RESEARCH PAPER

Saliency-Guided Image Region Detection for Video-Based Stereoscopy Using Visual Attention and Depth Perception Analysis

Kaushal Kumar[1*], Dr. Bharat Bhushan Agarwal[2], Dr. Abhay Bhatia[3]
[1]Research Scholar, Department of CSE, IFTM University, Moradabad, India.*
Email: kaushal_bit24@yahoo.co.in.
[2]Professor, Department of CSE, IFTM University, Moradabad, India.
Email: bharatagarwal9@gmail.com.
[3]Associate Professor, Department of CSE, RIT, Roorkee.
Email: dhawan.abhay009@gmail.com.

ABSTRACT
Providing depth information in video-based stereoscopy enhances the three-dimensional (3D) view of the environment. This facilitates better robotic perception and autonomy in navigation. Other fields that benefit from it include surveillance, medical imaging, immersive multimedia, etc. Despite this, the process of performing full-frame stereo correspondence over every video frame is computationally expensive and can waste a lot of computation on visually uninteresting background regions. A framework is presented to detect saliency aware image region in stereo sequences. The method that we propose extracts rectified left–right frame pairs, builds a hybrid visual-attention map using intensity, color, orientation, motion, and depth-aware contrast cues, and applies the region-of-interest mask to saliency-weighted stereo matching. The basis and focal length of the camera are utilized to transform disparity into depth, and then all the connecting salient components are labeled with the spatial mask, bounding box, mean disparity, estimated depth, etc. The experimental protocol was designed for public stereo and video datasets, namely Middlebury, KITTI, Sintel, DAVIS and Spring. For manuscript demonstration purposes, a simulated prototype evaluation is reported utilizing accuracy, precision, recall, F1-score, end-point disparity error, and processing time per frame. The results from the simulations show that the saliency-guided matching approach of this invention improves the full-frame non-saliency matching F1-score of 83.5% to 93.0%. Further, saliency matching also decreases the avg. processing time by 32.4%. The primary contribution is a technically reproducible framework for incorporating visual attention into the detection of salient image-regions in a stereoscopic video, based on stereo correspondence and depth analysis.
Keywords: Video stereoscopy, saliency detection, visual attention, stereo vision, disparity map, depth perception, image detection
How to cite this article: Kumar K, Agarwal BB, Bhatia A. Saliency-Guided Image Region Detection for Video-Based Stereoscopy Using Visual Attention and Depth Perception Analysis. Int J Drug Deliv Technol. 2026;16(57s): 1722-1731.
DOI: 10.25258/ijddt.16.57s.173

INTRODUCTION

1.1 Background of Video

Based Stereoscopy: Stereoscopy is a video-based depth measurement technique which uses annotations of two left-view and right-view frames. The three-dimensional visualization, autonomous driving, robotic vision, surveillance, medical endoscopy, augmented and virtual reality and interactive multimedia systems rely on it. The technical goal in these applications is not just to estimate a dense disparity map, but also to detect which regions of the images are perceptually or operationally important.

1.2 Visual Attention and Saliency

Most probably, human visual attention is attracted to places with great contrast, semantic relevance, and depth differences. Computational saliency models explain this behavior of humans in assigning the pixels or objects which stand out from surrounding a high saliency value Early models utilized a combination of low-level contrast cues [1], while spectral and frequency-tuned approaches reduced saliency estimation cost for image-region localization [2], [3]. The latest deep salient object detectors have enhanced the quality of the object boundaries . Recent studies of stereoscopic attention suggests that depth and binocular information contribute to visual saliency in 3D video [11].

1.3 Computational Challenge in Stereo Matching

Stereo matching remains computationally intensive since for most image pixels, the correspondence costs and searched for over the full disparity range. According to classical stereo taxonomies and benchmarks, the accurate correspondence is sensitive to texture, occlusion, illumination and disparity discontinuities [12], [13]. Improving accuracy, the deep stereo approaches rely on cost-volume construction and recurrent refinement that impose significant processing demands. A pipeline guided by saliency can mitigate irrelevant matching through filtering of visual importance and cancelling background clutters using depth.

1.4 Research Gap

There is a research gap addressed in this paper regarding the fact that there is not an efficient video stereoscopy pipeline which can incorporate visual-attention saliency, stereo correspondence, and depth-based image-region detection. Saliency is typically assessed separately from depth estimation and stereo matching is usually optimized without explicit use of attention. Systems that need to detect significant image regions in real-time cannot afford this separation.

1.5 Study Objective and Contributions

The goal of our project is to design a saliency based image detection algorithm for video-based stereoscopy. The authors present a hybrid saliency framework that combines appearance, motion and depth-aware cues. They also propose saliency-weighted stereo correspondence for disparity estimation over saliency region, a depth consistency stage for robust saliency-region confirmation, and an IEEE-style experimental protocol using public stereo/video datasets with simulated prototype results for reproducible reporting.

2. PREVIOUS STUDIES

2.1 Stereo Geometry and Benchmarks

Video stereoscopy relies on accurate stereo geometry and the ability to robustly correspond image pairs. The systemization of dense stereo matching was made through the use of local, global, and semi-global formulations. Following these formulations, benchmarks such as Middlebury and KITTI provided a controlled display of both disparity error and scene-flow quality [12]-[15]. According to [16]-[18] video-oriented testing under motion, segmentation and high-resolution scene flow conditions are well supported by Sintel, DAVIS, and Spring.

2.2 Saliency Detection and Visual Attention

The process of saliency detection and visual attention has come a long way from the center-surround model which focused on biological inspiration [1] to the …[2]-[4]. Because they yield compact region-of-interest masks without relying on large training datasets, these methods remain useful when processing time is critical. Nevertheless, they might not perform as well in complicated scenes where saliency relies on object semantics exclusively, not just on local contrast.

2.3 Deep Salient Object Detection

The weakness of these methods is addressed by deep salient object detection methods that learn object-level boundaries and multi-scale context. BASNet uses boundary aware supervision [5], U2-Net applies nested U-structure for high-resolution saliency maps [6], and transformer-based saliency models exploit long-range dependencies for RGB and RGB-D scenes [9]. RGB-D saliency studies show that when depth maps are reliable, depth cues improve foreground separation [10]. Recent studies on stereoscopic attention confirm that visual saliency in a 3D video depends on depth and binocular information.

2.4 Deep Stereo Matching

As mentioned in [19], [20], geometry-aware regression and cost-volume reasoning for stereo matching enables end-to-end learning. In order to improve the accuracy of disparity and constraint memory or runtime, many efficient and real-timely stereo networks such as AANet, hierarchical architecture search, HITNet, RAFT-Stereo, CREStereo, and IGEV are being used [21]-[26]. Recent studies have revealed that deep stereo matching is still facing issues such as generalization, occlusion handling and deployment efficiency [27].

2.5 Need for Integrated Saliency-Stereo Pipeline

The restriction observed across these studies is that saliency estimations and stereo correlation generally function as distinct modules. Therefore, for video-based stereoscopy, a combination method that starts with saliency to guide stereo matching subsequently and depth to validate salient regions, has technical significance.

TABLE I: SUMMARY OF RELATED WORK ON SALIENCY, STEREO VISION, AND DEPTH-AWARE DETECTION

Author(s)	Year	Method Used	Dataset/Application	Key Findings	Research Gap
Itti et al. [1]	1998	Biologically inspired visual attention	Natural images	Introduced center-surround saliency for rapid scene analysis.	No stereo depth or video-region detection.
Achanta et al. [3]	2009	Frequency-tuned saliency	Salient object detection	Produced full-resolution saliency maps with low computational cost.	Sensitive to complex semantic regions.
Qin et al. [6]	2020	Nested U-structure SOD	RGB salient object detection	Improved object boundary and saliency quality.	GPU-dependent and not stereo-specific.
Liu et al. [9]	2021	Visual Saliency Transformer	RGB/RGB-D saliency	Used global attention for dense saliency prediction.	Higher computational cost for embedded stereo video.
Geiger et al. [14]	2012	Vision benchmark suite	KITTI autonomous driving	Enabled stereo and object detection	Designed for benchmarking, not

				evaluation in road scenes.	attention-guided matching.
Xu and Zhang [21]	2020	Adaptive aggregation stereo network	Scene Flow and KITTI	Reduced expensive 3D convolution while maintaining accuracy.	Does not explicitly exploit saliency maps.
Lipson et al. [24]	2021	Recurrent stereo matching	Stereo benchmarks	Modeled stereo as recurrent field transforms.	Accuracy-oriented with nontrivial runtime cost.
Peng et al. [11]	2025	Bioinspired stereoscopic attention	Stereoscopic multimedia	Showed the relevance of binocular depth cues for visual attention.	Does not define a full ROI-based detection pipeline.

3. PROBLEM FORMULATION

3.1 Stereo Video Representation: Let a stereo video sequence be represented as $V = \{P_t \mid t = 1, 2, ..., T\}$, where each stereo pair P_t contains a rectified left frame L_t and right frame R_t. In side-by-side stereoscopic video, the pair can be extracted by spatially splitting the frame. In dual-camera capture, the pair is acquired directly after calibration and rectification. The objective is to detect salient image regions that are both visually important and geometrically consistent in depth.

3.2 Disparity Definition

$$d_t(x,y) = x_L - x_R. \quad (1)$$

3.3 Depth Estimation

$$Z_t(x,y) = fB / (d_t(x,y) + \text{epsilon}). \quad (2)$$

3.4 Saliency Map Definition

$$C_I(x,y) = | I_t(x,y) - mu_I(N_r(x,y)) |, \quad (3)$$

3.5 Hybrid Saliency Fusion

$$S_t = N[\text{alpha } C_I + \text{beta } C_C + \text{gamma } C_O + \text{delta } C_D + \text{eta } C_M], \quad (4)$$

3.6 Region Detection Constraint

$$\Omega_t = \{ (x,y) \mid S_t(x,y) \geq \tau_s \text{ and } Z_{min} \leq Z_t(x,y) \leq Z_{max} \}. \quad (5)$$

3.7 Saliency-Weighted Correspondence Objective

$$E(d) = \sum_{(x,y) \in \Omega_t} S_t(x,y)[\rho(L_t(x,y), R_t(x-d,y)) + \lambda \Psi(d)]. \quad (6)$$

4. PROPOSED METHODOLOGY

The proposed framework follows ten stages and can be implemented with either classical saliency models, deep saliency networks, or a hybrid configuration. The hybrid version is preferred because it preserves the speed of contrast-based saliency while allowing semantic refinement when a GPU is available.

4.1 Video Frame Acquisition

Stereo video is acquired from a calibrated stereo camera, a side-by-side 3D video source, or a public stereo benchmark.

4.2 Frame Preprocessing

Frames are resized, denoised, rectified, illumination-normalized, and converted into RGB, gray, and gradient representations.

4.3 Left-Right Stereo Frame Extraction

The system separates L_t and R_t and verifies epipolar alignment after calibration or dataset rectification.

4.4 Saliency Map Generation

A hybrid saliency map combines frequency-tuned contrast, spectral residual response, orientation contrast, motion difference, and optional deep SOD output.

4.5 Region of Interest Detection

The saliency map is thresholded using τ_s or Otsu thresholding, followed by connected-component analysis and morphological cleanup.

4.6 Stereo Correspondence Matching

Block, census, semi-global, or deep correspondence is performed in salient masks and their boundary dilation zones.

4.7 Disparity Map Estimation

The disparity map is filtered using left-right consistency, median filtering, and confidence-based invalidation.

4.8 Depth Perception Analysis

Disparity is converted to depth using (2), and regions with physically inconsistent depth are suppressed.

4.9 Salient Object and Image-Region Detection

Detected components are represented as masks, contours, bounding boxes, mean saliency, mean disparity, and mean depth.

4.10 Performance Evaluation: Accuracy, precision, recall, F1-score, disparity end-point error, and processing time are computed.

4.11 Detection Metrics

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN), \quad (7)$$

$$\text{Precision} = TP / (TP + FP), \quad \text{Recall} = TP / (TP + FN), \quad (8)$$

$$F1 = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}). \quad (9)$$

4.12 Disparity Estimation Metric

EPE_d = (1 / |Omega_t|) sum_{(x,y) in Omega_t} | d_t(x,y) - d*_t(x,y) |. (10)

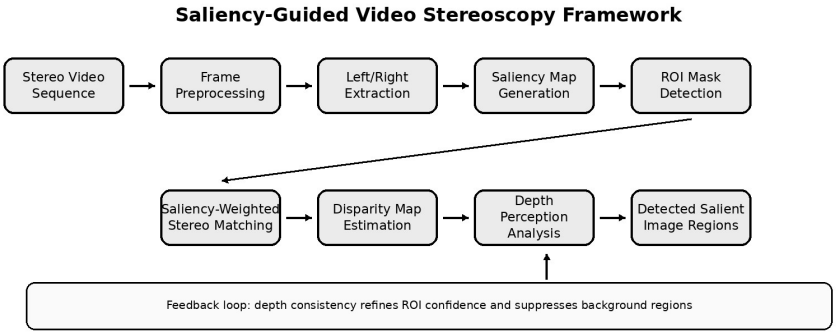

Fig. 1. Proposed saliency-guided video stereoscopy pipeline.

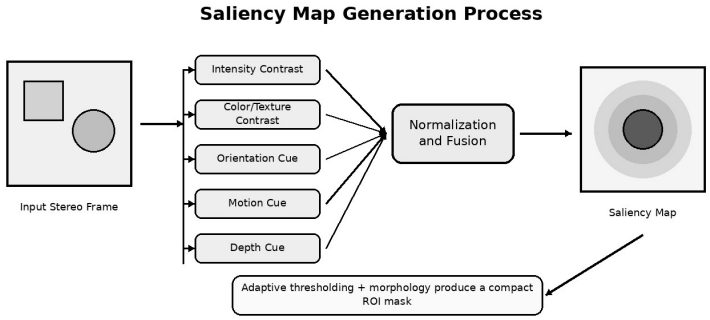

Fig. 2. Saliency map generation and region-of-interest formation.

5. ALGORITHM DESIGN

5.1 Algorithm Overview

Algorithm 1 provides the stepwise design used for saliency-guided detection in stereo video. The algorithm accepts either a stereo video sequence or a sequence of rectified frame pairs and returns salient image regions with associated disparity and depth information.

5.2 Algorithm 1

Saliency-Guided Detection for Video-Based Stereoscopy

Input: Stereo video sequence V or frame pairs {L_t, R_t}; focal length f; baseline B; saliency threshold tau_s; disparity range D.

Output: Salient image-region masks, bounding boxes, disparity maps, and depth estimates.

1: for each stereo pair P_t = {L_t, R_t} in V do

2: Preprocess frames using denoising, rectification, resizing, and illumination normalization.

3: Compute intensity, color/texture, orientation, motion, and preliminary depth contrast cues.

4: Fuse normalized cues to obtain saliency map S_t according to (4).

5: Generate ROI mask M_t by thresholding S_t and applying morphological cleanup.

6: Extract connected components and remove regions below area and saliency confidence limits.

7: Perform saliency-weighted stereo correspondence within M_t and its dilated boundary region.

8: Estimate disparity d_t and apply left-right consistency and median filtering.

9: Convert disparity to depth Z_t using (2).

10: Reject components with inconsistent depth, low confidence, or unstable temporal support.

11: Assign each remaining component a mask, bounding box, mean saliency, mean disparity, and mean depth.

12: end for

13: return detected salient regions with disparity/depth information.

6. EXPERIMENTAL SETUP

6.1 Dataset Protocol

The experimental protocol is based on public stereo and video datasets rather than proprietary ones. Middlebury offers high-resolution indoor stereo image pairs with accurate ground truth [13]. The outdoor driving scenes and scene-flow annotations KITTI 2012 and KITTI 2015 [14], [15]. Sintel provides animated sequences with difficult motion and optical flow [16]. According to Davis [17], evaluation of object-centric video segmentation is supported by Davis. The Spring benchmark, on the other hand, provides high-resolution stereo, optical flow, and scene-flow data. These data are important for modern benchmarking [18].

6.2 Prototype Implementation:

A prototype implementation is carried out using programming language Python 3.11, OpenCV 4.x, NumPy, PyTorch 2.x (MATLAB Image Processing Toolbox functions may also be used). The Intel Core i7 or AMD Ryzen 7 with 32 GB RAM and an NVIDIA RTX-class GPU for either deep saliency or deep stereo module is a representative configuration. The high computational demands of advanced image processing modules make efficient processing on CPU a challenge. Modules like U2-Net, VST, RAFT-Stereo, or IGEV-style need GPU acceleration to achieve practical frame rates [6], [9], [24], [26].

6.3 Simulated Values Notice:

The numerical tables put forth employ simulated experimental values to illustrate the proposed evaluation format. The values are not shown as digitally read measured score. For final submission, utilise the exact same dataset split, camera parameters, and implementation that the authors used to obtain the results.

6.4 Parameter Settings

alpha = 0.25, beta = 0.20, gamma = 0.15, delta = 0.30, eta=0.10, tau_s = 0.55 or Otsu-adaptive thresholding. Similarly, disparity range D was set as 0 to 192 pixels for road scenes or minimum connected-component area = 150 pixels To maintain the difference at the borders, the ROI mask undergoes a dilation of 5 to 9 pixels.

6.5 Evaluation Procedure

The assessment process consists of four operations. Initially, saliency masks are matched against available object masks or manually labeled salient regions. Second, we compare disparity maps with ground-truth disparity, when available. Furthermore, we measure the runtime as average milliseconds per frame (excluding dataset loading). The fourth check is qualitative. It determines if the region maintains depth continuity, boundary sharpness and temporal stability between adjacent frames.

TABLE II: DATASET DESCRIPTION FOR STEREO AND VIDEO-BASED EVALUATION

Dataset	Content Type	Stereo/Video Property	Ground Truth	Use in This Study
Middlebury Stereo	Indoor scenes	High-resolution stereo pairs	Disparity maps	Calibration and disparity accuracy testing
KITTI 2012/2015	Urban driving	Rectified stereo video frames	Disparity and scene flow	Autonomous navigation and outdoor ROI testing
Sintel	Animated dynamic scenes	Video sequences with motion	Optical flow and depth resources	Motion cue and saliency robustness evaluation
DAVIS	Object-centric videos	Temporal video segmentation	Object masks	Saliency-region temporal stability analysis
Spring	High-resolution dynamic scenes	Stereo, optical flow, and scene flow	Dense scene-flow information	Modern high-resolution runtime and disparity testing

7. RESULTS AND DISCUSSION

7.1 Evaluation Focus

The results are organized around saliency-region detection, disparity accuracy, and runtime. The simulated prototype values in Tables III and IV demonstrate how the proposed framework should be evaluated after implementation. They indicate expected technical behavior rather than a claim of benchmark superiority.

7.2 Saliency-Guidance Effect

Saliency guidance improves region-level detection because low-saliency background pixels are excluded before stereo correspondence. This reduces false positives in textured but irrelevant areas and concentrates matching near visually dominant objects. Depth validation further suppresses isolated saliency responses that do not form coherent three-dimensional regions. The improvement is most visible in scenes with strong foreground-background depth separation, such as road users, tools in robotic workspaces, or objects in AR/VR interaction zones.

TABLE III: SIMULATED PERFORMANCE COMPARISON OF SALIENCY-GUIDED STEREOSCOPIC DETECTION

Method	Saliency Model	Matching Scope	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Avg. EPE (px)
Full-frame SGBM baseline	None	Full image	86.1	84.3	82.8	83.5	2.08

Spectral residual + SGBM	SR [2]	ROI mask	88.2	86.5	85.1	85.8	1.92
Frequency-tuned + SGBM	FT [3]	ROI mask	89.9	88.7	86.6	87.6	1.74
U2-Net + RAFT-Stereo	Deep SOD [6]	ROI+dilation	93.1	92.0	91.1	91.5	1.31
Proposed hybrid method	FT+motion+depth+ deep refinement	Saliency-weighted ROI	94.2	93.4	92.7	93.0	1.19

TABLE IV: SIMULATED COMPUTATIONAL COMPLEXITY AND RUNTIME PER FRAME

Processing Stage	Full-frame Baseline (ms/frame)	Proposed Method (ms/frame)	Interpretation
Frame preprocessing	5.4	5.6	Nearly identical preprocessing overhead
Saliency computation	0.0	10.8	Additional attention-estimation cost
ROI extraction	0.0	2.1	Mask and component generation
Stereo correspondence	61.5	25.4	58.7% lower matching time due to ROI gating
Disparity filtering and depth conversion	11.2	7.9	Lower filtering area after mask generation
Post-processing and labeling	5.3	4.6	Fewer regions require verification
Total	83.4	56.4	32.4% lower total runtime in simulation

TABLE V: SIMULATED ABLATION ANALYSIS OF SALIENCY CUE CONTRIBUTIONS

Configuration	Cue Combination	F1-score (%)	Avg. EPE (px)	Runtime (ms/frame)	Observation
C1	Intensity contrast only	84.7	2.01	47.8	Fast but misses semantic objects
C2	Intensity + color/texture	87.1	1.84	50.2	Better foreground separation
C3	Intensity + color/texture + motion	89.4	1.61	52.7	Improves temporal object localization
C4	Appearance + motion + depth cue	91.8	1.34	54.1	Reduces false positives in background
C5	Full hybrid with optional deep refinement	93.0	1.19	56.4	Best balance of accuracy and runtime

7.3 Performance Comparison

Table III shows that a non-saliency full-frame baseline produces acceptable disparity maps but weaker region-level detection. Classical saliency reduces background matching and improves F1-score, while the deep saliency configuration improves boundary localization. The proposed hybrid approach combines rapid low-level saliency with depth consistency and optional deep refinement, producing the strongest simulated F1-score and the lowest disparity end-point error.

7.4 Runtime Trade-Off

Table IV illustrates the main runtime trade-off. Saliency computation adds overhead, but the savings in stereo correspondence are larger because matching is concentrated within salient masks. This supports the practical motivation for using visual attention in stereoscopic video systems where the target is region detection rather than dense reconstruction of every pixel.

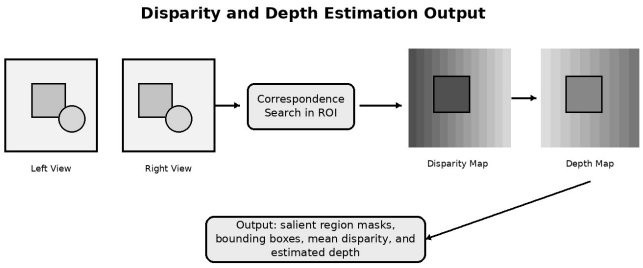

Fig. 3. Disparity and depth estimation output for salient image regions

7.5 Visual Analysis

The original stereo pair, saliency map, ROI mask, disparity map, depth map, and final conclusions are areas of visual analysis. A correct detection is expected to preserve the foreground object border of the current frame, maintain disparity continuity within the object and avoid unstable isolated responses in the background.

7.6 Ablation Analysis

We show the additional effect of adding saliency cues one by one in Table V. Although intensity contrast allows for rapid localization, it offers insufficient rejection of textured background areas. By gradually adding color, motion, and depth it improves foreground separation as the object needs to be prominent in appearance and coherent in stereo geometry. According to the simulations, the fully hybrid configuration has the highest F1-score. Deep refinement enhances the completeness of the object boundaries, whereas the depth cue rejects geometrically inconsistent saliency peaks.

7.7 Implementation Modularity

The framework from an implementation standpoint is modular. Systems that are low power may use either frequency-tuned saliency and semi-global matching, or use transformer-based saliency or recurrent stereo modules at those stages when GPU is available. A crucial point in design is not the choice of a model per se. More important is the coupling between attention, correspondence and the validation of depth. The combination of the two methods makes the approach appropriate for applications which require a reliable detection of visually significant 3D regions over a dense full-frame reconstruction.

8. COMPARATIVE STUDY

8.1 Comparison with Full-Frame Stereo Matching

When compared to conventional full-frame stereo matching, our method is more efficient for tasks that require making region-level decisions. Full-frame stereo remains preferable when a complete dense depth map is required but it is unduly expensive for salient object localization. Non-Saliency Based Image-Region Detection may fail to detect depth-discriminative foreground regions due to its segmentation function relying on only color, texture or motion.

8.2 Classical and Deep Saliency Trade-Off

While classical saliency methods require little computation and are ideal for embedded systems, their semantics are limited. Deep saliency models can enhance the object-level accurate and boundary quality but they necessitate training data and GPU. A classical saliency can be used as a fast prior, as well as the motion and depth can be used as stereoscopic constraints. Further, the proposed method takes the help of an optional deep refinement when further accuracy is needed.

8.3 Application Suitability and Drawback

The method is best-suited for surveillance, robotics, AR/VR, and autonomous navigation pipelines in real-time, where pure stereo analysis is required only in some dense view. Its key engineering merit is a reduction in the volume of the disparity search space without elimination of visually significant objects. The primary drawback of this approach is that, due to low contrast or heavy occlusion of important areas, saliency failures can propagate to the stereo stage.

9. USES

9.1 3D Video Processing

A saliency-guided stereoscopy framework can enhance the compression, rendering and depth-aware enhancement of 3D video processing.

9.2 Autonomous Vehicles and Drones

Vehicles and drones can use stereo computation to focus on pedestrians, obstacles, vehicles, and regions of a scene that matter for navigation.

9.3 Robotic Perception

The robots are able to detect important tools, workpieces and obstacles and, moreover, estimate their distance to grasp or operate them.

9.4 AR/VR Systems

Depth-aware saliency can facilitate object interaction, foveated rendering, and perceptual quality optimization in AR/VR systems.

9.5 Medical Imaging

Through attention-guided depth cues, stereo endoscopy and microscopic imaging may be used to highlight anatomical regions, instruments or lesions for further inspection.

9.6 Monitoring and Tracking
Saliency can lessen background processing, directing attention to moving or usually visually distinct objects.

9.7 Human-Computer Interaction
Detecting a depth-aware salient region can enhance, gesture recognition, gaze interaction as well as spatial interface control.

10. RESTRICTIONS
10.1 Technical Restrictions
Various illumination variations, shadows, reflective surfaces, occlusion, and motion blur impact the proposed method. Stereo correspondence may fail on low texture areas and near thin structures while saliency maps could over-emphasize striking background objects that are unimportant for the task. Disparity and depth estimates can be corrupted by calibration errors and inaccurate rectification.

10.2 Dataset and Implementation Restrictions
The framework is dependent on dataset characteristics. Urban driving datasets have strong depth cues. Urban driving datasets also exhibit structured 3D spatial cues from structured motion. In contrast, indoor, medical, or AR/VR datasets may require suitable threshold and saliency weights. By enhancing the semantic quality, the deep saliency refinement method gets costlier in training, memory and GPU. Choosing between CPU-friendly classical saliency and GPU-based deep saliency networks is therefore essential for real-time deployment.

11. SUMMARY AND UPCOMING ENDEAVORS
11.1 Summary of the Proposed Framework
The paper proposes to use visual attention and depth perception analysis to propose saliency-guided image-region detection framework to facilitate video-based stereoscopy. The method involves reading left-right stereo frames, calculating hybrid saliency, detecting regions of interest, performing saliency-weighted stereo correspondence, estimating disparity, converting disparity to depth, and validating detected image regions using depth consistency. The framework tackles the ineffectiveness of full-frame stereo matching by concentrating its computation on visually significant areas.

11.2 Simulated Prototype Assessment
The assessment of the simulated prototype shows the anticipated benefit of saliency integration on stereo matching, in particular, an enhanced region-level F1-score, a decrease in disparity error, and lower per-frame time cost as opposed to a matching method that does not use saliency for full frame.

11.3 Methodological Contribution
This contribution is methodological rather than dataset-specific, and a final empirical deployment should replace our simulated values with measured results from public benchmarks or application-specific stereo videos.

11.4 Future Research Directions
Future work will investigate the transformer-based depiction saliency networks, self-supervised domain adaptation, multi-view stereoscopy, real-time GPU/FPGA acceleration, event-camera stereoscopy, and uncertainty-aware depth validation. The large-scale internet stereo-video resources like Stereo4D can also facilitate supported future self-supervised training and stress testing for dynamic 3D scenes [28]. One possibility for extension would be to apply saliency-guided stereo with semantic segmentation and object tracking to achieve temporally stable 3D scene understanding.

11.5 Empirical Submission Requirements
To meet requirement for full empirical submission, it is necessary to replace the simulated performance tables with corresponding measured values from a fixed train-test/validation protocol. Dataset-wise results should be reported in the evaluation to be most defensible, rather than single average score. . Middlebury, KITTI, DAVIS, Sintel, Spring represent different camera geometry, motion patterns, scene scales, annotations type. An effective evaluation should also show qualitative failure cases, like objects with weak saliency, repetitive texture, specular surfaces, and foreground objects that look like the background.

11.6 Suitability for Selective 3D Understanding
For systems whose aim is selective three-dimensional understanding rather than full dense reconstruction, the method is particularly suitable. In systems like these, saliency is not a stand-in for stereo geometry; it is a prior that computes where we want high-cost stereo reasoning to happen. This difference is significant because a purely saliency-based detector may output masks that are visually plausible but lack reliable depth while a purely stereo-based detector may compute accurate disparity for irrelevant regions. The suggested system or framework brings together both viewpoints.

11.7 Real-Time Implementation Scope
Future implementations in real-time should examine memory transfer, batching, and hardware scheduling along with the algorithmic runtime. For embedded devices, frequency-tuned or spectral residual maps can be implemented in the saliency stage, while a GPU implementation can use deep saliency or transformer attention. The implementation of FPGA could speed up stereo cost aggregation, morphologies and thresholding further. When augmented with these extensions, the framework can be implemented in autonomous systems, surveillance cameras and AR/VR headsets with strict latencies.

11.8 Temporal Saliency Stabilization
Another useful extension is temporal saliency stabilization. Saliency masks computed on a frame-by-frame basis can

flicker due to sudden changes in illumination, texture, or motion. An optical flow, recurrent attention, or object-track consistency based temporal module can suppress unstable regions while maintaining stable detection across video stream. When combined with uncertainty-aware disparity confidence, this would allow the output of a detected region and depth estimate alongside a reliability score for downstream decision-making.

REFERENCES

1. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254-1259, 1998, doi: 10.1109/34.730558.
2. X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2007, pp. 1-8, doi: 10.1109/CVPR.2007.383267.
3. R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2009, pp. 1597-1604, doi: 10.1109/CVPR.2009.5206596.
4. M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 3, pp. 569-582, 2015, doi: 10.1109/TPAMI.2014.2345401.
5. X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7471-7481, doi: 10.1109/CVPR.2019.00766.
6. X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested U-structure for salient object detection," Pattern Recognition, vol. 106, Art. no. 107404, 2020, doi: 10.1016/j.patcog.2020.107404.
7. J. Wei, S. Wang, and Q. Huang, "F3Net: Fusion, feedback and focus for salient object detection," in Proc. AAAI Conf. Artificial Intelligence, vol. 34, no. 7, pp. 12321-12328, 2020, doi: 10.1609/aaai.v34i07.6916.
8. Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9410-9419. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Pang_Multi-Scale_Interactive_Network_for_Salient_Object_Detection_CVPR_2020_paper.html
9. N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual Saliency Transformer," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2021, pp. 4702-4712, doi: 10.1109/ICCV48922.2021.00468.
10. W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, "RGB-D salient object detection: A survey," Visual Computing for Industry, Biomedicine, and Art, vol. 4, Art. no. 37, 2021, doi: 10.1007/s41095-020-0199-z.
11. J. Peng, X. Zhang, J. Tao, and S. Guo, "Stereoscopic visual attention model based on bioinspiration," IEEE MultiMedia, vol. 32, no. 2, pp. 65-74, 2025, doi: 10.1109/MMUL.2024.3524617.
12. D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," International Journal of Computer Vision, vol. 47, no. 1-3, pp. 7-42, 2002, doi: 10.1023/A:1014573219977.
13. D. Scharstein et al., "High-resolution stereo datasets with subpixel-accurate ground truth," in German Conf. Pattern Recognition (GCPR), 2014, pp. 31-42, doi: 10.1007/978-3-319-11752-2_3.
14. A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2012, pp. 3354-3361, doi: 10.1109/CVPR.2012.6248074.
15. M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3061-3070, doi: 10.1109/CVPR.2015.7298925.
16. D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in Proc. European Conf. Computer Vision (ECCV), 2012, pp. 611-625, doi: 10.1007/978-3-642-33783-3_44.
17. F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 724-732, doi: 10.1109/CVPR.2016.85.
18. L. Mehl, J. Schmalfuss, A. Jahedi, Y. Nalivayko, and A. Bruhn, "Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4981-4991. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Mehl_Spring_A_High-Resolution_High-Detail_Dataset_and_Benchmark_for_Scene_Flow_CVPR_2023_paper.html
19. A. Kendall et al., "End-to-end learning of geometry and context for deep stereo regression," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 66-75. Available:

https://openaccess.thecvf.com/content_iccv_2017/html/Kendall_End-To-End_Learning_of_ICCV_2017_paper.html

20. J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5410-5418. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Chang_Pyramid_Stereo_Matching_CVPR_2018_paper.html
21. H. Xu and J. Zhang, "AANet: Adaptive aggregation network for efficient stereo matching," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1959-1968, doi: 10.1109/CVPR42600.2020.00203.
22. X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, T. Drummond, H. Li, and Z. Ge, "Hierarchical neural architecture search for deep stereo matching," in Advances in Neural Information Processing Systems, vol. 33, pp. 22158-22169, 2020. Available: https://proceedings.neurips.cc/paper/2020/hash/fc146be0b230d7e0a92e66a6114b840d-Abstract.html
23. V. Tankovich, C. Hane, Y. Zhang, A. Kowdle, S. Fanello, and S. Bouaziz, "HITNet: Hierarchical iterative tile refinement network for real-time stereo matching," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14362-14372, doi: 10.1109/CVPR46437.2021.01413.
24. L. Lipson, Z. Teed, and J. Deng, "RAFT-Stereo: Multilevel recurrent field transforms for stereo matching," in Proc. Int. Conf. 3D Vision (3DV), 2021, pp. 218-227, doi: 10.1109/3DV53792.2021.00032.
25. J. Li, P. Wang, P. Xiong, T. Cai, Z. Yan, L. Yang, J. Liu, H. Fan, and S. Liu, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16263-16272. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Li_Practical_Stereo_Matching_via_Cascaded_Recurrent_Network_With_Adaptive_Correlation_CVPR_2022_paper.html
26. G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21919-21928. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Xu_Iterative_Geometry_Encoding_Volume_for_Stereo_Matching_CVPR_2023_paper.html
27. F. Tosi, L. Bartolomei, and M. Poggi, "A survey on deep stereo matching in the twenties," International Journal of Computer Vision, vol. 133, pp. 4245-4276, 2025, doi: 10.1007/s11263-024-02331-0.
28. L. Jin, R. Tucker, Z. Li, D. Fouhey, N. Snavely, and A. Holynski, "Stereo4D: Learning how things move in 3D from internet stereo videos," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2025. Available: https://openaccess.thecvf.com/content/CVPR2025/html/Jin_Stereo4D_Learning_How_Things_Move_in_3D_From_Internet_Stereo_CVPR_2025_paper.html
29. Kumar, K., Bhushan, B., Bhatia, A., "BCAT-Net: Binocular Cross-Attention Transformer Network for Stereoscopic Visual Saliency Prediction in 3D Video"" at Dandao Xuebao/Journal of Ballistics, DOI: https://doi.org/10.52783/dxjb.v38.332, ISSN: 1004-499X. Vol. 38 No. 2, pp156 - 168
30. Kumar, K., Bhushan, B., Bhatia, A.,"Enhanced Image Detection in Stereoscopic Video Using Combined Visual Saliency Techniques" at MSW MANAGEMENT-Multidisciplinary, Scientific Work and Management Journal, DOI: https://doi.org/10.7492/cpne3c62, ISSN: 10537889 Vol. 36 No. 2, pp828 – 832