Self-Supervised Contrastive Video-Speech Representation Learning for Ultrasound

Jianbo Jiao, Yifan Cai, Mohammad Alsharid, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble

University of Oxford

Embedding Space

ϕ

ϕ

Background noise

Non-narrative random talk

Other un-related sound

Self-supervised

Multi-modal data

Narrative audio & video

Affinity-aware self-paced learning

Standard plane detection

Eye-gaze saliency prediction

Video

Speech

Fusion

Correlated?

Correlated?

Positive pair

Hard-positive pair

Negative pair

Hard-negative pair

Standard plane detection

Eye-gaze saliency prediction

Input

Ground Truth

Rand.Init.

Ours

more salient

less salient

Input

Ground Truth

Rand.Init.

Ours

more salient

less salient

A

ffinity-aware self-paced learning

Standard plane detection

Eye-gaze saliency prediction

Rand.Init.

Ours