-State Discrimination for Shot Boundary Detection

Tingyu Lin[1], Christian Stippel[1], Armin Dadras[1,2], Jakob Zenzmaier[3], Florian Kleber[1], Wolfgang Aigner[2], Robert Sablatnig[1]

[1]Computer Vision Lab, TU Wien [2]Institute of Creative\Media/Technologies, St. Pölten University of Applied Sciences [3]Ludwig Boltzmann Institute for Digital History

IDEA

A shot is a series of consecutive frames captured with the same camera setting, and it is the basic unit of film analysis. Shot boundary detection finds the frames where one shot ends and the next begins.

Most detectors rely on sudden visual change, so flashes, flicker, camera shake and film damage are often detected as boundaries. A true boundary changes the video persistently, whereas a pseudo-event is transient and the video returns to its previous state.

PERSIST therefore compares what the video shows before and after a sudden change, and marks a boundary only when the content stays different afterwards.

METHOD

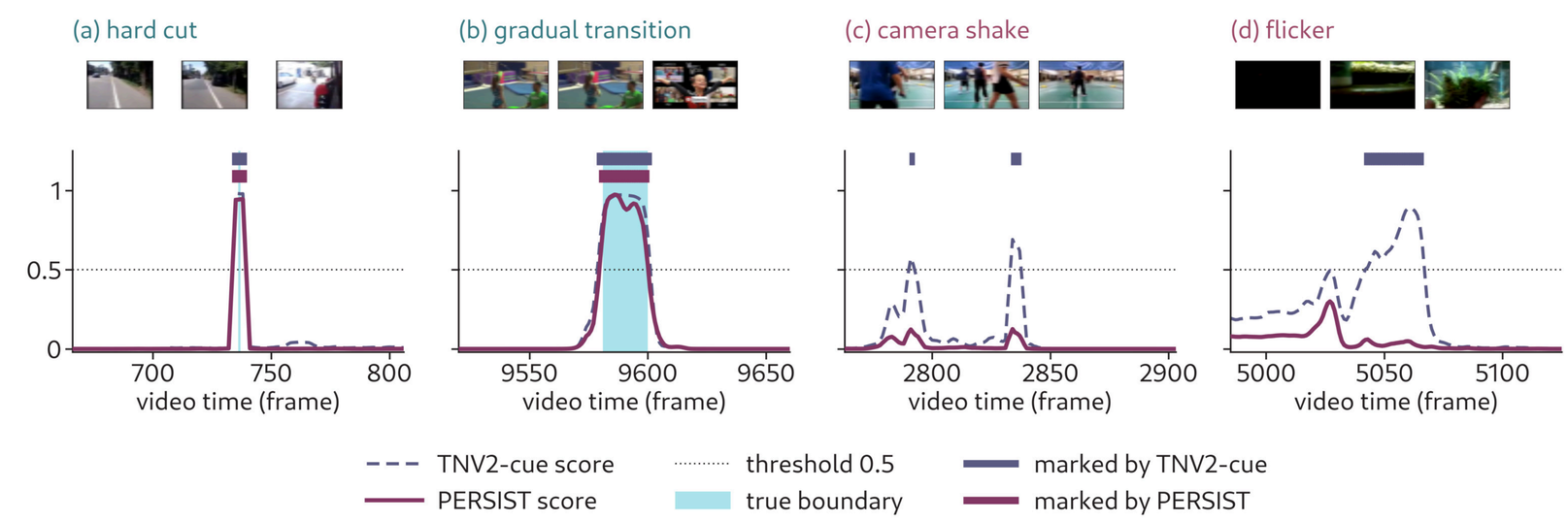
A fast pathway looks for sudden change. A slow pathway drives a latent state φ(t) that summarises the video at each moment.

Three gates then test each frame:

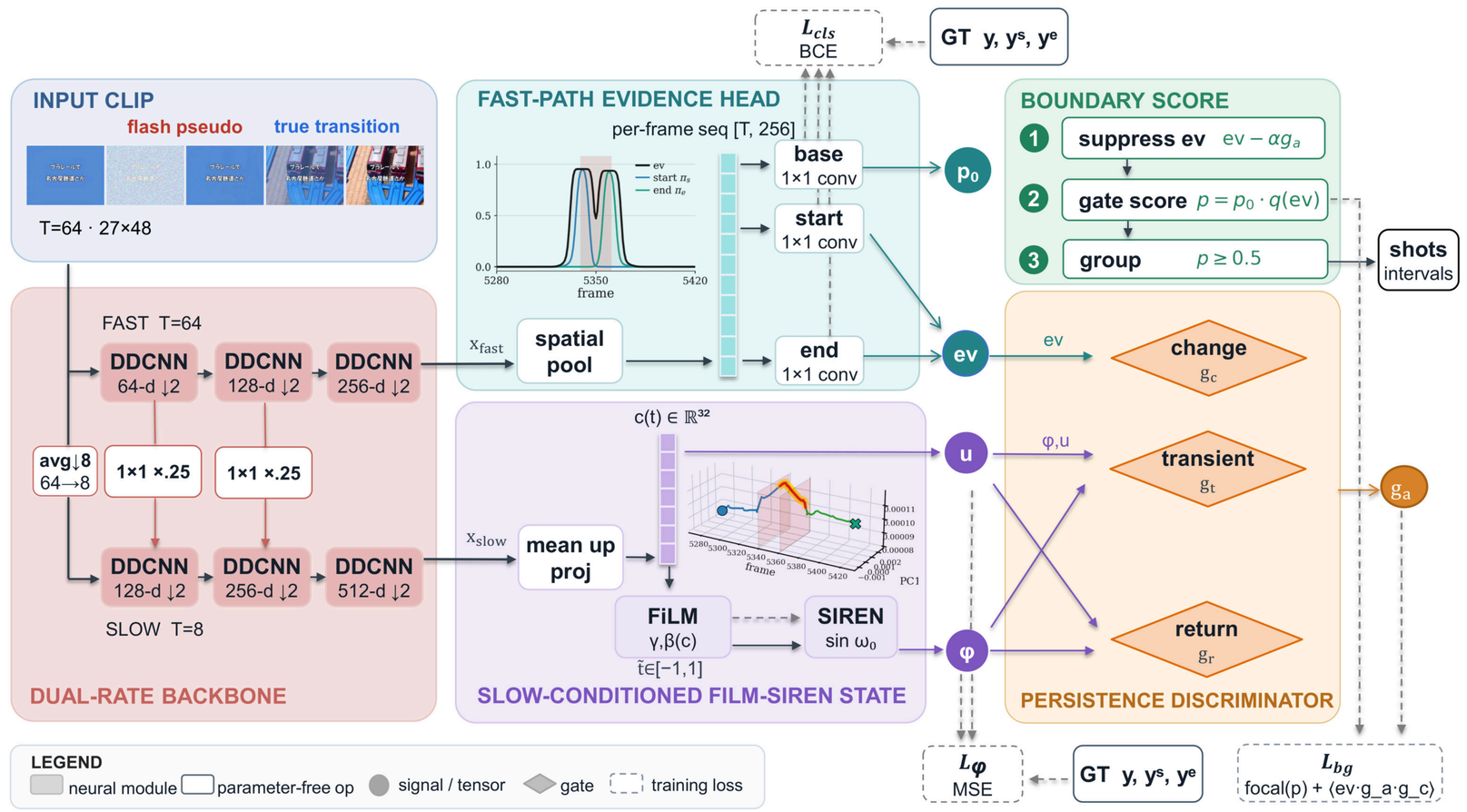
- Change g_c: is the local change strong?
- Transient g_t: does the observed signal jump more than φ?
- Return g_r: does φ go back to where it was?

Their product $g_a = g_c \cdot g_t \cdot g_r$ is high only if all three say yes. In training, the network learns to score such non-boundary frames low.

Both detectors mark the true boundaries, but only PERSIST ignores the pseudo-events.

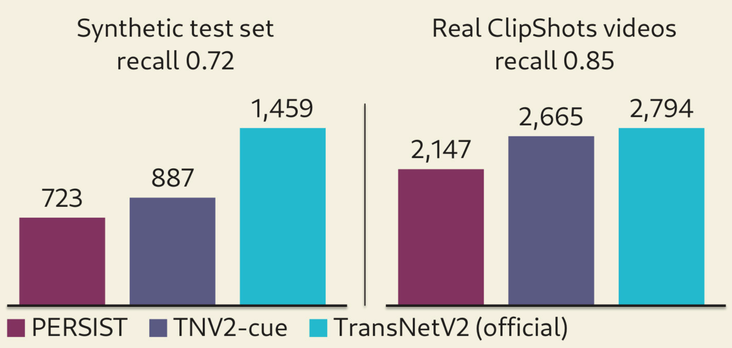


All four are real ClipShots videos. The dashed line is TNV2-cue, a baseline with the same backbone and training as PERSIST but with the cues of TransNetV2 instead of the persistence test.



FALSE ALARMS

PERSIST makes fewer false alarms.



All detectors are compared at the same recall, the share of true boundaries found. The synthetic test set adds a transition or pseudo-event to 2,727 unseen ClipShots clips.

On the synthetic test set, PERSIST makes about half the false alarms of TransNetV2 (723 vs 1,459), and fewer of every type:

−31%	−41%	−81%
flash	archival damage	text overlay

BENCHMARKS

PERSIST is highly competitive.

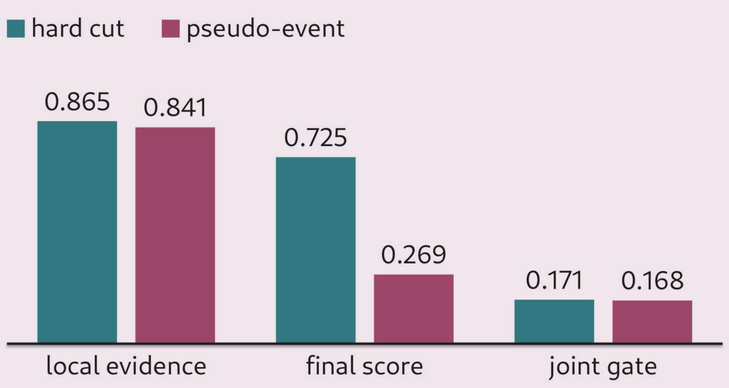
	TransNetV2		PERSIST	
F1	fixed	best	fixed	best
ClipShots	0.779	0.779	**0.791**	**0.794**
BBC	0.961	**0.969**	**0.966**	0.967
AutoShot	**0.799**	**0.830**	0.788	0.821
HIST-GST12	0.920	0.944	**0.923**	**0.946**

Higher F1 is better. “Fixed” uses the threshold 0.5, “best” the best threshold for each test set, and bold marks the better method.

The test sets range from web videos and a nature documentary to short videos and archival film. PERSIST equals or exceeds TransNetV2 in five of the eight columns and falls behind only on AutoShot, by about 0.01 F1.

INSIGHTS

PERSIST separates hard cuts from pseudo-events.



Averages over ClipShots test frames; pseudo-events here are non-boundary frames with high local evidence.

Hard cuts and pseudo-events look almost equally sudden but PERSIST gives hard cuts a 2.7 times higher final score. The joint gate barely differs between them at test time, so this difference comes from what the network learned during training.

TAKE AWAY

- A sudden change alone does not mean a new shot: what separates a true boundary from a pseudo-event is whether the change lasts.
- By testing this, PERSIST roughly halves the pseudo-event false alarms of TransNetV2 at the same recall and stays highly competitive on four benchmarks.

Code and model weights
github.com/linty5/PERSIST

tylin@cvl.tuwien.ac.at
This poster is presented at BMVC 2026, Lancaster, UK.

This research was funded in whole or in part by the Austrian Science Fund (FWF) [10.55776/DFH37], doc.funds.connect “Visual Heritage: Visual Analytics and Computer Vision Meet Cultural Heritage”.