LF driven saliency detection

1 April, 2025

STUDY ON VISUAL PERCEPTION OF LF CONTENT

Five rendering methods tested based on 3 scenarios :

- **All-in-focus** (everything sharp).
- **Region-in-focus** (only certain areas at specific depths are sharp).
- **Focal-sweep** (focus transitions over time from foreground to background or vice versa).

Analysis:

- Qualitative: Heatmaps were generated to observe differences in gaze patterns across different rendering types.
- Quantitative: Entropy analysis of fixation maps assessed how dispersed or concentrated viewers' attention was.

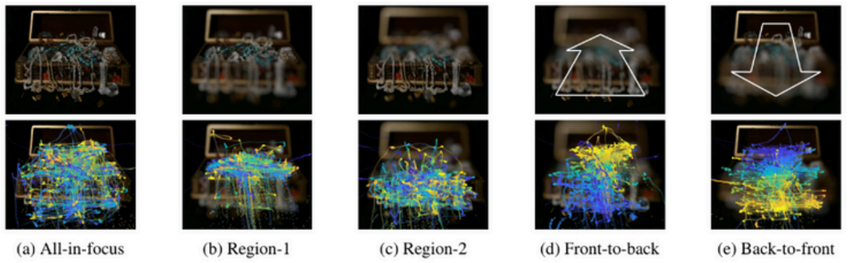


Figure 1: Scanpaths of Treasure light field renderings. Each continuous chain represents a participant, and colour mapping shows the passage of time which starts with blue. Yellow is the most recent time instant.

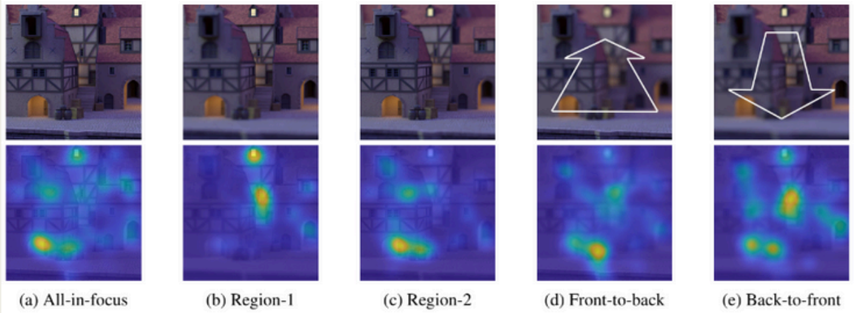


Figure 2: Medieval light field rendered three ways all-in-focus, front-to-back and back-to-front overlayed with a heatmap, generated from all participants and averaged over entire video.

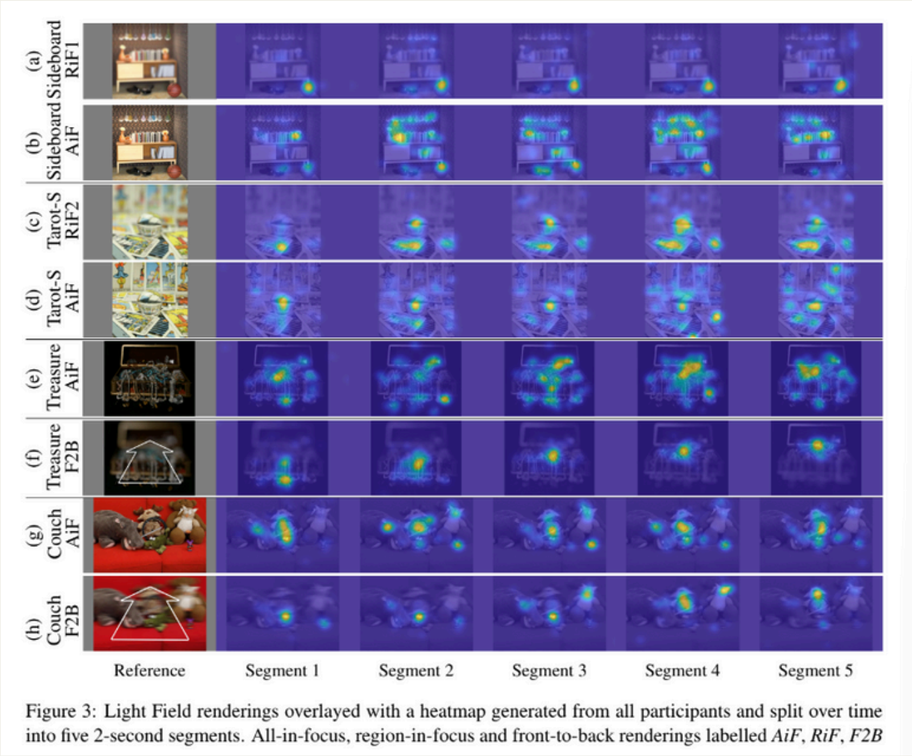


Figure 3: Light Field renderings overlayed with a heatmap generated from all participants and split over time into five 2-second segments. All-in-focus, region-in-focus and front-to-back renderings labelled *AiF*, *RiF*, *F2B*

✅ADVANTAGES

- **Novelty** (first analysis of how focus variation affects attention in LF)
- **Diverse dataset** (multiple sources → generalizability)
- **Realistic setup** (advanced eye tracker for accurate collection)

❌LIMITATIONS

- **Small sample size** (only 21 participants)
- **Fixed viewing conditions** (conducted on a 2D screen → not fully represent attention in LF displays)
- **Limited rendering types** (focuses on refocusing effects → not explore other LF interactions)
- **Entropy analysis constraints** (entropy-based analysis → not fully model complex gaze behaviors)

2014 - SALIENCY DETECTION ON LIGHT FIELD

- **Focusness** → how in-focus a region is (DCT + harmonic variance analysis)
- **Depth** Estimation → determines background (background prior) and foreground likelihood (based on focusness distributions)
- **Objectness** → ensures detected salient regions correspond to whole objects rather than fragmented parts

- Combine focusness measure + location prior → extract the bg & fg salient candidates.
- combine bg prior + contrast-based saliency detection → detecting saliency candidates
- Use objectness → weights for combining the saliency candidates from the all-focus img & from focal stack as final saliency map

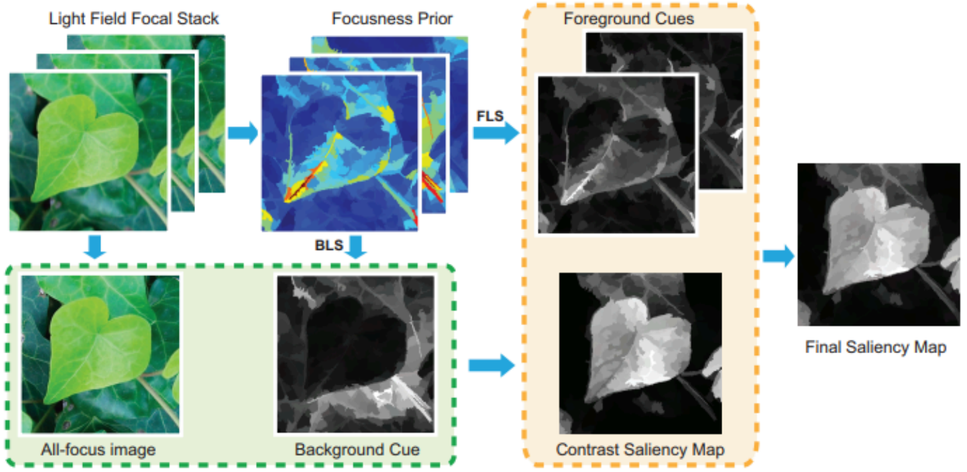
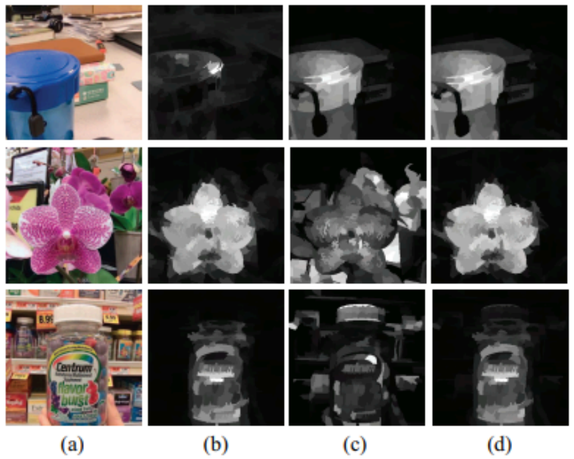


Figure 3. Processing pipeline of our saliency detection algorithm for light fields.



(a) (b) (c) (d)

Figure 7. Saliency detection using different cues. (a) All-focus images; (b) Detected saliency using focusness cues; (c) Detected saliency using color contrast. (d) Saliency results by combining (b) and (c).

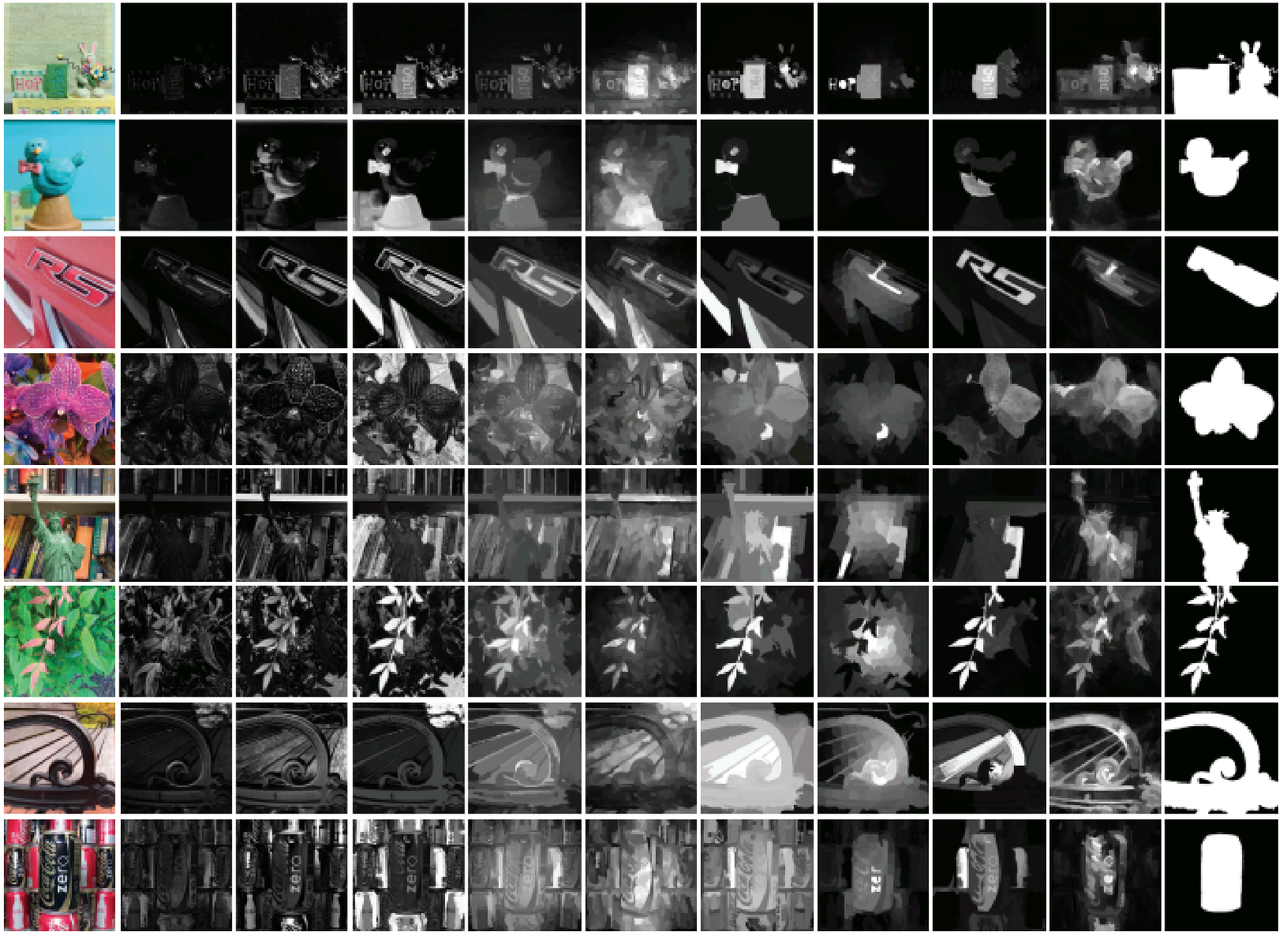


Figure 5. Visual Comparisons of different saliency detection algorithms vs. ours on our light field dataset.

✅ADVANTAGES

- Handles **challenging scenarios** (similar foreground/background, cluttered backgrounds, and occlusions)
- Incorporates **depth information** (unlike conventional 2D saliency methods that rely solely on color and texture)
- **More robust object segmentation** (using objectness cues rather than pixel-wise contrast)

❌LIMITATIONS

- Limited by **LF camera quality** (Lytro → small FoV & lower resolution wrt traditional cameras)
- **Depth estimation challenges** (LFs → depth information, accuracy → depends on scene composition)
- **Ignore explicit use of depth data** associated with salient regions (focusness & objectness to select foreground saliency candidates)

2015-SALIENCY DETECTION W/ A DEEPER INVESTIGATION OF LF

- Computes **saliency contrast** based on LF depth and color extracted from the all-focus image.
- **L2-norm metric** to measure contrast → improve separation between foreground/background.
- **SLIC algorithm** segments the all-focus image into super-pixels → ensuring:
  - Edge consistency preservation
  - Compact and uniform regions
- focal slices at different depth levels → compute **focusness maps**
  - Selects the background slice (evaluate focus distributions)
- **Saliency optimization algorithm** → cleaner and more precise detection

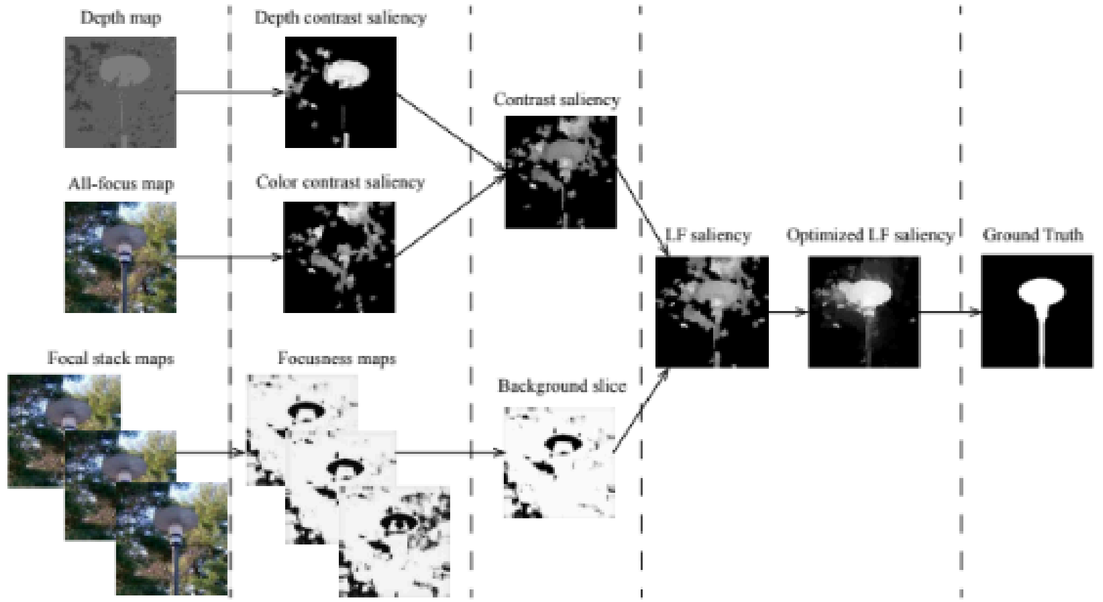


Figure 2: Pipeline of our proposed approach for light field saliency detection.

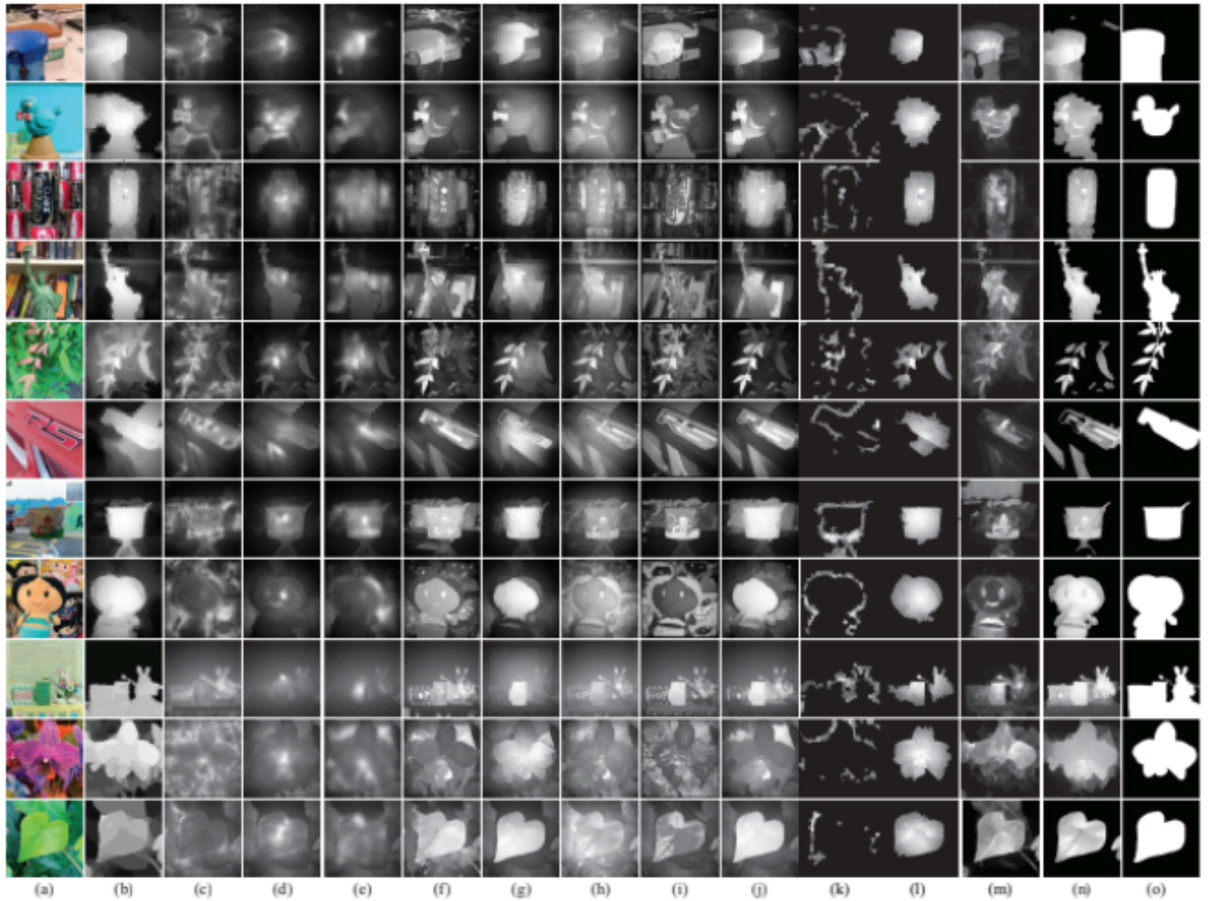


Figure 7: Visual comparisons of our approach and 2D/3D extended methods. (a) all-focus image; (b) depth map; (c) CNTX_D; (d) CovSal_D; (e) Tavakoli_D; (f) GS_D; (g) GBMR_D; (h)SF_D; (i) TD_D; (j) wCtr*_D; (k) DVS_Bg; (l) ACSD_Bg; (m) LFS; (n) Ours; (o) GT.

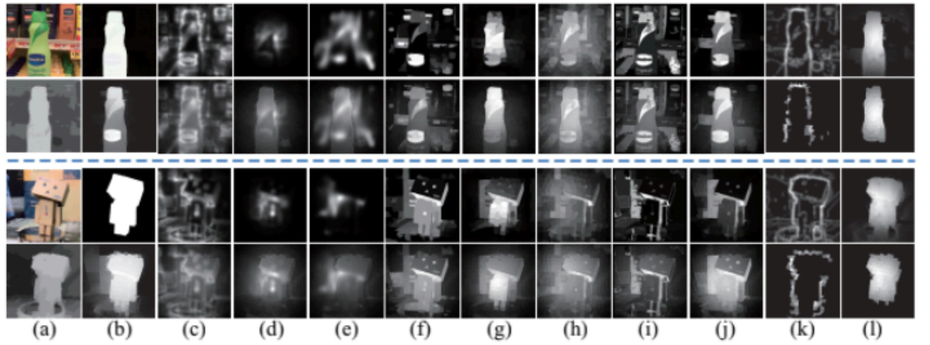


Figure 6: Visual comparisons of 10 state-of-the-art 2D/3D saliency detection models and their light field-extended versions for two examples. (a) all-focus image (Top) and depth map (Bottom); (b) GT (Top) and ours (Bottom); (c) CNTX; (d) CovSal; (e) Tavakoli; (f) GS; (g) GBMR; (h)SF; (i)TD; (j) wCtr*; (k) DVS; (l) ACSD.

✅ADVANTAGES

- **Improved saliency detection** compared to methods based only on color or depth
- Combines multiple visual cues (color, depth, and focusness) for **more accurate segmentation**
- **Enhances existing 2D/3D saliency models** by incorporating LF depth contrast and focusness-based background priors
- **Facilitates saliency estimation** by computing a background prior (on the focusness map of a focal slice) & local prior (object-background separation) even for challenging scenarios.
- Comparison with 2014 paper:
  - This method achieves a 4-7% **performance improvement** across multiple evaluation metrics.
  - **Higher accuracy**
  - **Better foreground object detection**

❌LIMITATIONS

- Current **LF cameras** → limited depth resolution
- **Higher computational cost** compared to traditional 2D saliency methods
- Current LF cameras → limited depth relief (affects performance)

2017 - SALIENCY DETECTION ON LF

- **Color** Cue → Derived from the all-in-focus image
- **Depth** Cue → Extracted from the depth map
- **Flow** Cue → Computed from the focal stack and multiple viewpoints (focusing flow and viewing flow)
- **Location** Prior → A multiplicative weighting factor enhancing saliency maps
- A **random-search-based strategy** → assigns optimal weights to the cues
- **Graph regularization** → ensures that adjacent superpixels take similar saliency values (refining the final saliency map)

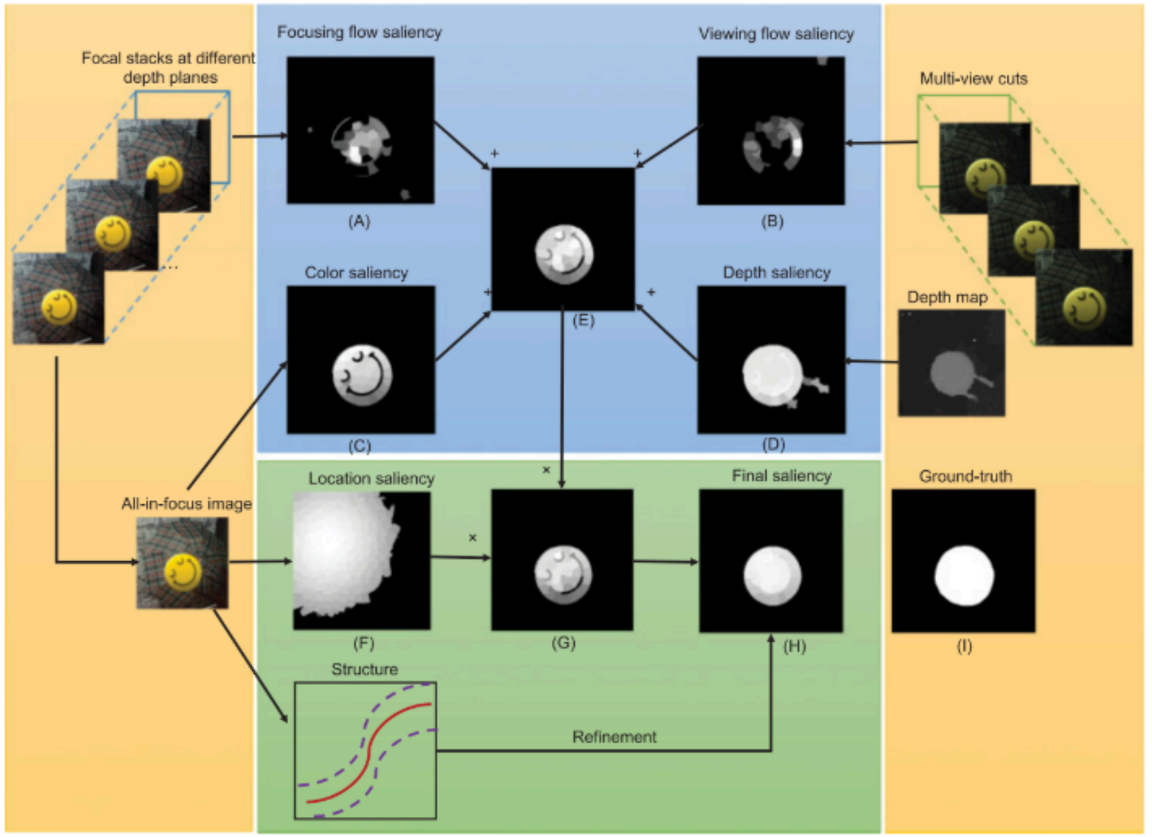


Fig. 1. An illustration of our multi-cue light-field saliency detection model. First, we generate the all-in-focus image, depth map, focal stacks, and multi-view cuts (images) from the corresponding light field captured by a Lytro camera. Light-field flow is estimated from focal slices and multiple viewpoints as a complement to the other information. We measure saliency by computing weighted local contrast, and the prior saliency map is refined by employing a structure cue to obtain the final saliency map.

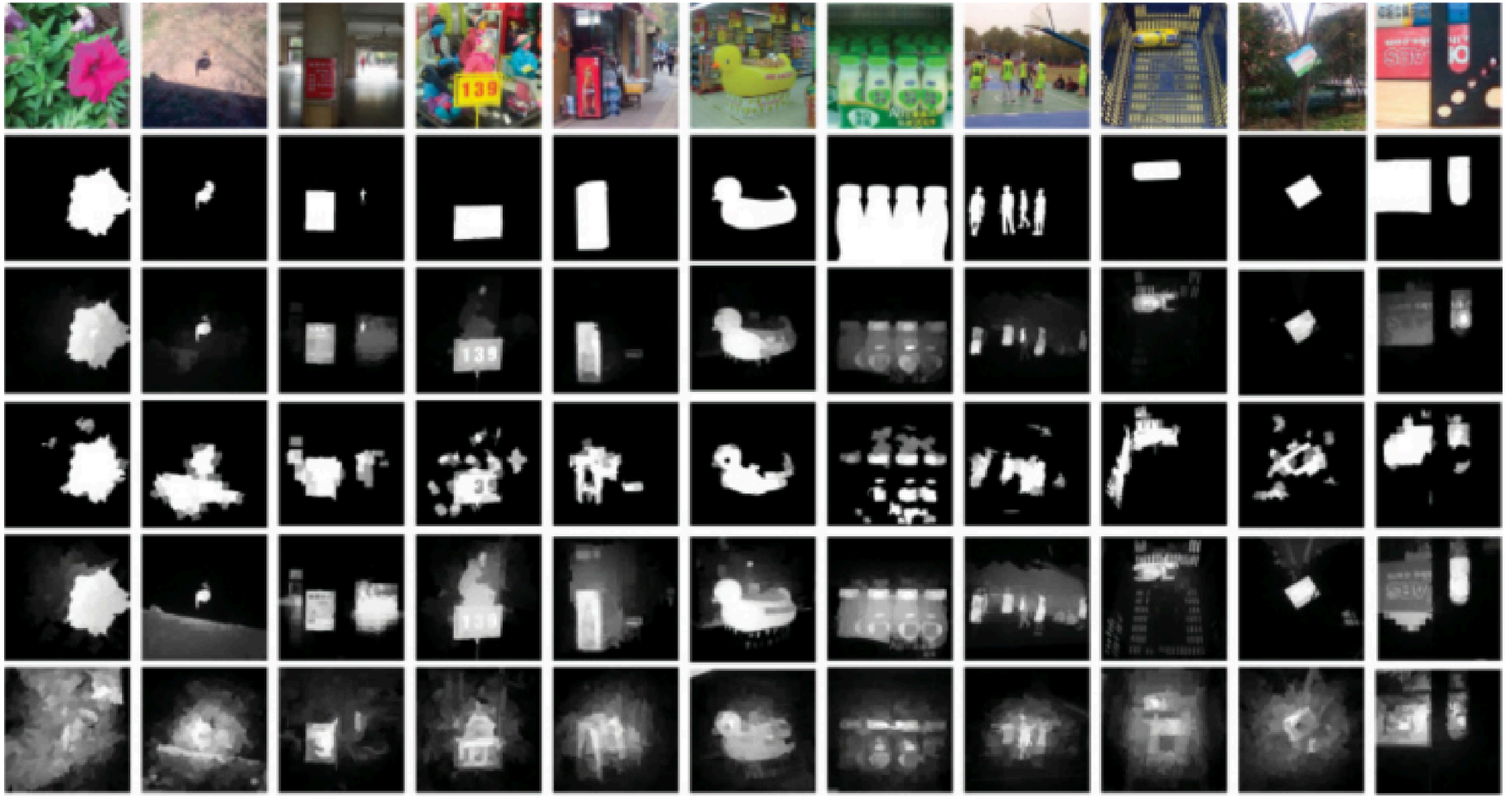


Fig. 6. Saliency detection results of different methods on the HFUT-Lytro dataset. From top to bottom: all-in-focus images, ground-truth maps, and saliency maps obtained by our approach, WSC [29], DILF [73], and LFS [30].

✅ADVANTAGES

- **Precision and F-measure Improved** by around 6% over the second-best method (DILF)
- **Lower MAE** wrt state-of-the-art methods
- Combine multiple cues (color, depth, flow, location) → **improve accuracy**
- **Faster** than WSC and LFS
- **Handles challenging cases** (cluttered backgrounds, varying lighting, and occluded objects)
- It outperforms state-of-the-art methods in **precision** and **robustness**

❌LIMITATIONS

- Slightly **higher computation time** wrt DILF.
- **Depth Cue Sensitivity** Less effective when depth estimation is inaccurate.
- No Training-Based Optimization: **Performance could** further **improve** with a learning-based approach.

2018 - PDNet

Dual-stream network designed for salient object detection in RGB-D images

Final saliency map → combine features from:

1. **Master Network** (RGB-based)
   - A fully convolutional encoder-decoder (VGG-16/VGG-19 as pre-trained encoder on RGB-based saliency datasets)
   - Feature extractor: Convolution layers → transform input into hierarchical feature representations
   - Shape restorer: Deconvolution layers → recover object details from the background
2. **Subsidiary Network** (Depth-based)
   - Extracts depth features independently (not treated depth as a fourth channel)
   - Encodes depth cues and integrates them into the master network via convolution layers.

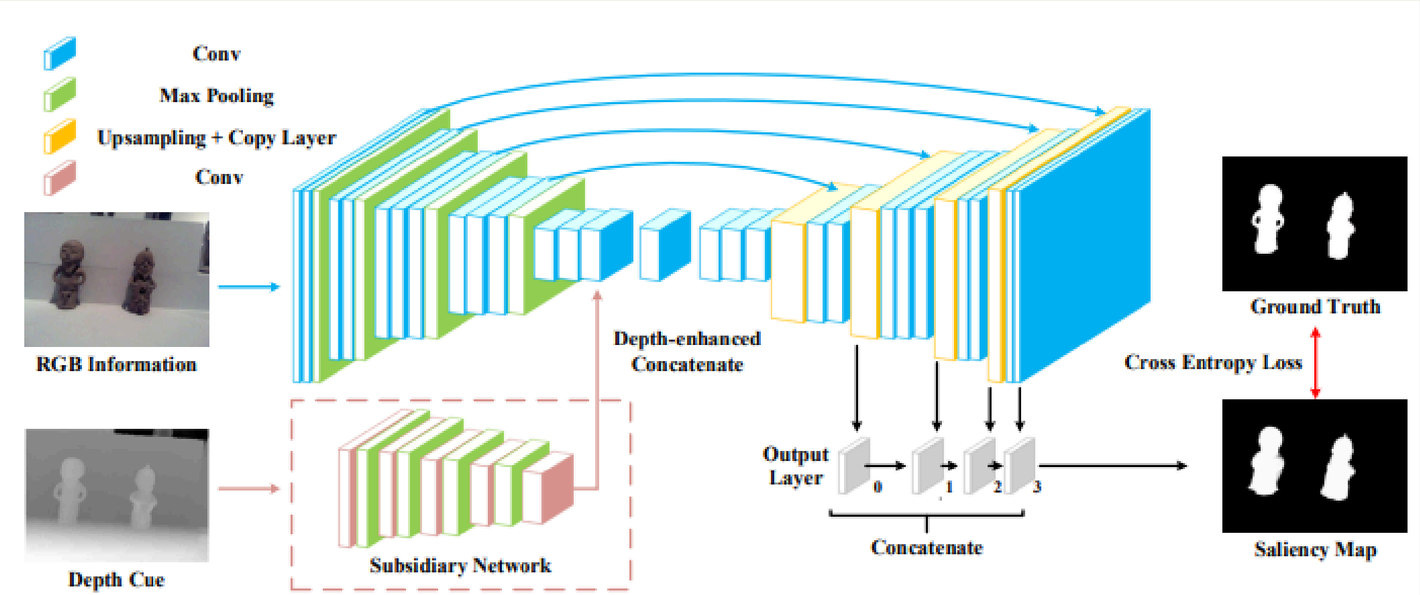


Fig. 1. The framework of the PDNet.

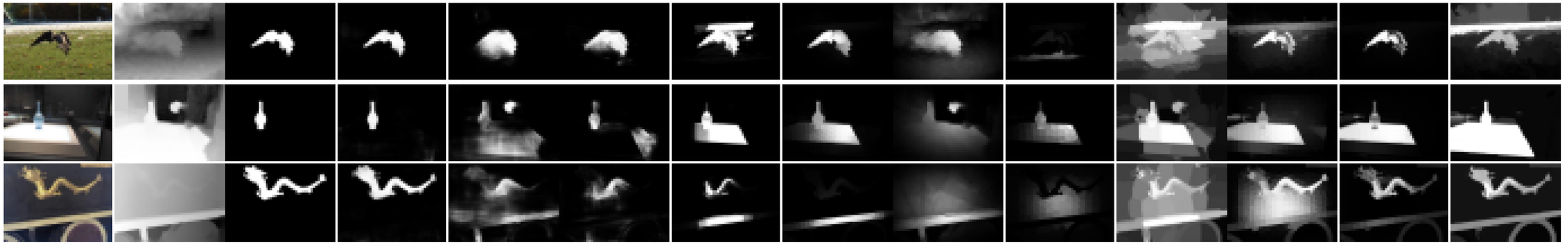


Image
Depth
Ground Truth
PDNet
CTMF17
DF17
TPF17
SE16
ACSD15
NLPR14
DES14
BSCA15
LIP15
HS16

✅ADVANTAGES

- Depth information is processed independently → **better feature extraction**
- Feature fusion → **improve object boundary accuracy**
- **Better saliency detection** in **complex scenes** wrt heuristic saliency methods.

❌LIMITATIONS

- Requires two networks → **increasing computational cost**
- **Dependent on depth quality** → low-resolution depth maps reduce effectiveness.
- **Not real-time** due to the complex fusion mechanism.

2019 - AFNet

Two-stream CNN → processes **RGB + depth images separately** → individual saliency maps → fusion module

1. **Two-Stream CNNs**
   - Each modality (RGB, Depth) → its own CNN (based on VGG-16)
2. **Saliency Fusion Module**
   - Learns a switch map → adaptively combine the RGB and depth saliency maps
3. **Loss Functions**
   - Edge-Preserving Loss → enhances object boundary sharpness and spatial coherence

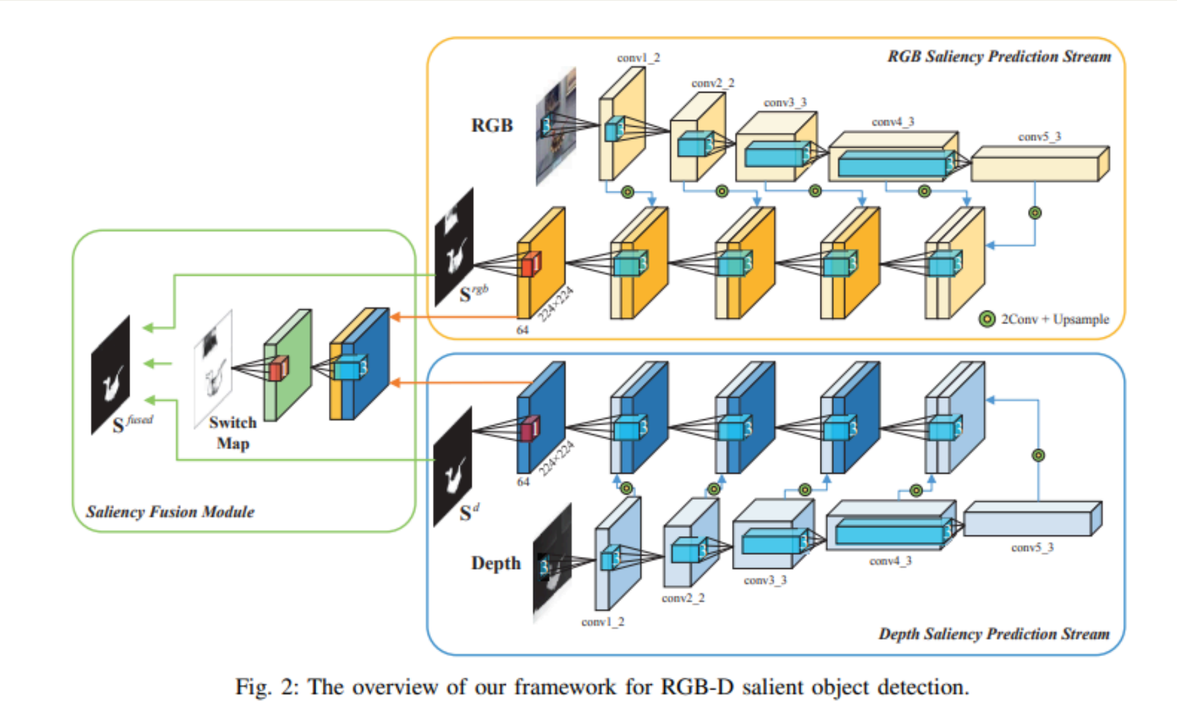


Fig. 2: The overview of our framework for RGB-D salient object detection.

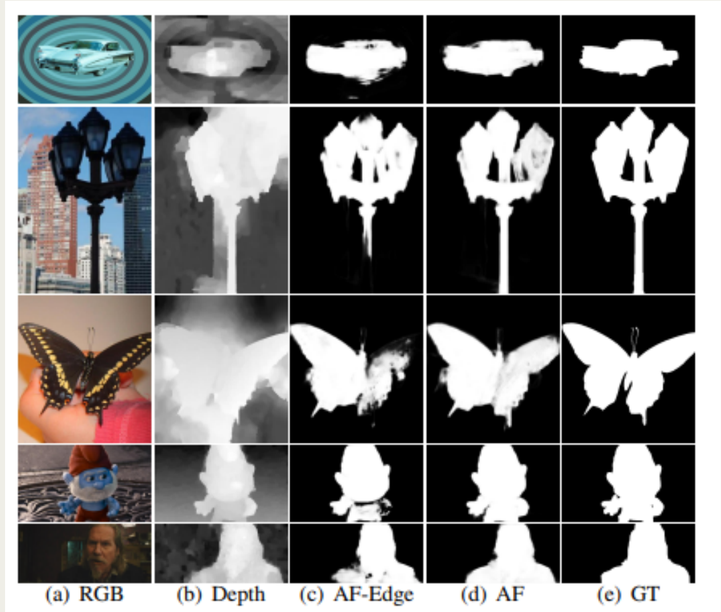


Fig. 3: Comparison of predictions with and without the edge-preserving loss.

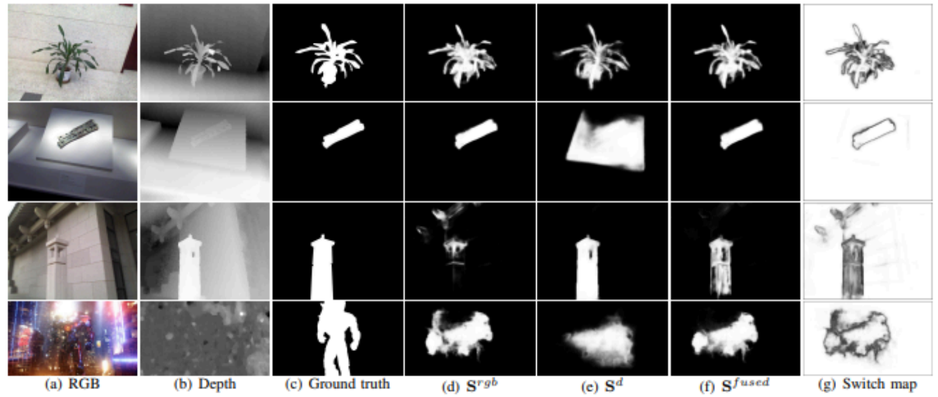


Fig. 1: Typical scenarios in RGB-D saliency object detection. Here, $\mathbf{S}^{rgb}$ denotes the result obtained by our RGB saliency prediction stream, $\mathbf{S}^{d}$ is the result from our depth saliency prediction stream, and $\mathbf{S}^{fused}$ is the final saliency detection result. Switch map is the map learned in our network for adaptive fusion.

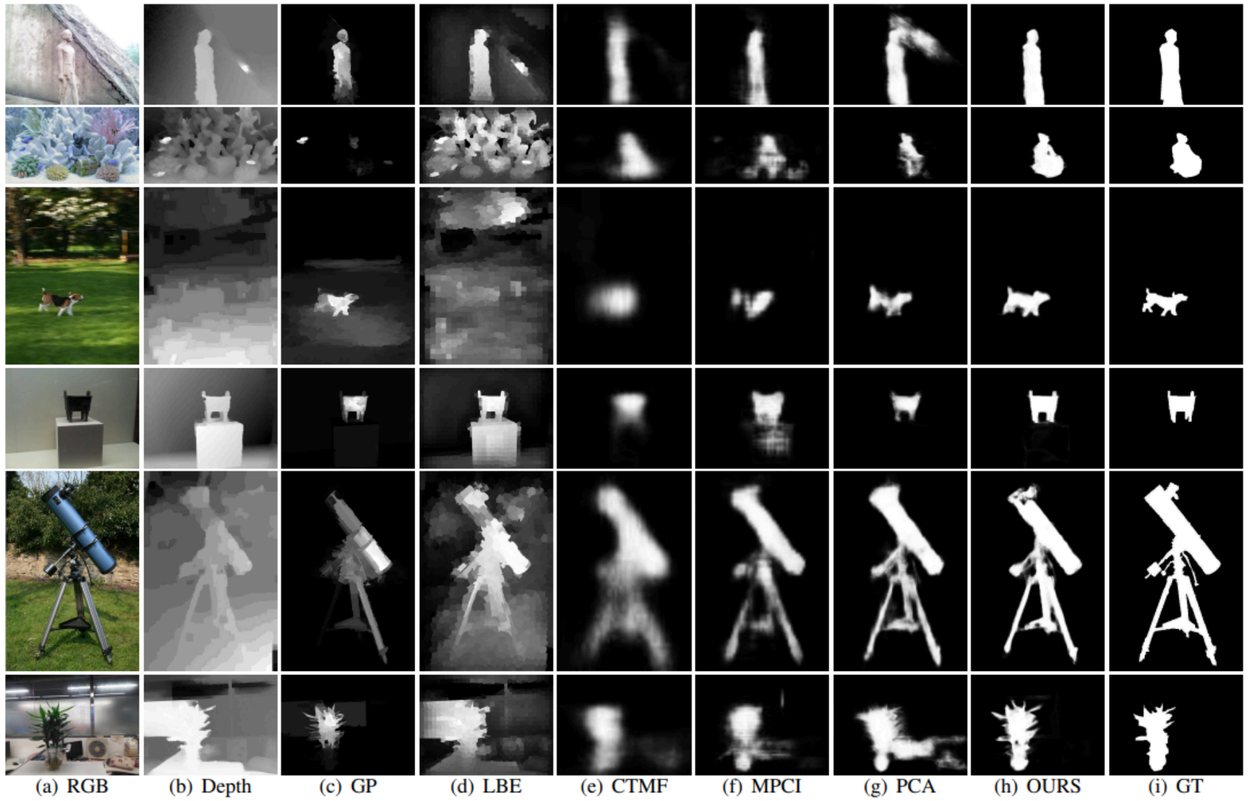


Fig. 6: Visual comparison of saliency maps.

✅ADVANTAGES

- **Superior Performance** → outperforms state-of-the-art methods across all datasets.
- **Better Boundary Preservation** → Reduces blur and enhances spatial coherence.
- **Adaptive Fusion** → The switch map dynamically selects the best modality instead of simple concatenation or fixed fusion.
- Edge Preservation → The loss function ensures **sharper boundaries** and **improved object separation**.
- Consistently **better accuracy** across multiple datasets

❌LIMITATIONS

- **Fails in challenging cases** → if an object is indistinguishable (RGB and depth) → detection is weak
- **Dependency on depth quality: performance is affected by low-quality or noisy depth data**
- **The two-stream architecture a**nd fusion module **increase computational demand**

2019 - DEEP LEARNING FOR LF SALIENCY DETECTION

- **RCNN** (Recurrent CNN) + **Attention Mechanism**
  - Uses convolutional features from each slice of the focal stack.
  - ConvLSTM → adaptively integrates information across slices.
  - Focuses on "good" slices (clear foreground, blurred background).
- **Adversarial Training** for robustness and improving model generalization
  - Introduces adversarial examples
- **Dual-Stream CNN Architecture** that process
  a. focal stack
  b. all-focus images
  - Recurrent attention mechanism ensures optimal fusion of slices.

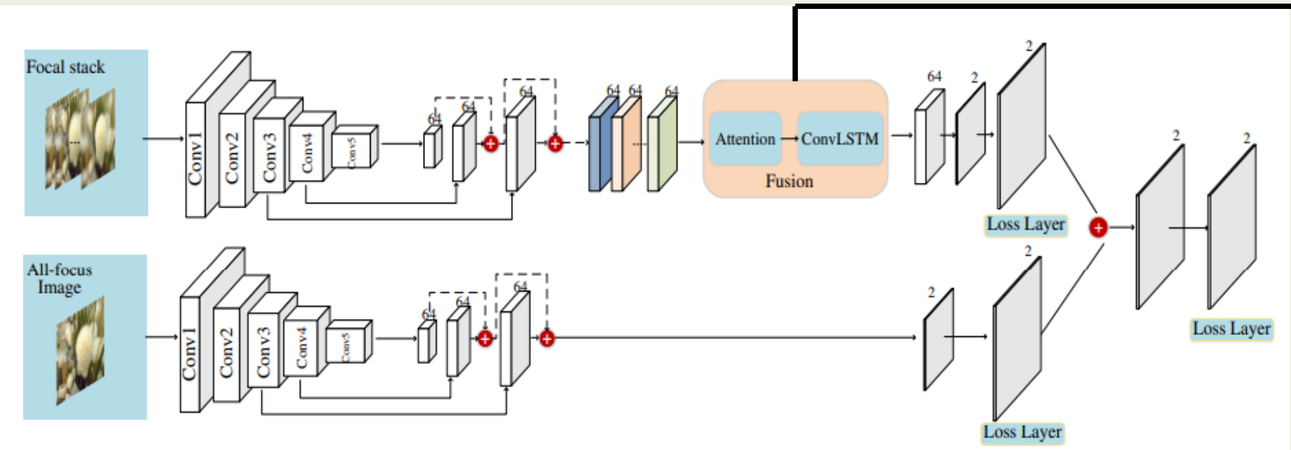


Figure 5. The overall two branch network based on the weighted late integration of the focal stack.

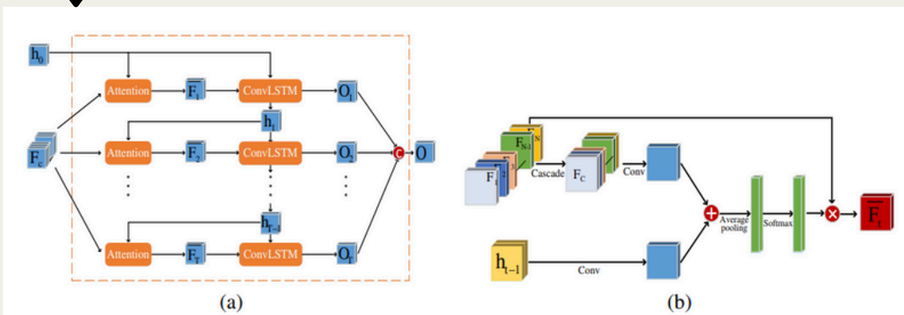


Figure 6. The structure of the recurrent attention network is shown in (a) and (b) denotes the attention subnet.

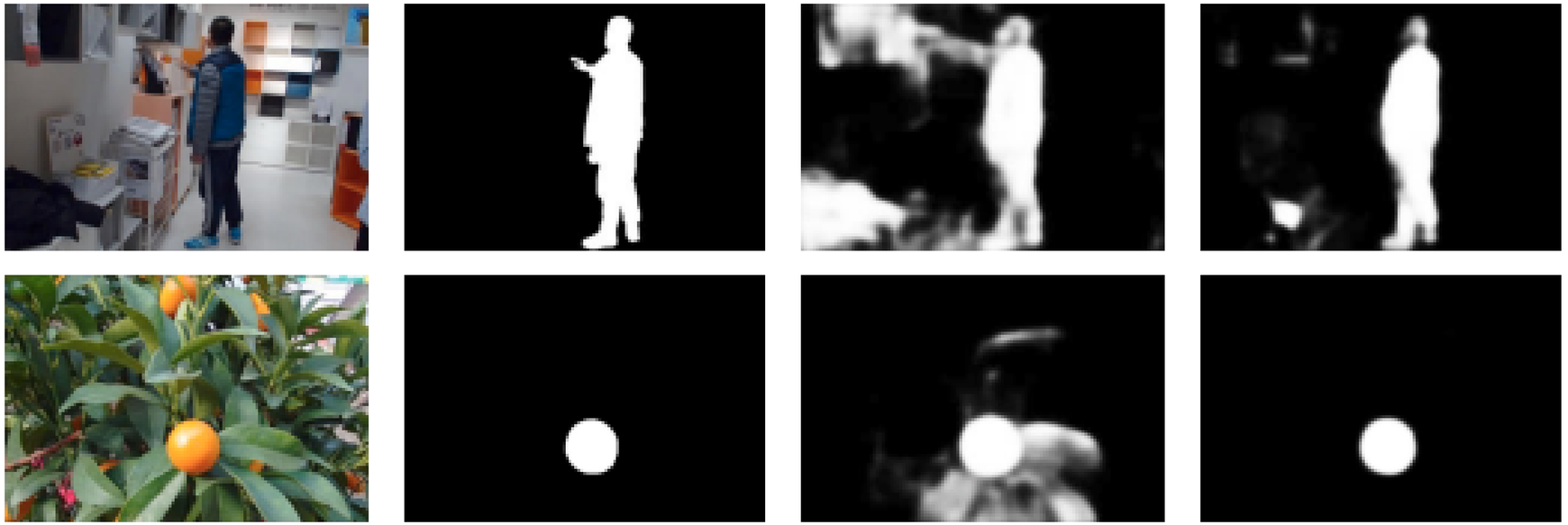


Figure 7. Examples with and without adversraial examples.

✅ADVANTAGES

- **Better Performance** wrt traditional methods
- **Robustness**: Adversarial training improves model resistance to perturbations.
- The **Recurrent Attention Network** effectively integrates focal slices → **improving saliency detection**
- ConvLSTM learns spatial dependencies across slices → **enhancing feature representation**

❌LIMITATIONS

- Recurrent architectures and adversarial training **increase training time**
- Requires **high-quality LF datasets** with per-pixel ground truth.
- Model may still be **sensitive to sophisticated attacks**

2019 - DEEP LF-DRIVEN SALIENCY DETECTION FROM A SINGLE VIEW

Two interconnected sub-tasks:

- **LF synthesis** → generates high-quality 4D LFs views from a single view using depth-based warping approach (capturing rich geometric details)
- **LF-driven saliency detection** → use synthesized LF to extract comprehensive saliency features & integrates them through a multi-view attention module (for precise saliency mapping)
- **MVAM** (Multi-view attention module) → multi-view saliency maps are warped back to the central view using depth maps.

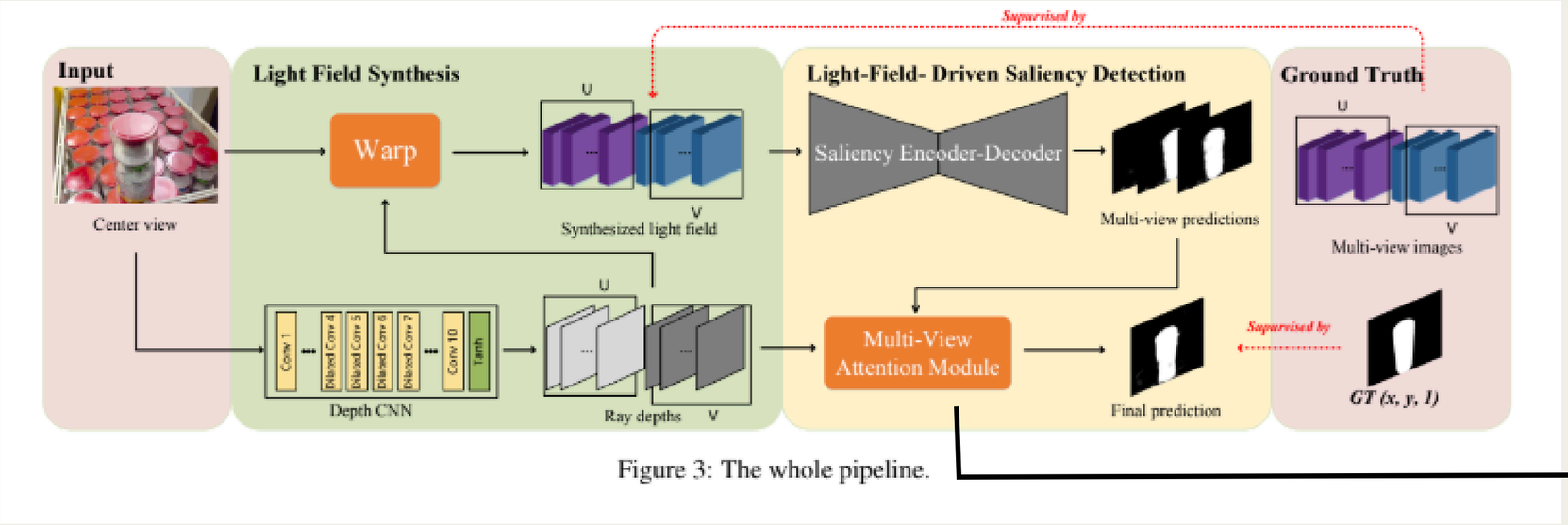


Figure 3: The whole pipeline.

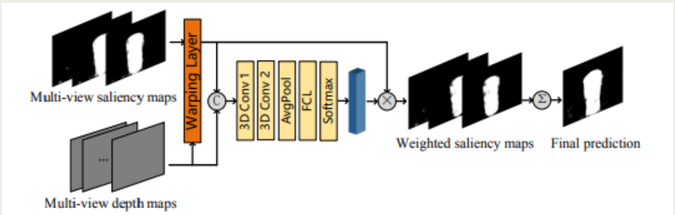


Figure 5: Details of multi-view attention module.

Figure 6: Visual comparsion of saliency maps on the proposed dataset.

✅ADVANTAGES

- MVAM aggregates multi-view saliency maps → **more accurate saliency predictions**
- **LF rendering + multi-view attention** → enhances saliency detection while maintaining efficiency by **reducing redundant LF information**
- **Outperforms** the state-of-the-art
- capable of capturing salient objects in **challenging scenes**

❌LIMITATIONS

- **computational demand**

➡ SOON sequel in 2022

2020 - LF SALIENCY DETECTION WITH DEEP CNN

- CNNs to model complex relationships between pixels and saliency LF images
- **MAC** (Model Angular Changes) **Blocks** → process micro-lens images effectively → capture angular variations inherent in LF data
- **VGG-16** for patch-based processing → divides LF images into patches & process them using this for patch-classification

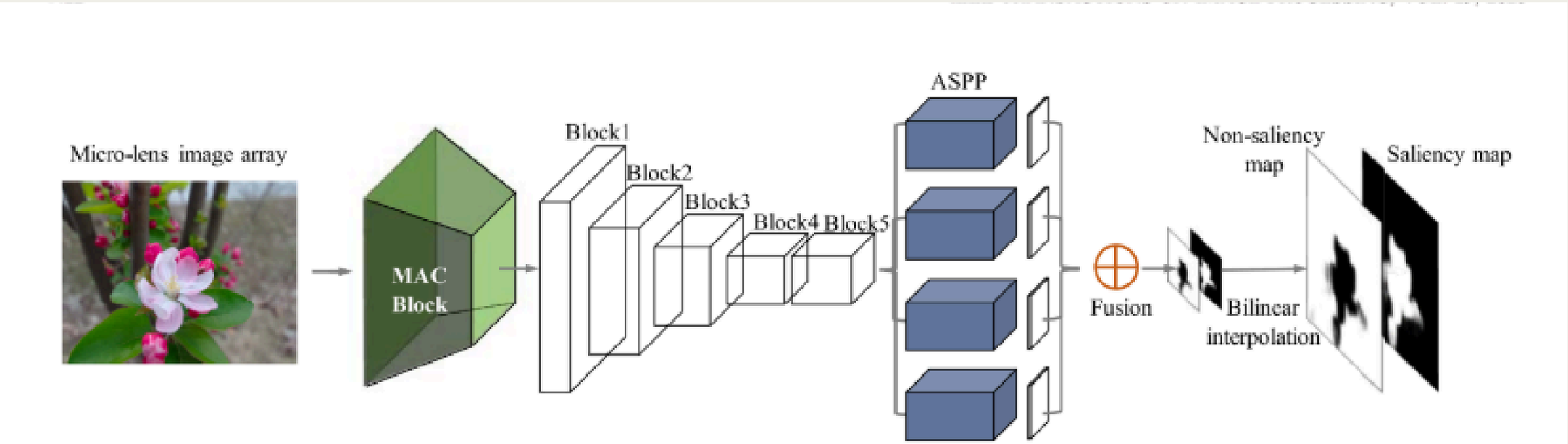


Fig. 2. Architecture of our network. The MAC building block converts the micro-lens image array of light fields into feature maps, which are processed by a modified DeepLab-v2 backbone model.

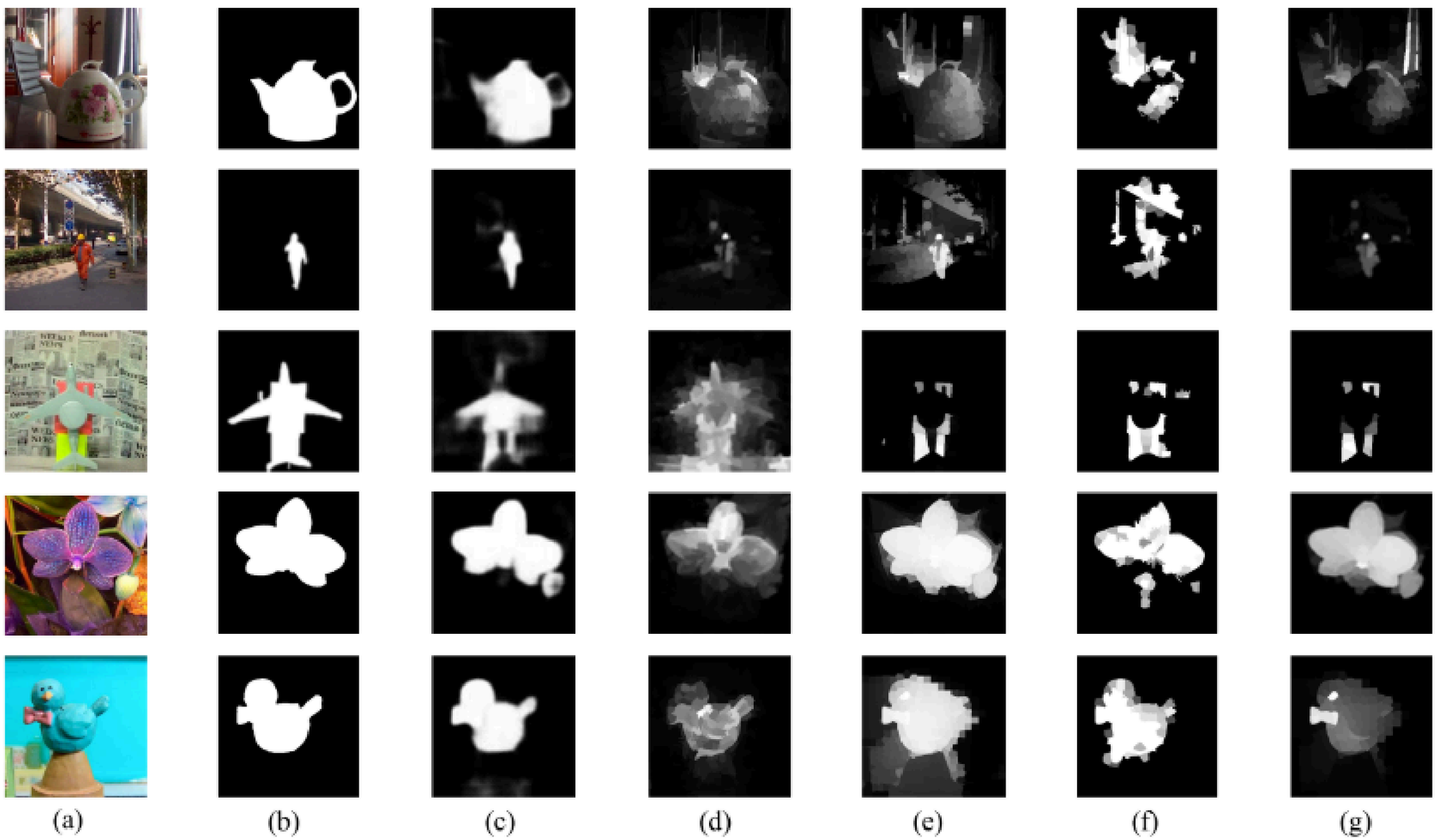


Fig. 14. Visual comparison of our best MAC block variant (Ours) and state-of-the-art methods on three datasets. (a) Central viewing/all-focus images. (b) Ground truth maps. (c) Ours. (d) LFS [36]. (e) DILF [19]. (f) WSC [20]. (g) Multi-cue [2]. The first five samples are taken from the proposed Lytro Illum dataset, the middle three samples are taken from the HFUT-Lytro dataset, and the last two samples are taken from the LFSD dataset.

✅ADVANTAGES

- Using **patch-based processing** enhances the network's ability to focus on local features → **improve saliency detection accuracy**
- Strong generalization across various LF datasets → **robustness** & **adaptability**
- handles **challenging scenarios**

❌LIMITATIONS

- **doesn't outperform all other methods**
- Performance strongly **relies on availability & quality of LF datasets**
- **computational complexity**

2022 - EXPLORING SPATIAL CORRELATION FOR LF SALIENCY DETECTION: EXPANSION FROM A SINGLE VIEW

End-to-end CNN framework:

- **LF synthesis from a single view** → central view expanded into an array of angular views (adjacent views: small differences (subtle parallax)) → 4 side views (efficiently repr. original LF)
- **LF-driven saliency detection:**
  - **Spatial Feature Extractor** → captures spatial info from multi-view images
  - **DSU** (Direction-Specific Unit) → capture spatial correlations bt views along horizontal & vertical directions
  - **Cascaded decoder** → refines saliency map (integrating multi-level features)

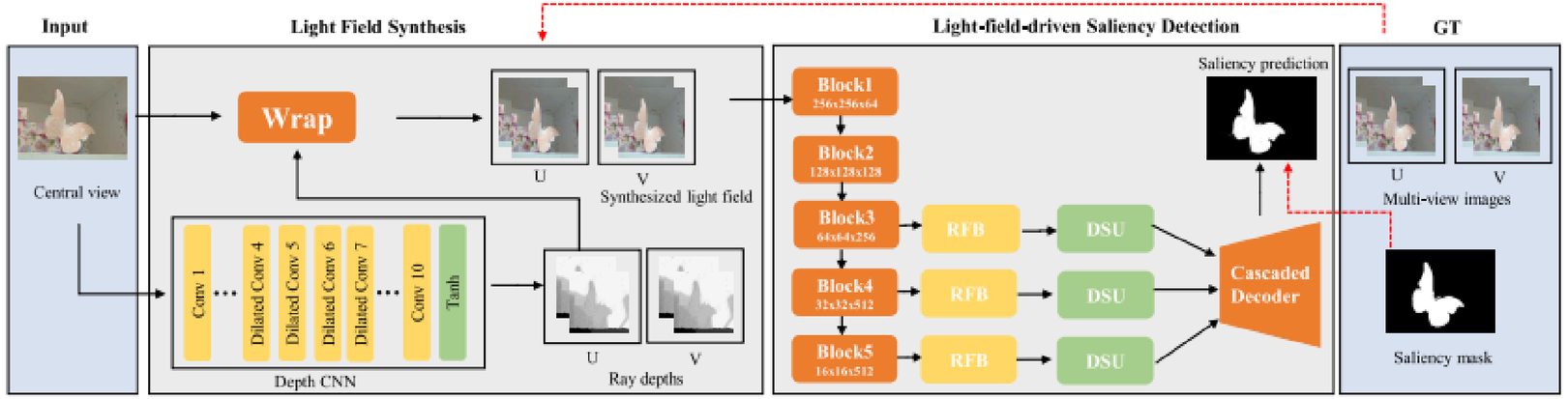


Fig. 2. Whole pipeline consists of two sub-networks: light field synthesis network and light-field-driven saliency detection network. The first network is designed to generate light field data. The second network contains three parts: spatial feature extractor, direction-specific screening unit and cascaded decoder.

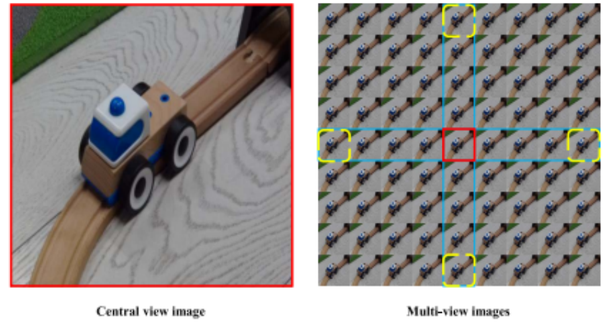


Fig. 3. Illustration of angular view selection in DUTLF-V2. Blue solid rectangles represent all 16 angular views along the horizontal and vertical directions, yellow dashed rectangles represent 4 side views.

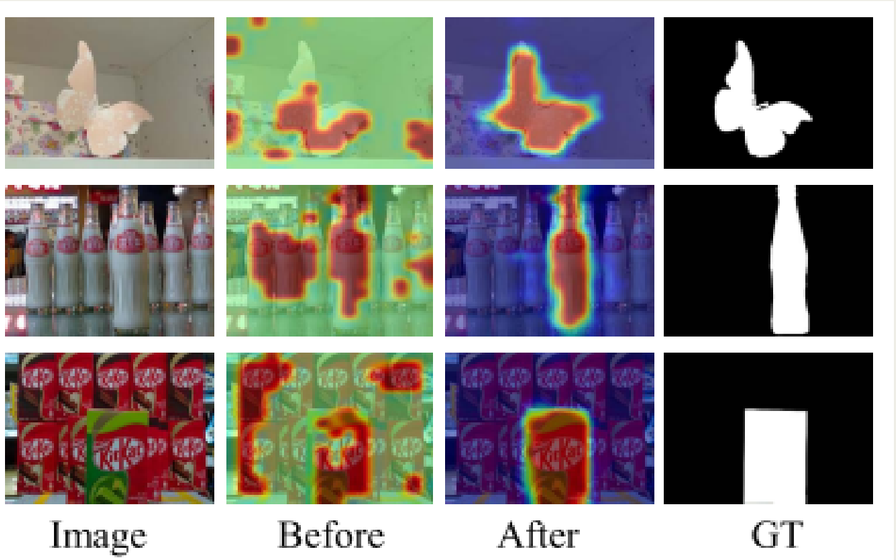


Fig. 8. Visualization for feature maps of several challenging scenes before and after being processed by DSU.

Fig. 5. Visual comparisons of the proposed method with top-ranking CNNs-based methods in some challenging scenes, including low contrast, complex background, transparent object and multiple objects.

✅ADVANTAGES

- outperforms the **2019** version**:**
  - **more spatially consistent** saliency maps (especially complex scenes)
  - **faster**
  - 2019 → used 16 views (redundancy + high computational cost). 2022 → uses only 4 side views: **enhance speed-accuracy trade-off** & **reduce computational load**
  - 2022 → utilize spatial parallax → **superior saliency detection** (challenging scenes)
- **better accuracy & robustness** compared to 2D,3D,4D methods & 2019 version

❌LIMITATIONS

- even if faster than 2019 → still **behind real-time requirements** (due to additional data processing)

2023 - A THOROUGH BENCHMARK AND A NEW MODEL FOR LF SALIENCY DETECTION

- **PKU-LF dataset** → with complex scenes like underwater & high-quality labeling (annotations: scribbles, bounding boxes, object-level, edge information)
- **STSA** (Symmetric Two-Stream Architecture)
  - **FIM** (Focalness Interweavement Module) → leverage structural characteristics of focal stacks (~ ConvLSTM)
  - **PDM** (Partial Decoder Module) → multi-modal + multi-scale features without losing details (segmentation)

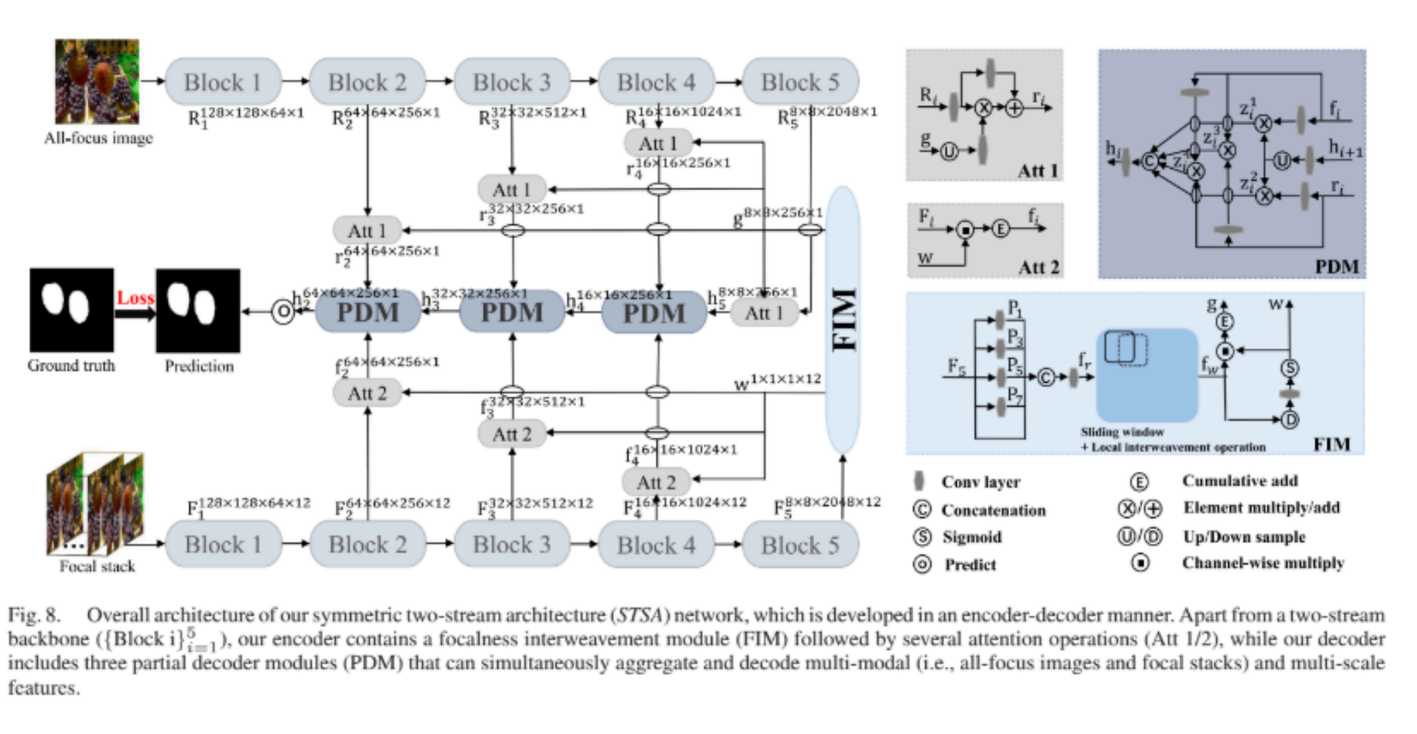


Fig. 8. Overall architecture of our symmetric two-stream architecture (*STSA*) network, which is developed in an encoder-decoder manner. Apart from a two-stream backbone ($\{\text{Block i}\}_{i=1}^{5}$), our encoder contains a focalness interweavement module (FIM) followed by several attention operations (Att 1/2), while our decoder includes three partial decoder modules (PDM) that can simultaneously aggregate and decode multi-modal (i.e., all-focus images and focal stacks) and multi-scale features.

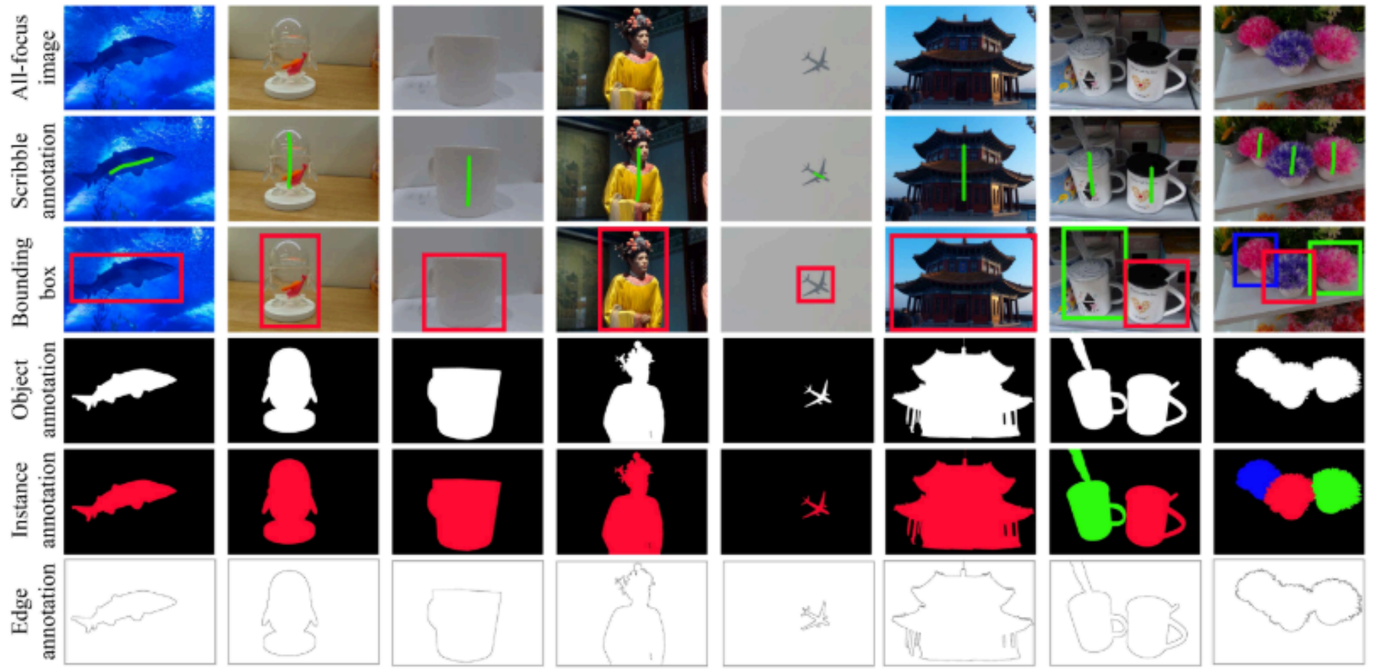


Fig. 3. Rich annotations of the proposed PKU-LF dataset including scribble annotations, bounding boxes, object-/instance-level annotations, and edge annotations. It is worth mentioning that the samples also indicate the complexity of our dataset, e.g., underwater scenes (1st column), transparent objects (2nd column), similar foregrounds and backgrounds (3rd column), low illumination (4th column), extremely small/large objects (5th and 6th columns), and multiple objects (7th and 8th columns).

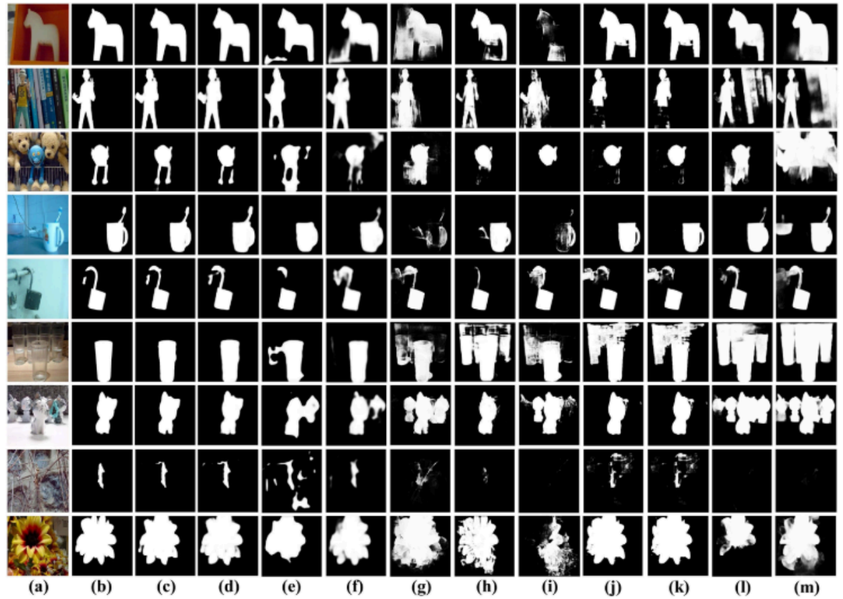


Fig. 12. Qualitative comparisons of the proposed *STSA* network with nine top-ranking methods. (a) All-focus images. (b) Ground truths. (c) Ours with the compound loss. (d) Ours with the cross-entropy loss. (e) ERNet [23]. (f) MoLF [21]. (g) BBS [17]. (h) SSF [71]. (i) UCNet [1]. (j) HDF [74]. (k) ATSA [16]. (l) MINet [13]. (m) GCPA [14].

✅ADVANTAGES

- **FIM reduces computational complexity** → more practical for real-world applications
- **unexplored scenarios** (underwater & high-resolution scenes)

❌LIMITATIONS

- **computational demanding**
- to **verify if** it’s able to **generalize** other than on PKU-LF dataset