visuAAL

Privacy-Aware and Acceptable Video-Based Technologies and Services for Active and Assisted Living

Context Recognition for the Application of Visual Privacy

ESR 14

University of Alicante

18/06/2024

Kooshan Hashemifard

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 861091".

Universitat d'Alacant
Universidad de Alicante
Project Coordinator

RWTH AACHEN UNIVERSITY

Stockholm University

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

TU WIEN

Motivation

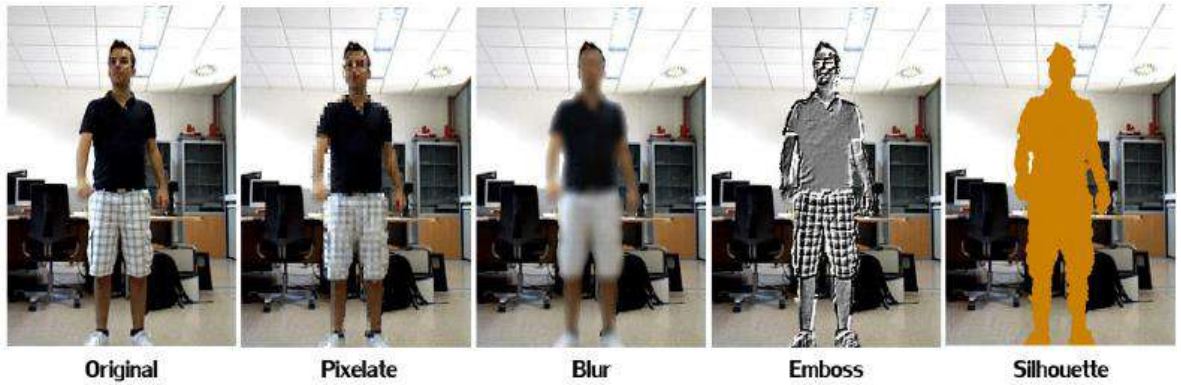
- Demographic changes
- Burden to care personnel and facilities
- Damage to autonomy, self-esteem and spirit
- Ambient-assisted living (AAL) and sensors
- Video-based technology
- The most directed and natural way to record events
  - Pros: Provide richer information
  - Cons: Easy to interpret by unauthorized viewers and privacy issues

Research project: Context recognition for the application of visual privacy

Privacy | Intelligibility in the video

- Find a visualization method to understand what is happening and at the same time preserve privacy

- Previous work introduced privacy-by-context:
  - Level-based visualization
  - Selected according to the context

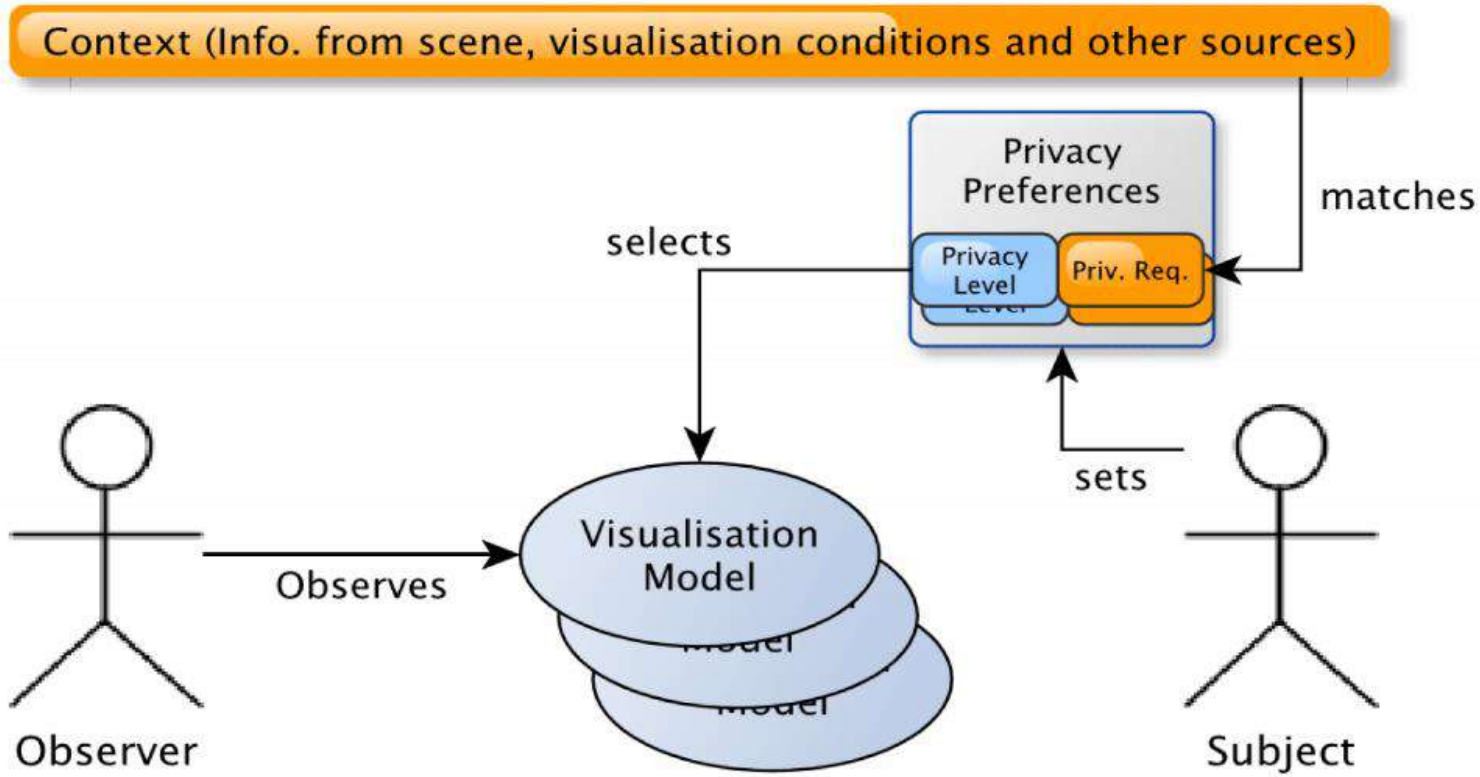


Objective: Automatic selection of visualization

Context Variables

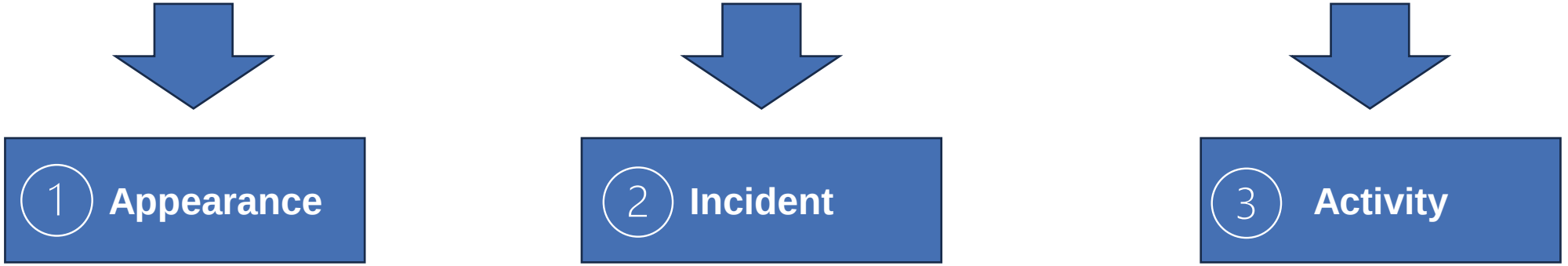
Context is modeled by the following variables:

- Activity
- Incident
- Appearance
- Observer
- Place, ...



Research Questions

- This leads to the following research questions, in line with the objective:
  - - Can we automatically estimate the context that may affect visualization preference?
  - - Can we estimate relevant activities?
  - - Can we estimate relevant events, such a fall?
  - - Can we estimate degree of nudity?
  - - Can we integrate them into a privacy-by-context approach so that visual privacy adapts in real time?



1 Appearance

Nudity and AAL

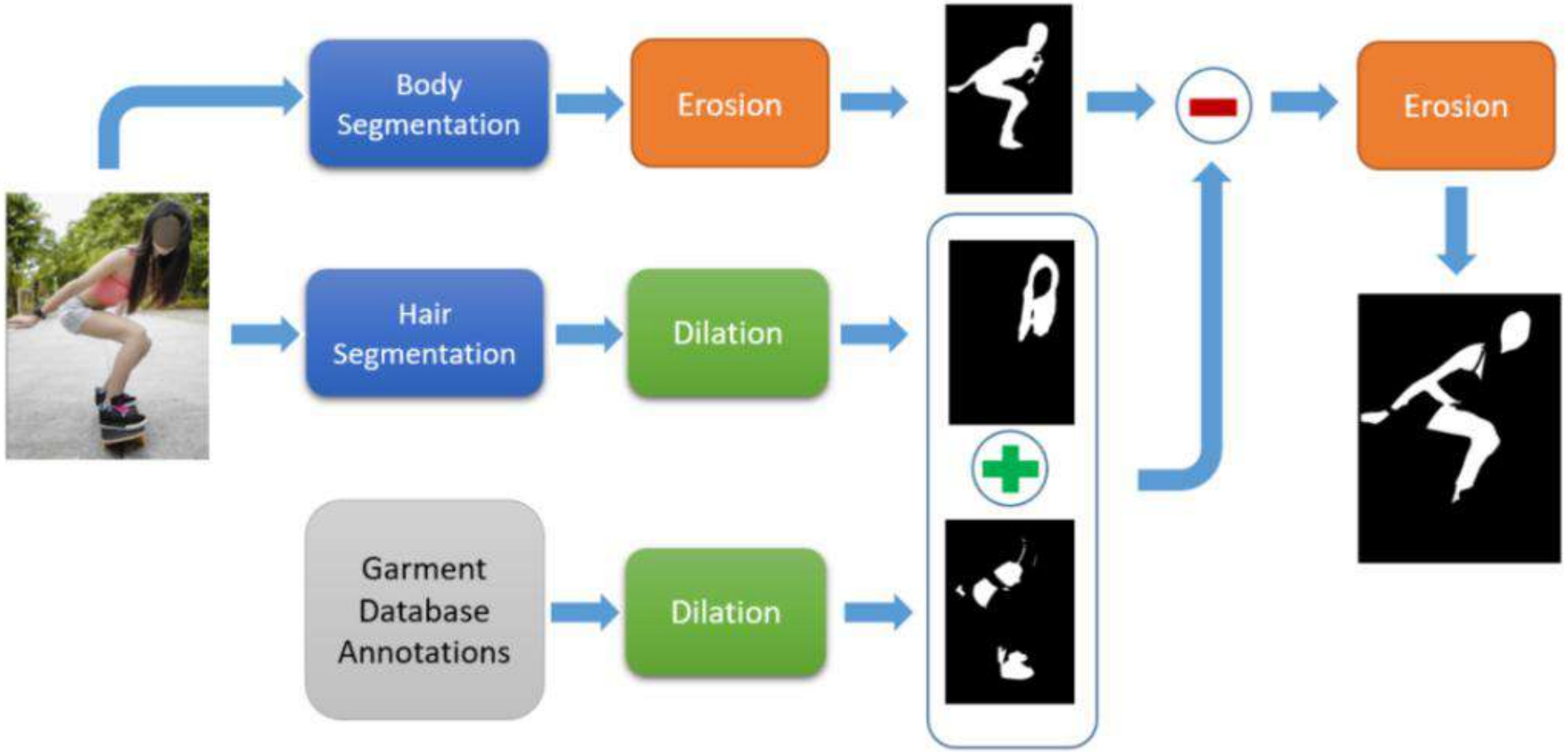
- Serious concerns about person appearance recorded in a video
- Appearance detection can be interpreted as nudity detection in the context of AAL
- Skin Exposure Detection plays a crucial role in nudity detection
- Estimate the degree of nudity

Appearance Detection in Literature

- Obscene image Classification
  - Naïve approach
  - No level-based classification
  - Subjective
  - Fail to generalize the problem
  - Limited standard datasets

- Garment Detection
  - Detect different clothing and wearables
  - Market analysis and finding trends
  - Does not necessarily represent the coverage level
  - Cannot cover all types of garments

1- Dataset

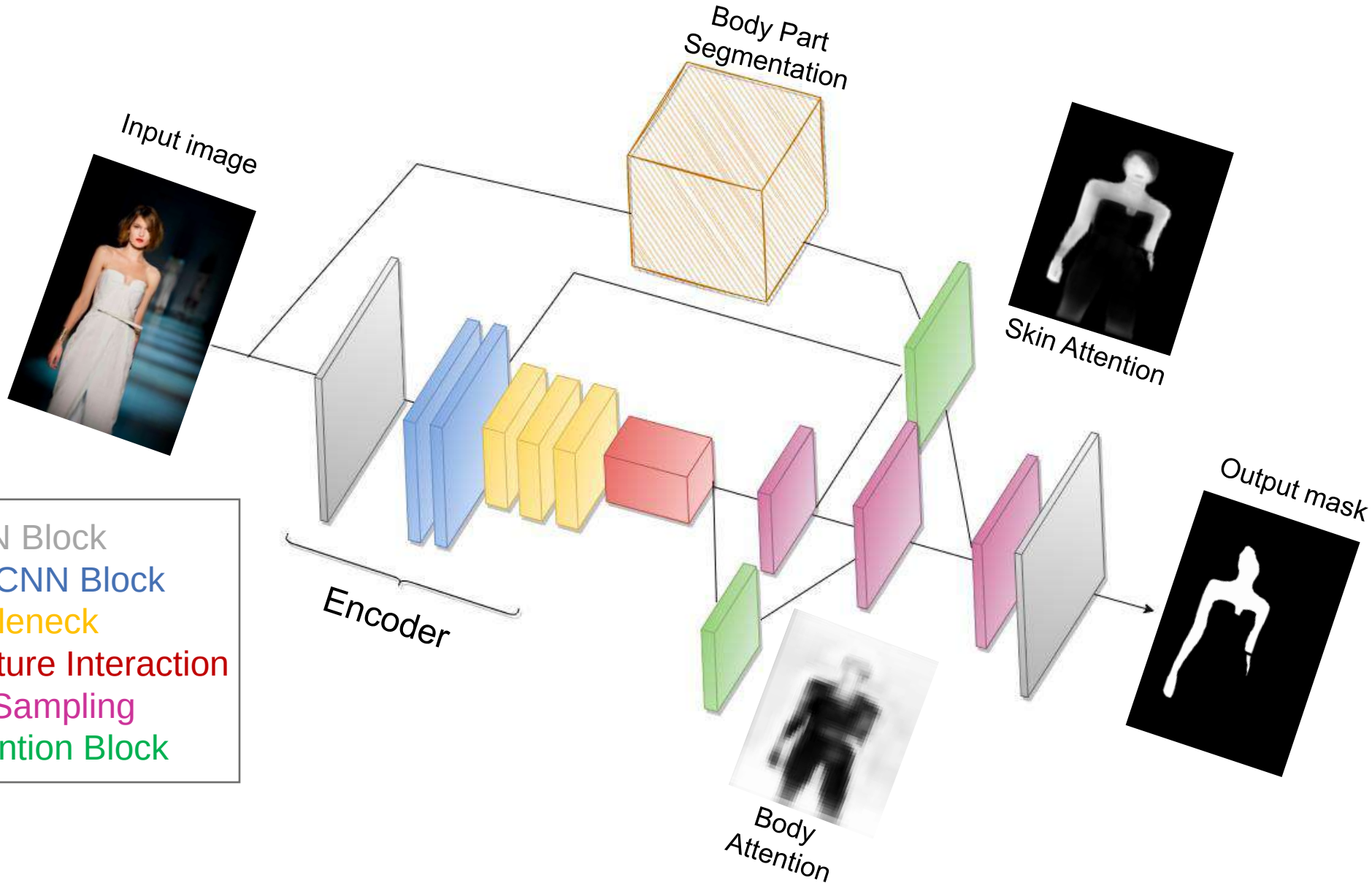

Body Segmentation
Erosion
Erosion
Hair Segmentation
Dilation
Garment Database Annotations
Dilation

1- Dataset

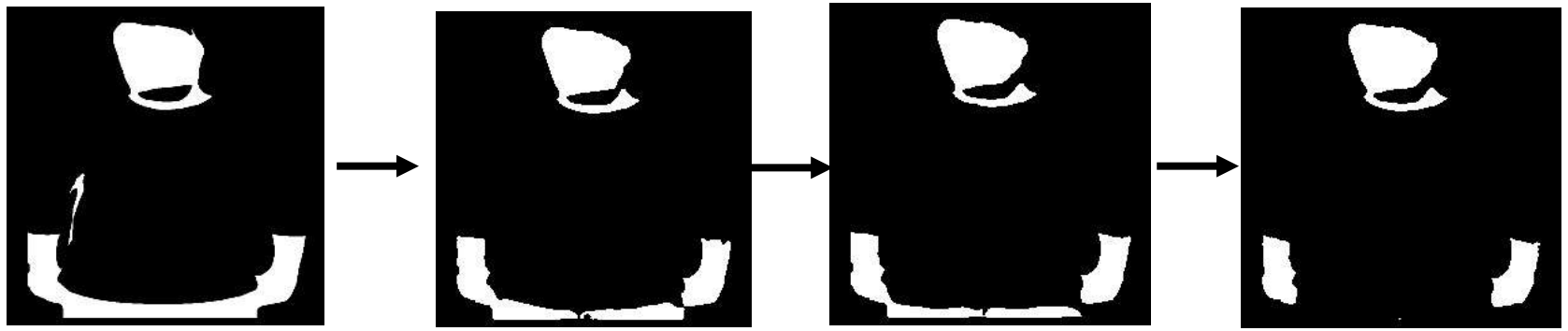
Validation

Body Erosion	Garment Dilation	Hair Dilation	Skin Erosion	Precision	Recall	F1score	CDR
5	1	3	1	93.45%	93.05%	93.25%	98.92%
3	1	3	3	92.28%	94.49%	93.37%	98.92%
3	1	1	7	91.94%	94.72%	93.31%	98.91%
5	1	1	7	93.11%	93.27%	93.19%	98.90%
7	5	5	5	95.72%	88.07%	91.74%	98.72%
7	7	7	7	96.05%	85.94%	90.71%	98.58%
1	1	1	1	90.44%	95.8%	93.04%	98.85%
0	1	3	0	90.78%	95.56%	93.11%	98.86%

2- Architecture



3- Training

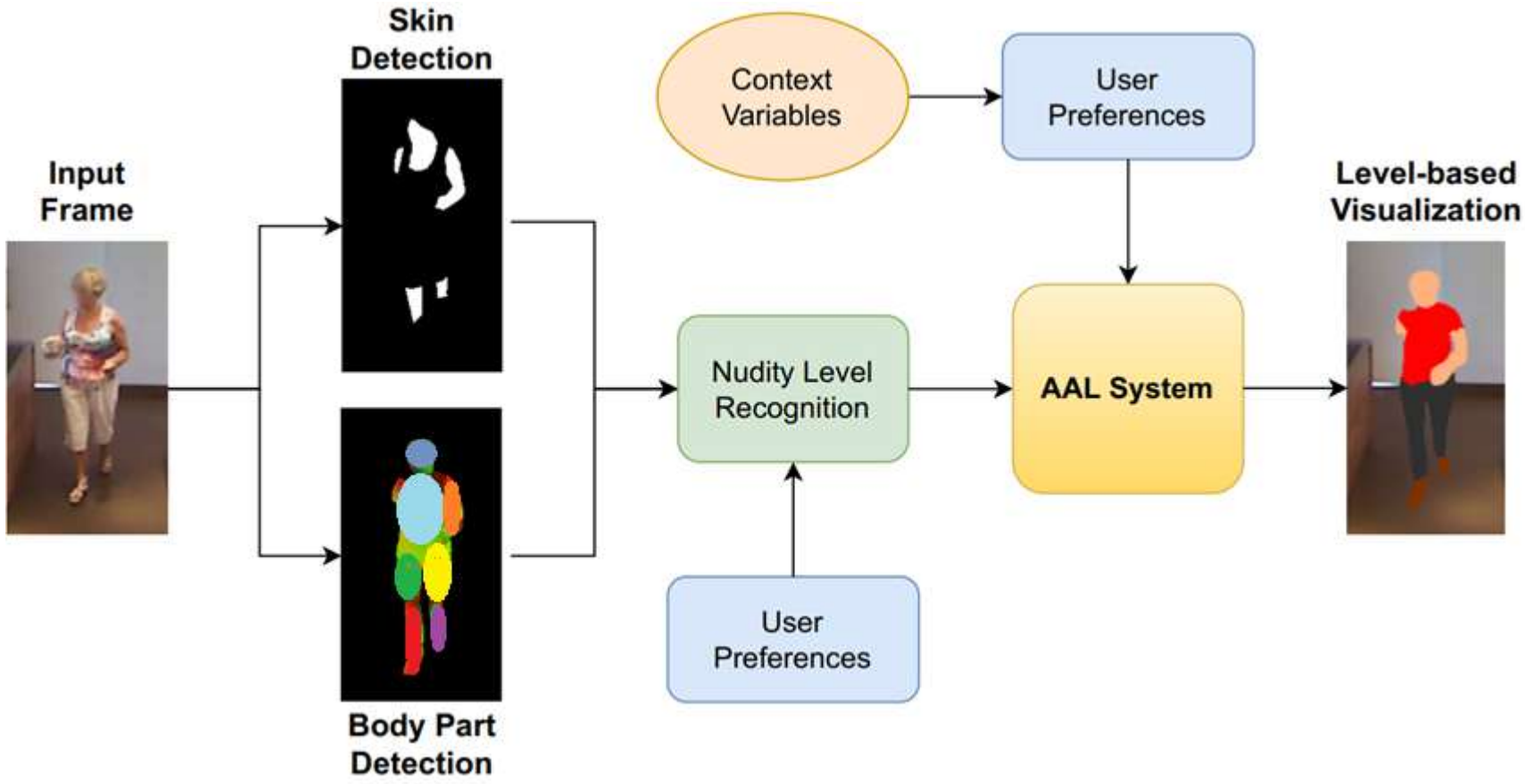


4- Evaluation

Method	Precision	Recall	F1-score	CDR	DSC	Param.
SegNet	80.71%	80.12%	80.29%	97.75%	78.82%	30M
UNet	82.66%	85.34%	83.83%	98.01%	82.38%	30M
Basic FCN	70.34%	84.88%	76.80%	97.11%	74.40%	10M
DSNet	85.80%	85.08%	85.40%	98.35%	84.14%	8M
HLNet	76.50%	79.86%	78.01%	97.43%	76.08%	1.2M
Proposed Method	**88.30%**	**85.91%**	**86.96%**	**98.45%**	**86.44%**	1M
Base Network	76.20%	78.22%	77.19%	97.24%	75.88%	1M
With Body Attention	78.11%	82.08%	80.04%	97.81%	78.92%	1M
With Skin Attention	86.62%	83.11%	84.82%	98.18%	83.33%	1M

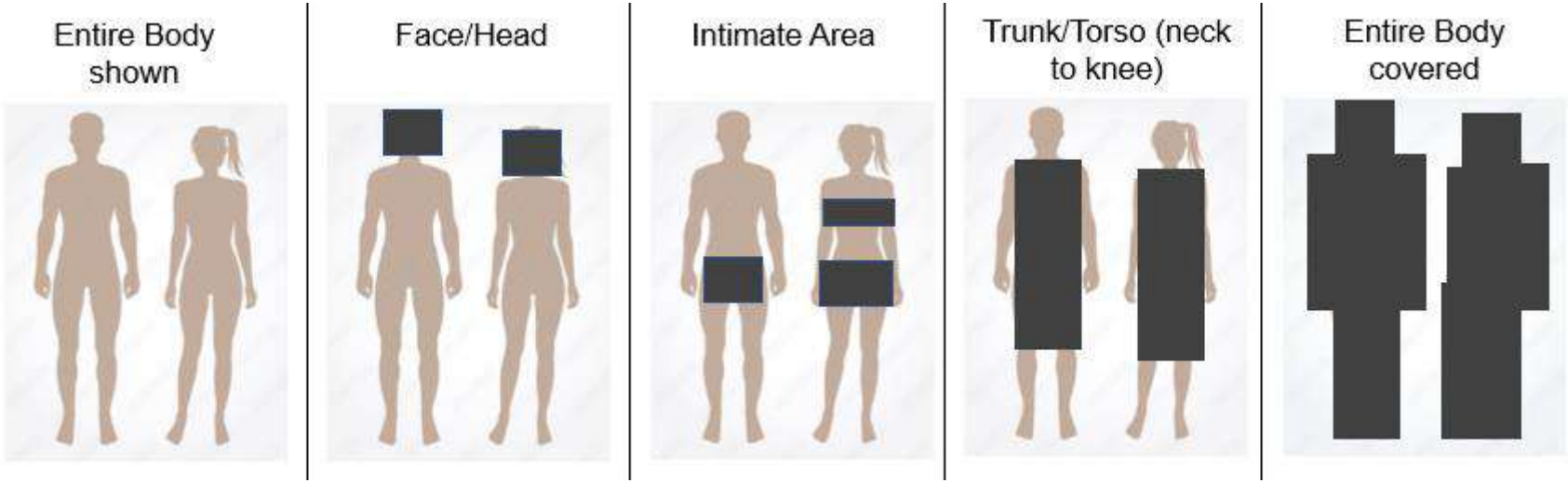
5- Nudity Recognition

Proposed Appearance Detection in AAL



Nudity Level

Nudity Level	Description
1	*Completely covered*
2	*Covered torso (neck to knee)*
3	*Covered intimate areas*
4	*Covered faces*
5	*Full body or exposed intimate areas*



2 Incident

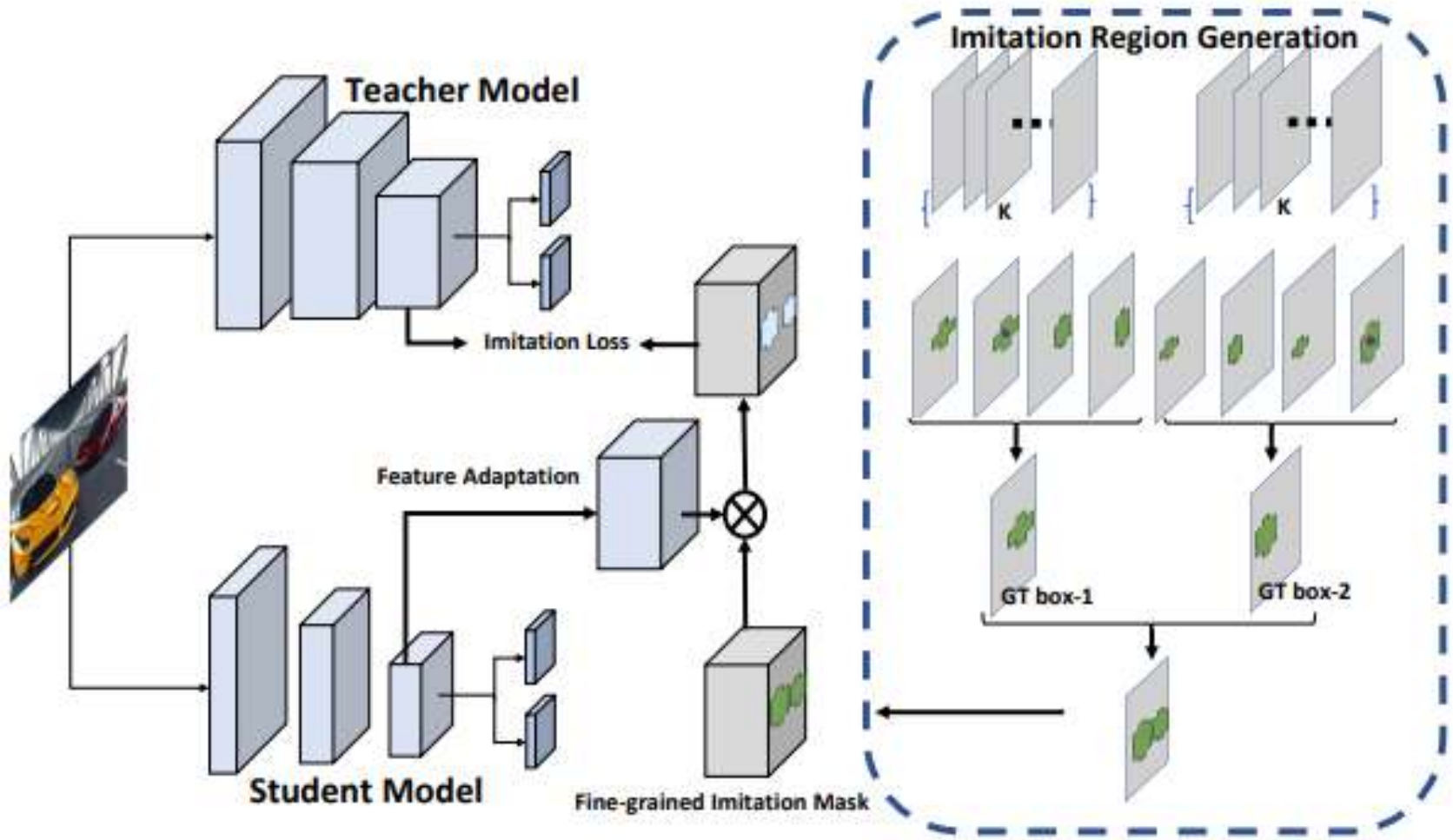
Transfer Learning

- End-to-end **Object Detection** approach with 2 classes: fallen and upright individuals
- Top layer of the object detector network was replaced with two output neurons for each status
- Pre-trained parameters retained, except for the last layer
- Fine-tune on E-FPDS dataset, contains 6982 images captured in indoor environments, with 5023 instances of falls and 2275 instances of non-falls in various scenarios, including variations in pose, size, occlusions, and lighting

Knowledge Distillation

- State-of-the-art CNN-based networks can be computationally expensive and difficult to deploy on smaller devices, especially in real-time and multi-tasking scenarios
- Knowledge distillation has emerged to directly learn compact models by transferring knowledge from a large model (teacher network) to a smaller one (student network)
- Reducing computational costs without sacrificing validity
- Fine-grained Feature Imitation method
  - Local features in the object region and near its anchor location contain important information and are more crucial for the detector and how the teacher model tends to generalize
  - The smaller student detector is trained by using both ground truth supervision and feature response imitation on object anchor locations from the teacher networks

Knowledge Distillation



Evaluation In-the-wild with Luxonis

- model’s performance in real-world settings after conversions
- videos, captured using the Luxonis OAK-D camera
- 38 videos recorded from various angles and labeled

Evaluation In-the-wild with Luxonis

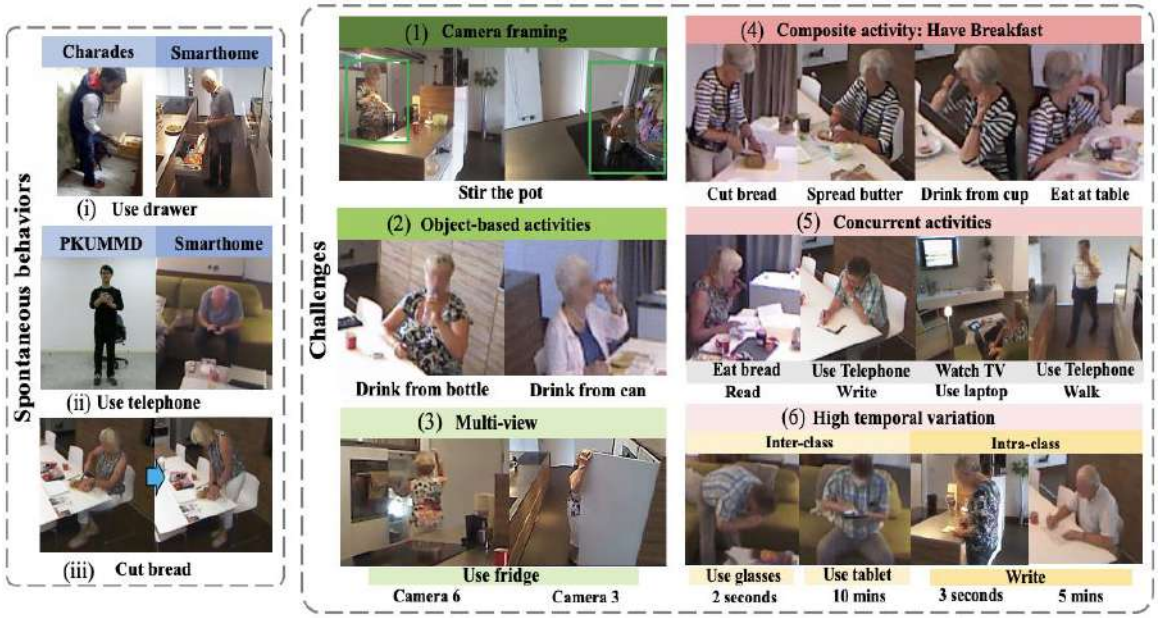
Model	Precision	Recall	mAP50	mAP50-95
YOLOv6 L	99.12%	98.64%	99.45%	65.58%
YOLOv6 M	94.95%	98.21%	98.64%	65.33%
YOLOv6 S	92.71%	95.22%	98.10%	62.67%
YOLOv5 L	93.48%	88.36%	94.89%	58.12%
YOLOv5 S	83.67%	86.44%	93.37%	46.77%
YOLOv5 S-v6 L	94.46%	91.83%	95.04%	57.29%

3 Activity

Background & Tools

Toyota Smarthome Dataset

- The subjects are senior people in the age range 60-80 years old.
- 35 Activities from daily living:
  - Composite Activities: cooking, cleaning, making breakfast
  - Elementary Activities: laydown, watch tv, use laptop, reading, phone call, take pill
  - Object-based Activities: drink from bottle/can



- Data Modalities available: RGB, depth, skeleton
- Data Modalities use:
  - RGB for scene details, depth maps for 3D structural information, skeletal data for 3D joint locations.

Multi-modal approaches integrate different modalities for a comprehensive understanding of complex actions.

Deep Learning Solution for HAR

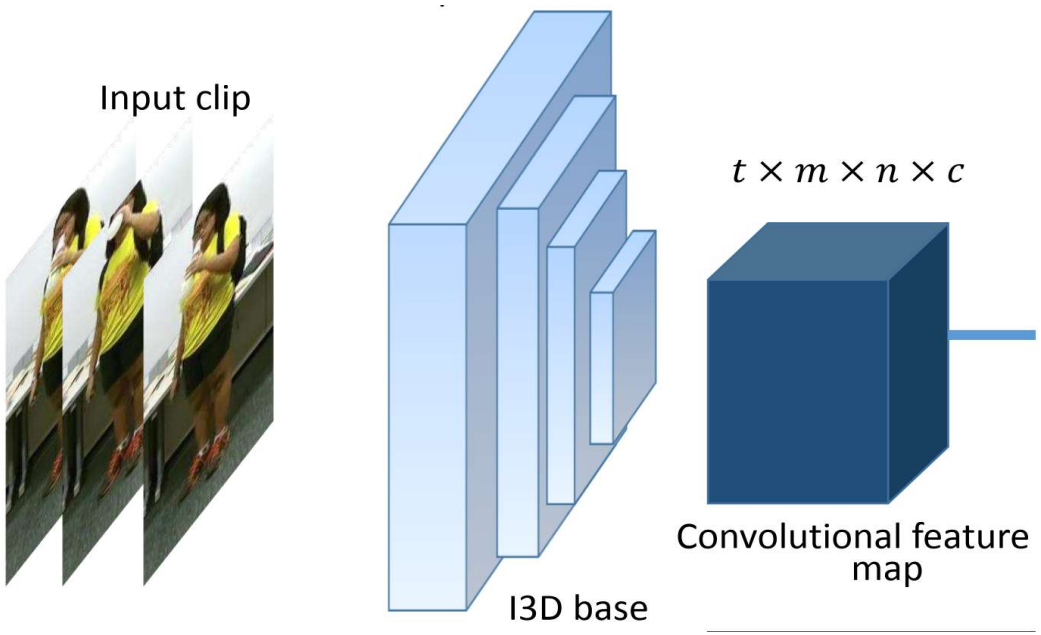
- **Spatial Information:**
  - Utilization of CNN**.** CNNs' ability to extract useful and discriminative features
- **Temporal Information:**
  - Existing deep architectures encode temporal information with limited solutions.
  - Challenges in acquiring both local and global variations of temporal features.
  - RNNs and time series models, higher dimension CNNs, newer Transformer architecture.
- **Transformers in HAR**
  - New encoder-decoder architecture using attention mechanism.
  - Success in natural language processing tasks; now applied to images and video recognition.
  - Video as a sequence of images, akin to language processing (image frames as words).
  - Not restricted to sequential processing; attention mechanism provides context for any position.
  - Relatively new approach, Increasing research focus on transformers for action recognition in recent years.

Approach

1- Backbone

CNN-based Network for rich feature extraction

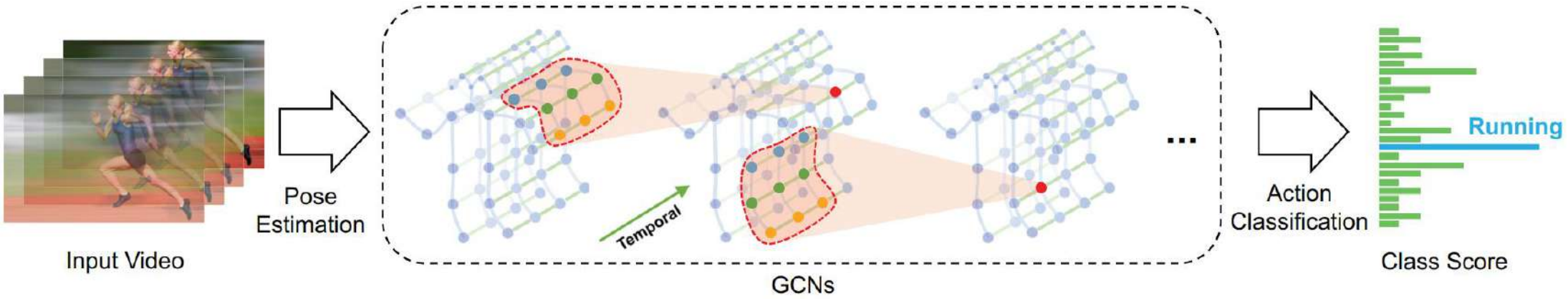
- 3D CNN Networks: I3D Net
- Time-distributed 2D CNN Networks: EfficientNet, DenseNet, InceptionNet, …



Approach

2- Pose Information

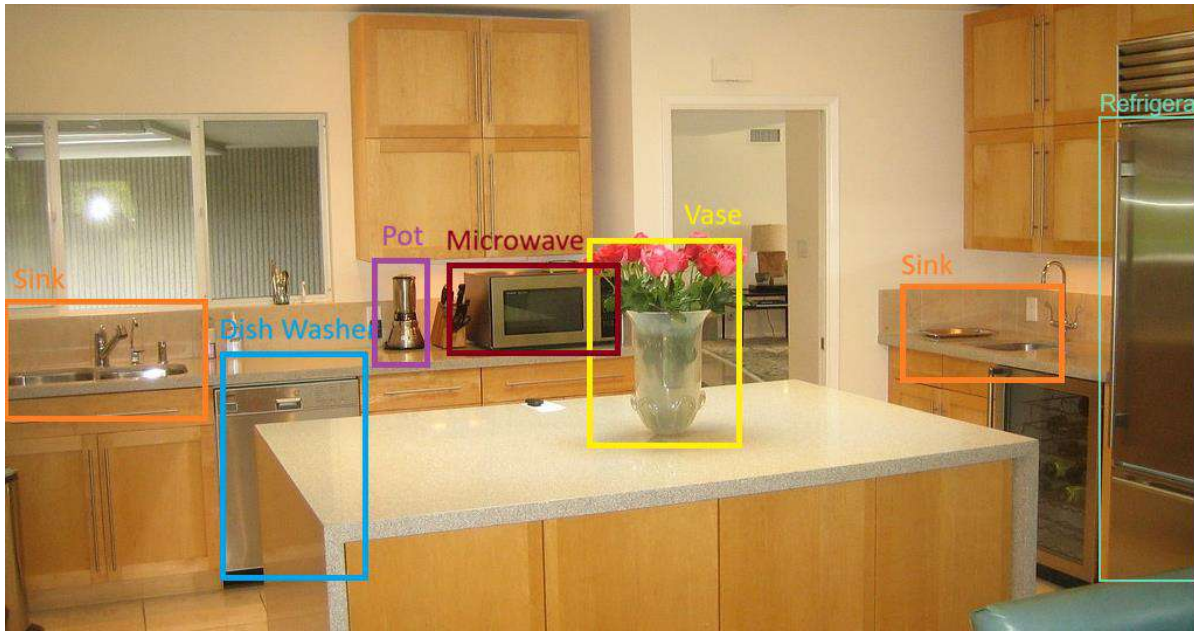
- The RGB image includes rich scene and object information, but does not encode long-range temporal information
- Pose info is more robust to changes in appearance, clothing, or lighting condition
- View invariance: Relationship between body joints coordinates can tackle Cross-view (CV) problem.
- Pose as Graph: joints as Vertexes and bones as Edges
- Approach: GCN and Body Joints Graph



Approach

3- Reinforcing Feature extraction with Object Detection

- Auxiliary information about Daily Activities
  - House Objects can play an import role for a given activity
  - Location information
- Advances in Object Detection
- House objects datasets

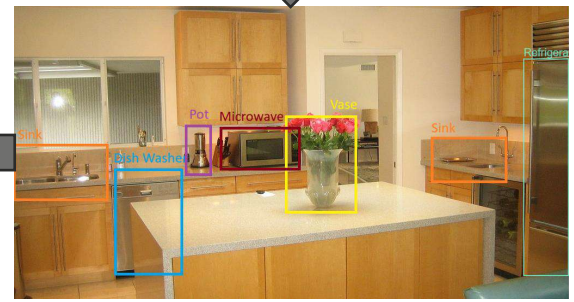


Approach

Object Groups of Interests

- **Person**
- **Kitchen Furniture:** Stove, Refrigerator, Oven, Microwave, Sink
- **Living room Furniture:** TV, Sofa, Table
- **Cutlery:** Spoon, Knife, Dish, Glass
- **Food**
- **Bathroom Furniture:** Toilet, Shower, Tube
- **Electronics:** Laptop, Cell phone, Tablet
- **Bedroom Furniture:** Bed, Wardrobe

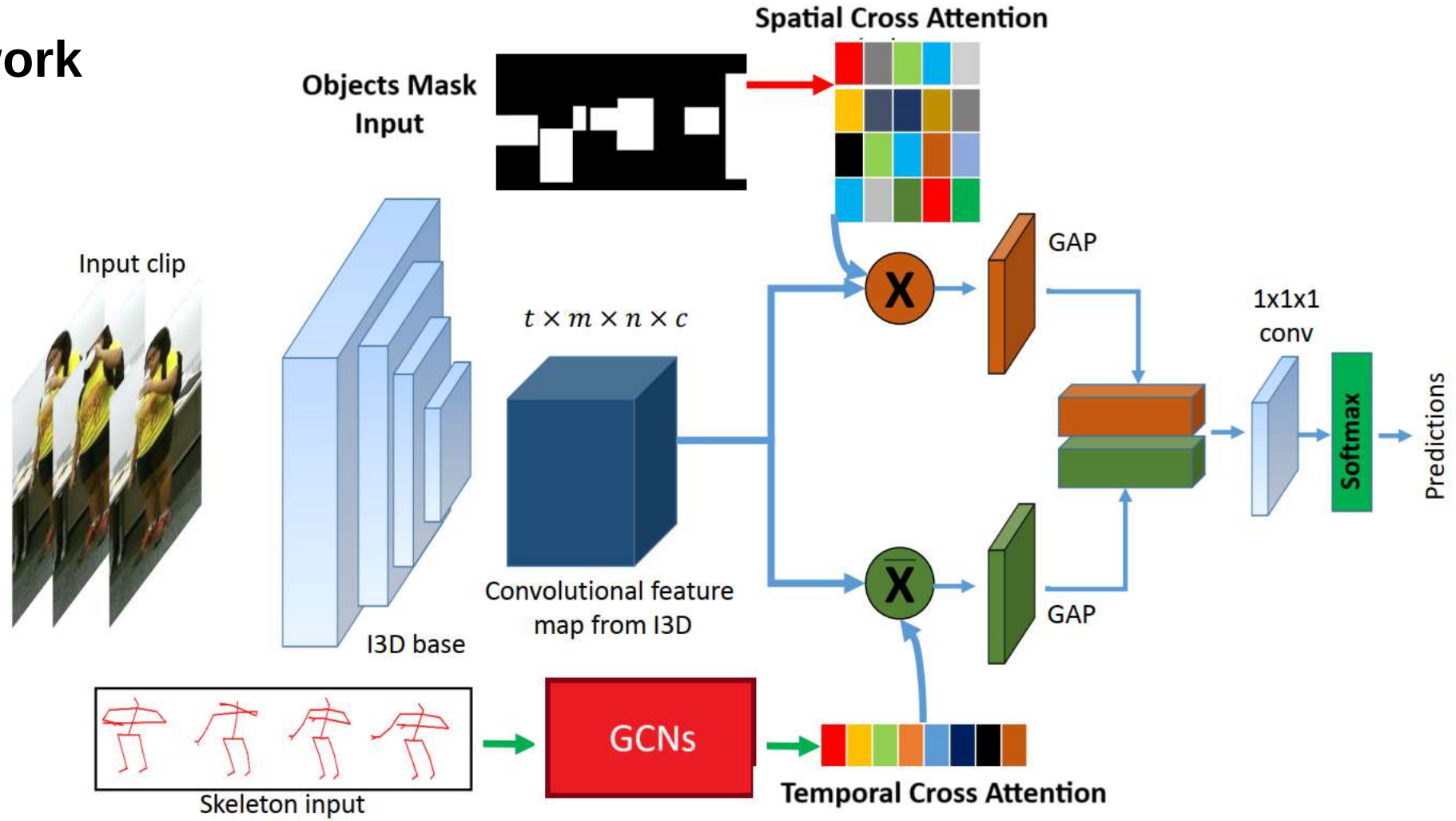
Object Detector
YOLO v8



visuAAL

Bringing All Together

Proposed Network

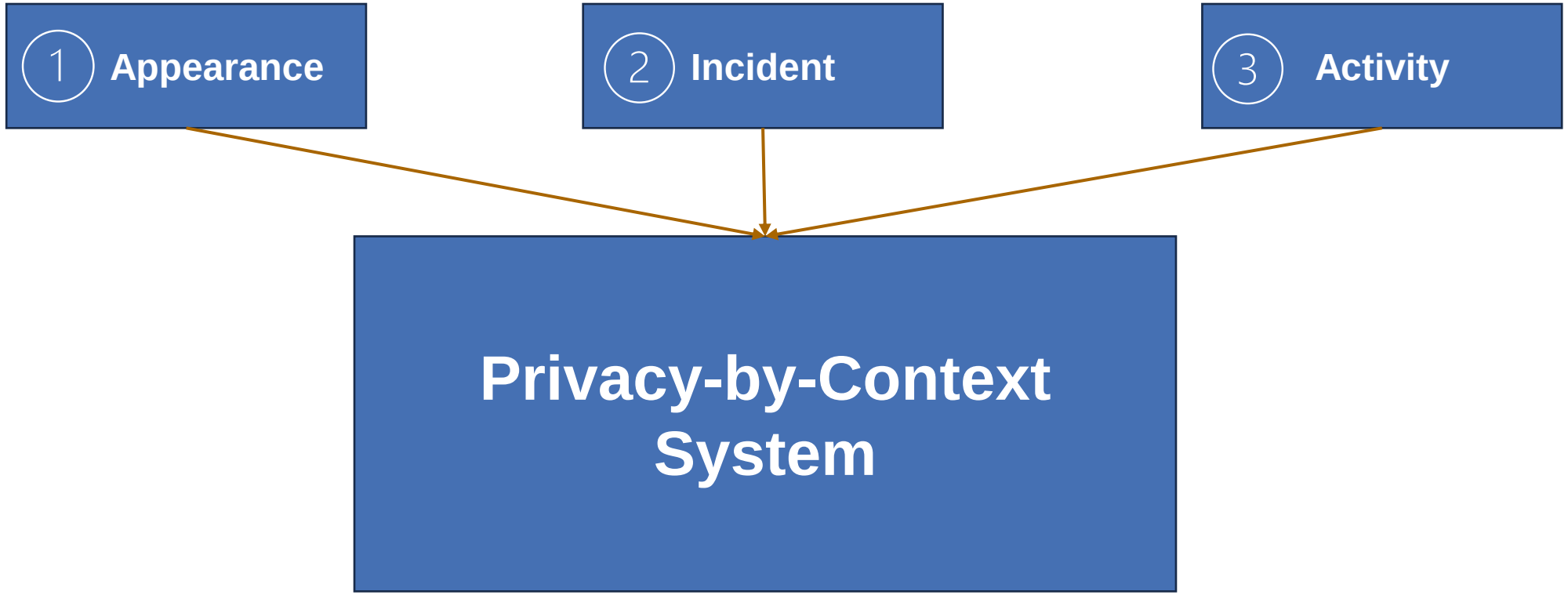


Evaluation & Comparison

Methods	RGB	Skeleton	Pre-train	CS	CV1	CV2
I3D	x		K400	53.4	34.9	45.1
AssembleNet++	x		K400	63.6	-	-
NPL	x		K400	-	39.6	54.6
Separable STA	x	x	K400	54.2	35.2	50.3
VPN	x	x	K400	60.8	43.8	53.5
LSTM		x	-	42.5	13.4	17.2
ST-GCN		x	-	53.8	15.5	51.1
2 s-AGCN		x	-	60.9	22.5	53.5
MS-G3D Net		x	-	61.1	17.5	59.4
UNIK		x	-	63.1	22.9	61.2
ViA		x	-	64	35.6	65.4
VIT Spatial + GCN Temporal (I3D backbone)	**x**	**x**	**Smarthome**	**66.14**	**41.12**	**61.2**
Without Temporal Pose	**x**	**x**	**Smarthome**	**63.58**	**37**	**55**

Top-view Activity Recognition

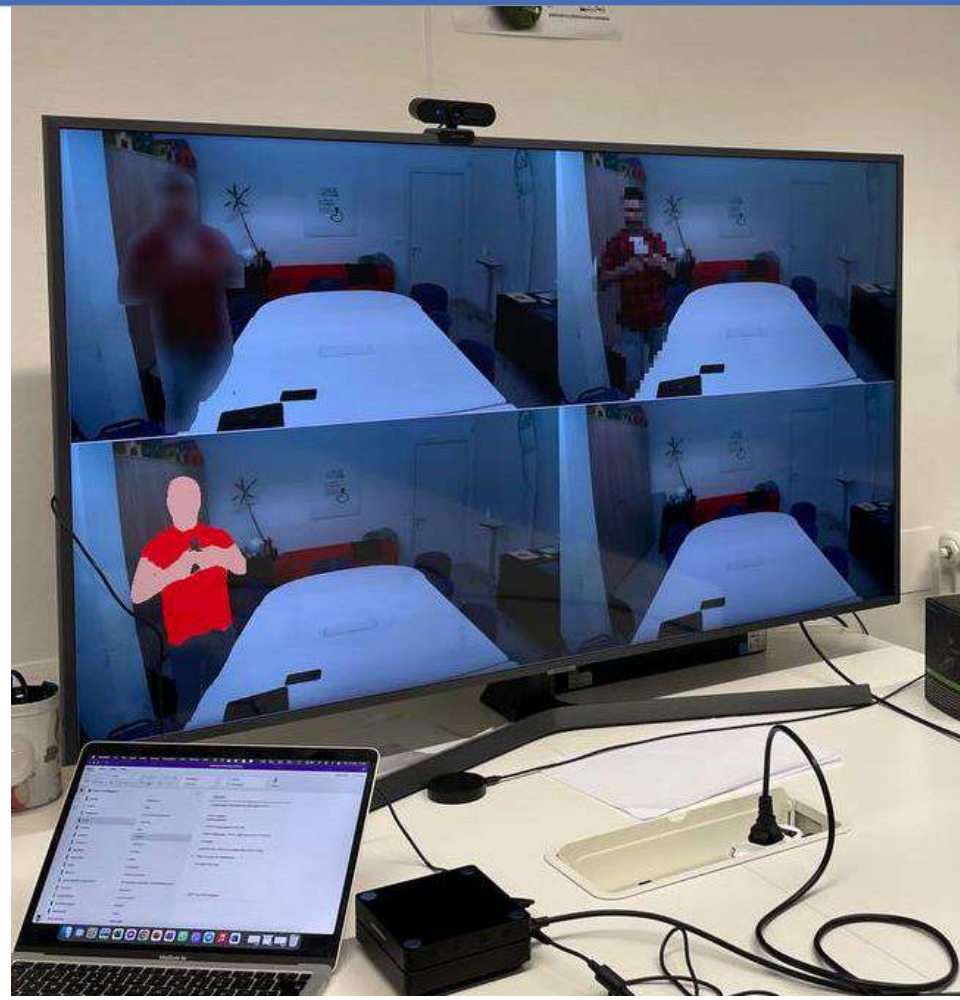
- 29 labeled classes of Activity
- More than 850 videos
- RGB and pose data
- Evaluating SOTA methods

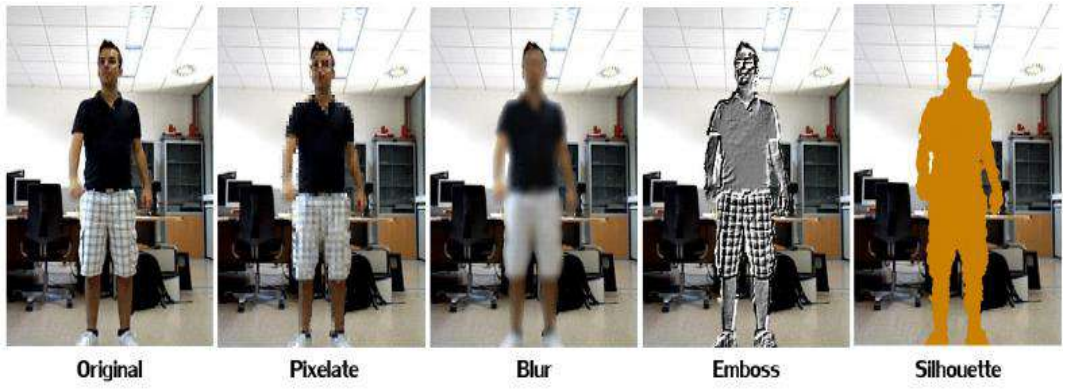

1 Appearance
2 Incident
3 Activity
Privacy-by-Context System

Privacy Filters

Visualization Demo:

- Real-time illustration of different visualization
- Deployed on Jetson Xavier NX





A Step Closer to Privacy by Context System

- Activity Detection ✓
- Incident Detection ✓
- Appearance Detection ✓
- Visualization Filters ✓

Thank you!

Kooshan Hashemifard

University of Alicante

k.hashemifard@ua.es

Challenges

Evaluation Protocols

- **Cross-Subject (CS):** same camera view, performed by different person
- **Cross-View (CV):** different camera view, performed by the same person

Background & Tools

Human Activity Recognition (HAR)

Action recognition

involves identifying and classifying specific actions or movements.

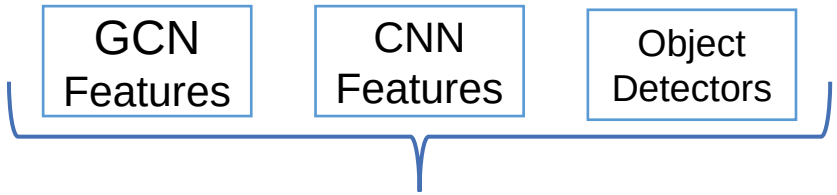
- Examples: walking, jumping, specific gestures like waving or handshaking
- Applications: human-computer interaction, sports analysis, and robotics

Activity recognition

is a broader concept that involves identifying and understanding a sequence of actions or interactions over a certain period to infer the overall activity or behavior.

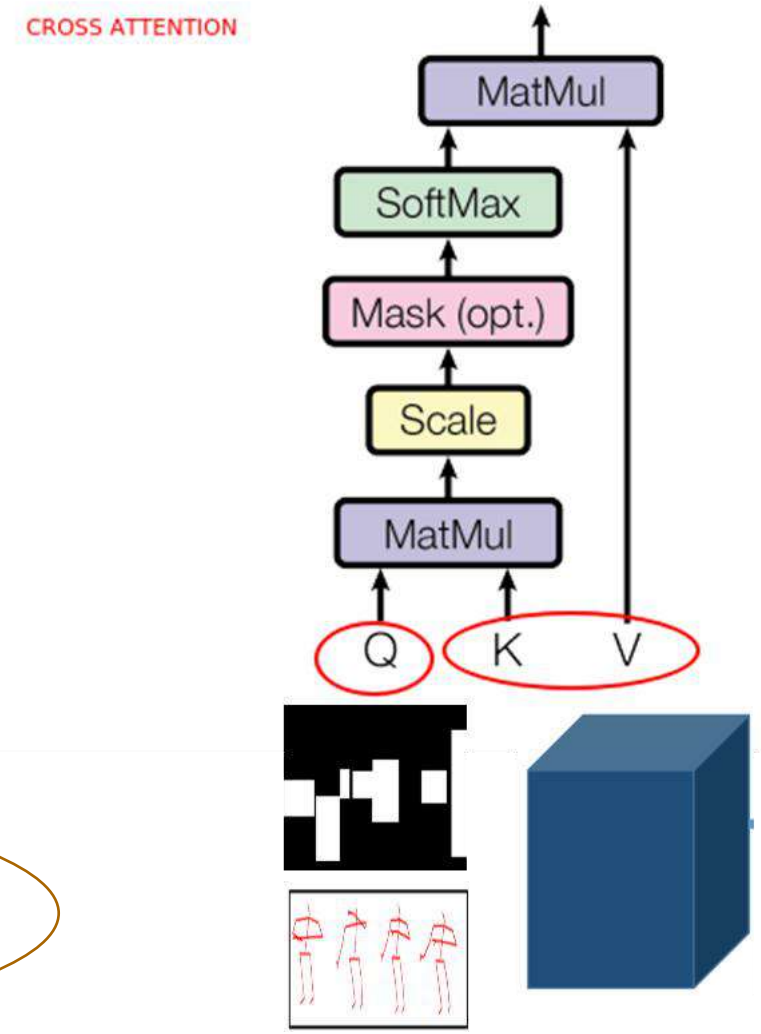
- Examples: cooking, playing basketball, working at a desk
- Applications: healthcare monitoring, smart homes, and context-aware computing

Approach



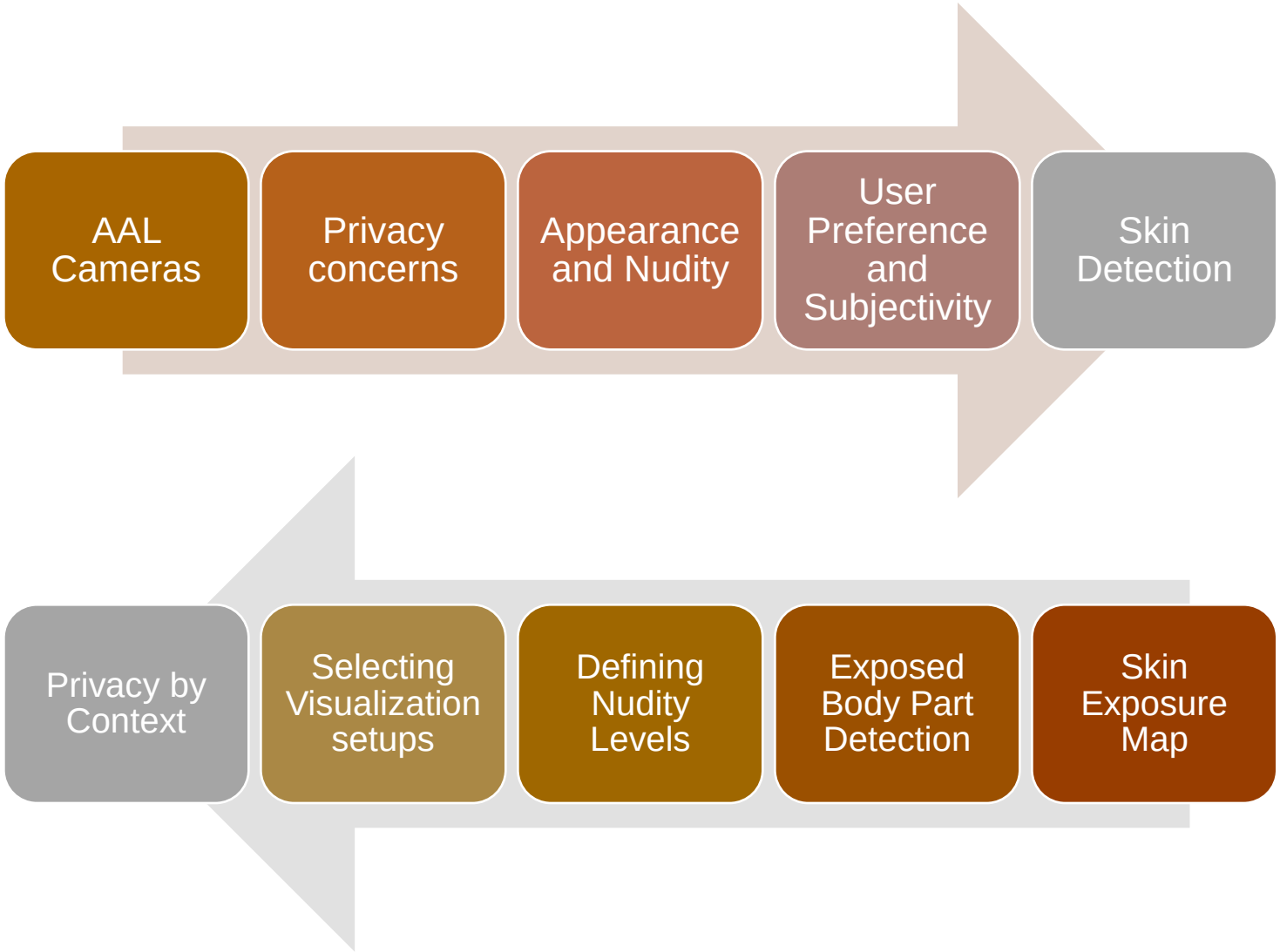
Multi-Source Information Integration Methods:

- **Feature Concatenation**: Combining features by concatenating them along a specific dimension.
- **Feature Fusion**: Integrating features through a fusion process, which can involve mathematical operations like addition, averaging, etc.
- **Cross Attention:** Focusing on relevant parts of input data through attention mechanisms, allowing the model to emphasize important information during merging.



visuAAL

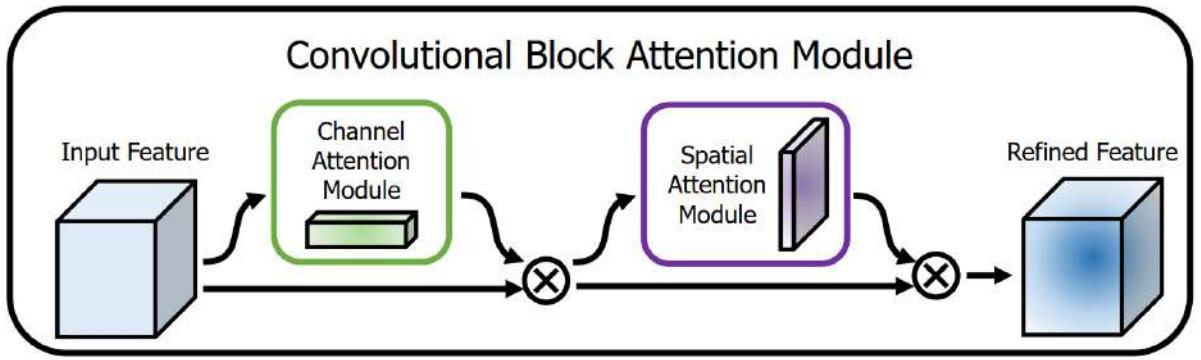
Adaptive Privacy by Appearance

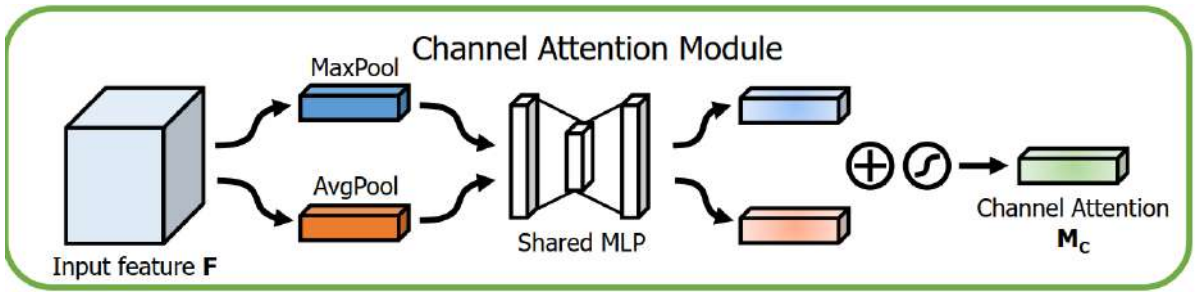


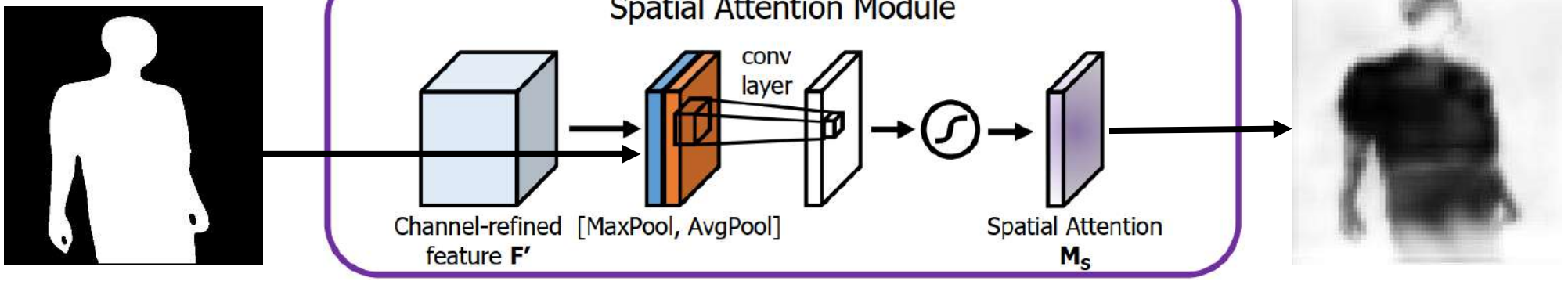
Attention Modules

Body Attention

$$\mathbf{F}' = \mathbf{M_c}(\mathbf{F}) \otimes \mathbf{F},$$
$$\mathbf{F}'' = \mathbf{M_s}(\mathbf{F}') \otimes \mathbf{F}',$$

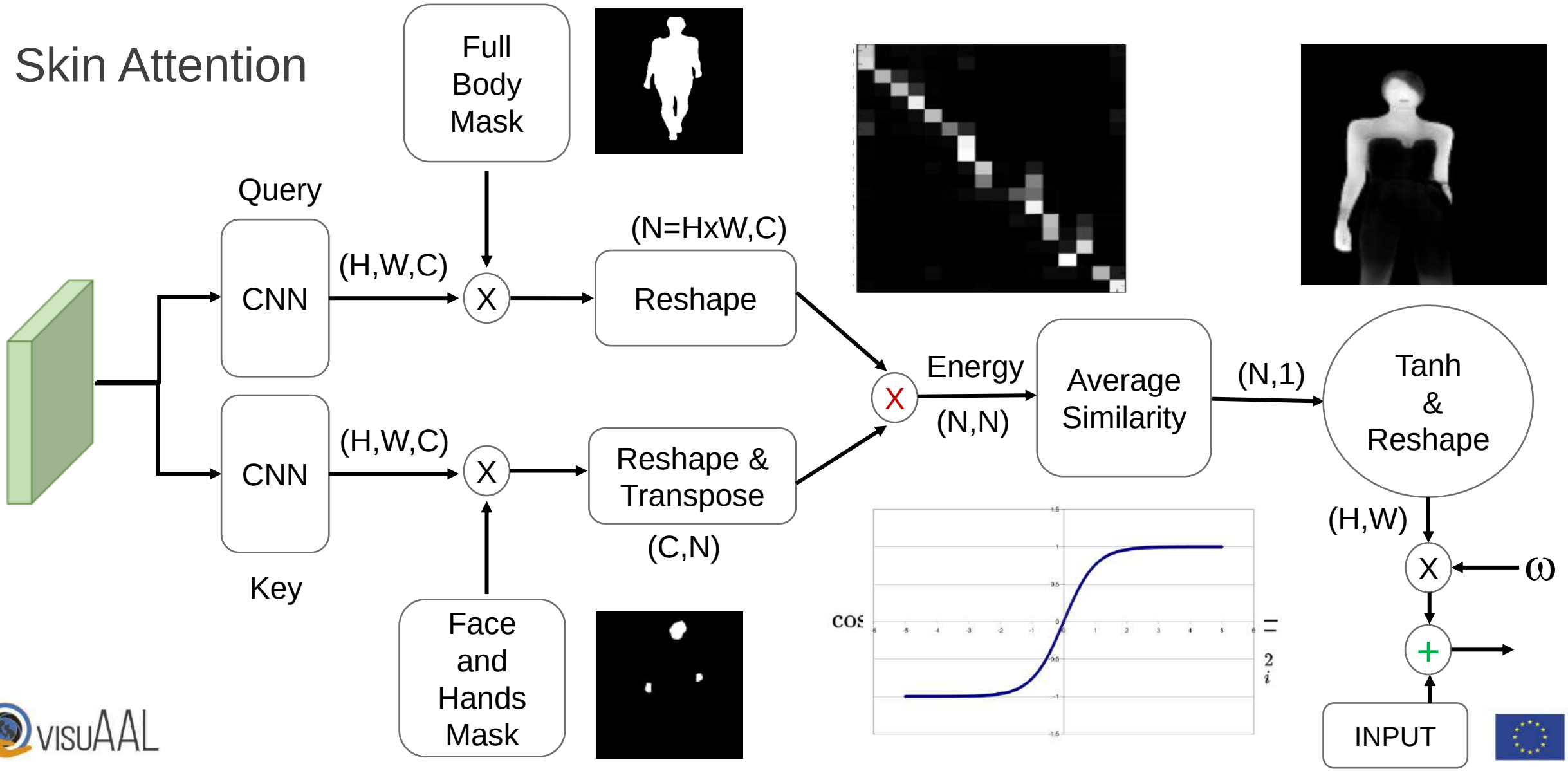






Attention Modules

Skin Attention



visuAAL