Panoramic Vision Transformer for Saliency Detection in 360° Videos

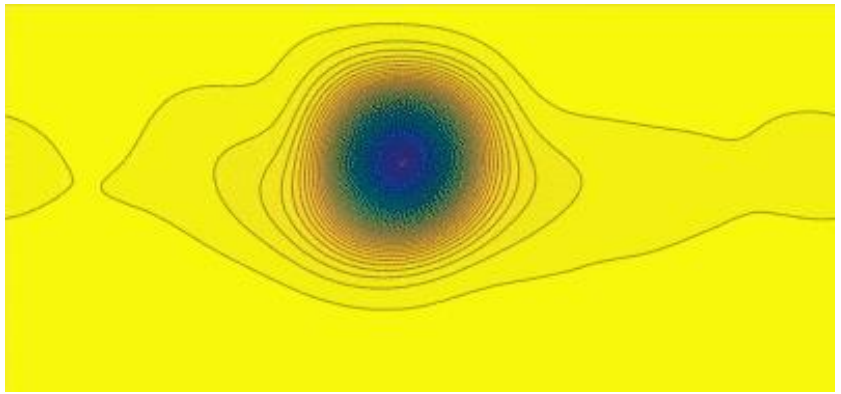
Heeseung Yun, Sehun Lee, Gunhee Kim

SEOUL NATIONAL UNIV.
VISION & LEARNING

ECCV
TEL AVIV 2022

360° Video Saliency Detection

- One of the core problems in 360° video understanding
  - Captures whole surroundings of a scene instead of specific predetermined context
- Directly related to practical scenarios
  - 360° video summarization (*i.e.*, cinematography)
  - Dynamic rendering for VR



Chen et al. Enhancing High-resolution 360 Streaming with View Prediction. 2017.

360° Video Saliency Detection

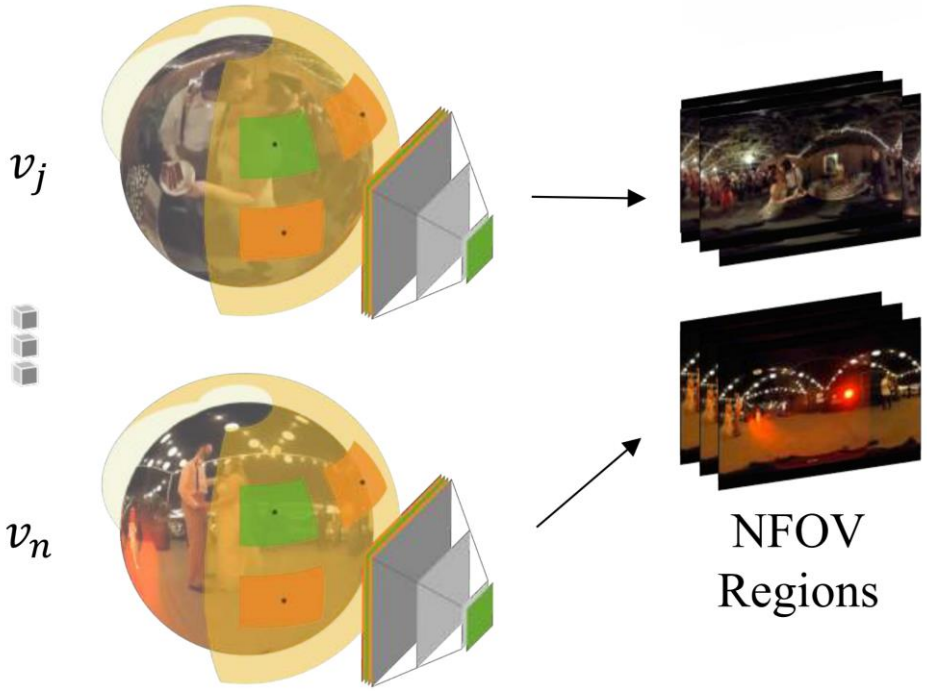
- A ~~simple~~ tricky question
  - “Which direction to watch if you were in the scene?”

- Problem 1. architecture design for panoramic videos
  - Taking distortion & discontinuity into account

- Problem 2. modeling “saliency” in panoramic videos
  - Often ambiguous and subjective

Architecture Design for 360° Videos

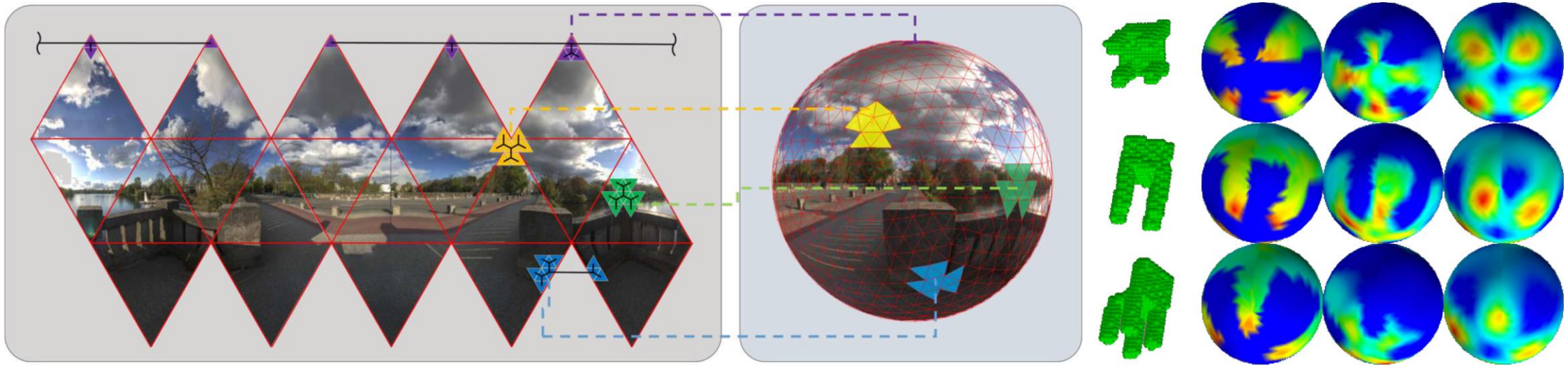
- Ignoring geometric property
  - 😐 Simple
  - 😕 Dense NFoV projection is not scalable



Lee et al. A Memory Network Approach for Story-based Temporal Summarization of 360 Videos. In CVPR 2018.

Architecture Design for 360° Videos

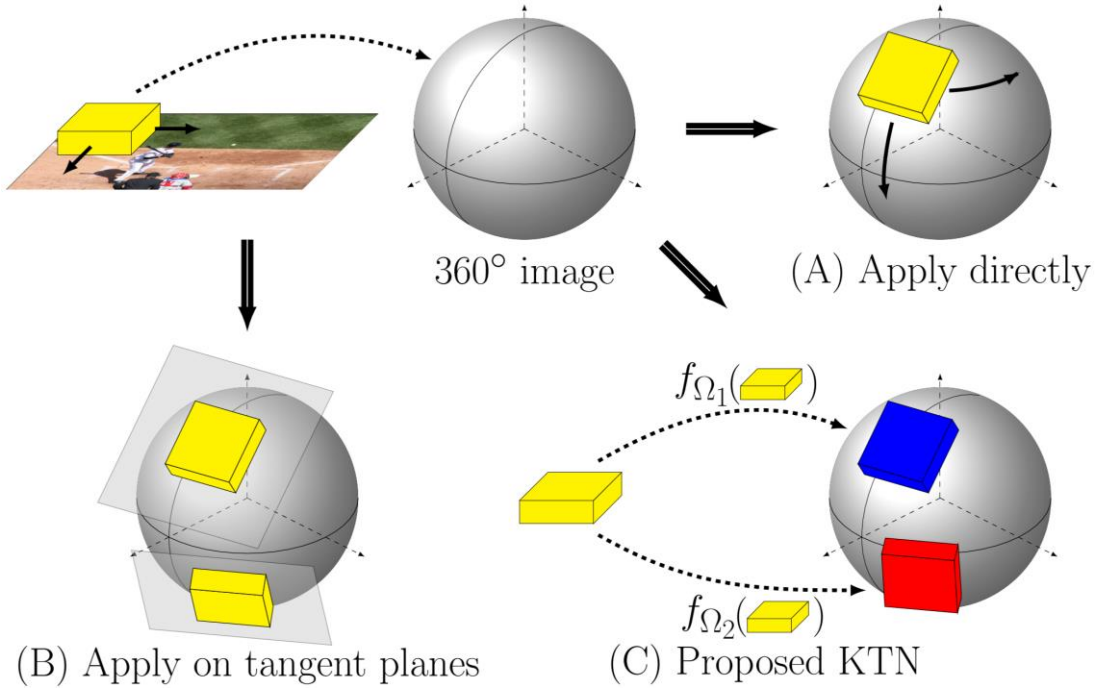
- Using designated architecture
  - 😀 Well-tailored for spherical input format
  - 😕 **Not transferrable**



Esteves et al. Learning SO(3) Equivariant Representations with Spherical CNNs. In ECCV 2018.
Lee et al. SpherePHD: Applying CNNs on a Spherical PolyHeDron Representation of 360 Images. In CVPR 2019.

Architecture Design for 360° Videos

- Additional modules & training for geometric adaptation
  - 😀 Transferrable architecture
  - 😕 Training overhead, layerwise error accumulation



Su et al. Kernel Transformer Networks. In CVPR 2019.

Architecture Design for 360° Videos

- Previous approaches
  - Ignoring geometric property
  - Using designated architecture
  - Additional modules & training for geometric adaptation

- **Solution**: Local patch-based modeling
  - 😀 Geometry-aware approximation
  - 😀 Transferrable from most Vision Transformer (ViT) variants
  - 😀 No additional modules or training for geometric adaptation
  - 😀 Applicable to various 360° video formats

Modeling Saliency in Panoramic Videos

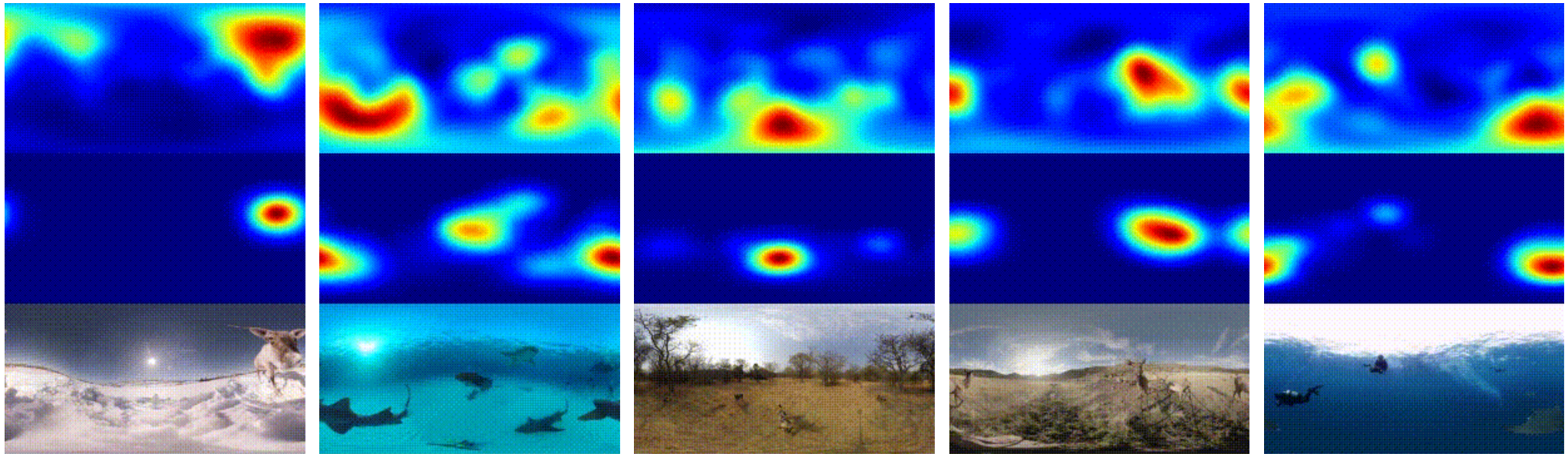
- Saliency itself is a longstanding question in CV
  - Saliency as self-information, anomaly, class activation, etc.

- *"Which direction to watch if you were in the scene?"*
  - Ambiguity and subjectivity
  - Supervised learning usually inapplicable

- **Solution**: leverage features from NFoV domain
  - 😀 Rich and readily available pretrained knowledge
  - 😀 Spatio-temporal feature consistency suffices for 360° video saliency

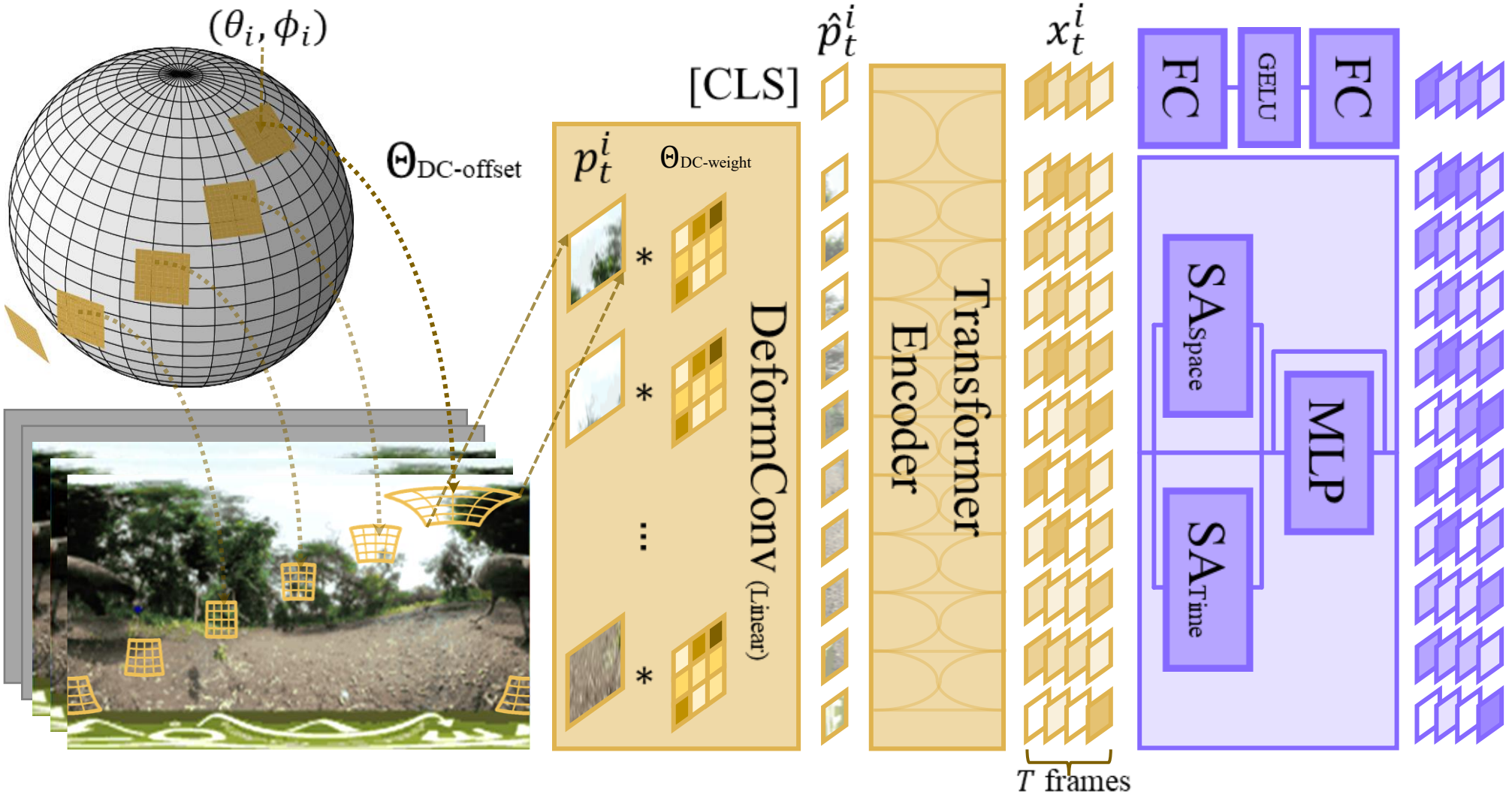
PAVER

- **Pa**norami**c** **V**ision Transform**er** for 360° videos
  - First to adopt ViT to encode omnidirectional imagery
  - No additional module, trivial overhead, video format-agnostic
  - Do not require class activation, optical flow for saliency
  - Competitive results in 360° video saliency detection & quality assessment



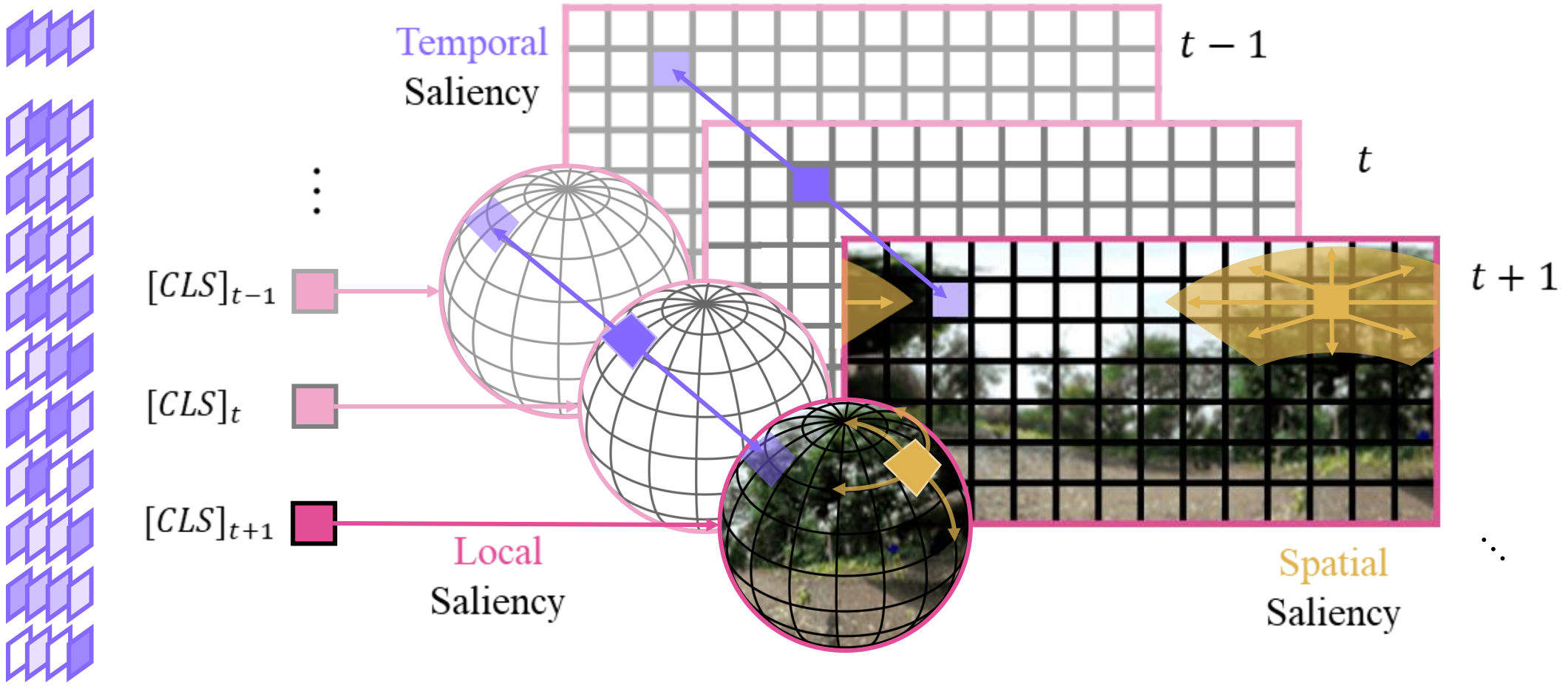
PAVER

- **Pa**noramic **V**ision Transform**er**: Encoder



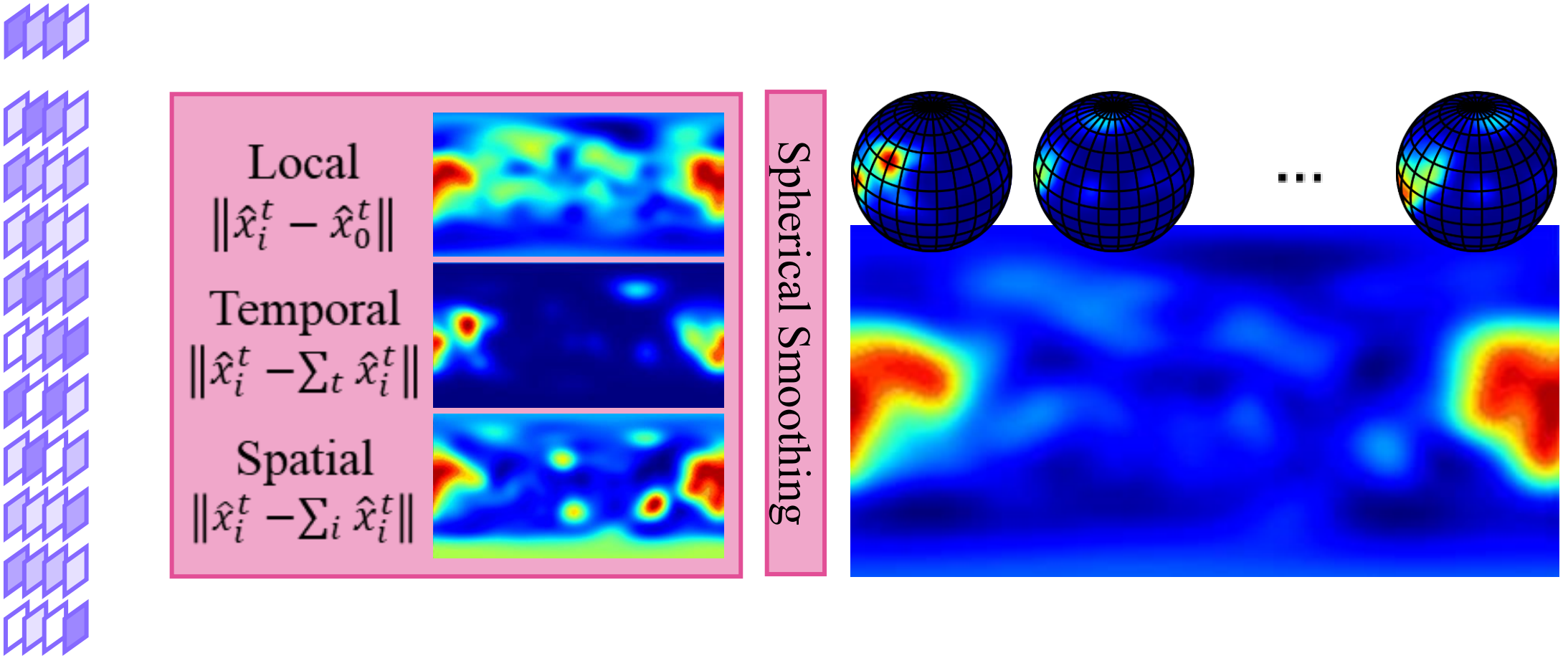
PAVER

- **Pa**norami**c** **V**ision Transform**er**: Decoder



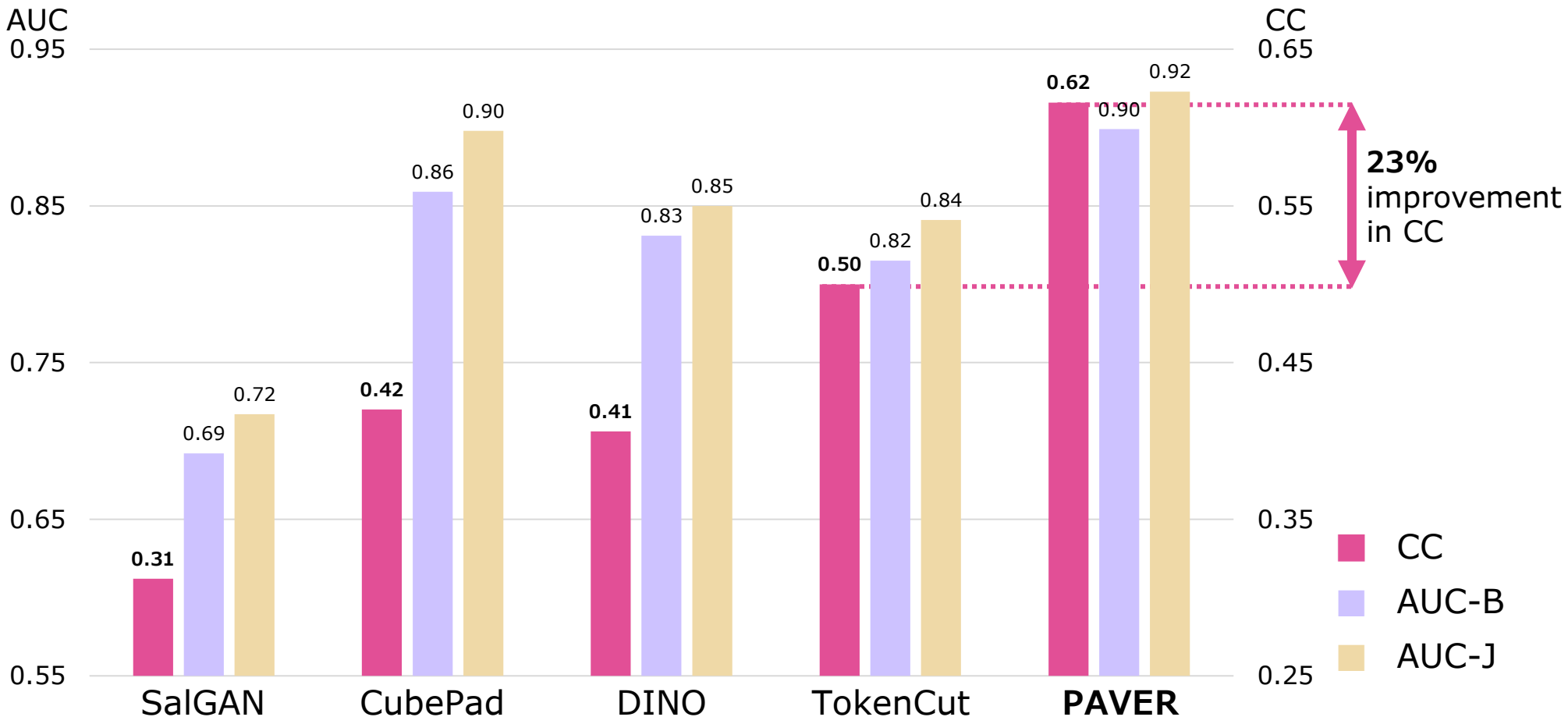
PAVER

- **Pa**noramic **V**ision Transform**er**: Decoder



Experiments

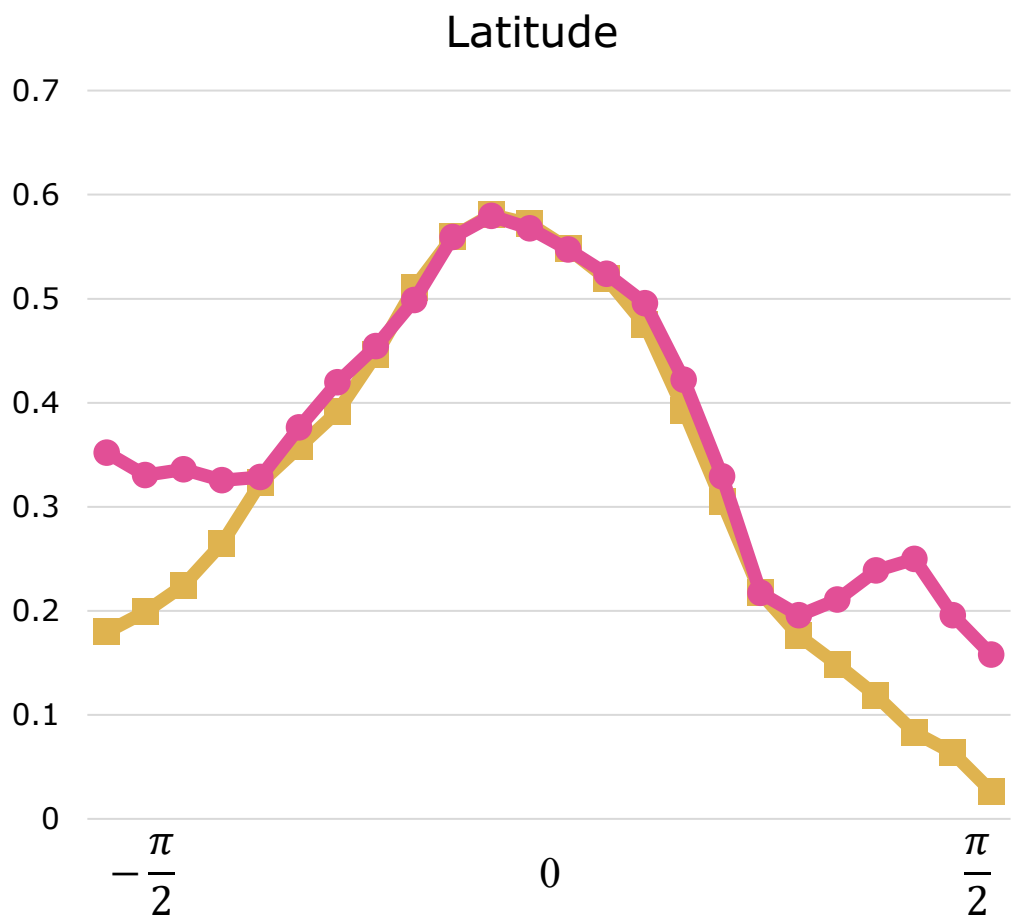
- 360° video saliency detection (Wild360)

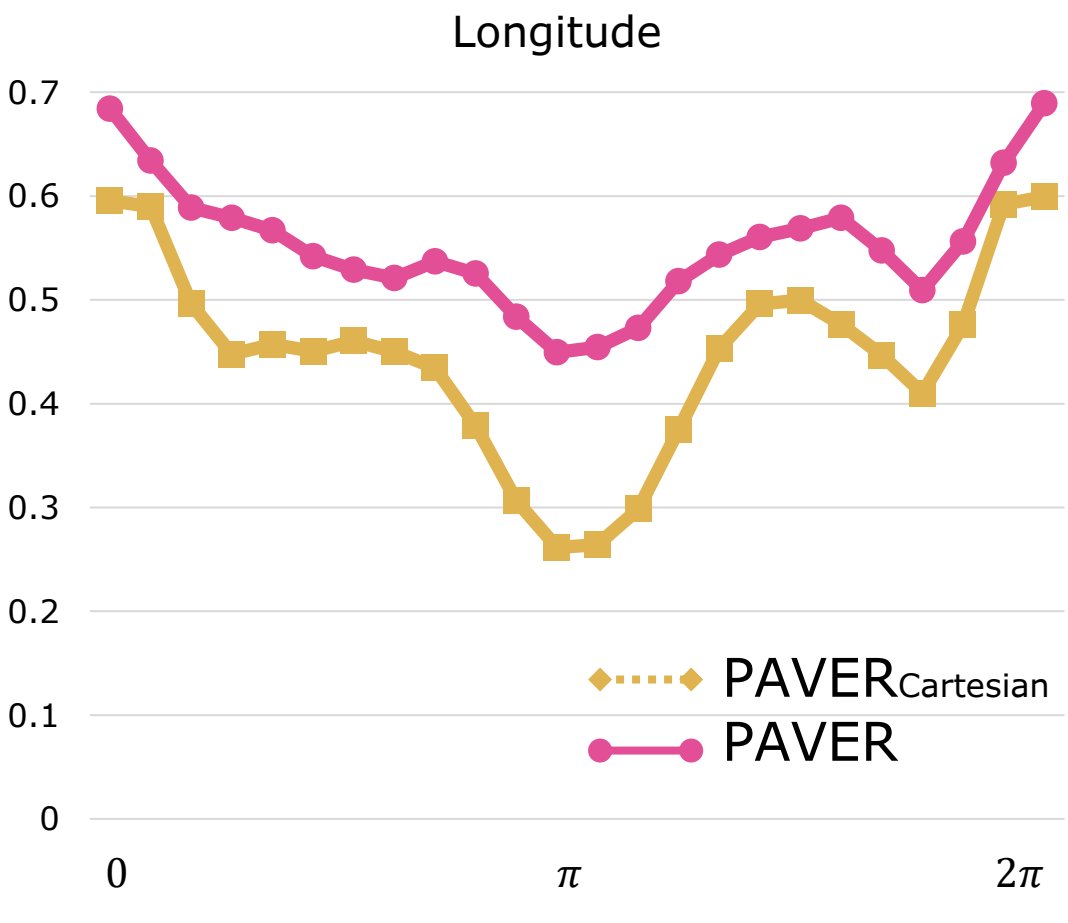


Cheng et al. Cube Padding for Weakly-Supervised Saliency Prediction in 360 Videos. In CVPR 2018.

Experiments

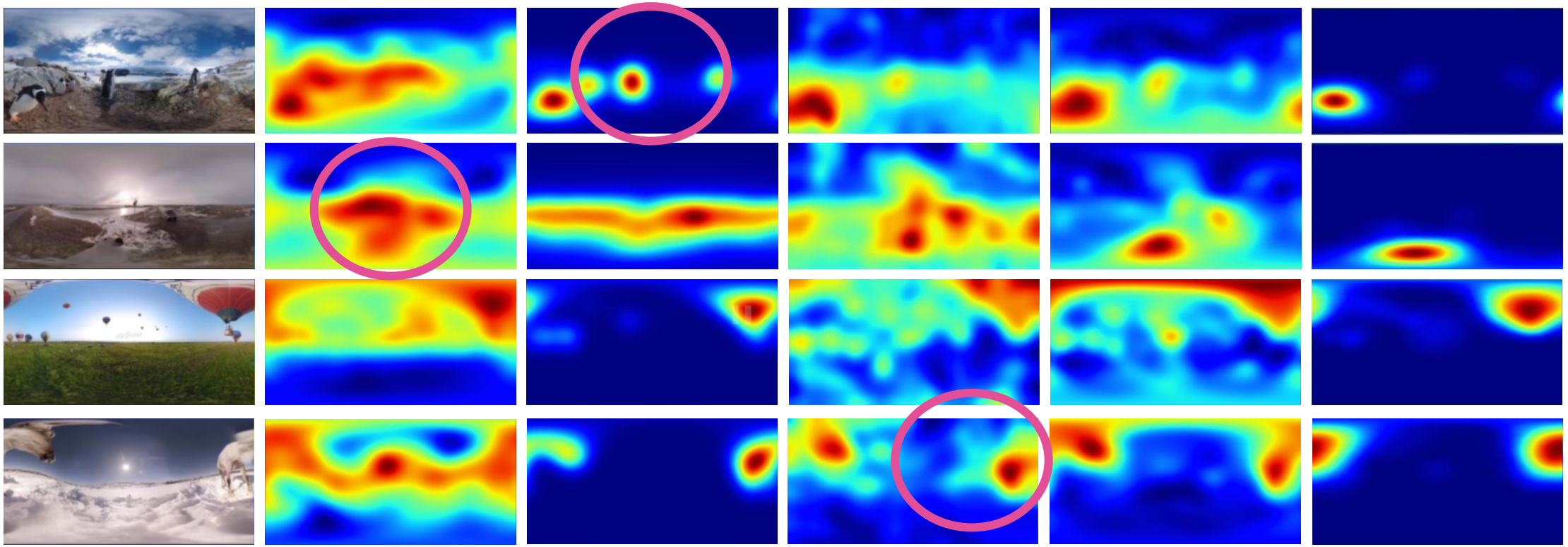
- 360° video saliency detection (Wild360)





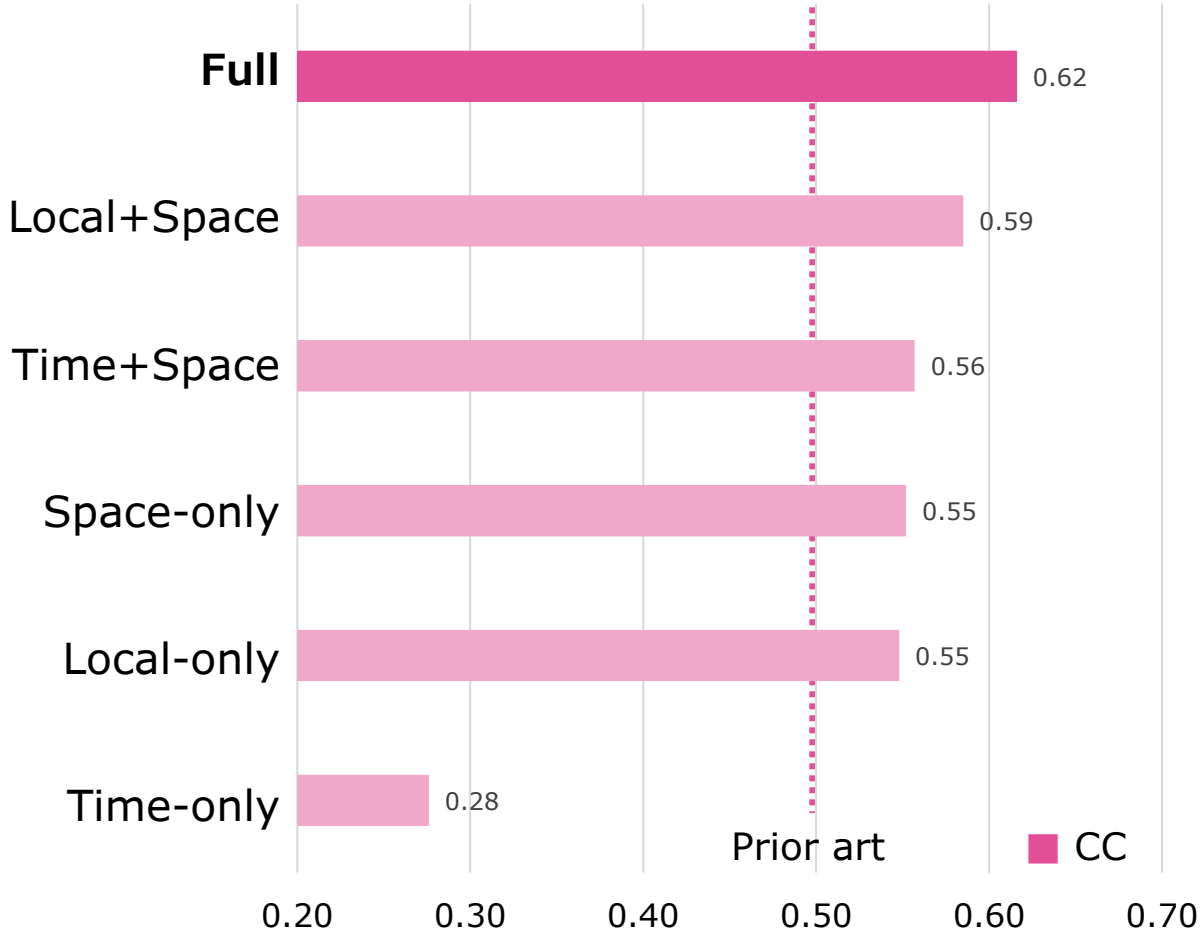
Experiments

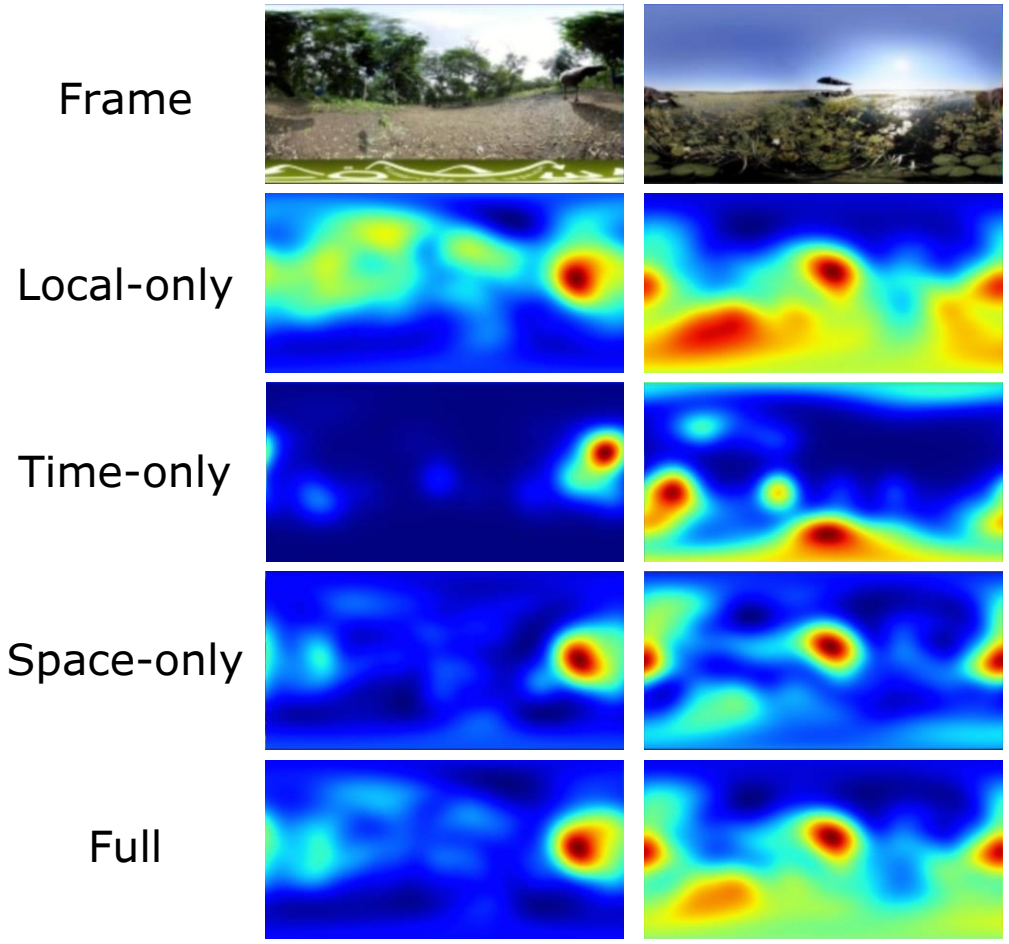
- 360° video saliency detection (Wild360)



Experiments

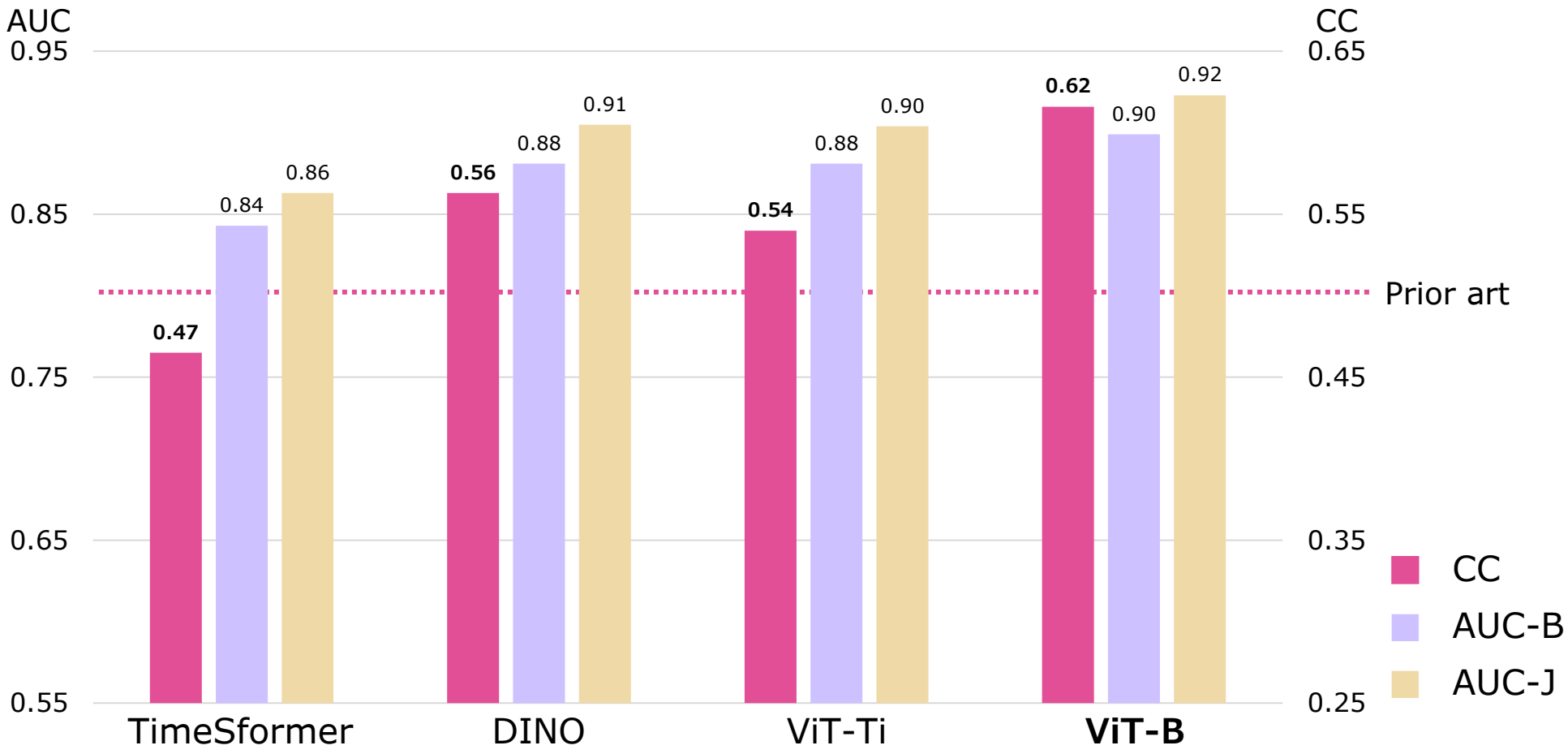
- Saliency score decomposition





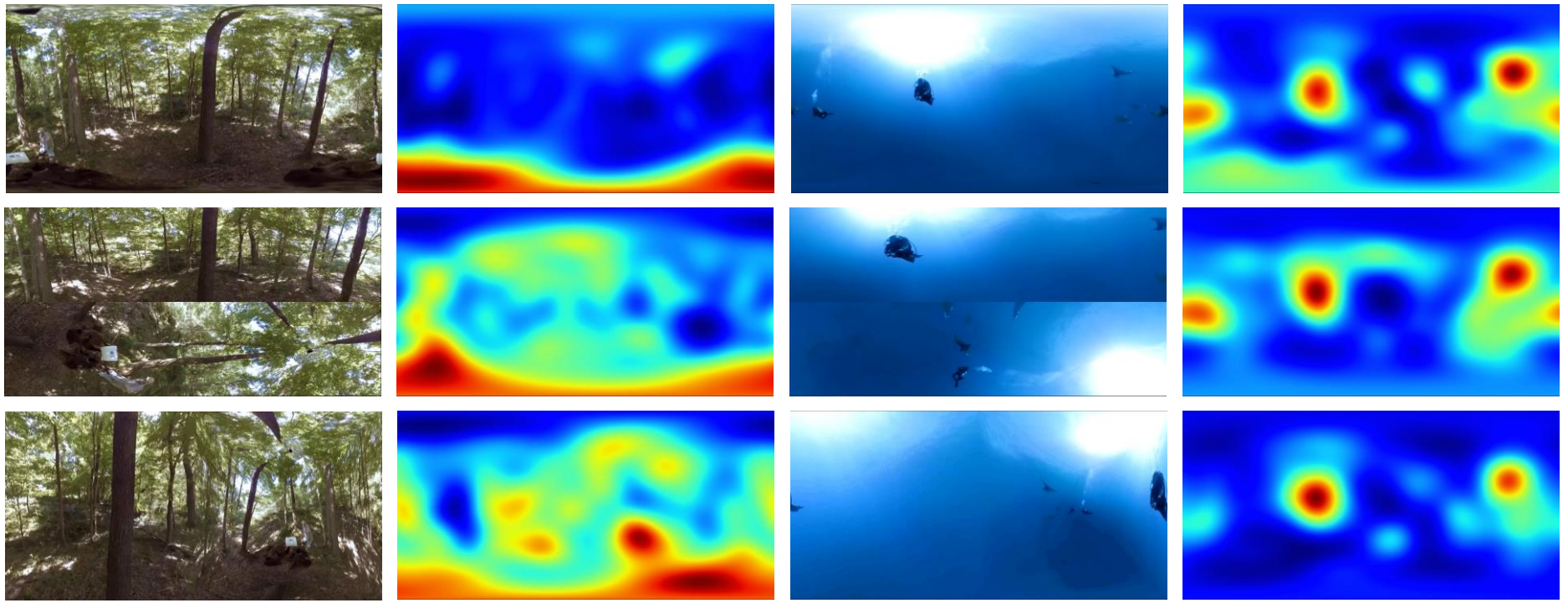
Experiments

- Transfer from different pretrained knowledge



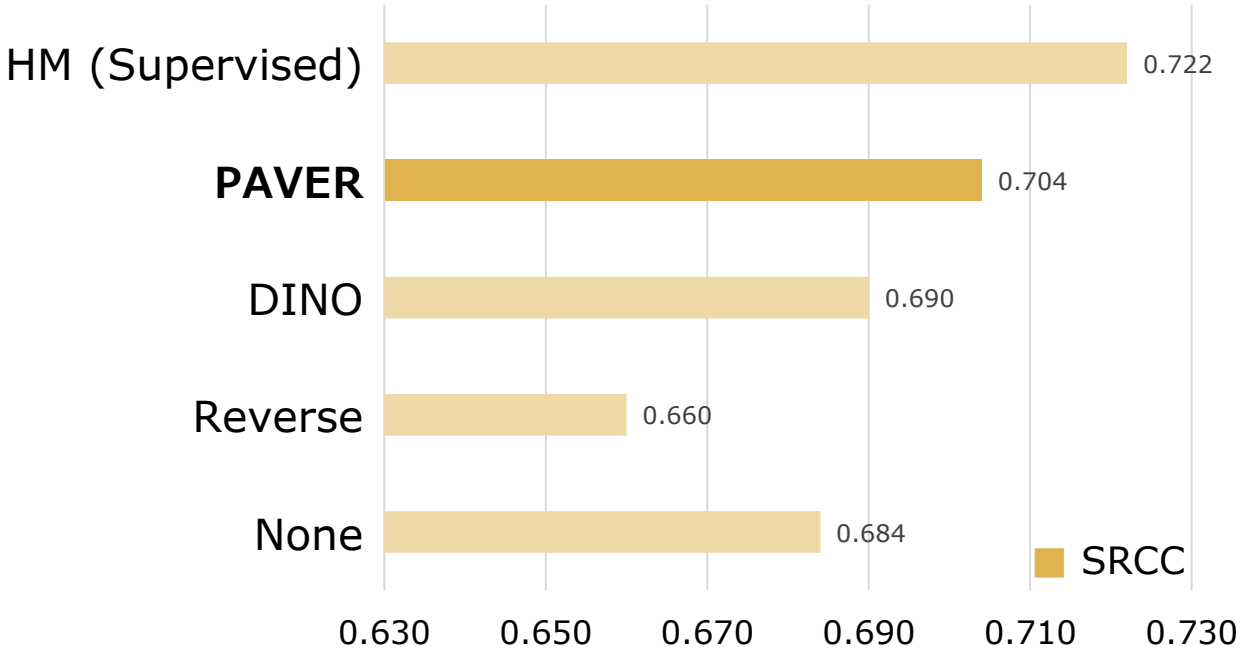
Experiments

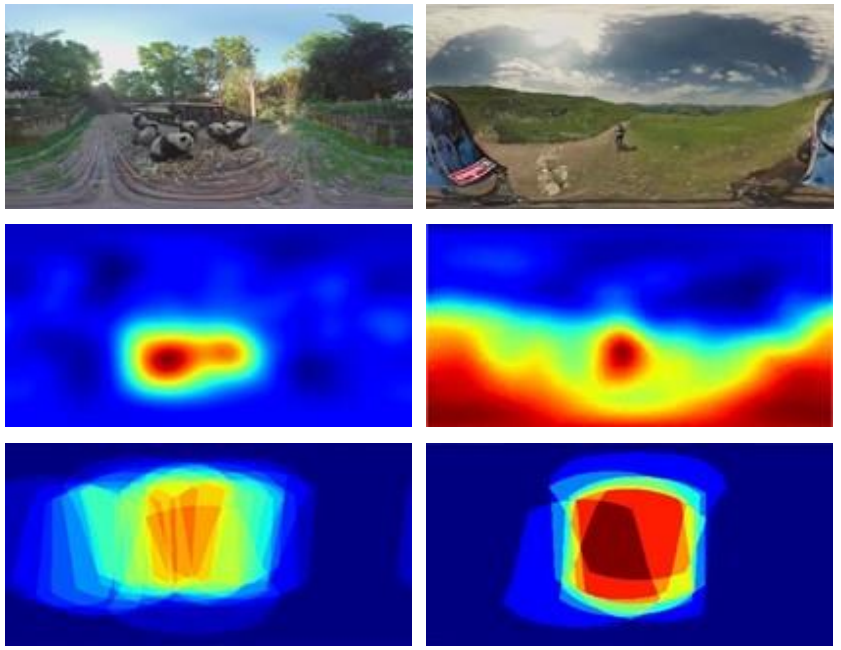
- Applying to various 360° video formats



Experiments

- Video quality assessment of 360° videos (VQA-ODV)
  - Relevant to rendering quality control in VR pipeline





Li et al. Bridge the Gap Between VQA and Human Behavior on Omnidirectional Video. In ACM MM 2018.

Concluding Remarks

- Local patch-based architecture for 360° videos
  - First attempt to adopt ViT to encode omnidirectional imagery
  - No additional module, trivial overhead
- Spatio-temporal consistency of local patch features for measuring saliency
  - Leverage pretrained knowledge
  - Independent of class activation, optical flow, etc.
- Competitive results in 360° video saliency detection & quality assessment

Thank You

Code available at 👇
https://github.com/hs-yn/**PAVER**