Salient Object Segmentation in 360° images/videos and light field

Yi Zhang

THÈSE DE DOCTORAT DE

L'INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : Signal, Image et Vision

Par

Yi ZHANG

Salient Object Segmentation in 360° images/videos and light field

Thèse présentée et soutenue à Rennes, le 2 Novembre 2022
Unité de recherche : Institut d'Electronique et des Technologies du numéRique (IETR) - UMR CNRS 6164
Thèse N° : 22ISAR 25 / D22 - 25

Rapporteurs avant soutenance :

Olivier LEZORAY	Professeur, Université de Caen, France
Jenny BENOIS	Professeur, Université de Bordeaux, France

Composition du jury :

Président :	Olivier LEZORAY	Professeur, Université de Caen, France
Examinateurs :	Aljosa SMOLIC	Professeur, Haute École de Lucerne, Suisse
	Ying FU	Professeur, Institut de Technologie de Pékin, Chine
	Wassim HAMIDOUCHE	Maître de conférences, INSA de Rennes, France
Dir. de thèse :	Olivier DEFORGES	Professeur, INSA de Rennes, France

Table of contents

Acknowledgment

I would like to thank all colleagues of IETR-VADDER, who have been providing me with their selfless kindness during my PhD. Specially, I would like to thank Prof. Olivier Deforges, my director of the thesis, who plays the roles of my tutor/father/friend and shows his visionary that supports me with my researches.

Second, I would like to thank senior researchers including Dr. Jing Zhang from Australian National University and Prof. Geng Chen from Northwestern Polytechnical University, who are willing to cooperate with me on interesting topics closely related to the thesis, and to share with me their excellent techniques towards conducting outstanding researches.

Importantly, I would like to thank the Chinese Scholar Council for its generous financial supports during my PhD.

Finally, I would like to thank my family and friends for showing me unconditonal kindness and for giving me their endless love.

Résumé en français

La vision humaine se compose généralement de deux phases, c'est-à-dire une vision de bas niveau et une vision de haut niveau. Plusieurs capteurs d'yeux humains saisissent les lumières réfléchies par les environnements environnants. Les neurones transfèrent ensuite les informations saisies par les capteurs au cortex visuel où les caractéristiques de vision de bas niveau (*e.g.*, bord, couleur, forme, profondeur, couleur, orientation et mouvement) sont garanties.

Les caractéristiques de bas niveau codées sont ensuite transmises à d'autres régions fonctionnelles du cerveau humain où des caractéristiques de haut niveau sont produites. Les fonctionnalités de haut niveau sont ensuite utilisées pour servir de base à la naissance de la conscience (*e.g.*, la reconnaissance d'objets). En fait, le succès de ce système hiérarchique de vision humaine est dû à un mécanisme essentiel tout au long du processus de transmission des caractéristiques, à savoir le système d'attention visuelle, qui médiatise la sélection des informations importantes de manière ascendante et descendante.

D'autre part, l'apprentissage en profondeur a dominé le domaine de la vision par ordinateur au cours des dernières années, en raison de l'essor des sources de calcul (*e.g.*, les unités de traitement graphique), de la naissance d'ensembles de données de pré-formation à grande échelle (*e.g.*, ImageNet [1]), d'une capacité d'apprentissage exceptionnelle des réseaux de neurones à convolution profonde (*e.g.*, VGGs [2]) et d'une large application de méthodologies d'optimisation adaptative (*e.g.*, l'optimiseur Adam [3]). Le succès des réseaux de neurones à convolution profonde pour des tâches telles que la classification d'images [1] et la détection d'objets [4] doit à leurs architectures constituées de couches neuronales hiérarchiques. Selon une étude de visualisation de réseau neuronal convolutif telle que [5], les cartes de caractéristiques des couches neuronales inférieures correspondent à des caractéristiques de vision de bas niveau telles que les coins et les bords, tandis que les cartes de caractéristiques de haut niveau montrent les apparences d'objets à partir d'images données. Malgré les progrès réalisés pour imiter le système visuel humain, la faible capacité de généralisation et le fonctionnement interne inexplicable des algorithmes d'apprentissage en profondeur actuels, les empêchent d'être directement transférés à différentes tâches difficiles. Dans les cas généraux, il existe plusieurs ensembles de données de référence avec des annotations spécifiques et des réseaux de neurones profonds avec des architectures et des composants exclusivement conçus pour des tâches de vision par ordinateur particulièrement difficiles.

En tant que tendance en plein essor de l'apprentissage en profondeur et de ses applications réussies pour les tâches de vision par ordinateur, la modélisation de l'attention humaine basée sur l'apprentissage en profondeur a attiré l'attention croissante de la communauté au cours des dernières

années.

Le contexte

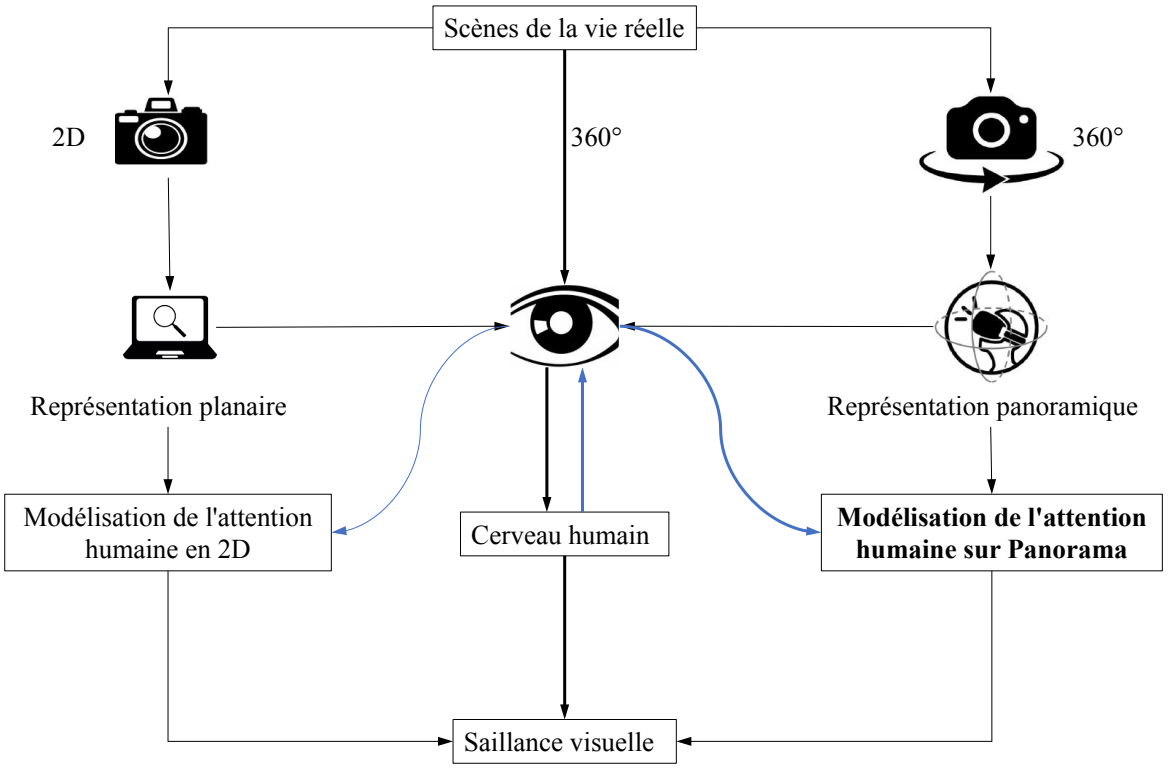


Fig. 1 Une illustration des relations entre l'attention visuelle réelle, la modélisation de l'attention visuelle traditionnelle basée sur 2D et la modélisation de l'attention visuelle basée sur des images/vidéos panoramiques à 360°. Avec des caméras à 360° et des écrans montés sur la tête, la détection de saillance visuelle basée sur un panorama à 360° est potentiellement capable de mieux imiter le comportement du système visuel humain dans des scènes réelles, par rapport au scénario 2D. Les flèches noires indiquent le flux d'informations. Les flèches bleues représentent le retour d'attention.

Comme le montre la fig. 1, les recherches actuelles liées à la modélisation de l'attention humaine sont soit basées sur deux dimensions (2D) soit sur la réalité virtuelle[1] images et vidéos. Généralement, la différence entre la modélisation de l'attention visuelle basée sur la 2D et la réalité virtuelle est double:

i. Les images/vidéos 2D sont collectées avec des caméras normales qui ne sont capables d'enregistrer que des scènes réelles observées à partir de fenêtres locales contenant un contexte limité. En particulier, les caméras VR possèdent un champ de vision de 360°×180° (Fig. 2) et sont capables d'enregistrer tout le contexte de scènes réelles. Par conséquent, par rapport à la modélisation de l'attention humaine basée sur 2D, la modélisation de l'attention basée sur 360° est basée sur des données contenant beaucoup plus d'indices visuels, possédant ainsi le potentiel d'imiter une attention

[1] Dans cette thèse, nous utilisons les termes de réalité virtuelle, 360°, panoramique et omnidirectionnelle indifféremment.

visuelle humaine plus réaliste.

ii. Les comportements visuels humains de l'observation d'un écran d'ordinateur sont différents de ceux de l'observation d'environnements immersifs avec des visiocasques. Par conséquent, la vérité terrain (*e.g.*, le mouvement des yeux) des tâches liées à la modélisation de l'attention 2D/VR a tendance à effectuer des distributions différentes. Par exemple, les attentions basées sur la 2D sont biaisées au centre tandis que celles basées sur la réalité virtuelle sont biaisées par l'équateur.

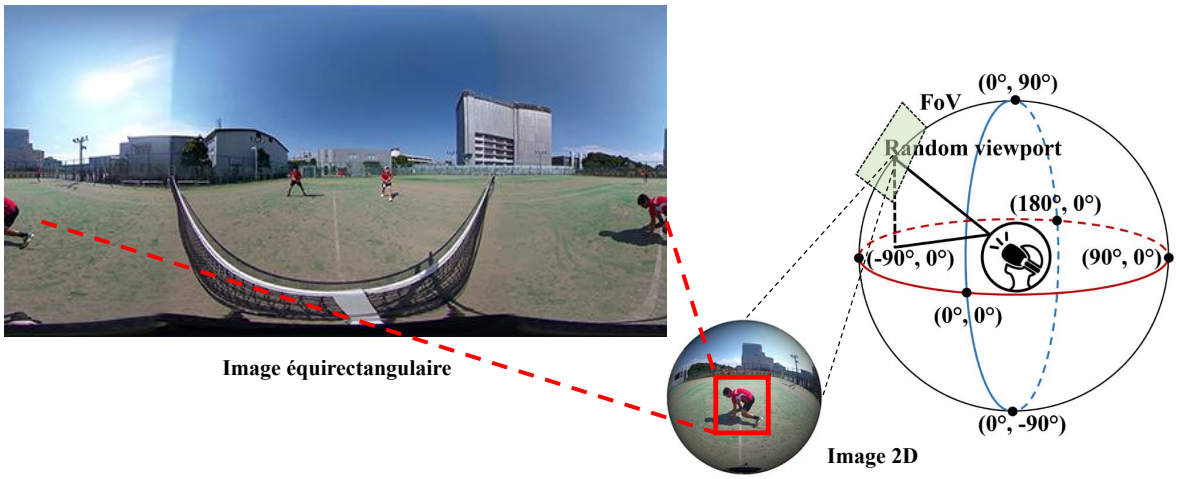


Fig. 2 Une comparaison entre la réalité virtuelle et la représentation 2D. La caméra VR capture les scènes réelles observées à partir d'un champ de vision de 360°×180°. Les images omnidirectionnelles collectées sont généralement représentées sous forme d'images équirectangulaires. La caméra normale fournit des images 2D enregistrant des scènes avec un contexte limité observé à partir de fenêtres locales.

Au cours des dernières années, des ensembles de données de référence à grande échelle basés sur la 2D et la réalité virtuelle et des modèles d'apprentissage en profondeur ont été proposés pour faire progresser le domaine de la modélisation de l'attention visuelle humaine.

Modélisation de l'attention humaine dans le domaine 2D. Les premiers ensembles de données de référence tels que MIT300 [6] et CAT2000 [7] ont collecté des données sur les mouvements de l'œil humain en menant des expériences de suivi oculaire basées sur des images statiques 2D. Ces ensembles de données basés sur des images[2] ont publié des centaines ou des milliers d'images représentant plusieurs catégories de scènes réelles, avec des cartes de fixation par image reflétant les distributions de l'attention humaine. Dans ce cas, des méthodes d'apprentissage en profondeur telles que DeepGaze [8], SALICON [9] et DeepFix [10] ont utilisé des réseaux de neurones convolutifs pour apprendre la correspondance entre les distributions d'images d'entraînement et les distributions de fixations humaines. Avec une supervision des fixations humaines, ces méthodes peuvent être entraînées pour finalement représenter les régions de haute saillance sur des images non vues (ensemble de données de test). Des ensembles de données ultérieurs tels que DHF1K [11] et LEDOV [12] ont collecté 500 à 1K courtes vidéos représentant diverses scènes de la vie quotidienne (*e.g.*, événements sociaux et sports) pour étudier les comportements visuels humains dans un scénario dynamique. Des

[2] `https://saliency.tuebingen.ai/datasets.html`

recherches telles que OM-CNN [13] et ACLNet [11] ont appliqué ConvLSTM [14] (un type de réseau neuronal récurrent largement utilisé) pour extraire les caractéristiques spatio-temporelles d'images vidéo consécutives, afin de prédire la dynamique humaine fixations.

En fait, le mécanisme d'attention visuelle humaine est non seulement capable de guider le regard humain, mais aussi d'aider l'homme à reconnaître les objets importants pour la tâche de vision telle que la compréhension de la scène. Des recherches récentes [15, 16] ont ainsi étudié une tâche relativement nouvelle, *i.e.*, salient object segmentation (*a.k.a.* détection d'objets saillants ou modélisation de l'attention visuelle au niveau de l'objet), qui vise à segmenter finement les objets saisissant la majeure partie de l'humain attentions. Comme le développement d'ensembles de données salient object segmentation à grande échelle tels que MSRA10K[3], DUT-O [17], PASCAL-S [18] , HKU-IS [19], DUTS [20] et SOC [21] (qui fournissent tous des masques binaires au niveau des pixels étiquetés manuellement comme vérité terrain), des dizaines de méthodes d'apprentissage en profondeur [16] ont il a été proposé de mener à bien votre tâche de manière entièrement-/faiblement-/non-supervisée. En outre, des ensembles de données salient object segmentation basés sur la vidéo récemment établis tels que VOS [22] et DAVSOD [23] fournissent des étiquettes pixel par pixel d'objets saillants parmi des milliers d'images vidéo, permettant ainsi salient object segmentationbasé sur la vidéo. Il convient de mentionner que divers modules d'attention [24] ont été proposés pour faciliter la modélisation de l'attention humaine au niveau de l'objet.

Modélisation de l'Attention Humaine en Panorama 360° Considérant le potentiel de la modélisation de l'attention à 360° pour imiter l'attention humaine réelle dans des scènes de la vie réelle, et la possibilité d'acquérir une grande quantité d'images et de vidéos à 360° en utilisant des caméras VR grand public telles que la série Insta360 ONE, Ricoh Theta Z1 et GoPro Max , plusieurs jeux de données tels que [25–28] ont été proposés ces dernières années, pour la modélisation statique ou dynamique de l'attention humaine au niveau de la fixation. Il convient de noter que ces ensembles de données ne fournissent que des données sur les mouvements de la tête ou des yeux comme vérité de terrain, ne pouvant donc pas refléter strictement l'attention humaine sur des cibles saillantes spécifiques. Des recherches récentes [29–32] ont exploré la détection d'objets dans des vidéos à 360°. Cependant, ces méthodes ont été proposées pour la détection de boîtes englobantes et formées pour détecter tous les objets dans des scènes à 360°, ne pouvant donc pas être utilisées pour explorer la modélisation de l'attention humaine au niveau de l'objet dans des environnements immersifs. En règle générale, salient object segmentation basé sur des images/vidéos à 360° est encore un domaine ouvert, sans aucun jeu de données ou méthode de référence proposé avant l'année 2019.

Objectifs&Contributions

L'objectif principal de cette thèse est la modélisation de l'attention visuelle au niveau de l'objet dans des environnements immersifs de réalité virtuelle[4] via des techniques d'apprentissage en profondeur. En effet, cette tâche n'est pas seulement étroitement liée à diverses tâches classiques de vision par ordinateur telles que la classification d'images [1], la détection d'objets [4], la segmentation

[3] `https://mmcheng.net/msra10k/`

[4] *a.k.a.* images, vidéos ou scènes dynamiques audiovisuelles à 360°

d'instances [33], la segmentation sémantique [34] et la comprendre [35], mais joue également un rôle important dans les applications potentielles s'adaptant à des scénarios réels tels que le post-traitement de photos, la navigation, la conduite autonome, la réalité augmentée et le robot humanoïde.

Pour remplir une solide thèse de doctorat, cette thèse démêle l'objectif principal de la modélisation de l'attention humaine au niveau de l'objet basée sur le panorama en trois tâches progressives difficiles, *i.e.*, **i** modéliser l'attention visuelle humaine au niveau de l'objet dans le champ lumineux, avec de nouvelles techniques d'apprentissage en profondeur basées sur l'attention; **ii** modélisation de l'attention visuelle au niveau de l'objet dans un panorama statique à 360° avec un nouveau modèle d'apprentissage en profondeur basé sur l'attention; **iii** modéliser l'attention audiovisuelle au niveau de l'objet dans un panorama dynamique à 360°, qui imite l'attention humaine réelle dans des scènes réelles et possède un potentiel pour des applications réelles.

Ainsi, les contributions de cette thèse se résument comme suit:

Un résumé. Un aperçu systématique des méthodes de pointe pour la modélisation de l'attention humaine au niveau de la fixation et au niveau de l'objet dans les domaines 2D et 360°. Un résumé complet sur les modèles d'attention de pointe dans le domaine de la vision par ordinateur. Les aperçus sont présentés dans Chapter 2.

Ensembles de données&benchmarks basés sur des images/vidéos à 360°. Un nouvel ensemble de données de référence et des études de référence complètes vers une image à 360° salient object segmentation, qui est détaillée dans Chapter 3. Il convient de noter que nous incluons à la fois la vérité terrain au niveau de l'objet et de l'instance au niveau des pixels dans notre nouvel ensemble de données proposé. De plus, un nouvel ensemble de données de référence et des études de référence complètes vers salient object segmentation basé sur la vidéo à 360°. En particulier, nous considérons à la fois des signaux audio et visuels pour construire l'ensemble de données, imitant ainsi mieux le scénario du monde réel. De plus, cette partie des travaux est incluse dans Chapter 3.

Nouvelles méthodologies vers le champ lumineux salient object segmentation. Le champ lumineux salient object segmentation est un domaine relativement nouveau et, étant similaire à 360° salient object segmentation, il est d'une grande importance pour les applications industrielles de réalité augmentée. À cette fin, nous explorons les méthodes de champ lumineux salient object segmentation de pointe et proposons en outre de nouveaux modèles d'apprentissage en profondeur. Les composants clés de nos nouveaux modèles incluent divers mécanismes d'attention pour la fusion de caractéristiques multimodales. Les travaux sont détaillés dans Chapter 4.

Nouvelles méthodologies vers panoramique salient object segmentation. Pour combler davantage le vide du domaine du 360° salient object segmentation, nous proposons respectivement de nouveaux modèles de référence pour mener salient object segmentation en images et vidéos 360°. La ligne de base basée sur l'image à 360° tire parti des mécanismes d'attention pour une fusion efficace des fonctionnalités basée sur des repères visuels multi-vues à 360°. La ligne de base basée sur la vidéo utilise à la fois des repères auditifs et visuels pour repérer les cibles parmi les images d'une séquence donnée. De nombreux résultats qualitatifs/quantitatifs ont été obtenus pour vérifier l'efficacité ainsi que la robustesse des méthodes proposées. Veuillez vous référer à Chapter 5 pour plus de détails sur les travaux.

Une navigation biref

Par conséquent, le chapitre suivant passe en revue divers types d'ensembles de données et de méthodologies de référence représentatifs liés à votre tâche. En outre, sur la base de nos observations sur ces méthodes récentes de pointe, nous résumons en outre les modèles d'attention de base utilisés non seulement dans salient object segmentation mais également dans les tâches générales de vision par ordinateur, afin d'établir des bases théoriques et empiriques solides pour les travaux suivants de la thèse. Le troisième chapitre présente de nouveaux ensembles de données de référence et des études approfondies concernant la modélisation de l'attention humaine au niveau de l'objet à 360°. Le quatrième chapitre détaille nos travaux vers le champ lumineux salient object segmentation. Le cinquième chapitre présente en outre nos travaux visant à créer de nouvelles bases pour la réalisation de salient object segmentation en images et vidéos à 360°, respectivement.

En conclusion, cette thèse a réussi à segmenter des objets saillants à la fois en panorama 360° et en champ lumineux. De nouveaux ensembles de données et des lignes de base basées sur l'attention ont été proposés pour une segmentation efficace des objets saillants dans les scènes panoramiques statiques et dynamiques. En outre, de nouveaux modèles d'attention ont également été proposés pour une segmentation précise des objets saillants. En tant que l'un des points de départ de la détection de cibles saillantes immersive basée sur le multimédia, nous espérons que cette thèse pourra inspirer des idées pour de futures recherches dans les domaines de la segmentation d'objets, de la VR/AR, de l'apprentissage audiovisuel et de l'apprentissage multimodal.

Chapter 1

Introduction

Human vision generally consists of two phases, *i.e.*, low-level vision and high-level vision. Multiple sensors of human eyes grasp the lights reflected by the surrounding environments. Neurons then transfer the information grasped by sensors to visual cortex where low-level vision features (*i.e.*, edge, color, shape, depth, color, orientation and motion) are generated. The coded low-level features are then conveyed to other functional regions of human brain where high-level features are produced. The high-level features are then used to serve as the foundation of the birth of consciousness (*e.g.*, object recognition). In fact, the success of this hierarchical human vision system owes to an essential mechanism throughout the whole process of feature transmission, namely visual attention system, which mediates the selection of important information in a bottom-up and top-down manner.

On the other hand, deep learning has been dominating the field of computer vision during the past years, owing to the boom of computational sources (*e.g.*, graphics processing units (GPUs)), birth of large-scale pre-training datasets (*e.g.*, ImageNet [1]), outstanding learning ability of deep convolutional neural networks (*e.g.*, VGGs [2]) and wide application of adaptive optimization methodologies (*e.g.*, Adam optimizer [3]). The success of deep convolutional neural networks for tasks such as image classification [1] and object detection [4] owes to their architectures consisting of hierarchical neural layers. According to convolutional neural network visualization study such as [5], the feature maps of bottom neural layers correspond to low-level vision features such as corner and edge, while high-level feature maps show the appearances of objects from given images. Despite the progresses on mimicking human visual system, poor generalization ability and inexplicable inner working of current deep learning algorithms, both prevent them from being directly transferred to different challenging tasks. In general cases, there are several benchmark datasets with specific annotations, and deep neural networks with exclusively designed architectures and components for particular challenging computer vision tasks.

As the booming trend of deep learning and its successful applications for computer vision tasks, deep learning based human attention modeling has been appealing increasing attention from the community during the past years.

1.1 Context

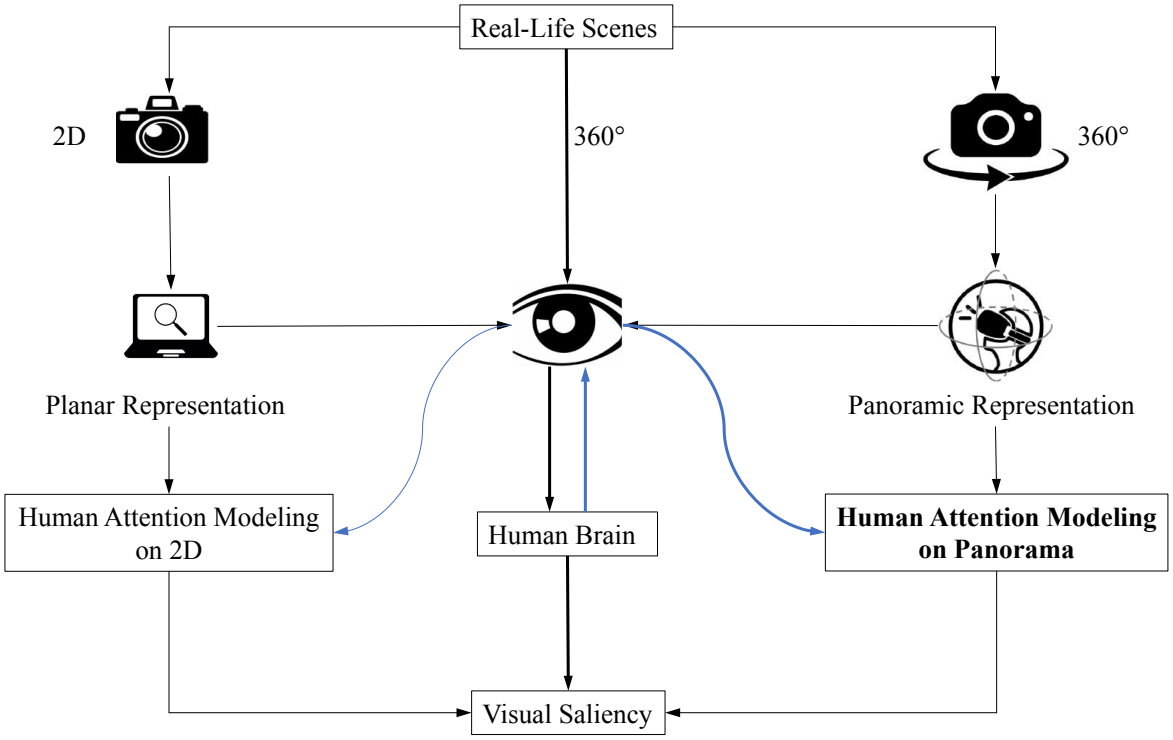


Fig. 1.1 An illustration of the relationships between real visual attention, traditional 2D-based visual attention modeling and 360° panoramic images/videos-based visual attention modeling. With 360° cameras and head-mounted displays, 360° panorama based visual saliency detection is potentially able to better mimic the behavior of human visual system in real-life scenes, when compared to 2D scenario. The black arrows denote information flow. The blue arrows represent attention feedback.

As shown in Fig, 1.1, current human attention modeling related researches are either based on two dimensional (2D) or VR[1] images and videos. Generally, the difference between 2D and VR based visual attention modeling is twofold:

i. 2D images/videos are collected with normal cameras which are only capable of recording real-life scenes observed from local viewports containing limited context. Specially, VR cameras own a field-of-view (FoV) of 360°×180° (Fig. 1.2) and are able to record the whole context of real-life scenes. Therefore, compared to 2D based human attention modeling, 360° based attention modeling is based on data containing much more visual cues, thus owning the potential of mimicking more realistic human visual attention.

ii. The human visual behaviors of observing computer screen are different when compared to those of observing immersive environments with head-mounted displays. Therefore, the ground truth (*e.g.*, eye movement) of 2D/VR attention modeling related tasks tend to perform different distributions. For instance, 2D based attentions are center-biased while VR based ones are equator-biased.

[1] In this thesis, we use the terms of VR, 360°, panoramic and omnidirectional interchangeably.

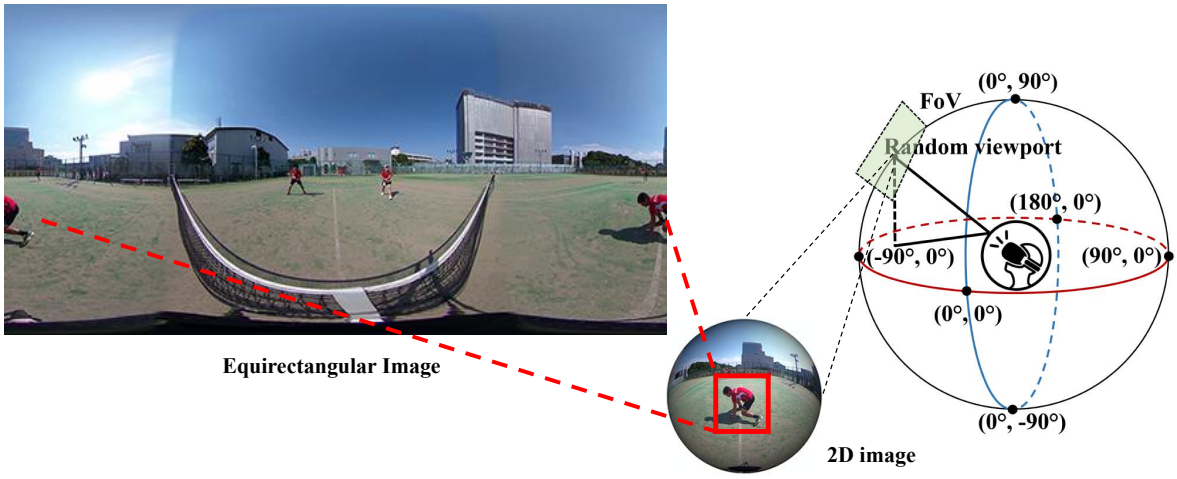


Fig. 1.2 A comparison between VR and 2D representations. VR camera captures the real-life scenes observed from a FoV of 360°×180°. The collected omnidirectional images are usually represented as equirectangular images. Normal camera provides 2D images recording scenes with limited context observed from local viewports.

During the past few years, both 2D and VR based large-scale benchmark datasets and deep learning models have been proposed to advance the field of human visual attention modeling.

Human Attention Modeling in 2D Domain. Early benchmark datasets such as MIT300 [6] and CAT2000 [7] collected human eye movement data by conducting eye-tracking experiments based on 2D static images. These image based datasets[2] released hundreds or thousands of images representing several real-life scene categories, with per-image fixation maps reflecting human attention distributions. In this case, deep learning methods such as DeepGaze [8], SALICON [9] and DeepFix [10] used convolutional neural networks to learn the mapping between the distributions of training images and the distributions of human fixations. With a supervision of human fixations, these methods are able to be trained to finally depict the regions of high saliency on unseen images (testing set). Later datasets such as DHF1K [11] and LEDOV [12] collected 500 to 1K short videos representing various daily-life scenes (*e.g.*, social events and sports) to investigate human visual behaviors in dynamic scenario. Researches such as OM-CNN [13] and ACLNet [11] applied ConvLSTM [14] (a type of widely used recurrent neural network) to extract spatial-temporal features of consecutive video frames, to predict dynamic human fixations.

In fact, human visual attention mechanism is not only able to guide where human look, but also to aid human in recognizing the objects important for vision task such as scene understanding. Recent researches [15, 16] thus investigated a relatively new task, *i.e.*, salient object segmentation (*a.k.a.* salient object detection or object-level visual attention modeling), which aims at finely segmenting the objects grasping most of the human attentions. As the development of large-scale salient object segmentation datasets such as MSRA10K[3], DUT-O [17], PASCAL-S [18], HKU-IS [19], DUTS [20] and SOC [21] (which all provide manually labeled pixel-wise binary masks

[2] `https://saliency.tuebingen.ai/datasets.html`
[3] `https://mmcheng.net/msra10k/`

as ground truth), dozens of deep learning methods [16] have been proposed to conduct salient object segmentation in fully/weakly/non-supervised manners. Besides, recently established video based salient object segmentation datasets such as VOS [22] and DAVSOD [23] provide pixel-wise labels of salient objects among thousands of video frames, thus enabling video based salient object segmentation. What is worth mentioning is that, various attention modules [24] have been proposed to facilitate the modeling of object-level human attention.

Human Attention Modeling in 360° Panorama Considering the potential of 360° attention modeling in mimicking real human attention in real-life scenes, and feasibility of acquiring large amount of 360° images and videos by using consumer-level VR cameras such as Insta360 ONE series, Ricoh Theta Z1 and GoPro Max, several datasets such as [25–28] have been proposed in the past few years, for static or dynamic fixation-level human attention modeling. It is worth noting that, these datasets provide only head or eye movement data as ground truth, thus not being able to strictly reflect human attention to specific salient targets. Recent researches [29–32] explored object detection in 360° videos. However, these methods were proposed for bounding box detection and trained to detect all objects in 360° scenes, thus not being able to be used to explore object-level human attention modeling in immersive environments. Generally, 360° image/video based salient object segmentation is still an open area, without any benchmark datasets or methods proposed before 2019.

1.2 Objectives&Contributions

The main focus of this thesis is the object-level visual attention modeling in VR immersive environments[4] via deep learning techniques. Indeed, this task is not only closely related to various classical computer vision tasks such as image classification [1], object detection [4], instance segmentation [33], semantic segmentation [34] and scene understanding [35], but also plays an important role in potential applications adapting to real-life scenario such as photo post-processing, navigation, self-driving, augmented reality and humanoid robot.

To fulfill a solid PhD dissertation, this thesis disentangles the main objective of panorama-based object-level human attention modeling into three progressive challenging tasks, *i.e.*, **i** modeling object-level human visual attention in light field, with newly propose attention-based deep learning techniques; **ii** modeling object-level visual attention in static 360° panorama with new attention-based deep learning model; **iii** modeling object-level audio-visual attention in dynamic 360° panorama, which mimics real human attention in real-life scenes and owns potential for real-life applications.

Therefore, the contributions of this thesis are summarized as follows:

Reviews. A systematical overview of state-of-the-art methods towards fixation-level and object-level human attention modeling in both 2D and 360° domains. A thorough review about state-of-the-art attention models in the field of computer vision. The overviews are presented in Chapter 2.

360° image-/video-based datasets&benchmarks. A new benchmark dataset and comprehensive benchmark studies towards 360° image salient object segmentation, which is detailed in Chapter 3. It is worth noting that we include both the object-/instance-level pixel-wise ground truth in our newly proposed dataset. In addition, a new benchmark dataset and comprehensive benchmark studies towards 360° video-based salient object segmentation. Specially, we consider both audio and visual cues to construct the dataset, thus better mimicking the real-world scenario. Also, this part of works is included in Chapter 3.

New methodologies towards light field salient object segmentation. Light field salient object segmentation is a relatively new area and, being similar to 360° salient object segmentation, is of great importance for industrial augmented reality applications. To this end, we explore the state-of-the-art light field salient object segmentation methods and further propose new deep learning models. The key components of our new models include varying attention mechanisms for multi-modal feature fusion. The works are detailed in Chapter 4.

New methodologies towards panoramic salient object segmentation. To further fill the blank of the field of 360° salient object segmentation, we respectively propose new baseline models to conduct salient object segmentation in 360° images and videos. The 360° image-based baseline takes advantage of attention mechanisms for effective feature fusion based on 360° multi-view-based visual cues. The video-based baseline uses both auditory and visual cues to spot targets among frames of a given sequence. Extensive qualitative/quantitative results have been conducted to verify the effectiveness as well robustness of the proposed methods. Please refer to Chapter 5 for details of the works.

[4] *a.k.a.* 360° images, videos or audio-visual dynamic scenes

1.3 Outline of the thesis

Therefore, the next chapter reviews various types of representative salient object segmentation related benchmark datasets and methodologies. Besides, based on our observations towards these recent state-of-the-art methods, we further summarize basic attention models used in not only salient object segmentation but also general computer vision tasks, to establish solid theoretical and empirical foundations for the following works of the thesis. The third chapter presents new benchmark datasets and comprehensive studies regarding 360° object-level human attention modeling. The forth chapter details our works towards light field salient object segmentation. The fifth chapter further introduces our works towards building new baselines for conducting salient object segmentation in 360° images and videos, respectively.

As a conclusion, this thesis has successfully achieved salient object segmentation in both 360° panorama and light field. New datasets and attention-based baselines have been proposed for effective segmentation of salient objects in both static and dynamic panoramic scenes. Besides, new attention models have also been proposed for accurate segmentation of salient objects in light field. As one of the starting points of immersive multimedia-based salient target detection, we hope this thesis is able to inspire ideas for future researches in the fields of object segmentation, VR/AR, audio-visual learning and multi-modal learning.

Chapter 2

Background

2.1 Introduction

Salient object segmentation (*a.k.a.* salient object detection) has been continually grasping attention from the computer vision community in the past decades [15, 16]. As shown in Fig. 2.1, commonly seen image&video segmentation tasks including instance segmentation [4] where all instance-level entities are pixel-wisely outlined, semantic segmentation [34, 36] where all image/video pixels are annotated with specific object-level labels, panoptic segmentation [37] where all image/video pixels are annotated with specific instance-level labels, and generic object segmentation [38] as shown in the second row of Fig. 2.2 where all foreground and background objects are annotated. Being different to above traditional segmentation tasks, salient object segmentation aims to finely segment the objects constantly grasping visual attention, thus being regarded as an interdisciplinary area of human perception and object segmentation. On the other hand, the task of salient object segmentation is also closely related to saliency prediction (*a.k.a.* fixation prediction) [11], where specific regions appealing human attention are detected (The third row of Fig. 2.2). Importantly, the task of salient object segmentation focuses on the regions that are not only salient but also explainable from a perspective of cognitive vision.

The following sections introduce the related tasks in details, and discuss the connections between some of these tasks and our main focus, *i.e.*, salient object segmentation in 360° panoramic images and videos. Specifically, this chapter first reviews the classical human attention modeling task, *i.e.*, saliency prediction (*a.k.a.*, fixation-level attention modeling). Further, this chapter overviews various derivative tasks, which have recently witnessed a prosperous development of salient object segmentation community. Besides, this chapter also overviews current state-of-the-art attention models in the field of general computer vision. Finally, we coolect the formulations of current widely used salient object segmentation metrics. Note the metrics are used for a variety of downstream tasks related to salient object segmentation (*e.g.*, light field/panoramic salient object segmentation). The aim of these reviews is to provide context towards the works included in the following chapters, thus establishing a solid foundation for the thesis.

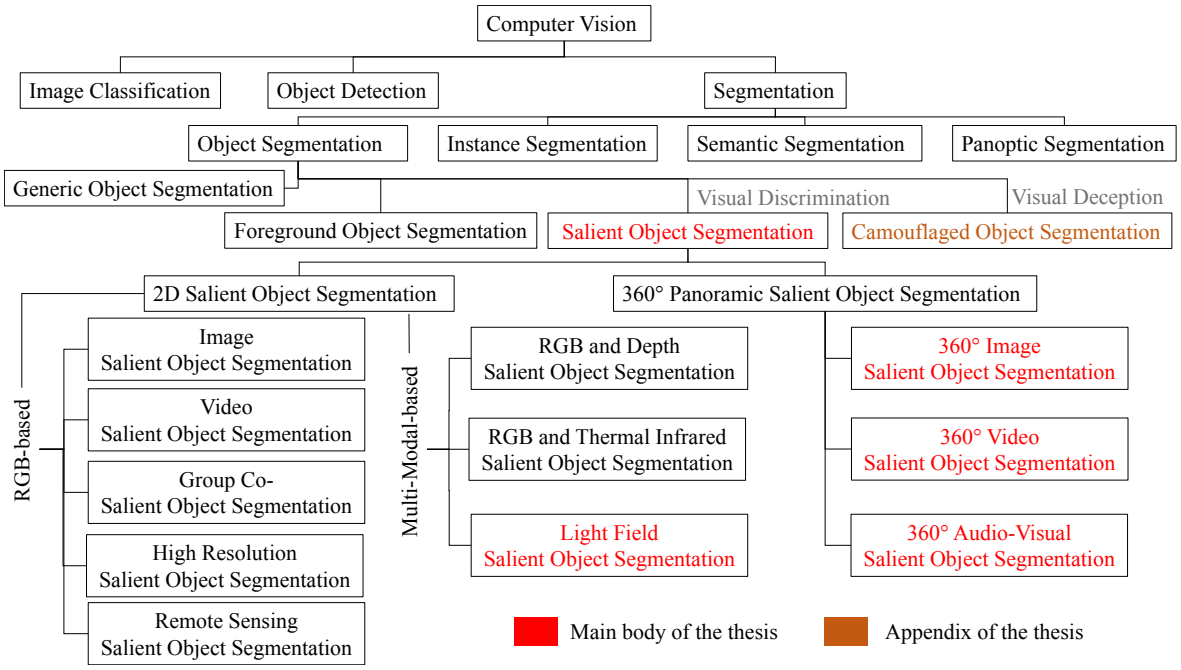


Fig. 2.1 The background of the task of salient object segmentation. Besides, the main focus of this thesis is highlighted.

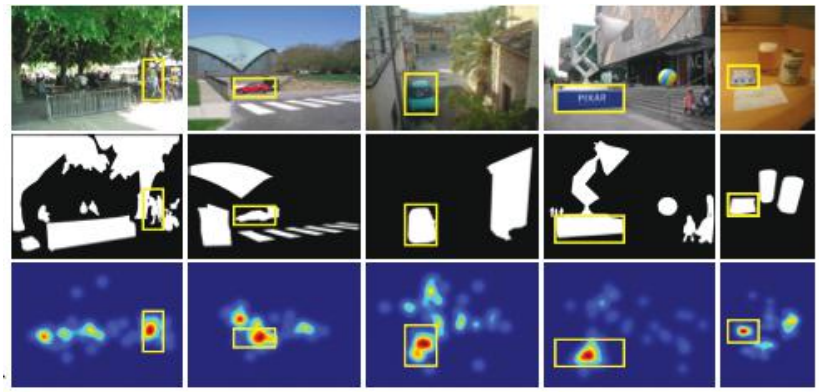


Fig. 2.2 A comparison of the tasks of salient object segmentation, generic object segmentation and saliency prediction. The salient objects are annotated with bounding boxes at the first and the second row. The saliency maps are listed at the third row. This figure is cited from [39].

2.2 Saliency prediction

The community has witnessed a significant development of visual saliency prediction methodologies, thanks to the booming trend of deep learning and large-scale annotations (*e.g.*, fixations gained by conducting eye-tracking experiments supported by widespread consumer-level Head-Mounted Displays (HMDs)). This section reviews the widely used datasets in recent two decades and advanced deep learning methods in the last few years.

2.2.1 Saliency prediction in 2D images/videos

2D image/video-based saliency prediction has been attracting attention from the research community during the past years (*e.g.*, basic information regarding the widely used 2D images/videos saliency prediction benchmark datasets can be found at MIT/Tübingen Saliency Benchmark[1]). This section briefly reviews recent development towards 2D-based saliency prediction from the aspects of widely used datasets and representative methodologies.

Dataset	Citation	Images	Observers	Tasks	Durations	Extra Notes
CAT2000	Ali Borji, Laurent Itti. CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research [CVPR 2015 workshop on "Future of Datasets"]	**4000** images from 20 categories *size:* 1920x1080px 1 dva* ~ 38px	**24** per image (120 in total) *ages:* 18-27	free viewing	5 sec	This dataset contains two sets of images: train and test. Train images (100 from each category) and fixations of 18 observers are shared but 6 observers are held-out. Test images are available but fixations of all 24 observers are held out. *eyetracker:* EyeLink1000 (1000Hz)
EMOtional attention dataset (EMOd)	Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Juan Xu, Mohan Kankanhalli, Qi Zhao. Emotional Attention: A Study of Image Sentiment and Visual Attention [CVPR 2018] (Spotlight)	**1019** emotion-eliciting images, including 321 emotion-evoking pictures selected from IAPS *size:* 1024x768px	**16** *ages:* 21-35	free viewing	3 sec	The images cover a diversity of emotional contents, arousing various sentiments, such as happiness, excitement, awe, disgust, and fear. Object-level and image-level annotations, code, and CNN models for saliency prediction are available on the project page. *eyetracker:* Eyelink 1000 (1000Hz)
Task-driven Webpage Saliency Dataset	Quanlong Zheng, Jianbo Jiao, Ying Cao, and Rynson W.H. Lau. Task-driven Webpage Saliency [ECCV 2018]	**200**(71 for task I, 120 for task II, 9 for task III) *size:* > 224x224px	**24**	information browsing, form filling, shopping	-	The web pages contain contents of varying densities, from little text to heavy text and from few images to many images.. *eyetracker:* Tobii T60
FIGRIM Fixation Dataset	Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, Aude Oliva. Intrinsic and Extrinsic Effects on Image Memorability [Vision Research 2015]	**2787** natural scenes from 21 different indoor and outdoor scene categories *size:* 1000x1000px 1 dva ~ 33px	**15** average per image (42 in total) *ages:* 17-33	memory task	2 sec	Includes: annotated (LabelMe) objects and memory scores for 630 of the images *eyetracker:* EyeLink 1000 (500 Hz)
Eye Fixations in Crowd (EyeCrowd) data set	Ming Jiang, Juan Xu, Qi Zhao. Saliency in Crowd [ECCV 2014]	**500** natural indoor and outdoor images with varying crowd densities *size:* 1024x768px 1 dva ~ 26px	**16** *ages:* 20-30	free viewing	5 sec	The images have a diverse range of crowd densities (up to 268 faces per image). Annotations available: faces labelled with rectangles; two annotations of pose and partial occlusion on each face. *eyetracker:* Eyelink 1000 (1000Hz)
Fixations in Webpage Images (FiWI) data set	Chengyao Shen, Qi Zhao. Webpage Saliency [ECCV 2014]	**149** webpage screenshots from in 3 categories. *size:* 1360x768px 1 dva ~ 26px	**11** *ages:* 21-25	free viewing	5 sec	Text: 50, Pictorial: 50, Mixed:49 *eyetracker:* Eyelink 1000 (1000Hz)
VIU data set	Kathryn Koehler, Fei Guo, Sheng Zhang, Miguel P. Eckstein. What Do Saliency Models Predict? [JoV 2014]	**800** natural indoor and outdoor scenes *size:* max dim: 405px 1 dva ~ 27px	**100,22,20,36** *ages:* 18-23	explicit saliency judgement, free viewing, saliency search, cued object search	until response, 2 sec, 2 sec	*eyetracker:* Eyelink 1000 (250Hz)
Object and Semantic Images and Eye-tracking (OSIE) data set	Juan Xu, Ming Jiang, Shuo Wang, Mohan Kankanhalli, Qi Zhao. Predicting Human Gaze Beyond Pixels [JoV 2014]	**700** natural indoor and outdoor scenes, aesthetic photographs from Flickr and Google *size:* 800x600px 1 dva ~ 24px	**15** *ages:* 18-30	free viewing	3 sec	A large portion of images have multiple dominant objects in the same image. Annotations available: 5,551 segmented objects with fine contours; annotations of 12 semantic attributes on each of the 5,551 objects *eyetracker:* Eyelink 1000 (2000Hz)
VIP data set	Keng-Teck Ma, Terence Sim, Mohan Kankanhalli. A Unifying Framework for Computational Eye-Gaze Research [Workshop on Human Behavior Understanding 2013]	**150** neutral and affective images, randomly chosen from NUSEF dataset	**75** *ages:* undergrads, postgrads, working adults	free viewing, anomaly detection	5 sec	Annotations available: demographic and personality traits of the viewers (can be used for training trait-specific saliency models) *eyetracker:* SMI RED 250 (120Hz)
Salient360	Yashas Rai, Jesus Gutierrez, and Patrick Le Callet. A dataset of head and eye movements for 360 degree images. Proceedings of the 8th ACM on Multimedia Systems Conference, 2017.	**60** 360 deg images from 5376x2688px 18332x9166px	**40** *ages:* 19-52	free viewing	25 sec	This dataset was acquired based on the images displayed on the head mounted display (HMD) Oculus-DK2. Eye gaze data was captured from a Sensomotoric Instruments (SMI) sensor in the HMD, which transmitted eye-tracking data binocularly at 60Hz. Images come from different categories such as, cityscapes, small rooms, scenes containing human faces, great halls and natural landscapes. *eyetracker:* SMI Sensor in HMD (60Hz)

Fig. 2.3 Statistics of widely used 2D image saliency prediction datasets. This figure is cited from [40].

Image Datasets: The widely used benchmark datasets for image-based saliency prediction including MIT300 [41], CAT2000 [7], SALICON [42] and iSUN [43]. It is worth noting that all these datasets provide per-image fixation map as ground truth, to enable the training of fully supervised deep learning algorithms. Key information in terms of these image datasets is shown in Fig. 2.3. It is also worth

[1] `https://saliency.tuebingen.ai/datasets.html`

Dataset	Citation	Videos	Observers	Tasks	Durations	Extra Notes
Coutrot Database 1	Antoine Coutrot, Nathalie Guyader. How saliency, faces, and sound influence gaze in dynamic social scenes [JoV 2014] Antoine Coutrol, Nathalie Guyader. Toward the introduction of auditory information in dynamic visual attention models [WIAMIS 2013]	60 videos *size:* 720x576px	72 *ages:* 20-35	free viewing	average: 17 sec	4 categories: one moving object, several moving objects, landscapes, people having a conversation. Each video has been seen in 4 auditory conditions. *eyetracker:* EyeLink 1000 (1000 Hz)
Coutrot Database 2	Antoine Coutrot, Nathalie Guyader. An efficient audiovisual saliency model to predict eye positions when looking at conversations [EUSIPCO 2015]	15 videos *size:* 1232x504px	40 *ages:* 22-36	free viewing	average: 44 sec	Videos of 4 people having a meeting. Each video has been seen in 2 auditory conditions (with and without the original soundtrack). *eyetracker:* EyeLink 1000 (1000 Hz)
SAVAM	Yury Gitman, Mikhail Erofeev, Dmitriy Vatolin, Andrey Bolshakov, Alexey Fedorov. Semiautomatic Visual-Attention Modeling and Its Application to Video Compression [ICIP 2014]	41 videos *size:* max dim: 1920px, other dim: 1080px	50 *ages:* 18-56	free viewing	average: 20 sec	Left and right stereoscopic views available for all sequences. Nevertheless, only the left view was demonstrated to observers. *eyetracker:* SMI iViewXTM Hi-Speed 1250 (500Hz)
Large-scale eye-tracking database of videos (LEDOV)	Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, Zulin Wang. DeepVS: A Deep Learning Based Video Saliency Prediction Approach [ECCV 2018]	538 videos	32 *ages:* 20-56	free viewing	average: 12 sec	This dataset includes diverse content, containing a total of 179,336 frames, 6,431 seconds, and 5,058,178 fixations. The diverse content refers to the daily action, sports, social activity and art performance of human, and the videos of animal and man-man objects are also included. All videos are at least 720p resolution and 24 Hz frame rate. *eyetracker:* Tobii TX300 (1000Hz)
Dynamic Human Fixation (DHF1K)	Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, Ali Borji. Revisiting Video Saliency: A Large-scale Benchmark and a New Model [CVPR 2018]	1000 videos size: 640x360px	17 *ages:* 20-56	free viewing	average: 20 sec	It comprises a total 1,000 video sequences with 582,605 frames (covering a wide range of scenes, motions, activities) with total duration of 19,420 seconds. The dynamic stimuli were displayed on a 19 inch display (resolution 1440 x 900). *eyetracker:* SMI RED 250 (250Hz)
DepthAware Video Saliency Dataset	George Leifman, Dmitry Rudoy, Tristan Swedish, Eduardo Bayro-Corrochano, Ramesh Raskar. Learning Gaze Transitions from Depth to Improve Video Saliency Estimation [ICCV 2017]	54 videos	91 *ages:* 20-67	free viewing	average > 20 sec	This dataset includes video sequences of static and dynamic scenes, acquired by static and dynamic sensors, doors and outdoors. Videos vary in durations ranging from 25 to 200 seconds. They are converted to a 30 fps, resulting in approximately 100K frames across all videos. *eyetracker:* Gazepoint GP3 Eye Tracker (60Hz)
MUltiple-Face Videos with Eye Tracking fixations (MUFVET-I)	Yufan Liu, Songyang Zhang, Mai Xu, and Xuming He. Predicting Salient Face in Multiple-face Videos [CVPR 2017]	65 videos	39 *ages:* 20-49	free viewing	varying 10-20 sec	This is the first eye tracking database for multiple-face videos. In total, 1,252,822 fixations of all 39 subjects on 65 videos are obtained. All videos in MUFVET are with either indoor or outdoor scenes, selected from Youtube and Youku, and they are all encoded by H.264. *eyetracker:* Tobii X2-60 (60Hz)
MUltiple-Face Videos with Eye Tracking fixations (MUFVET-II)	Yufan Liu, Songyang Zhang, Mai Xu, and Xuming He. Predicting Salient Face in Multiple-face Videos [CVPR 2017]	100 videos	36 *ages:* 20-55	free viewing	varying 10-20 sec	It includes 1,737,826 fixations acquired from all 36 subjects in this dataset,. *eyetracker:* Tobii TX300
Sal-360	Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. Saliency Detection in 360 Videos [ECCV 2018]	104 videos	27 *ages:* 20-24	free viewing	avergae > 20 secs	Videos were selected from the Sports-360 dataset. The video contents involve five sports (basketball, parkour, BMX, skateboarding, and dance), and the duration of each video is between 20 and 60 seconds. Each video is watched by at least 20 volunteers. The total time used for data collection is about 2000 minutes. *eyetracker:* '7invensun a-Glass' eye tracker embedded into the HMD
DR(eye)VE	Andrea Palazzi, Davide Abati, Simone Calderara, Francesco Solera, and Rita Cucchiara. Predicting the Driver's Focus of Attention: the DR(eye)VE Project [ArXiv 2018]	74 videos	8 *ages:* 20-40	driving	5 min	Videos were recorded in different contexts, both in terms of landscape (downtown, countryside, highway) and traffic condition, ranging from traffic-free to highly cluttered scenarios, in diverse weather conditions (sunny, rainy, cloudy) and at different hours of the day (both daytime and night). It contains 555,000 frames. A frontal camera acquired the scene at 720p/30fps, and users pupils were tracked. *eyetracker:* SMI ETG 2w Eye Tracking Glasses (ETG)- 60Hz

Fig. 2.4 Statistics of widely used 2D video saliency prediction datasets. This figure is cited from [40].

noting that, so far the biggest image-based saliency prediction dataset, *i.e.*, SALICON, owns about 10K images of training set.

Video Datasets: The commonly used benchmark datasets for video-based saliency prediction including DIEM [44], HOLLYWOOD-2 [45], UCF-Sports [45], DHF1K [11], LEDOV [12], Coutrot [46], MUFVET [47] and salient-KITTI [48]. As shown in Fig. 2.4, DHF1K is so far the largest video dataset, where about 1K videos are included. Besides, as the development of deep learning based multi-modal learning, audio-visual saliency prediction dataset such as Coutrot [46] has been proposed.

Methodologies: During the past years, convolutional neural networks (CNNs) have been the mainstream of framework designing in the field of saliency prediction [40]. In this case, this section summarizes the representative CNNs-based methods proposed in the last few years. Early method such as SalEMA [49] used VGG-based encoder [2] to extract spatial features of given video frames and applied ConvLSTM [14] to model temporal information among video frames. DINet [50] took advantage of dilated CNNs to expand receptive field of the framework. STRA-Net [51] designed residual attention block to conduct feature refinement based on ConvGRU [14] framework. DeepUSPS [52] applied hand-crafted method to generate pseudo-labels for the self-supervision based saliency prediction framework. Sal-DCNN [53] designed multiple decoders corresponding to multi-domain features. UAVD [54] analyzed attention maps of different layers of VGG-based CNNs. St-Net [56] proposed

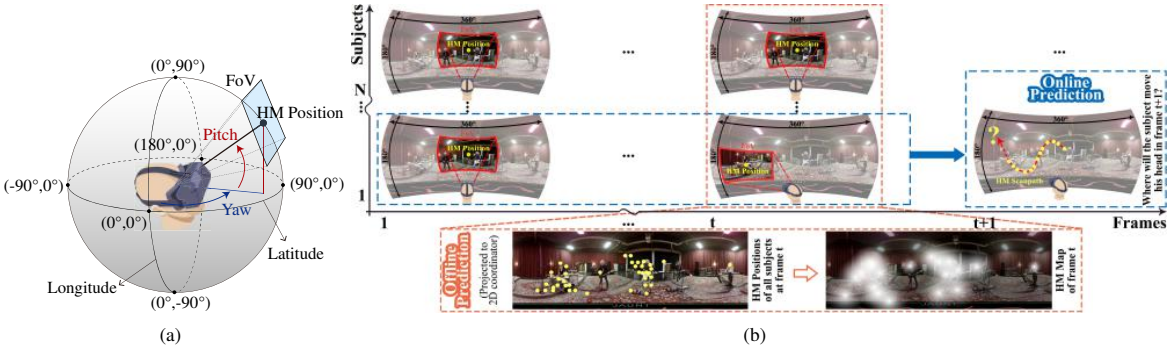
Table 2.1 Summary of recent saliency prediction methods. I/V/AV-Sal means image/video/audio-visual-based saliency prediction methods, respectively.

Method	Modality	Year	Publication	Key Words
SalEMA [49]	V-Sal	2019	BMVC	ConvLSTM, SOTA on DHF1K.
DINet [50]	I-Sal	2019	TMM	Dilated convolution, inception module.
STRA-Net [51]	V-Sal	2019	TIP	ConvGRU, residual attentive learning, attention.
DeepUSPS [52]	I-Sal	2019	NeurIPS	Self-supervised learning, pseudo-labels generation.
Sal-DCNN [53]	I-Sal	2019	AAAI	Noise-added phase spectrum, multi-domain decoder.
UAVD [54]	I-Sal	2019	CVPR	Hierarchical feature visualization.
DAVE [55]	AV-Sal	2019	arXiv	3D ResNet for audio-visual encoding.
St-Net [56]	V-Sal	2020	TIP	spatial-temporal feature fusion, LSTM.
SalSAC [57]	V-Sal	2020	AAAI	ConvLSTM attention module, SOTA on DHF1K.
FastSal [58]	I-Sal	2020	ICPR	MobileNetV2, distillation networks.
STAViS [59]	AV-Sal	2020	CVPR	SOTA on Controt, bilinear layer for audio-visual fusion.
SF-Net [60]	AV-Sal	2020	ECCV	Salient face detection, SOTA on MUFVET, LSTM.
MMS [61]	AV-Sal	2020	TIP	Cross-modal kernel canonical correlation analysis.
DeepGaze [62]	I-Sal	2021	ICCV	Out-of-domain prediction, complementarity analysis.
STANet [63]	AV-Sal	2021	CVPR	Class activation mapping, weakly-supervised learning.
DAVNet [64]	AV-Sal	2021	ICIP	Feature pyramid module.
AViNet [65]	AV-Sal	2021	IROS	SoundNet block, trilinear interpolation, 3D CNNs.
GASP [66]	AV-Sal	2021	IJCAI	Attention mechanism, recurrent gated multi-modal unit.
HD2S [67]	V-Sal	2021	IJCV	Domain adaptive learning, domain-specific learning.
WeakFix [68]	I-Sal	2022	TIP	Weakly supervised learning, object proposal, attention.
EEEA-Net [69]	I-Sal	2022	TII	Knowledge distillation, pseudo-labels.

an attention-aware ConvLSTM to mine the temporal features of inputting sequences. HD2S [67] added conspicuity modules to fuse multi-scale features extracted from CNNs. WeakFix [68] modeled visual attention competition mechanism via softmax based attention modules. EEEA-Net [69] created a teacher-student framework via pseudo-knowledge distillation. Besides, as the development of multi-modal learning, combining the auditory and visual cues (Table. 2.1) to gain more realistic human attention modeling has become a new trend in the saliency prediction community. Among these methods, a variety of attention-based modules (STAViS [59], STANet [63], GASP [66], AViNet [65]) have been proposed. DAVE [55] is the pioneer work which simply concatenated the audio and visual features extracted from separate 3D CNNs. STAViS [59] further designed a deeply supervised attention module to facilitate the audio-visual fusion. MMS [61] used cross-modal kernel canonical correlation to quantify audio based saliency maps. STANet [63] designed three attention modules for the fusion of spatial-temporal features, spatial-audio features and spatial-temporal-audio features, respectively. Most recently, GASP [66] proposed gate attention for multi-modal late fusion. AViNet [65] explored bilinear based fusion for the features extracted from 3D CNNs based visual encoder and 1D CNNs based SoundNet [70]. Please note that a detailed statistics of recent audio-visual saliency prediction methods is presented in Table. 2.2.

Table 2.2 Detailed comparison of audio-visual saliency prediction methods. † indicates non-deep learning audio encoding method. CNNs denotes convolutional neural networks.

Method	Video Type	Year	Audio Type	Audio Encoder	Audio pre-traning dataset
DAVE [55]	2D	ArXiv'19	Mono and stereo sound	3D CNNs	DAVE [55]
STAViS [59]	2D	CVPR'20	Mono sound	1D CNNs [70]	Flickr [70]
SF-Net [60]	2D	ECCV'20	Mono and stereo sound	3D CNNs	MVVA [60]
†MMS [61]	2D	TIP'20	Mono sound	‡	‡
STANet [63]	2D	CVPR'21	Mono and stereo sound	2D CNNs	AVE [71]
DAVNet [64]	2D	ICIP'21	Mono sound	1D CNNs [70]	Flickr [70]
AViNet [65]	2D	IROS'21	Mono sound	1D CNNs [70]	Flickr [70]
GASP [66]	2D	IJCAI'21	Mono and stereo sound	3D CNNs	DAVE [55]
†PO-AVS [72]	360°	AI'20	Ambisonics	‡	‡
AVS360 [73]	360°	VCIP'20	Mono sound and ambisonics	3D CNNs	DAVE [55]
†360-SSSL [74]	360°	MVA'21	Ambisonics	‡	‡



(a) (b)

Fig. 2.5 An introduction of 360° panoramic saliency prediction (this figure is cited from [28]). (a) A subject freely explores 360° content with Head-Mounted Display (HMD). (b) Fixations of all subjects shown in 360° video frames.

2.2.2 Saliency prediction in 360° images/videos

As the popularization of 360° cameras such as Insta360 series[2], Ricoh Theta Z1 and GoPro Max, it becomes easy to obtain large-scale 360° panoramic images and videos (with the field-of-view (FoV) of 360°×180° illustrated in Fig. 2.5), which are able to be displayed with Head-Mounted Displays (HMDs). Therefore, in the recent few years, several 360° image and video datasets (as shown in Table. 2.3) have been established for saliency prediction in panorama. The wide FoV gives panoramic saliency prediction a huge potential to mimic human visual attention in real-life daily scenes, making it a popular topic not only in computer vision society but also wide interdisciplinary area of virtual reality, augmented reality and mixed reality industries. Table. 2.3 collects detailed information regarding the eye-tracking experiments of each of the widely used 360° panoramic saliency prediction datasets.

As the development of large-scale 360° saliency prediction benchmark datasets such as VR-scenes [76] and PVS-HMEM [28], deep learning methods have enjoyed increasing attention from the community. Some most recently proposed representative methods are collected in Table. 2.4. Specif-

[2] https://www.insta360.com/fr/

Table 2.3 Summary of 360° saliency prediction datasets.

Dataset	Citation	Images/Videos	Observers	Tasks
Salient!360 [25]	Rai, Yashas, etc. A dataset of head and eye movements for 360 images. MMSys '18, 2017. ACM.	98 images indoor/outdoor 3840×1920 without audition	63	Free viewing with HMDs with eye tracker
SaliencyVR [26]	Sitzmann, Vincent, etc. Saliency in VR: How do people explore virtual environments? TVCG, 2018.	22 images 14/8 indoor/outdoor 8196×4092 without audition	169	Free viewing with HMDs with eye tracker
Salient!360V2 [75]	E. J. David, etc. A dataset of head and eye movements for 360 videos. MMSys '18, 2018. ACM.	19 videos 9/10 indoor/outdoor 3840×1920 without audition	57 age 19-44 25F/32M	Free viewing with HMDs with eye tracker
VR-scene [76]	Y. Xu, Y. Dong, J. Wu, Z. Sun, etc. Gaze prediction in dynamic 360° immersive videos CVPR 2018	208 videos indoor/outdoor 3840×1920 with audio	45 age 20-24 20F/25M	Free viewing with HMDs with eye tracker
360saliency [77]	Z. Zhang, Y. Xu, J. Yu, and S. Gao Saliency Detection in 360° Videos ECCV 2018	104 videos Sports scenes from [29] 3840×2160 without audio	27 age 20-24	Free viewing with HMDs with eye tracker
VQA-ODV [78]	C. Li, M. Xu, X. Du, and Z. Wang. Bridge the gap between vqa and human behavior on omni-video: A large-scale dataset and a deep learning mode ACM MM 2018	60 videos various scenes no more than 8K without audio	221 age 19-35 78F/143M	Free viewing with HMDs with eye tracker
PVS-HMEM [28]	Mai, Xu, etc. Predicting Head Movement in Panoramic Video: A Deep Reinforcement Learning Approach TPAMI 2018	76 videos various scenes ranking from 3K to 8K without audio	58	Free viewing with HMDs with eye tracker
AVP-360 [79]	Fang-Yi Chao, etc. Audio-visual perception of omnidirectional video for virtual reality applications ICMEw 2020	15 videos music/conversation 3840×1920 with audio	45	Free viewing with HMDs with eye tracker

ically, Cube360 [80] developed cube-padding technique to facilitate the 360° saliency prediction by using 2D CNNs. SalGAN360 [81] proposed cube-map based augmentation technique to improve the CNNs' performance on 360° saliency prediction benchmarks. MT-DCNN [82] applied ConvLSTM to learn temporal features for viewport alignment. SalGCN proposed spherical graph CNNs to encode 360° images. SalFOOL [84] explored the robustness of current metrics for 360° saliency model evaluation. SalGAIL [85] applied generative adversarial learning module to learn the reward of the deep reinforcement learning module, to predict fixations and trajectories. ATSal [86] designed attention stream to model global saliency. MultiVUS [87] further applied self-attention mechanism to fuse and refine the features extracted from augmented equirectangular images. Most recent, ScanGAN360 [88] developed conditional generative adversarial neural network to predict scanpath and fixations. SPVP360 [89] used spherical CNNs to conduct viewport localization.

It is worth noting that, recent methods such as PO-AVS [72], AVS360 [73] and 360-SSSL [74] have combined both audio and visual cues of the 360° videos reflecting real-life daily scenes. These

Table 2.4 Summary of recent 360° saliency prediction methods. PI/PV/PAV-Sal means image/video/audio-visual-based 360° panoramic saliency prediction methods, respectively. † denotes non-deep learning method.

Method	Modality	Year	Publication	Key Words
Cube360 [80]	PV-Sal	2018	CVPR	Cube-padding, weakly-supervised.
SalGAN360 [81]	PI-Sal	2018	ICMEw	GAN, cube map.
MT-DCNN [82]	PV-Sal	2020	TMM	Viewport localization, object detection, multi-task learning.
SalGCN [83]	PI-Sal	2020	ACM MM	Spherical graph convolutional neural networks.
AVS360 [73]	PAV-Sal	2020	VCIP	3D CNNs, audio energy map, bottleneck fusion.
SalFOOL [84]	PV-Sal	2021	ICCV	360° metric robustness, KL-divergence.
SalGAIL [85]	PI-Sal	2021	TIP	Generative adversarial imitation learning.
ATSal [86]	PV-Sal	2021	ICPR	Attention mechnism, SOTA on Salient!360.
MultiVUS [87]	PI-Sal	2021	ICCV	Mutual information learning, contrastive learning.
†PO-AVS [72]	PAV-Sal	2021	AI	Multi-sensory integration, proto-objects.
†360-SSSL [74]	PAV-Sal	2021	MVA	MFCC, late fusion of audio-visual information.
ScanGAN360 [88]	PI-Sal	2022	TVCG	Generative model. scanpath prediction, multi-task learning.
SPVP360 [89]	PV-Sal	2022	TOMM	Spherical convolutional neural networks, video multi-cast.

methods took advantage of multi-modal inputs to advance saliency prediction towards multimedia applications, also to mimic more realistic human attention which is indeed influenced by both audio and wide FoV based visual information. Considering the prospective researches regarding audio-visual learning, Table. 2.2 further summarizes the details of recent representative audio-visual saliency prediction frameworks. So far, the mainstream of audio-visual models' structures are still based on CNNs. Besides, an interesting finding is that methods (*e.g.*, [59, 73]) taking advantage of both audio and visual cues tend to show better performance on widely used public 2D/360° video saliency detection benchmarks, respectively.

In conclusion, the task of 360° image/video based saliency prediction has experienced a boost. However, the objective of the task dose not explain human attention towards object-level recognition. In other words, current 2D/360° saliency prediction methods do not convey the concept of obejct-level saliency, thus being far from AR/VR applications where the detection of objects that grasp human attention are important.

2.3 Salient object segmentation in 2D RGB domain

Based on the attributes/modalities of inputting data, 2D RGB based salient object segmentation (*a.k.a.* salient object detection (SOD)) can be classified into five categories, *i.e.*, image-based salient object segmentation (I-SOD), video-based salient object segmentation (V-SOD), group (co-) salient object segmentation (Co-SOD), high-resolution salient object segmentation (HR-SOD) and remote sensing salient object segmentation (RS-SOD). The aim of the above mentioned classifications is to make the salient object segmentation methods to adapt to specific application scenario. Therefore, different types of salient object segmentation methods may focus on different challenges in terms of object segmentation. For instance, image or video based salient object segmentation methods use merely RGB 2D data to predict the finely structure of salient objects in commonly seen 2D images/videos. Group salient object segmentation aims at simultaneously locating the objects (belonging to specific category) in a group of images, thus emphasizing the intrinsic features representing specific object categories, and are robust to appearance changes of each identical object category. Further, high-resolution salient object segmentation methods use high-resolution images as inputs and take advantage of both local and global spatial information for finely segmentation of large objects. On the contrary, remote sensing salient object segmentation focuses on salient object segmentation in photos captured by remote sensors. Therefore, multiple extremely small salient objects at low resolution may be collected in remote sensing salient object segmentation datasets. The commonly used datasets and representative methodologies of each category of 2D RGB based salient object segmentation are detailed in the following sections.

2.3.1 Image-based salient object segmentation

Table 2.5 Summary of widely used image-based salient object segmentation datasets. #Img: The number of images/video frames. #GT: The number of object-level pixel-wise masks. Obj.-Level = Object-Level Labels. Ins.-Level = Instance-Level Labels. Fix. GT = Fixation Maps.

Dataset	Publication	#Img	#GT	$\texttt{min}(W,H)$	$\texttt{max}(W,H)$	Obj.-Level	Ins.-Level	Attribute	Fix. GT
ECSSD [90]	CVPR'13	1,000	1,000	139	400	✓			
DUT-O [17]	CVPR'13	5,168	5,168	139	401	✓			✓
PASCAL-S [18]	CVPR'14	850	850	139	500	✓			✓
HKU-IS [19]	CVPR'15	4,447	4,447	100	500	✓			
DUTS [20]	CVPR'15	15,572	15,572	100	500	✓			
ILSO [91]	CVPR'17	1,000	1,000	142	400	✓	✓		
SOC [21]	ECCV'18	6,000	6,000	161	849	✓	✓	✓	

The widely used datasets for image-based salient object segmentation are shown in Table. 2.5. As shown in Table. 2.5, the commonly used image salient object segmentation datasets are relatively small when compared to widely used image classification datasets such as ImageNet-1K [1] which collects nearly 14 million 2D RGB images, since gathering manually labeled pixel-wise ground truth for each of the salient objects is surely a time-consuming and laborious process. The largest image salient object segmentation dataset, *i.e.*, DUTS [20], contains about 10K images for training and 5K images for testing. Although the task of salient object segmentation only focuses on object-level

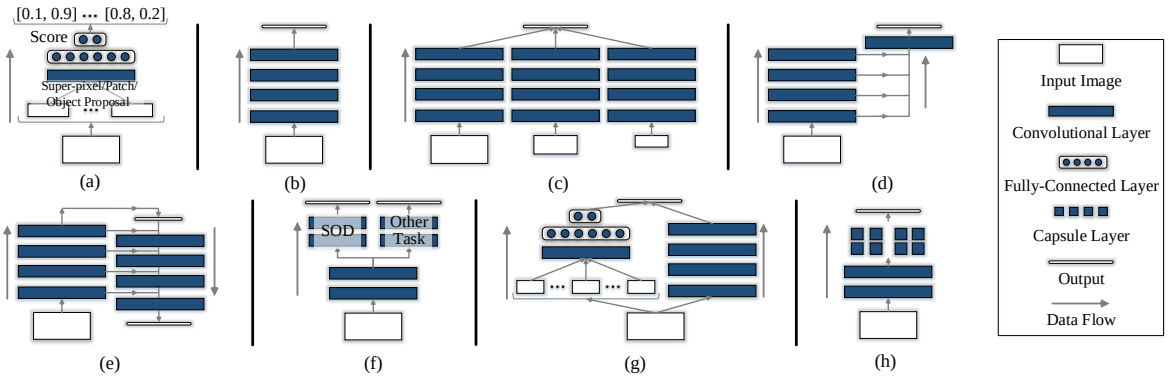


Fig. 2.6 Commonly seen structures of image-based salient object segmentation methods, this figure is cited from [16]. (a) denotes multi-layer perception architecture. (b)-(f) are all convolutional neural network based frameworks. Specifically, (b) single stream model w/o hierarchical decoder, (c) Siamese encoder, (d) side-out fusion model, (e) U-Net like structure w/ top-down and bottom-up multi-scale fusion guidance, (f) Multi-task based decoder ensembles. (g) multi-supervision based multi-branch structure. (h) capsule based decoder.

binary segmentation, datasets such as ILSO [91] and SOC [21] also provide instance-level pixel-wise masks as ground truth for salient instance segmentation. And it is worth noting that only two datasets (*i.e.*, [17,18]) defined salient objects according to fixations gained by conducting eye-tracking experiments. The other datasets directly regard the main objects located in/near the image center as the salient ones. For detailed statistics of image salient object segmentation datasets, please refer to Table. 2.5.

As for image salient object segmentation methodologies, there are hundreds of deep learning methods proposed during the past decade. The key concepts regarding most recent methods are collected in Table. 2.6 and Table. 2.7. The extra review regarding the previous image salient object segmentation methodologies are detailed in [15, 16].

Generally, recent image salient object segmentation methods are almost deep learning based and are varying in terms of not only learning paradigms (*i.e.*, fully-/weakly-/un-supervised learning) but also backbone structures (*e.g.*, ResNets [142] and vision transformers [143]). Besides, image salient object segmentation methods tend to focus on the designing of feature decoding stage (Fig. 2.6), in order to improve model performance on multiple segmentation benchmarks. Specifically, the classical state-of-the-art methods such as PoolNet [140], EGNet [141], MINet [126] all applied U-Net [144] like encoder-decoder structure for the task. F3Net [122] designed side-out fusion mechanism to facilitate the fusion of hierarchical features from backbone network. More recent methods such as CSF [129], SAMNet [105] proposed light-weight architectures containing less parameters however showing comparable performance to the regular U-Net like models. Besides pursuing state-of-the-art performance, most recently proposed image salient object segmentation method such as UPL [92] explored the uncertainty of model predictions thus advancing the field of salient object segmentation towards explainable deep learning. Method such as DCFD [93] explored intrinsic features from a perspective of causal inference, thus advancing the salient object segmentation towards robust modeling. In addition, energy-based model such as SalCNet [94] tried to combine the energy-based gener-

Table 2.6 Summary of recent 2D image based salient object segmentation methods in years of 2021-2022.

Method	Year	Publication	Key Words
UPL [92]	2022	AAAI	Consistency based uncertainty estimation, pseudo labels.
DCFD [93]	2022	AAAI	Confounding biases, de-confounded training.
SalCNets [94]	2022	AAAI	Energy-based model, latent variable model.
NSS [95]	2022	AAAI	Adaptive flood filling, transformer bottleneck.
TRACER [96]	2022	AAAI	Masked-edge attention module, explicit edge loss.
RCSBNet [97]	2022	WACV	Recursive contour-saliency Blocks.
MBAB [98]	2022	TOMM	Detail modeling and body filling as sub-tasks.
NSAL [99]	2022	TMM	Noise-robust adversarial learning framework/
DRNet [100]	2022	TCSVT	Progressive dual attention mechanism.
SCWS [101]	2021	AAAI	Saliency structure consistency loss.
PFS [102]	2021	AAAI	Feature shrinking module, pyramid structure.
LGSL [103]	2021	AAAI	Knowledge review network with attention-based sampler.
GDC [104]	2021	AAAI	BiLSTM encoder, reader-aware topic modeling.
SAMNet [105]	2021	TIP	Attention, multi-scale, light-weight architecture.
MesSal [106]	2021	CVPR	Mesh saliency, 2D-to-3D correspondence.
MSFNet [107]	2021	ACM MM	Multi-scale fusion.
CCNet [108]	2021	TIP	Decomposition of edge and skeleton priors.
VST [109]	2021	ICCV	Cross modality transformer.
DHQNet [110]	2021	ICCV	High-resolution refinement module.
iNAS [111]	2021	ICCV	Integral search space.
SCA [112]	2021	ICCV	Semantic scene context refining module.
MFNet [113]	2021	ICCV	Multiple pseudo labels filter.
SSL [114]	2021	TIP	Structure similarity loss, purification module.
CTD [115]	2021	ACM MM	Complementary trilateral decoding.
GVT [116]	2021	NeurIPS	Transformer, energy-based genenrative model.
DSR [117]	2021	NeurIPS	Graph neural network, inductive bias.
PSG [118]	2021	TIP	Loss function for multi-scale supervision.
DACNet [119]	2021	TIM	Dense attention mechanisms based feature steering.

ative models and state-of-the-art segmentation architectures for salient object segmentation modeling.

What is worth mentioning is that, some methods listed in Table. 2.6 and Table. 2.7 own attention-based modules designed for feature fusion or feature refinement at the hierarchical decoding layers. For instance, early method AFNet [137] built attention feedback modules between each of the encoding-decoding layers of a U-Net like framework. GateNet [132] proposed gated attention unit based on Sigmoid function. The gated units were added to the paths between each of the current decoder layers and their previous layers. SAGD [124] used channel attention [145] and spatial attention [146] to refine the features before the multi-stage fusion at the decoder. Most recently, VST [109] used transformer layers [143] at the bottleneck of the encoder-decoder framework and thus acquiring improved model performance. DRNet [100] established residual learning module between each pair of attention maps (original map and its reverse), thus benefiting the model's ability of distinguishing the salient regions from non-salient ones. NSS [95] used transformer layers to learn features based on weak supervision of manually labeled points.

Table 2.7 Summary of recent 2D image based salient object segmentation methods in years of 2019-2020.

Method	Year	Publication	Key Words
PFPNet [120]	2020	AAAI	Multiple feature polishing modules.
GCPANet [121]	2020	AAAI	Global context flow module.
F3Net [122]	2020	AAAI	Cascaded feed-back decoder.
ADA [123]	2020	AAAI	Adversarial network for multi-spectral saliency detection.
SAGD [124]	2020	AAAI	Attentional convGRU, recurrent local attention.
ScrSOD [125]	2020	CVPR	Scribble annotation, scribble boosting.
MINet [126]	2020	CVPR	Multi-scale interactive module.
ITNet [127]	2020	CVPR	Light-weight two-stream framework.
LDF [128]	2020	CVPR	Decoupled multi-supervisions.
CSF [129]	2020	ECCV	Res2Net backbone, cross-stage fusion.
nRef [130]	2020	ECCV	Cross domain learning.
NoiseSal [131]	2020	ECCV	Noise label as auxiliary supervision.
GateNet [132]	2020	ECCV	Gate attention mechanism.
DFI [133]	2020	TIP	Object, edge and skeleton learning.
CAGNet [134]	2020	PR	Multi-scale feature extraction.
VPL [135]	2020	TCyb	Light-weight structure, hierarchical perception.
APL [136]	2020	NeurIPS	Adversarial pace learning.
AFNet [137]	2019	CVPR	Boundary priors, attention feedback module.
BASNet [138]	2019	CVPR	Edge priors.
CPD [139]	2019	CVPR	Cascaded partial decoder.
PoolNet [140]	2019	CVPR	Pyramid pooling module.
EGNet [141]	2019	ICCV	Edge guidance.

In conclusion, various image salient object segmentation methodologies are able to introduce basic techniques such as segmentation networks, and experience towards assigning training/testing settings to the 360° panoramic salient object segmentation. However, image salient object segmentation methods are all limited to 2D images representing scenes within a local viewport, from a perspective of omnidirectional vision. Considering the 360°×180° FoV, 360° image processing is indeed more challenging however necessary for the development of AR/VR applications where true human perception must be mimicked.

2.3.2 Video-based salient object segmentation

Table. 2.8 shows detailed statistics in terms of the commonly used video-based salient object segmentation (V-SOD) datasets, including SegTrackV2 [147], FBMS [148], MCL [149], ViSal [150], DAVIS [151], UVSD [152], VOS [22] and DAVSOD [23]. Early datasets such as FBMS [148], MCL [149] and ViSal [150] provide no more than 1K annotated video frames, while later datasets such as DAVSOD [23] provides more than 20K consecutive video frames with manual labels. It is worth noting that all listed V-SOD datasets provide pixel-wise binary masks as ground truth for V-SOD task. Besides, only VOS [22] and DAVSOD [23] strictly followed the guidance of fixations to annotate the salient objects, for the video frames contained in these two datasets own more complex context and multiple foreground/background objects. According to the VOS [22] and DAVSOD [23],

Table 2.8 Summary of video-based salient object segmentation datasets. #Img: The number of images/video frames. #GT: The number of object-level pixel-wise masks. Obj.-Level = Object-Level Labels. Ins.-Level = Instance-Level Labels. Fix. GT = Fixation Maps. Attr. = Attributes.

Dataset	Publication	#Img	#GT	$\min(W,H)$	$\max(W,H)$	Obj.-Level	Ins.-Level	Attr.	Fix. GT
SegTrack V2 [147]	ICCV'13	1,065	1,065	212	640	✓			
FBMS [148]	TPAMI'13	13,860	720	253	960	✓			
MCL [149]	TIP'15	3,689	463	270	480	✓			
ViSal [150]	TIP'15	963	193	240	512	✓			
DAVIS2016 [151]	CVPR'16	3,455	3,455	900	1,920	✓		✓	
UVSD [152]	TCSVT'16	3262	3262	240	877	✓			
VOS [22]	TIP'18	116,103	7,467	312	800	✓			✓
DAVSOD [23]	CVPR'19	23,938	23,938	360	640	✓	✓	✓	✓

Table 2.9 Summary of recent video based salient object segmentation. UVOS = unsupervised video object segmentation. SVOS = semi-supervised video object segmentation. RVOS = referring video object segmentation. VSOD = video salient object detection.

Method	Pub.	Task	Encoder	Decoder	Notes
TransAOT [153]	arXiv'22	SVOS	Swin-B [154]	FPN	Swin transformer.
YOFO [155]	AAAI'22	RVOS	ResNet50 [142]	BERT	Image&language learning.
SITVOS [156]	AAAI'22	SVOS	ResNet50/18	STM [157]	Transformer, Siamese.
EFS [158]	AAAI'22	UVOS	ResNet50	STM [157]	S-measure in model.
RPCM [159]	AAAI'22	SVOS	ResNet101	DeepLabV3	Uncertainty estimation.
CANet [160]	WACV'22	UVOS	ResNet101	PANet [161]	Contrastive learning.
WSV [162]	CVPR'21	VSOD	ResNet50	ConvLSTM	Scribble.
DCFNet [163]	ICCV'21	VSOD	ResNet101	Convs	Context sensitive.
FSNet [164]	ICCV'21	UVOS	ResNet50	PPM [165]	Optical flow.
TranspNet [166]	ICCV'21	UVOS	ResNet50/101	UNet [144]	Sinkhorn module [167].
STINet [168]	TIP'21	VSOD	Convs	UNet	Sinkhorn
AOT [169]	NeurIPS'21	SVOS	MobileNetV2 [170]	FPN [171]	Transformer
STCN [172]	NeurIPS'21	SVOS	ResNet50	STM [157]	L2 similarities.
TENet [173]	ECCV'20	VSOD	ResNet50	Convs	Excitation modules.
SEGCN [174]	TIP'20	VSOD	BASNet [138]	BASNet [138]	GCN
PCSA [175]	AAAI'20	VSOD	MobileNet	RFB [176]	Global attention.
MGA [177]	ICCV'19	VSOD	ResNet101/34	Convs	Optical flow, ASPP.
RCRNet [178]	ICCV'19	VSOD	ResNet50	Ref.Mod.	NER module.
SSAV [179]	CVPR'19	VSOD	ResNet50	Convs	ConvLSTM.
COSNet [180]	CVPR'19	UVOS	DeepLabv3 [181]	Convs	Co-attention.

fixation-guided judgment is one of the reliable evidences for defining salient objects, especially for images/videos representing challenging real-life scenes.

Video-based salient object segmentation methods' statistics are listed in Table. 2.9. Generally, video-based salient object segmentation methods and video object segmentation methods all aim at finely segmenting the visually salient objects among given video frames. Therefore, most of the recently proposed methods (Table. 2.9) were tested on both video based salient object segmentation and video object segmentation benchmarks.

And it is worth noting that, the definition of learning paradigms in the field of video object segmentation are different from the ones in the field of video salient object segmentation. For instance,

weakly-supervised video salient object segmentation methods are those which use a part of the pixel-wise ground truth or other types of ground truth (*e.g.*, scribble, depth) as supervision for the model training process, while weakly-supervised video object segmentation methods are those using the ground truth of the first frame of given sequence to support the model testing process. Further, fully-supervised video salient object segmentation methods can be fairly compared with un-supervised video object segmentation methods. Most recently, method such as YOFO [155] processes multi-modal inputs (language-based and visual-based) to conduct referring video object segmentation.

As for model architecture, being similar to above discussed image based salient object segmentation, recent video methods also apply CNNs (*e.g.*, ResNets [142]) or Transformers (*e.g.*, Swin-B [154]) to extract the features from inputting spatial-temporal cues. Obviously, effectively modeling the temporal information within sequences is one of the main challenges of the field. Methods such as RCRNet [178] and FSNet [164] applied classical optical flow priors to facilitate the task. Besides, ConvLSTM [14] was applied as the basic component of temporal module of methods such as SSAV [23] and WSV [162]. Importantly, video-based salient object segmentation methods such as COSNet [180], PCSA [175], DCFNet [163], WSV [162], SITVOS [156] and TransAOT [153] also used different attention mechanisms to improve model performance. Specifically, COSNet [180] invented co-attention mechanism to fuse and refine the useful features extracted from consecutive video frames. PCSA [175] proposed sequence-based global attention to enhance the learning of moving salient objects among video frames. DCFNet [163] proposed dynamic context-aware filtering module which consists of multiple dynamic filtering units and a Softmax scoring layer. WSV [162] built appearance-motion fusion module that consists of both channel attention [145] and spatial attention [146]. SITVOS [156] utilized multiple transformer layers to build a Siamese framework to learn global temporal features and local spatial features. TransAOT [153] designed a transformer-based association module at the bottleneck, to fuse and refine inter-frame spatial features.

In conclusion, video-based salient object segmentation focuses on the modeling of human attention in 2D videos, by taking advantage of both static and dynamic visual cues. However, the context within 2D videos is far from the one in real daily life, where multiple foreground and background objects are included in an immersive dynamic view (*e.g.*, 360° videos record the natural scenes containing global spatial-temporal context). Therefore, the state-of-the-art methodologies in 2D domain may fail in 360° domain. However, salient object segmentation in panoramic dynamic scenes still lacks of investigation.

2.3.3 Co-salient object segmentation

Group-based or co-salient object segmentation (CoSOD), as a specific branch of 2D RGB-based salient object segmentation, is appealing increasing attention from the community in the past few years. This type of methods pay attention to co-occurring salient objects among a group of given images containing totally different background scenes. Similar to image/video-based salient object segmentation, recent co-salient object segmentation methods also rely on large-scale image datasets with pixel-wise ground truth of salient objects. The commonly used datasets are concluded in Table. 2.10. Obviously, current co-salient object segmentation datasets are relatively small when compared to image/video salient object segmentation counterparts. The largest co-salient object segmentation dataset

Table 2.10 Summary of group co-salient object segmentation datasets. #Img: The number of images/video frames. #GT: The number of object-level pixel-wise masks. Obj.-Level = Object-Level Labels. Ins.-Level = Instance-Level Labels.

Dataset	Year	Publication	#Img	#GT	Obj.-Level	Ins.-Level	Group
MSRC [182]	2005	ICCV	240	240	✓		8
iCoseg [183]	2010	CVPR	643	643	✓		38
ImgPair [184]	2011	TIP	210	210	✓		105
CoSal2015 [185]	2015	CVPR	2,015	2,015	✓		50
WICOS [186]	2018	AAAI	364	364	✓		1
CoSOD3K [187]	2020	CVPR	3,316	3,316	✓	✓	160
CoCA [188]	2020	ECCV	1,295	1,295	✓	✓	80

Fig. 2.7 An illustration of group(co)-salient object segmentation. There are two groups of images with specific classes of co-salient objects (*e.g.*, gymnast and basketball), surrounded by similar or totally different background scenes. This figure is cited from [189].

Table 2.11 Summary of recent group co-salient object segmentation methods.

Method	Year	Publication	Key Words
UFO [190]	2022	arXiv	Multi-tasks, transformer, MLP, patch collaboration.
DCFM [191]	2022	CVPR	Democratic Prototype Generation Module.
GLNet [192]	2022	TCyb	Global local correspondence modeling.
MGF [193]	2021	AAAI	Graph convolutional network, multi-scale.
DeepACG [194]	2021	CVPR	Edge-enhanced module.
GCoNet [195]	2021	CVPR	Depth-wise correlation, group consensus.
CADC [196]	2021	ICCV	Consensus-aware kernel construction.
ICNet [197]	2020	NeurIPS	Normalized masked average pooling.
CoADNet [198]	2020	NeurIPS	Group-attentive semantic aggregation.

so far is CoSOD3K [187] collecting about 3K images with both object-level and instance-level pixel-wise ground truth.

Co-salient object segmentation methods aim at achieving robust features to represent a class of

objects with different appearances and surrounded with different background scenes (Fig. 2.7). Empirically, co-salient object segmentation methods learn to segment the visually salient objects almost centered (or nearly centered) at the given images, also pay attention to the mining of intrinsic features which represent a class of objects and are able to be robust to the changes of objects' appearances and surroundings. Therefore, when compared to traditional image-based salient object segmentation, a key challenge of co-salient object segmentation is to detect the co-salient objects as much as possible, and ignore rarely seen objects (*a.k.a.* objects do not show in all images within one group) as much as possible. To this end, relatively early method CoADNet [198] proposed group-attentive semantic aggregation module to extract the semantic features based on both local spatial cues of each of the images, and global spatial cues via self-attention mechanism [199]. ICNet [197] applied normalized masked average pooling to extract features representing intra-features of the given images. A correlation fusion module was then proposed to aggregate the intra-features for inter-frame saliency consistency modeling. Later GCoNet [195] and CADC [196] built attention consensus modules to refine the inter-frame features for better model performance. DCFM [191] was also inspired by self-attention [199] and thus proposing democratic feature enhancement module to refine the encoded features from a group of images.

2.3.4 High-resolution salient object segmentation

Table 2.12 Summary of recent high resolution salient object segmentation methods.

Method	Year	Publication	Key Words
PGNet [200]	2022	CVPR	Cross model grafting module, swin transformer, attention.
HRMod [201]	2019	ICCV	Global local fusion network.

As the popularization of new smartphones which are able to produce high-resolution (*e.g.*, 4K, 8K) images, there is an increasing demand for high-resolution image processing techniques in computer vision community. In this case, recognizing and finely segmenting the salient objects in high-resolution images is able to facilitate the development of new smartphone applications. Therefore, datasets such as HRSOD [201] and UHSD [200] have been recently established. Specifically, UHRSD [200] with a training set of 4,932 images and a testing set of 988 images. It is worth noting that each of the collected images are at 4K-8K resolutions. Besides, HRSOD [201] contains a training set of 1,610 images and 400 images for model testing. So far, the high-resolution salient object segmentation is still a new sub-area of salient object segmentation where only a few methods (*e.g.*, PGNet [200] and HRMod [201], with key concepts concluded in Table. 2.12) have been proposed.

2.3.5 Remote sensing salient object segmentation

Remote sensing salient object segmentation is a new branch of salient object segmentation where images are collected with remote sensors on air-crafts or satellites. Specially, remote-sensing image datasets include extremely small salient objects when compared to the ones in other salient object segmentation related datasets (Fig. 2.8). Being similar to other topics regarding RGB salient object

Table 2.13 Summary of recent remote sensing salient object segmentation methods.

Method	Year	Publication	Key Words
CorrNet [202]	2022	TGRS	Light-weight feature extraction subnet, cross-layer correlation.
ACCoNet [203]	2022	TCyb	Bifurcation-aggregation block.
ERPNet [204]	2022	TCyb	Edge prior, recurrent network.
DAFNet [205]	2020	TIP	Global context-aware attention.
LV-Net [206]	2019	TGRS	Nested connection in decoder.

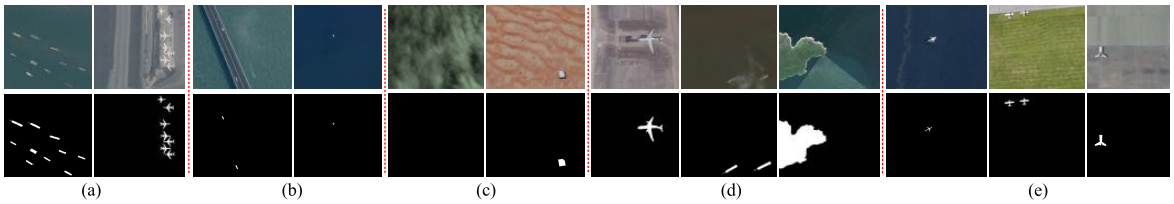


Fig. 2.8 An illustration of remote sensing salient object segmentation. (a)-(e) represent various challenging cases. This figure is cited from [205].

segmentation. Recent remote sensing salient object segmentation methods (Table. 2.13) focus on the designing of decoders' structures. LV-Net [206] designed a pyramid V-shape decoder to fuse the multi-stage features for finely segmentation of objects in optical remote sensing images. DAFNet [205] proposed hierarchical global context-aware attention modules, which were added to each of the encoder layers. ERPNet [204] proposed edge-aware position unit modules, which were then added to each of the decoding steps. ACCoNet [203] built attention based adjacent context coordination module to facilitate adjacent feature fusion. CorrNet [202] designed light-weight framework with feature enhanced modules consisting of cascaded channel and spatial attentions, to support the long skip connections between each of the encoding-decoding layers.

2.4 Salient object segmentation with 2D multi-modal data

As the recent development of depth/infra-red sensor based/light field cameras, it is easy to collect 2D images with depth/infra red information or a variety of light field modalities such as focal stacks and multi-view images [207]. Recent researches [207, 208] show that these augmented multi-modal data are able to boost the model performance for salient object segmentation. In the following sections, we illustrate the details of recent development of the fields of RGB-depth/infra-red salient object segmentation and light field salient object segmentation.

2.4.1 RGB-depth salient object segmentation

Table 2.14 Summary of RGB-depth salient object segmentation datasets. #Img: The number of images. #GT: The number of object-level pixel-wise masks.

Dataset	Year	Publication	#Img&#GT	Sensor	Resolution	Scene categories
STERE [209]	2012	CVPR	1,000	Sift flow	1,200×900	Various
GIT [210]	2013	BMVC	80	Microsoft Kinect	640×480	Indoor
DES [211]	2014	ICIMCS	135	Microsoft Kinect	640×480	Indoor
NLPR [212]	2014	ECCV	1,000	Microsoft Kinect	640×480	In/Out-door
NJUD [213]	2014	ICIP	1,985	FujiW3	1,213×828	Movie scenes
SSD [214]	2017	ICCVw	80	Optical flow	960×1,080	Movie scenes
DUT-RGBD [215]	2019	TIP	1,200	(not provided)	400×600	In/Out-door
SIP [216]	2020	TNNLS	929	Huawei Mate10	992×744	human-centered

RGB-depth salient object segmentation is a task where models use depth information (Fig. 2.9) as auxiliary information to facilitate locating and segmenting the salient objects in given 2D RGB images. Current RGB-depth salient object segmentation datasets' scales range from about 0.1k to no more than 2K (Table. 2.14). As images collected with different cameras (*e.g.*, Microsoft Kinect, Huawei Mate 10) tend to be varying in terms of depth quality, recent researches (Table. 2.15) always test their proposed methods on five or more datasets to clarify the model effectiveness and robustness.

Specifically, recent methods (*e.g.*, SPNet [217], CMIM [218], CDNet [219], DSAM [220], DSNet [221], DepthNet [222]) still largely rely on attention mechanisms to implement the RGB and depth information fusion. For instance, SPNet [217] designed channel attention based cross-modal feature enhancement module to support the feature fusion between RGB and depth encoding branches. CMIM [218] applied dual attention module [223] to aid the refinement of features from mutual information regularizer. CDNet [219] proposed new dynamic scheme to fuse the features extracted from original and estimated depth maps with channel attention mechanism. DSAM [220] decomposed the original depth map to different types of depth-based priors. A depth sensitive module was then proposed to gain useful features based on the decomposed depth priors. Most recently, DSNet [221] built attention consistency module to facilitate stable training of the proposed teacher-student framework.

Fig. 2.9 An illustration of RGB-depth salient object segmentation. (a)-(i) are the examples of commonly used datasets for RGB-depth salient object segmentation. For each of the datasets, the ground truth, depth map and the given image are listed from right to left. This figure is cited from [208].

2.4.2 RGB-thermal salient object segmentation

RGB-thermal salient object segmentation is a relatively newly proposed task in the computer vision community. The task focuses on utilizing both 2D RGB images and thermal maps collected by infrared sensors, to conduct salient object segmentation. As shown in Fig. 2.10, the thermal maps are able to provide complementary information regarding the saliency judgments. Specifically, as shown in Fig. 2.10, (a) reflects the situation where thermal maps highlight the salient objects while RGB ones do not, and (b) vice versa. So far, the commonly used RGB-thermal salient object segmentation dataset is VT series including VT821 [235], VT1000 [232] and VT5000 [236], which provide 821, 1000 and 5000 image pairs (2D RGB image and corresponding thermal infrared image), respectively.

There are only a few methods exclusively proposed for RGB-thermal salient object segmentation such as CRA [231], CGL [232], RFF [233] and APNet [234]. Similar to RGB-depth salient object segmentation methods, some of the RGB-thermal salient object segmentation models designed different attention-based multi-modal feature fusion modules to adapt to the task. For instance, re-

Table 2.15 Summary of recent RGB-depth salient object segmentation methods.

Method	Year	Publication	Key Words
DSNet [221]	2022	TIP	Teacher-student network, pseudo depth maps.
DepthNet [222]	2022	TIP	RGB-D correlation modeling, light weight.
ASB [224]	2022	TIP	Complementary edge information mining.
MobileSal [225]	2021	TAPMI	Inverted residual block, compact pyramid refinement.
DSU [226]	2021	ICLR	Depth-disentangled saliency update framework.
DSAM [220]	2021	CVPR	Depth sensitive attention module.
DCF [227]	2021	CVPR	Cross reference module.
RD3D [228]	2021	AAAI	3D Convs based encoder-decoder.
HAINet [229]	2021	TIP	Hierarchical alternate interaction module.
CDNet [219]	2021	TIP	Two-stage multi-modal feature fusion.
UTANet [230]	2021	TIP	Adaptive depth-error weights.
CMIM [218]	2021	ICCV	Cascaded learning framework, mutual info regularizer.
SPNet [217]	2021	ICCV	Multi-modal feature aggregation.

Table 2.16 Summary of recent RGB-thermal infrared salient object segmentation methods.

Method	Year	Publication	Key Words
CRA [231]	2019	TCSVT	Challenge-sensitive analysis, unified ranking model.
CGL [232]	2019	TMM	Collaborative graph, joint optimization.
RFF [233]	2021	TCSVT	Multi-scale, multi-modality, and multi-level fusion.
APNet [234]	2021	TETCI	Iterative adversarial learning, progressively guided optimization.

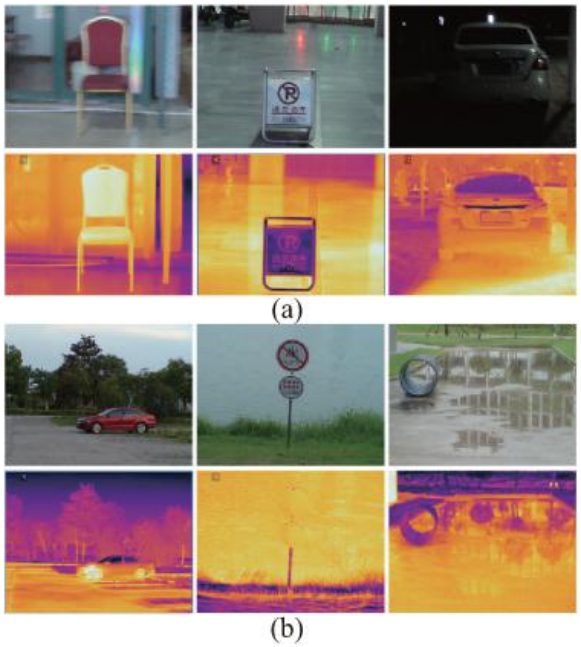


Fig. 2.10 An illustration of RGB-thermal salient object segmentation (this figure is cited from [232]). (a)/(b) show that RGB images and their corresponding thermal maps contain complementary visual cues.

cent RFF [231] built complementary weighting module to dynamically assign attention weights to RGB and thermal based features. Besides, APNet [234] applied channel and spatial attention in their proposed progressively guided optimization module, to extract features from semantic and spatial data.

2.4.3 Light field salient object segmentation

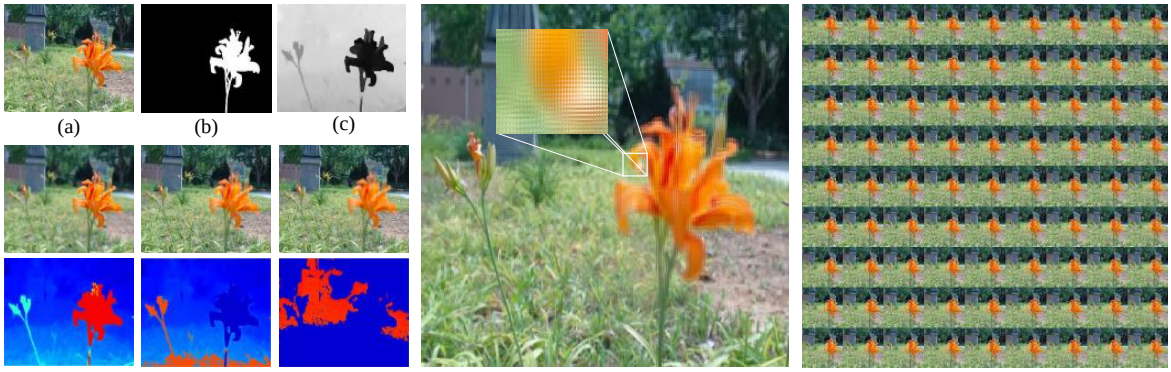


Fig. 2.11 An illustration of light field salient object segmentation (this figure is cited from [237]). (a)/(b)/(c) denote the given 2D RGB image, corresponding object-level pixel-wise ground-truth and depth map, respectively. (d) means focal stacks (top) and their focal regions (bottom). (e) and (f) denote the micro-lens based images and multi-view images, respectively.

Nowadays, it is convenient to collect different types of light field data with consumer-level light field cameras such as Lytro products. Fig. 2.11 shows an example illustrating the main types of light field modalities in widely used light field salient object segmentation datasets such as LFSD [238], HFUT [239], DUTLF-F [240], DUTLF-M [241], Lytro [242], DUTLFV2 [237] and CITYU [243]. The detailed statistics regarding these datasets are included in Table. 2.17. Most recently established datasets such as Lytro [242] and DUTLFV2 [237] provide all types of commonly seen light field data, including focal stacks, multi-view images, depth maps and micro-lens based images.

Table 2.17 Summary of light field salient object segmentation datasets. #Img: The number of images. #GT: The number of object-level pixel-wise masks.

Dataset	Year	Publication	#Img&#GT	Focal Stack	Multi-view	Depth	Micro Lens
LFSD [238]	2014	CVPR	100	✓	✓	✓	✓
HFUT [239]	2017	TOMM	255	✓	✓	✓	✓
DUTLF-F [240]	2019	ICCV	1,462	✓		✓	
DUTLF-M [241]	2019	IJCAI	1,580		✓		
Lytro [242]	2020	TIP	640	✓	✓	✓	✓
DUTLFV2 [237]	2021	arXiv	4,204	✓	✓	✓	✓
CITYU [243]	2021	TIP	817		✓	✓	✓

With the increasing public databases, extensive light field salient object segmentation methodologies have been proposed during the last few years (Table. 2.18). Relatively early method DLLF [240]

Table 2.18 Summary of recent light field salient object segmentation methods.

Method	Year	Publication	Key Words
DLLF [240]	2019	ICCV	Convs for light field, late fusion architecture.
DLSD [241]	2019	IJCAI	View-wise attention mechanism, light field synthesis.
MoLF [244]	2019	NeurIPS	ConvLSTM, focal stack, memory-oriented feature fusion.
ERNet [245]	2020	AAAI	Focal stack, teacher-student network.
LFNet [246]	2020	TIP	Integration of focusness, depths and objectness cues.
MAC [242]	2020	TIP	Micro lens, sampled sub-aperture images.
MTCNet [248]	2020	TCSVT	3D convs, multi-task decoder, edge prediction.
OBGNet [249]	2021	ACM MM	Epipolar plane images, occlusion extraction module.
DLGLRG [250]	2021	ICCV	Focal stack feature aggregation, reciprocative guidance.
GAGNN [243]	2021	TIP	Multi-scale graph networks.
SANet [251]	2021	BMVC	Complementary information learning, focal stack.
TCFANet [252]	2021	SPL	Multi-stream framework.
PANet [253]	2021	TCyb	Sharpness recognition module, multi-source learning module.
MGANet [254]	2021	ICMEw	Generative adversarial networks for light field data.
MEANet [247]	2021	N.Comp.	Multi-modal edge supervision.
DGENet [255]	2021	IVC	Recurrent global-guided focus module.

applied VGG-based hierarchical convolutional layers to extract the features from RGB image and its corresponding focal stacks, separately. The focal stack based features were refined by attention-convLSTM module before the multi-modal fusion process. DLSD [241] designed multi-view attention module to filter the useful features. MoLF [244] proposed memory-induced mechanism based on channel attention and ConvLSTM, to facilitate the feature extraction from focal stacks. Later ERNet [245] further used channel attention to support the knowledge distillation of the features extracted from focal stacks and RGB images. LFNet [246] also took advantage of both focal stacks and RGB images and used attention module to adjust the RGB-based features for efficient feature fusion in the following decoding stage. Most recently proposed MEANet [247] applied channel-spatial attention to facilitate edge priors based multi-branch supervisions for the training of the proposed framework.

2.5 Salient object segmentation in panorama

Table 2.19 Summary of 360° panoramic datasets. #Img: The number of images/video frames. #GT: The number of object-level pixel-wise masks (ground truth for SOD). Pub. = Publication. Obj.-Level = Object-Level Labels. Ins.-Level = Instance-Level Labels. Fix. GT = Fixation Maps. † denotes equirectangular (ER) images.

Dataset	Pub.	#Img	#GT	$\texttt{min}(W,H)$	$\texttt{max}(W,H)$	Obj.-Level	Ins.-Level	Fix. GT
F-360iSOD [256]	ICIP'20	107†	107	1,024	2,048	✓	✓	✓
360-SOD [257]	JSTSP'20	500†	500	512	1,024	✓		
360SSOD [258]	TVCG'20	1,105†	1,105	546	1,024	✓		

As the prosperous development of 360° saliency prediction and 2D salient object segmentation, one idea is to combine the advantages of both fields and introduce 360° salient object segmentation to the community. The task of 360° based image/video salient object segmentation is able to mimic the real human attention in static/dynamic immersive environment, thus advancing saliency prediction to cognitive vision by introducing object-level saliency judgments, also closing the gap between salient object segmentation and potential augmented/virtual reality applications where omnidirectional images based object-level saliency detection may play an important role (*e.g.*, AR glasses display rendered virtual objects based on real salient objects in 360° immersive environments).

As the main focus of this thesis, panoramic salient object segmentation has gained relatively rare attention from the community of computer vision, mainly due to the lack of large-scale datasets and comprehensive benchmark studies. As shown in Table. 2.19, F-360iSOD (ours) [256], 360-SOD [257] and 360SSOD [258] are the only datasets for 360° panoramic salient object segmentation. Besides, no dataset or method has been proposed for video-based salient object segmentation in the past years.

To fill the blank of 360° salient object segmentation researches, this thesis systematically works on image/video-based benchmark datasets (Chapter 3) and new baseline methodologies (Chapter 5).

2.6 State-of-the-art attention models

This section introduces the recent attention models widely used for deep learning based computer vision. It is widely known that human visual system largely depends on a specific mechanism, that is able to efficiently divert human attention towards salient objects and regions for effective scene understanding. Inspired by this physiological prior, current deep learning models added specific modules to mimic human attention mechanism. By doing so, they acquired similar ability of detecting key objects and scenes benefiting specific computer vision tasks, by adaptively putting more weight on specific sets of model features leading to better predictions.

In the following sub-sections, we detail the concepts and representative state-of-the-art works towards attention models, thus establishing solid theoretical and empirical foundations for the presentation of our works in subsequent sections.

2.6.1 Categories of attention models

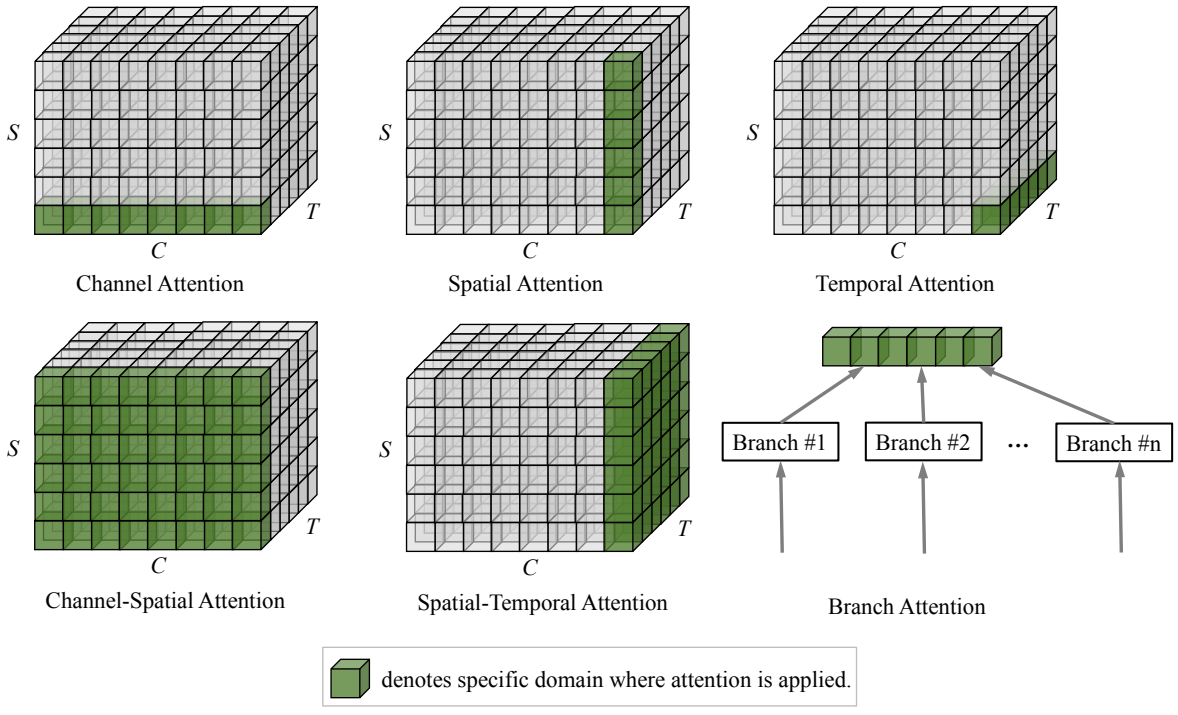


Fig. 2.12 An illustration of six commonly seen types of attention models (or, mechanisms) applied in current state-of-the-art deep learning methods in the field of computer vision. Based on different operation domains (*i.e.*, "channel – *C*", "spatial – *S*" and "temporal – *T*") of given feature maps, current state-of-the-art attention models can be classified into five categories, *i.e.*, channel-wise attentions, spatial attentions, temporal attentions, channel-spatial attentions, spatial-temporal attentions. Besides above attentions operated on single branch, branch attentions can be used to fuse and refine inter-branch-based features.

In the deep learning era, attention models have been developed and widely applied to frameworks

Table 2.20 Descriptions corresponding to each of the attention categories. Please refer to Fig. 2.12 for visualization of each type of attention.

Attention Category	Description	Representative works
Channel attention	Selecting key channels with attention mask across feature channel domain.	SENet [145], ECANet [259], *etc.*
Spatial attention	Selecting key spatial regions with attention mask across feature spatial domain.	GENet [260], Non-local [261], *etc.*
Temporal attention	Selecting key video frames with attention mask across feature temporal domain.	GLTR [262], TAM [263], *etc.*
Branch attention	Selecting key branches with attention mask across multiple branches of features.	Highway [264], Condconv [265], *etc.*
Channel-spatial attention	Selecting key features with attention mask(s) across both channel and spatial domains.	CBAM [146], Triplet [266], *etc.*
Spatial-temporal attention	Selecting key features with attention mask(s) across both spatial and temporal domains.	RSTAN [267], STA [268], *etc.*

consisting of convolutional (*e.g.*, resnets [142]) and/or transformer [143] layers. The aim of these attention models is to adaptively select the important features with specific functions, thus improving model performance upon specific benchmarks by a large margin.

General formulation. Generally, the attention models observe the following formulation:

$$Feat. = f(A(Feat.), Feat.), \tag{2.1}$$

where "*Feat.*" denotes the inputting features of deep learning networks. $A(\cdot)$ means an exclusively designed attention module that operates on specific domains of "*Feat.*". $f(\cdot)$ defines a specific function (which is usually adding and/or concatenation operations) that combines the attention-enhanced features "$A(Feat.)$" and original ones "*Feat.*".

In this case, based on the types of operation domains (channel, spatial and temporal) of given features from specific layers of deep learning networks, current state-of-the-art attention models can be classified into five categories, *i.e.*, channel attention, spatial attention, channel-spatial attention, temporal attention and spatial-temporal attention. Besides intra attentions, another mechanism, namely branch attention, which operates between branches of feature maps to model their inter-branch dependencies. The differences of these attention mechanisms are visualized in Fig. 2.12.

In addition, Table 2.20 further details the descriptions of these six types of attentions. An explanation towards six attention categories is as follows:

Channel attention denotes a type of attention modules that channel-wisely refine the given feature maps, by adaptively generating attention mask across all feature channels. The superiority of channel attention is that it is able to model the interdependencies between each of the channels of given

features. Owing to the outstanding effectiveness and usability, channel attention model such as SENet [145] has been one of the most widely used attention mechanisms (with a citation of more than 13K) during the past few years. Later works such as ECANet [259] and GSoP-Net [269] acquired further improvement based on the framework of SENet [145].

Spatial attention emphasizes the long-range dependencies modeling across spatial domain of given features. Inspired by the excitation module proposed by SENet [145], GENet [260] further designed a gather-excite module to implicitly build an attention mask, thus enabling the convolutional neural network to learn features representing contextual long-range dependencies. Besides, representative spatial attention models such as non-local network [261] and vision transformers [143] both highlight the superiority of their spatially long-range dependencies modeling abilities, especially when compared to traditional convolutional networks (*e.g.*, [2]).

Temporal attention is proposed for temporal feature refinement when inputs are sequential visual data such as video clips. In fact, not each of the video frames contribute equally to specific vision tasks such as dynamic person re-identification [262] and video recognition [263], temporal attention is thus used for temporal dependencies modeling to efficiently and effectively extract the key features.

Channel-spatial attention combines the advantages of both channel attention and spatial attention, thus proposing to refine given features based on both channel and spatial domains. A representative work is CBAM [146], which cascaded a channel-based module and a spatial-based module and thus gaining attention masks for feature refinement. As a result, the proposed channel-spatial module [146] has been widely used to refine convolutional network-based features (*e.g.*, [266, 270]). Besides cascaded channel-spatial attention, representative work such as Triplet network [266] directly generated attention mask across spatial-channel domains with a three-branch attention module.

Spatial-temporal attention is a type of attention mechanism that takes advantage of both intra-frame spatial features and inter-frame temporal features, to facilitate specific vision tasks such as action recognition [267] and dynamic person re-identification [268].

Branch attention is a relatively special class of attention mechanism that fuses and refines global features from multiple branches. Representative works include Highway network [264], Condconv [265], *etc.*

It is worth noting that, the decision of which type of attention to use may depend on the specific tasks.

2.6.2 Representative attention models

As shown in Fig. 2.13, early attention model such as RAM [271] used recurrent neural network to learn features from one local region at a time, then to select key features representing important locations. Being different to aforementioned spatial attention, *i.e.*, GENet [266], RAM [271] transformed given image to sequential patches and built attention mask across the patches, rather than building attention mask across spatial domain of holistic features. Later work such as Highway network [264] was inspired by LSTM [273] and proposed a "information highway" consisting of gate units to adaptively focus on key features learning.

Channel&Spatial Attention. As the success of classical "squeeze-excitation" framework (Fig. 2.14) proposed by SENet [145], multiple works such as GSoP-Net [269], SKNet [274], ECANet [259]

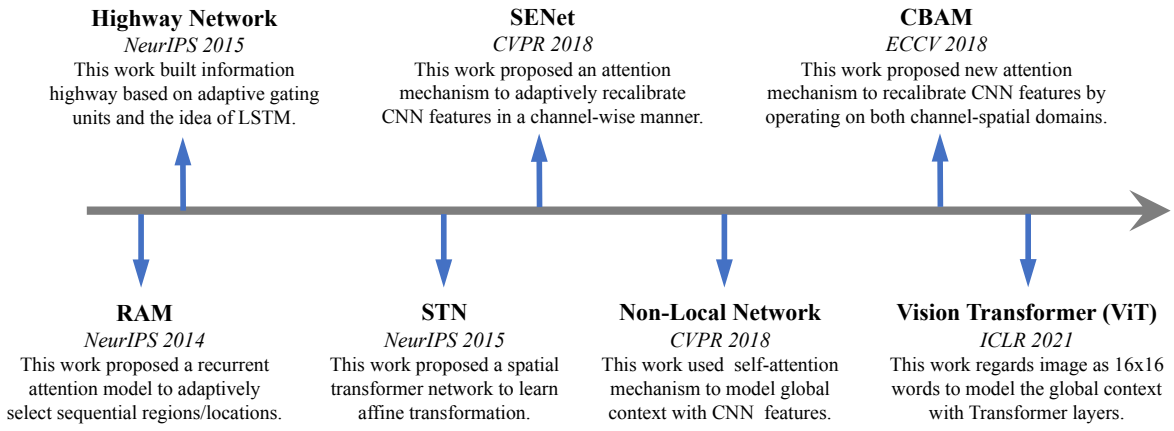


Fig. 2.13 The history towards a prosperous development of attention models in the field of computer vision. Due to the limited space along the timeline, we only summarize several representative methods (*i.e.*, RAM [271], STN [272], Highway Network [264], SENet [145], Non-Local Network [261], CBAM [146] and ViT [143]) in this figure. Please refer to Section 2.6.2 for detailed illustrations.

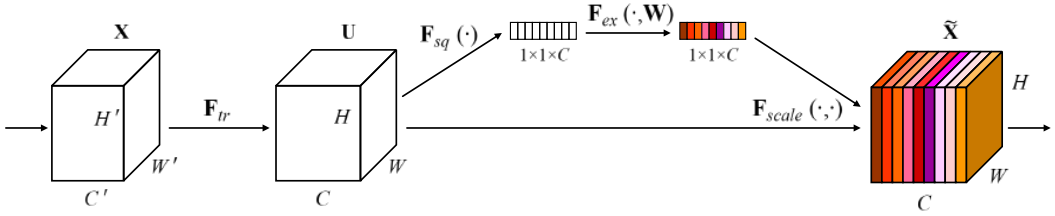


Fig. 2.14 Illustration of "squeeze-excitation" module in SENet [145]. Please note that this figure is cited from [145].

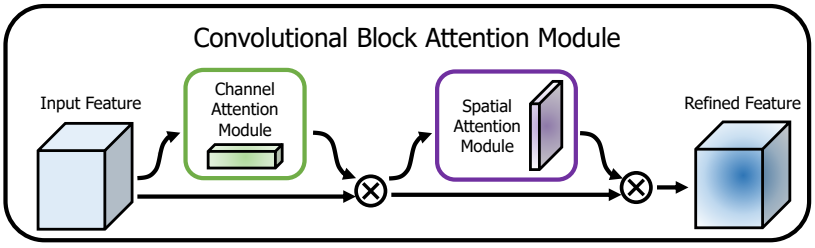
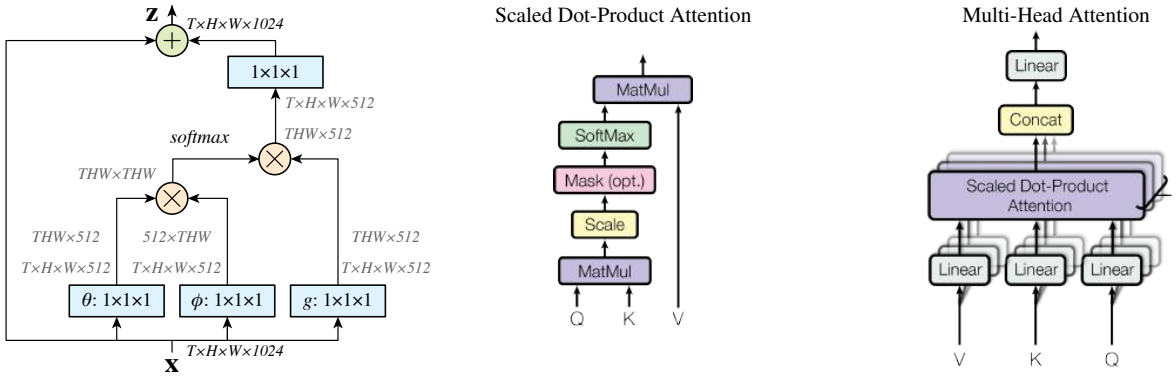


Fig. 2.15 Illustration of cascaded channel-spatial attention module in CBAM [146]. Please note that this figure is cited from CBAM [146].

and CBAM [146] followed SENet and conducted further improvements based on it. Specifically, Squeeze-excitation attention [145] emphasized the channel-wise effective features by squeezing the spatial features with an adaptive average pooling layer and by computing channel-wise attention using two fully-connected layers. GSoP-Net [269] replaced the squeeze module (*i.e.*, average pooling layer) of SENet with a proposed "global second-order" pooling layer. SKNet [274] further proposed a three-stage (*i.e.*, splitting, fusion and selection) attention mechanism, where the input features were split into multiple branches and convolved with different kernels. The processed features were then fused with squeeze-excitation attentions and summed as final output. Also based on squeeze-excitation



(a) (b)

Fig. 2.16 Illustration of self-attention mechanisms. (a) denotes the self-attention module used in non-local network (note that this sub-figure is cited from [261]). (b) compares the self-attention and multi-head self-attention mechanisms, the latter is further used in vision transformer [143] (note that this sub-figure is cited from [199]).

mechanism, ECANet [259] focused on computing local adjacent channel attention by replacing the two fully-connected layers of squeeze-excitation model with an 1D-convolutional layer. Besides above channel attention-based models, CBAM [146] used a large kernel (*e.g.*, 7×7) to further extract spatial attention based on channel-wise-refined features. Similarly, BAM [275] also applied both the channel and spatial attentions to feature refinement. However, it simply sums the attention matrix, rather than cascading the channel-/spatial-based ones as in CBAM (Fig. 2.15).

Self-Attention. Self-attention [199] (Fig. 2.16 (b) – "dot-product attention") is widely used in the fields of natural language processing and multi-modal learning. Self-attention is a type of operation where the input feature is first mapped to "query", "key" and "value" features via fully-connected layers, respectively. The final output feature is computed as the output of a dot product of "value" and the result of a dot product of "query" and "key".

Inspired by self-attention, especially the "query"-"key"-"value" mechanism, non-local network (Fig. 2.16 (a)) proposed non-local block to model global contextual correlations in spatial domain of features gained from convolutional layers. As the development of computational sources (*e.g.*, GPUs), large-scale multi-head self-attention modules (Fig. 2.16 (b)) have been applied to vision Transformers (*e.g.*, ViT [143]), which advances deep learning models towards better performance on multiple benchmarks in the fields of image classification, object detection and segmentation.

2.7 Evaluation for salient object segmentation

Following the common settings in the field of salient object segmentation, in this thesis, we apply four widely used metrics, *i.e.*, F-measure (F_β) [276], MAE ($\mathscr{M}$) [277], S-measure (S_α) [278] and E-measure (E_ϕ) [279], to evaluate all benchmark models and our proposed methods. Generally, S_α and E_ϕ are the recently proposed metrics. With the pixel-wise binary ground truth and output saliency maps, S_α quantifies the objects' structure similarities while the E_ϕ considers the similarities regarding both local details and global context. Besides, F_β and $\mathscr{M}$ focus only on the local per-pixel matches. Specifically,

MAE ($\mathscr{M}$) computes the mean absolute error between the normalized predicted saliency map $P \in [0,1]$ and the corresponding ground truth $G \in \{0,1\}$,

$$\mathscr{M} = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H} \mid G(i,j) - P(i,j) \mid, \tag{2.2}$$

where W and H denote the width and the height of the given image.

F-measure (F_β) computes both *Precision* and *Recall*, being formulated as:

$$F_\beta = \frac{(1+\beta^2)Precision\ Recall}{\beta^2 Precision + Recall}, \ \ with \ \ Precision = \frac{|P \cap G|}{|P|}, Recall = \frac{|P \cap G|}{|G|}, \tag{2.3}$$

where G is the ground truth and P denotes a binary mask converted from a predicted saliency map. Multiple P are computed by assigning different integral thresholds τ ($\tau \in [0,255]$) to the saliency map. The β^2 is set to 0.3 according to [276]. Note that we may report mean, adaptive or max F-measure scores during quantitative evaluation, to be consistent with the settings of previous benchmarks of specific tasks.

S-measure (S_α) evaluates the structural similarities between the prediction and the ground truth. The metric is defined as:

$$S = \alpha S_o + (1-\alpha) S_r, \tag{2.4}$$

where S_r and S_o denote the region-/object-based structure similarities, respectively. $\alpha \in [0,1]$ is empirically set as 0.5 to arrange equal weights to both region-level and object-level quantitative evaluation. [278].

E-measure (E_ϕ) is a cognitive vision-inspired metric evaluating both global and local similarities between two binary maps. The metric is defined as:

$$E_\phi = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H} \phi\left(G(i,j), P(i,j)\right), \tag{2.5}$$

where ϕ represents the enhanced alignment matrix [279]. Note that we may report mean, adaptive or max F-measure scores during quantitative evaluation, to be consistent with the settings of previous benchmarks of specific tasks.

2.8 Conclusion

In this chapter, we thoroughly reviewed the recent state-of-the-art works towards salient object segmentation in both 2D and 360° domains. A finding is that most recently proposed models tend to exclusively design attention modules to adapt to specific tasks (*e.g.*, mutual attention [180]). Besides, we also reviewed the basic attention models in the field of computer vision, which are essential components for the establishment of effective deep learning models upon multiple benchmarks (*i.e.*, light field/360° panoramic salient object segmentation).

Chapter 3

Datasets & benchmarks on 360° images and videos

3.1 Introduction

Chapter 2 provides general information towards the benchmark datasets (*e.g.*, citations, scales and annotation types), which have been widely used for multiple tasks related to salient object segmentation. This chapter carefully discusses the details of 2D RGB image and video salient object segmentation datasets, and highlights several key aspects (*e.g.*, data sources, datasets' scales, hierarchical annotations and salient objects' attributes) in terms of large-scale salient object segmentation dataset construction. This chapter then presents the details of two newly proposed 360° salient object segmentation datasets, *i.e.*, F-360iSOD[1] [256] and PAVS10K[2], which consider the summarized key issues (aspects) regarding large-scale salient object segmentation dataset construction.

It is worth noting that there was neither dataset nor benchmark exclusively designed for 360° salient object segmentation, before the year of 2019 when F-360iSOD was proposed. Since the year of 2019, only four datasets including 360-SOD [257], 360SSOD [258], F-360iSOD and PAVS10K were established for conducting salient object segmentation in 360° domain.

The rarity of 360° salient object segmentation datasaet is mainly due to a lack of manual large-scale pixel-wise annotations of salient objects contained in 360° panoramic images and videos. In fact, acquiring pixel-wise object-level and instance-level masks is an extremely time-consuming and laborious process. Besides, due to the annotators' preference and stochasticity in labeling the objects, an unavoidable systematical errors tend to exist in current large-scale dense annotations[3]. One may ask that, why not constructing datasets for the development of merely un-supervised and self-supervised methods? Indeed, the training of these types of models does not require large-scale annotations. However, salient object segmentation is so far a too challenging task for non fully supervised deep learning models. In fact, current un-/self-/weakly-supervised models can seldomly be compared

[1] Fixation-based salient object detection in 360° images – F-360iSOD

[2] dataset with about 10K pixel-wise annotations for the task of panoramic audio-visual salient object segmentation – PAVS10K

[3] Dense annotations denote pixel-level labels, rather than category labels in commonly seen image classification datasets.

to fully-supervised ones [16]. Besides, datasets with large-scale dense annotations can be easily used for the training of un-/self-/weakly-supervised models. Therefore, establishing large-scale datasets with thorough dense annotations is a necessity to fulfill the target of this thesis.

3.2 Key aspects for salient object segmentation datasets' construction

This section discusses several key aspects towards large-scale dataset construction based on a thorough review of previously proposed 2D RGB salient object segmentation datasets, to provide solid theoretical and empirical foundations for the construction of datasets used for salient object segmentation in 360°.

Based on Chapter 2, which presents general statistics of commonly used datasets for salient object segmentation in multiple domains, a qualified salient object segmentation dataset must include at least the following considerations from a perspective of theory:

- Appropriate protocol to ensure reasonable judgments towards salient objects (sometimes subjective experiments are needed to aid making definitions for the salient objects).
- Specific annotation protocol to ensure correct and high-quality manual annotations.

Besides theoretical principles (*a.k.a.*, protocols), empirical aspects are of equal importance for the successful construction of large-scale salient object segmentation datasets. These empirical aspects may include:

- The dataset must contain images and/or videos representing a variety of real-life scenes, where foreground/background objects varying in categories, sizes, appearances, shapes and other challenging attributes (such as occlusion and out-of-view) are included.
- As the task of salient object segmentation is a sub-branch of human attention modeling, the proposed salient object segmentation dataset must include a majority of images/videos where objects, that continually grasp human attention, exist.
- The dataset must contain large-scale images and/or videos with per-image annotations. According to the review of multiple types of salient object segmentation datasets in Chapter 2, the commonly applied datasets' scales tend to be varying. Generally, the representative datasets tend to include a few hundreds to several thousands of annotated salient objects.
- Dense annotations can be regarded as the most special feature of salient object segmentation datasets, based on the statistics presented in the Chapter 2. To contribute qualified datasets to the training/testing of fully supervised deep learning methods aiming at finely segmenting the salient objects, manually labelled per-pixel binary masks corresponding to each of the images and/or video frames within the proposed datasets must be included.
- Besides the pixel-wise labels, to facilitate following comprehensive benchmark studies, hierarchical annotations regarding scene/object categories and challenging objects' attributes are necessarily included in the proposed datasets.
- Detailed statistical analysis must be conducted to clarify the feasibility and complexity of the proposed dataset. For instance, to reflect the complexity of immersive real-life scenes, the majority of images and/or videos contained in the datasets are supposed to include salient objects possessing multiple challenging attributes.

The following sections illustrate the details regarding these key issues for large-scale salient object segmentation dataset construction in a progressive manner. Specifically, the establishment of a qualified salient object segmentation dataset can be decomposed to four progressive steps, *i.e.*, source data collection, protocol designing, annotation manufacturing and statistical analysis.

3.2.1 Sources

Content. Current salient object segmentation datasets directly collected 2D RGB images and videos from the Internet by searching a variety of key words describing specific objects and scenes commonly seen in the real life. These key words may include indoor/outdoor scenes, human-related occasions (such as dramas, concerts, conferences, travel, sports), various object categories (such as persons, instruments, electronics, animals), *etc*. Images/videos possessing these visually salient objects/scenes are able to provide a foundation for the development of data-driven methods which aim at mimicking human attention in real life.

Besides visually salient objects and scenes, recent datasets such as JOT [280] and SOC [21] emphasized the importance of non-salient objects (*e.g.*, obscure and cluttered objects shown in Fig. 3.1), which may play an important role in constructing balanced salient object segmentation datasets. In fact, salient objects are not necessarily seen in each of the real-life scenes. [21, 280] argued that current datasets, which exclude images without salient objects, are seriously unbalanced due to the selection bias during data collection process.

Complexity. The complexity of a dataset is usually proportional to the density of objects it contains. Early image-based salient object segmentation datasets [282, 283] collected images containing only one or two visually foreground objects surrounded by simple background. Recently proposed datasets such as SOD [284], ECSSD [90], DUT-O [17], PASCAL-S [18], HKU-IS [19] and ILSO [91] collected images with more challenging scenes where no more than four main objects appear. Most recently, dataset such as SOC [21] includes more challenging scenes with multiple foreground objects and cluttered background context.

As for video-based salient object segmentation, early datasets such as ViSal [150], UVSD [152] and DAVIS2016 [151] contain video frames with simple context consisting of only one or two spatially connected foreground objects. Most recent datasets such as VOS [22] and DAVSOD [23] collected more challenging dynamic scenes with four to five foreground objects per-frame.

The next section illustrates the common protocols used for gaining the salient objects out of cluttered foreground/background scenes.

3.2.2 Protocols

Datasets' protocols indicate the basic principles for judging and defining the salient objects. A reasonable protocol is of the most important theoretical component of a qualified salient object segmentation dataset. As the divergence of data conditions, protocols are largely different among the datasets.

In 2D domain, the visually salient objects usually indicate the foreground/main objects that constantly grasp human attention in static or dynamic scenes. Early datasets such as [150, 282, 283] tend

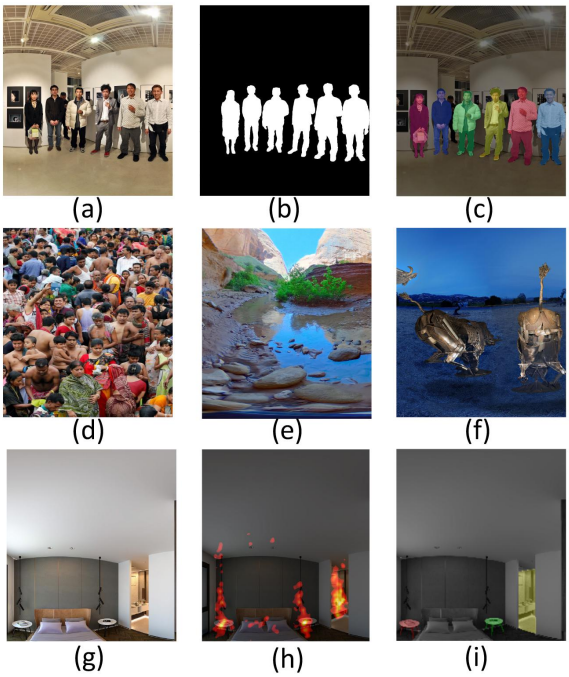


Fig. 3.1 A visualization example which illustrates partial key aspects of salient object segmentation construction. The first row indicates typical pixel-wise manual annotations in current salient object segmentation datasets, *i.e.*, object-level (binary) and instance-level masks. The second row shows specific real-life scenes (scattered people, pure background objects and meaningless foreground objects) tend not to be included in current most datasets. The third row shows salient objects defined with the guidance of fixations. This figure is cited from [281].

to confound the concepts of foreground and salient objects since they only contain simple scenes with one or two objects. However, later datasets such as [18, 22, 23, 91, 284] introduced specific protocols to aid filtering the salient objects out of multiple foreground objects. According to psychological research such as [285], human visual attention mechanism is able to support human to enumerate no more than five objects at one glimpse. To mimic real object-level visual attention mechanism in challenging salient object segmentation datasets and to facilitate datasets' annotations, several protocols based on either explicit subjective judgments or eye-tracking experiments, have thus been proposed.

Protocols based on explicit subjective judgments. Explicit subjective judgments based protocols directly leverage subjective opinions of multiple subjects as guidance for making definition towards salient objects. These strategies are useful for accurately defining salient objects in scenes with simple context (*e.g.*, salient persons in real-life daily scenes shown in Fig. 3.1). Some representative datasets include MSRA [282], SED [283], ASD [276], ECSSD [90], ViSal [150] and UVSD [152].

Specifically[4], a voting strategy was first proposed by MSRA-A and MSRA-B [282], where three and nine viewers were recruited, respectively. Each of the viewers was asked to annotate one object per image with a bounding box, the salient objects were then selected by implementing the majority

[4]The detailed introduction of these protocols is cited from [281].

rule among all viewers, which is defined as:

$$A_T = \frac{\sum_{x \in (p_x > T)} p_x}{\sum_x p_x}, \tag{3.1}$$

where A_T is the percentage of image pixels upon which salient intensity are above a empirical threshold T . Further, p_x is defined as:

$$p_x = \frac{1}{S} \sum_{s=1}^{S} m_x^s. \tag{3.2}$$

where S is the number of subjects, m_x^s is the binary mask labeled by the sth viewer, corresponding to the xth image.

Compared to the MSRA-A [282], Bruce-A [286] is a salient object segmentation dataset which contains 120 relatively images with multiple (≤ 4) visually salient foreground objects. 70 observers were employed to judge the salient objects. The labeling consistency between the observers is defined as:

$$C_k = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n=70} \frac{|a_{ik} \cap a_{jk}|}{|a_{ik} \cup a_{jk}|}. \tag{3.3}$$

where a_{ik} and a_{jk} are pixel-wise ground truth annotations of the ith and jth observers, corresponding to the kth image. The C_k is a value between 0 and 1. A better overlap of labels among observers leads to a higher C_k and vice versa. As a result, the C_k was reported to be relatively low when multiple foreground objects appear on the kth image, indicating the divergence between observers is significant. Further, 59 images with high labeling consistency ($C_k \geq 0.75$) were selected.

PASCAL-S [18] is another widely-used salient object segmentation dataset. Twelve subjects were involved to freely click on the fully segmented object regions. The final salient objects were selected based on their saliency ranking. Note that the saliency intensity of each object region is the total number of clicks it receives, divided by the number of subjects.

HKU-IS [19] defined salient objects also via labeling consistency metric. It directly excluded the images with low labeling consistency. Specifically, the labeling consistency of three annotators is defined as:

$$R = \frac{\sum_x (\Pi_{s=1}^3 a_x^{(s)})}{\sum_x 1(\sum_{s=1}^3 a_x^{(s)} \neq 0)}. \tag{3.4}$$

where $a_x^{(s)}$ is the binary saliency mask annotated by the sth subject over the xth image. R is the ratio of the pixels labeled as salient by all three subjects and the ones labeled by at least one of the three subjects. The images with $L > 0.9$ were then kept for further annotations. Finally, the salient objects were confirmed by using a majority principle (two out of three). The principle can be formulated as:

$$S_x = 1(\sum_{s=1}^{3} a_x^{(s)} \geq 2). \tag{3.5}$$

where S_x is the ground-truth salieny map, which may contain multiple salient objects over the xth image.

SOC [21] is a relatively newly proposed large-scale salient object segmentation dataset, which

includes a two-stage annotation procedure. At the first stage, five viewers were recruited to annotate salient objects without the constraint of number of the objects selected. At the second stage, only the images with high labeling consistency were reserved. The eligible images are the ones where a majority (≥ 3) of viewers annotated the identical objects[5]. The IoU is defined as:

$$IoU = \frac{Area \quad of \quad Overlap}{Area \quad of \quad Union}. \tag{3.6}$$

As for video-based salient object segmentation datasets, RSD [287] contains multiple salient objects with bounding box labels) in 62,356 video frames. 23 annotators were involved to construct the dataset. Quatitatively, the saliency $m_{i,k}$ of the kth object in the ith frame is defined as :

$$m_{i,k} = \frac{s_{i,k}}{\sum_j s_{i,j}}. \tag{3.7}$$

where $s_{i,k}$ is the number of annotators who selected the kth object in the ith frame, while $\sum_j s_{i,j}$ is the total number of annotations in the ith frame.

ViSal [150] and UVSD [152] are early video salient object segmentation datasets containing 17 and 18 dynamic scenes, respectively. In these datasets, the salient objects were simply defined as the visual foreground objects in each of the sequences.

Protocols based on eye-tracking experiments. As the collected scenes become more complex, high consistency between a few annotators is increasingly hard to achieve due the divergence of personal preference. To this end, fixations (the third row of Fig. 3.1) have been widely applied (*e.g.*, [22, 23]) to facilitate the annotation of salient objects.

Specifically[6], based on the eye-tracking experiments implemented in salient object segmentation datasets such as [18, 39, 286, 288–291], it has been proved that there is a consistency between fixations and explicit subjective judgments. Depending on which, [22, 23, 39] attempted to annotate salient objects with the guidance of fixations.

JUDD-A [39] is an early salient object segmentation dataset which applied the fixations to the annotation process of single salient object (per-image). At the first stage, multiple objects were annotated pixel-wisely by two observers. Then the objects with the region containing the highest fraction (compared to the other objects regions) of fixations were selected as salient objects.

Inspired by PASCAL-S [18], VOS [22] further annotated multiple salient objects (per-video) by applying fixation points. Instead of counting the number of fixations in separate frames, a so-called fixation density is defined and used to quantify the salient object annotation. Let $I_t \in V$ be a key frame presented at time t in a given video (V), while $O \in I_t$ be an annotated object. Note that $t_f \in T$ (the frames within short period (T) following the selected key frame (I_t)) is considered to solve the fixation sparsity in single frame. The fixation density at the region of O can then be defined as:

$$D(O) = \frac{1}{||O||} \sum_{t_f \in T} 1(\sum_{p \in O} D_{f,p} * exp(-\frac{(t_f - t)^2}{2\delta_t^2})), \tag{3.8}$$

[5]Identical objects are defined as objects owning > 0.8 of intersection over union (IoU) based on no more than three bounding boxes

[6]The detailed introduction towards these protocols is cited from [281].

where $D_{f,p}$ is formulated as:

$$D_{f,p} = exp(-\frac{(x_f - x_p)^2 + (y_f - y_p)^2}{2\delta_s^2}). \tag{3.9}$$

where δ_s is empirically set to 0.03 of video width (or height if it is larger than width), also δ_t is set to 0.1s. The salient objects in a given video were further defined by empirically thresholding their saliency scores on the whole video scale. The saliency score S is thus defined as:

$$S = \frac{\sum_{I_t \in V} \sum_{O \in I_t} D(O)}{\sum_{I_t \in V} \sum_{O \in I_t} 1}. \tag{3.10}$$

More recently, DAVSOD [23] also labeled multiple salient objects in each of the videos, by combining subjective annotation and fixation maps from external dataset [11]. About 5 viewers were recruited to freely annotate several objects pixel-wisely in each of the frames, with the fixation maps simultaneously displayed as reference. Importantly, the fixation maps contain smoothed saliency regions, rather than disconnected fixation points.

Based on the above subjective experiments based protocols, a typical procedure including the common steps of salient object segmentation dataset construction can be drawn (Fig. 3.2).

3.2.3 Annotations

Early datasets such as MSRA-A/B [282], RSD [287], STC [292] and DUT-O [17] provide only bounding box annotations. As the models are supposed to finely segment the salient objects from given images and videos, recent datasets such as SED1/2 [283], ASD [276], SOD [284], iCoSeg [183], MSRA5K [293], Infrared [288], ImgSal [289], CSSD [90], ECSSD [90], Bruce-A [286], THUR15K [294], JUDD-A [39], PASCAL-S [18], UCSB [290], OSIE [291], HKU-IS [19], ViSal [150], UVSD [152], XPIE [295], ILSO [91], DUTS [20], VOS [22], SOC [21], DAVSOD [23] and SIP [216], are able to provide manually labeled object-level annotations (*e.g.*, the first row of the Fig. 3.1). Furthermore, datasets such as ILSO [91], SOC [21], DAVSOD [23] and SIP [216] provide both object-level and instance-level pixel-wise labels (*e.g.*, the first row of the Fig. 3.1), to facilitate salient instance segmentation related tasks.

Statistics regarding the annotations' scales and types are presented in Table 3.1, Table 3.2 and Tabel 3.3.

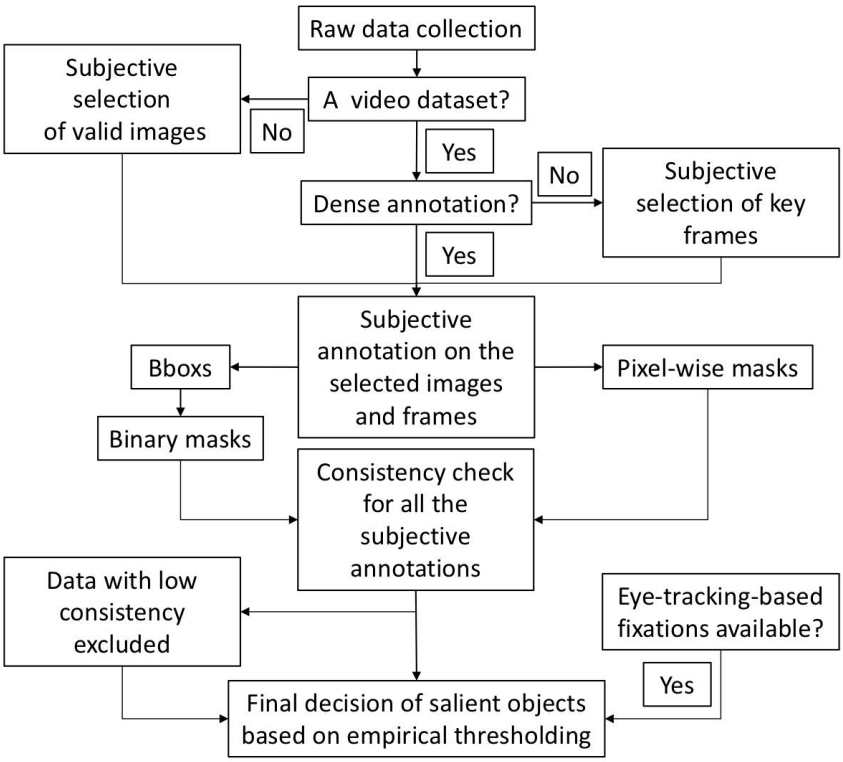


Fig. 3.2 A flowchart represents typical process of acquiring salient objects' annotations in current salient object segmentation datasets. Please note that "dense annotations" in this flowchart indicate per-image/video frame pixel-wise annotations, rather than merely per-pixel labels as mentioned in the text. This figure is cited from [281].

Table 3.1 An overview in terms of annotations of widely used 2D image/video salient object segmentation datasets (1/3). Pub. = publications. N.Images = number of images contained in the given image salient object segmentation dataset. N.Sequences = number of sequences included in the given video salient object segmentation dataset. T.Annotations = type of annotations in the given image/video dataset. N.Annotators = number of annotators involved in the dataset labeling process. N.Subjects = number of subjects recruited for conducting eye-tracking experiments. "-" denotes no information provided. This table is updated based on the statistics in [281], which collects partial statistical results from [15].

No.	Dataset	Year	Pub.	N.Images	N.Sequences	T.Annotations	N.Annotators	N.Subjects
1	**MSRA-A** [282]	2007	CVPR	20,000	-	Bounding Box	3	-
2	**MSRA-B** [282]	2007	CVPR	5,000	-	Bounding Box	9	-
3	**SED1** [283]	2007	CVPR	100	-	Pixel-wise Oject-Level	3	-
4	**SED2** [283]	2007	CVPR	100	-	Pixel-wise Oject-Level	3	-
5	**ASD** [276]	2009	CVPR	1,000	-	Pixel-wise Oject-Level	1	-
6	**RSD** [287]	2009	ICME	-	431	Bounding Box	23	-
7	**SOD** [284]	2010	TPAMI	300	-	Pixel-wise Oject-Level	7	-
8	**iCoSeg** [183]	2010	CVPR	643	-	Pixel-wise Oject-Level	1	-
9	**MSRA5K** [293]	2011	BMVC	5,000	-	Pixel-wise Oject-Level	1	-
10	**Infrared** [288]	2011	CVPR	900	-	Pixel-wise Oject-Level	2	15

Table 3.2 An overview in terms of annotations of widely used 2D image/video salient object segmentation datasets (2/3). Pub. = publications. N.Images = number of images contained in the given image salient object segmentation dataset. N.Sequences = number of sequences included in the given video salient object segmentation dataset. T.Annotations = type of annotations in the given image/video dataset. N.Annotators = number of annotators involved in the dataset labeling process. N.Subjects = number of subjects recruited for conducting eye-tracking experiments. "-" denotes no information provided. This table is updated based on the statistics in [281], which collects partial statistical results from [15].

No.	Dataset	Year	Pub.	N.Images	N.Sequences	T.Annotations	N.Annotators	N.Subjects
11	**STC** [292]	2011	J.CSB	-	32	Bounding Box	1	-
12	**ImgSal** [289]	2012	TPAMI	235	-	Pixel-wise Oject-Level	19	50
13	**CSSD** [90]	2013	CVPR	200	-	Pixel-wise Oject-Level	1	-
14	**ECSSD** [90]	2013	CVPR	1,000	-	Pixel-wise Oject-Level	5	-
15	**DUT-O** [17]	2013	CVPR	5,172	-	Bounding Box	5	5
16	**Bruce-A** [286]	2013	J.VR	120	-	Pixel-wise Oject-Level	70	20
17	**THUR15K** [294]	2014	J.VC	15,000	-	Pixel-wise Oject-Level	1	-
18	**JUDD-A** [39]	2014	TIP	900	-	Pixel-wise Oject-Level	2	15
19	**PASCAL-S** [18]	2014	CVPR	850	-	Pixel-wise Oject-Level	12	8
20	**UCSB** [290]	2014	J.V	700	-	Pixel-wise Oject-Level	100	8
21	**OSIE** [291]	2014	J.V	700	-	Pixel-wise Oject-Level	1	15

Table 3.3 An overview in terms of annotations of widely used 2D image/video salient object segmentation datasets (3/3). Pub. = publications. N.Images = number of images contained in the given image salient object segmentation dataset. N.Sequences = number of sequences included in the given video salient object segmentation dataset. T.Annotations = type of annotations in the given image/video dataset. N.Annotators = number of annotators involved in the dataset labeling process. N.Subjects = number of subjects recruited for conducting eye-tracking experiments. "-" denotes no information provided. This table is updated based on the statistics in [281], which collects partial statistical results from [15].

No.	Dataset	Year	Pub.	N.Images	N.Sequences	T.Annotations	N.Annotators	N.Subjects
22	**HKU-IS** [19]	2015	CVPR	4,447	-	Pixel-wise Oject-Level	3	-
23	**ViSal** [150]	2015	TIP	-	17	Pixel-wise Oject-Level	1	-
24	**UVSD** [152]	2016	TCSVT	-	18	Pixel-wise Oject-Level	1	-
25	**XPIE** [295]	2017	CVPR	10,000	-	Pixel-wise Oject-Level	2	-
26	**ILSO** [91]	2017	CVPR	1,000	-	Pixel-wise Object-Level Instance-Level	3	-
27	**DUTS** [20]	2017	CVPR	15,572	-	Pixel-wise Oject-Level	-	-
28	**VOS** [22]	2017	TIP	-	200	Pixel-wise Oject-Level	4	23
29	**SOC** [21]	2018	ECCV	6,000	-	Pixel-wise Object-Level Instance-Level	5	-
30	**DAVSOD** [23]	2019	CVPR	-	226	Pixel-wise Object-Level Instance-Level	5	-
31	**SIP** [216]	2020	TNNLS	1,000	-	Pixel-wise Object-Level Instance-Level	11	-

3.2.4 Statistical analysis

After acquiring large-scale image/video data and corresponding annotations, systematical statistical analysis should be conducted to show the difficulty and validity of the proposed datasets. Generally, a newly propposed dataset should be analyzed from the aspects of scales, scene/object categories, quality and diversity of pixel-wise annotations and scene/object attributes.

Specifically, as shown in Table 3.1, Table 3.2 and Table 3.3, current salient object segmentation datasets' scales range from several hundreds to 20K. Besides, most recently proposed image dataset [21] is able to include 80 categories of salient objects commonly seen in real-life daily scenes, while video dataset [23] provides about 70 classes of frequently seen realistic dynamic scenes. As the increasingly challenging scenes collected, most recently proposed datasets (*e.g.*, [21,23]) are able to manufacture high-quality and diverse annotations for defined salient objects, including bounding boxes and both object/instance-level pixel-wise masks. To facilitate comprehensive benchmark studies and inspire new models, recent datasets such as DAVIS2016 [151] and SOC [21] are able to provide annotations labeling specific attributes of salient objects (Fig. 3.3).

3.2.5 Discussion

This section reviews the recently proposed datasets for 2D RGB salient object segmentation and summarizes several key issues regarding large-scale salient object segmentation dataset construction. Particularly, there are three key issues[7] are of most important and should be emphasized for the construction of large-scale salient object segmentation dataset in 360° panorama.

High-quality pixel-wise labels. Labeling salient objects with pixel-wise masks is consistent with the prior knowledge in the field of psychology, that people tend to simultaneously pay attention to several disconnected semantic regions [296]. Besides, to simulate the human capability of distinguishing entities belonging to single object category, instance-level labels are also important and useful to a well established salient object segmentation dataset.

Balanced salient object segmentation dataset. As highlighted in [21], some of the image salient object segmentation datasets discarded images without salient objects, thus introducing selection bias to the process of dataset construction. Considering the importance of keeping images without salient objects, [281] emphasized three main principles for the judgement of non-salient objects. As shown in the second row of Fig. 3.1, the divergence among viewers tend to be significant when asked to choose the most salient person. Therefore, objects with crowded candidates of the same class tend to be non-salient ones. Besides, the natural objects such as rocks, sky belong to background. Further, the objects with complex shape and texture are recognized as non-salient objects. On the other hand, clear faces, people, animals, cars and text are commonly considered as salient objects [286].

Fixation-based salient object annotation. Recent video datasets [22,23] applied the fixation data to the salient object annotation task. However, since the thresholds are empirically fixed, the annotation methods may not be directly applied to other video salient object segmentation datasets. Using fixation data to effectively annotate salient objects is still an open issue. Future works are suggested to shift more attention towards fixation-based salient object segmentation.

[7] The detailed introduction towards these key issues is cited from [281].

ID	Description
BC	*Background Clutter*. The back- and foreground regions around the object boundaries have similar colors (χ^2 over histograms).
DEF	*Deformation*. Object undergoes complex, non-rigid deformations.
MB	*Motion Blur*. Object has fuzzy boundaries due to fast motion.
FM	*Fast-Motion*. The average, per-frame object motion, computed as centroids Euclidean distance, is larger than $\tau_{fm} = 20$ pixels.
LR	*Low Resolution*. The ratio between the average object bounding-box area and the image area is smaller than $t_{lr} = 0.1$.
OCC	*Occlusion*. Object becomes partially or fully occluded.
OV	*Out-of-view*. Object is partially clipped by the image boundaries.
SV	*Scale-Variation*. The area ratio among any pair of bounding-boxes enclosing the target object is smaller than $\tau_{sv} = 0.5$.
AC	*Appearance Change*. Noticeable appearance variation, due to illumination changes and relative camera-object rotation.
EA	*Edge Ambiguity*. Unreliable edge detection. The average ground-truth edge probability (using [11]) is smaller than $\tau_e = 0.5$.
CS	*Camera-Shake*. Footage displays non-negligible vibrations.
HO	*Heterogeneus Object*. Object regions have distinct colors.
IO	*Interacting Objects*. The target object is an ensemble of multiple, spatially-connected objects (*e.g.* mother with stroller).
DB	*Dynamic Background*. Background regions move or deform.
SC	*Shape Complexity*. The object has complex boundaries such as thin parts and holes.

(a)

Attr	Description
AC	***Appearance Change.*** The obvious illumination change in the object region.
BO	***Big Object.*** The ratio between the object area and the image area is larger than 0.5.
CL	***Clutter.*** The foreground and background regions around the object have similar color. We labeled images that their global color contrast value is larger than 0.2, local color contrast value is smaller than 0.9 with clutter images (see Sec. 3).
HO	***Heterogeneous Object.*** Objects composed of visually distinctive/dissimilar parts.
MB	***Motion Blur.*** Objects have fuzzy boundaries due to shake of the camera or motion.
OC	***Occlusion.*** Objects are partially or fully occluded.
OV	***Out-of-View.*** Part of object is clipped by image boundaries.
SC	***Shape Complexity.*** Objects have complex boundaries such as thin parts (*e.g.*, the foot of animal) and holes.
SO	***Small Object.*** The ratio between the object area and the image area is smaller than 0.1.

(b)

Fig. 3.3 Statistics in terms of objects' attributes. (a) is cited from [151]. (b) is cited from SOC [21].

3.3 A dataset for salient object segmentation in 360° images

Considering the key aspects (issues) concluded at the last section, we propose a new image dataset for fixation-based salient object segmentation in 360° panorama [256]. This section presents the details of this work.

3.3.1 Introduction

The panoramic image[8], which captures the content on the whole 360°×180° viewing range surrounding a viewer, plays an import role in VR/AR applications and distinguishes itself from traditional 2D image which covers only local viewport. Recently, civil Head-Mounted Displays (HMDs) have been developed to provide observers an immersive and interactive experience by allowing them to freely rotate their head and thus focusing on desired scenes and objects. Considering the fact that some salient parts of the 360° image attract more human attentions than the others [26], visual saliency prediction (*a.k.a.* fixation prediction) in panorama becomes one of the appealing issues in the field of computer vision and is considered as a key to explore human observation behavior in virtual environments. The fixation prediction and salient object segmentation are both closely related to the concept of visual saliency. Thanks to the accessibility of HMDs and eye trackers, image [25] and video (e.g., [76–78]) datasets have been constructed for the deep learning-based fixation prediction in panoramic content. However, [257] is the only research for 360° salient object segmentation, which does not use the fixations as a guidance for the salient object annotation.

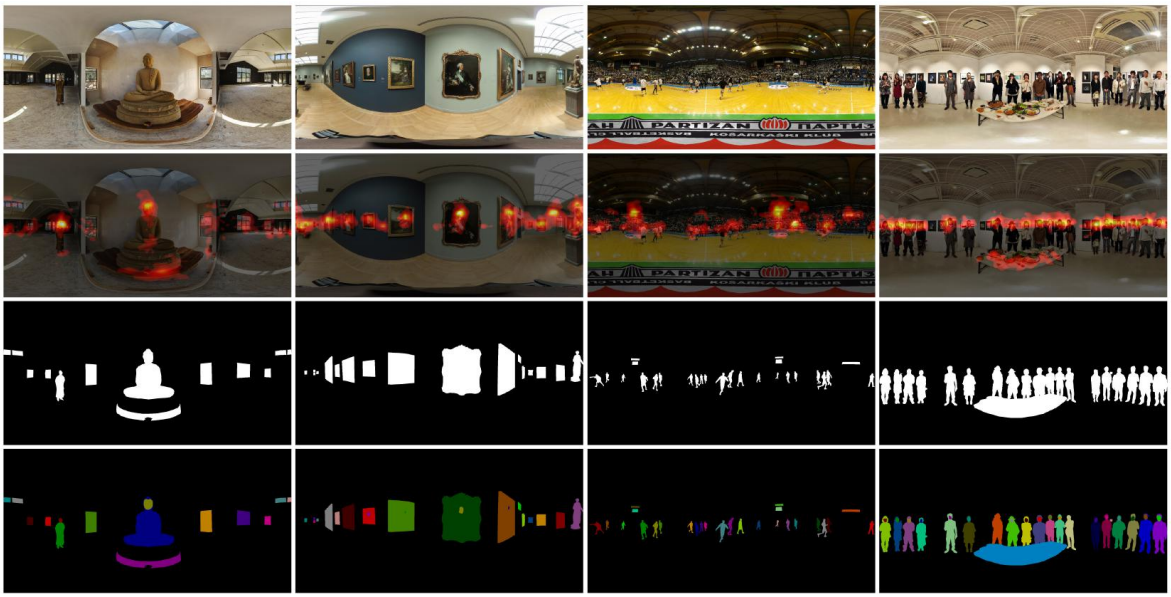


Fig. 3.4 Examples of the proposed fixation-based panoramic image dataset, *i.e.*, F-360iSOD. The first row shows four panoramic images presented as equirectangular image. The second row presents images overlapped with thresholded fixation maps. The third row denotes object-level ground truth for salient object segmentation. the fourth row indicates instance-level pixel-wise masks.

[8] In this thesis, panoramic, 360°, omnidirectional are used interchangeably.

As shown in Fig. 3.4, 360° images tend to own richer scenes and much more foreground objects compared to images collected in traditional 2D salient object segmentation datasets (e.g., [17–20,90]). Therefore, it is more challenging to differentiate the salient objects from the non-salient ones in panoramas. Preserving panoramic images with a few obvious foreground objects while discarding those ambiguous ones may bring selection bias to the dataset, thus being inefficient for exploring the real human attention behavior as viewing panoramic content. Based on the strong correlation between fixation prediction and explicit human judgements [39], and the successfully established fixation-based 2D salient object segmentation datasets [22, 23, 39], an intuition is that the salient objects in panoramas can also be manually annotated with the assistance of fixations, thus representing the real-world daily scenes. Thus, the main content of this section are:

- A fixation-based 360° image dataset (F-360iSOD) with both object-/instance-level pixel-wise annotations.
- A new benchmark includes six state-of-the-art 2D salient object segmentation models [121, 138–141, 297], evaluated by five widely used salient object segmentation metrics [276–279, 298].
- A discussion towards key issues for 360° image salient object segmentation dataset construction.

The uniqueness of the proposed F-360iSOD. As topic-related, there are two types of fixation-based panoramic datasets focusing on head movement prediction and eye movement prediction, respectively. Datasets such as 360-VHMD [27], VR-VQA48 [299] contain only head tracking data, while Salient!360 [25], Stanford360 [26], VQA-OV [78], VR-scene [76] and 360-Saliency [77] provide ground-truth eye fixations. Besides, 360-SOD [257] is a newly proposed omnidirectional image dataset for salient object segmentation. However, the salient objects are labeled based on pure explicit subjective judgements, rather than fixation-based guidance. Besides, the dataset does not provide instance-level ground truth or object category labels.

3.3.2 Dataset statistics

F-360iSOD contains 107 (52 indoor/55 outdoor) panoramic images with challenging real-world daily scenes, 1,165 salient objects (from 72 object classes) manually labeled with precise object-/instance-level masks.

Image collection. The F-360iSOD is a 360° image dataset with totally 107 panoramic images collected from Stanford360 [26] and Salient!360 [25] which contain 85 and 22 equirectangular images, respectively[9]. All the images of the proposed F-360iSOD are represented as equirectangular images with a medium resolution of 2048×1024 for convenient processing.

Salient object annotation. Inspired by 2D salient object segmentation datasets [22, 23, 39] where fixation data were used to aid the salient object annotation, an expert was asked to manually annotate (by tracing boundaries) the salient objects with both the object-/instance-level masks on the collected

[9] Stanford360 and Salient!360 are so far the only panoramic image datasets that provide eye movement based fixation data.

equirectangular images, under the guidance of fixation maps convoluted by a Gaussian with a standard deviation empirically set to 3.34° of visual angle [25] (note that each of the Gaussian-smoothed fixation maps is thresholded with an adaptive saliency value to keep the top one-10th of each self before shown to the annotator). The whole annotation process has been repeated three times to pass the quality check implemented by two other experts, before gaining the final ground truth. Besides, nine images without any salient object annotations are reserved in F-360iSOD, to avoid the common selection bias of 2D salient object segmentation datasets (as mentioned in "balanced datasets" at the last section), brought by an assumption that there is at least one salient object in each of the image. The total dataset's pixel-wise annotations are visualized in Fig. 3.5, Fig. 3.6, Fig. 3.7 and Fig. 3.8.

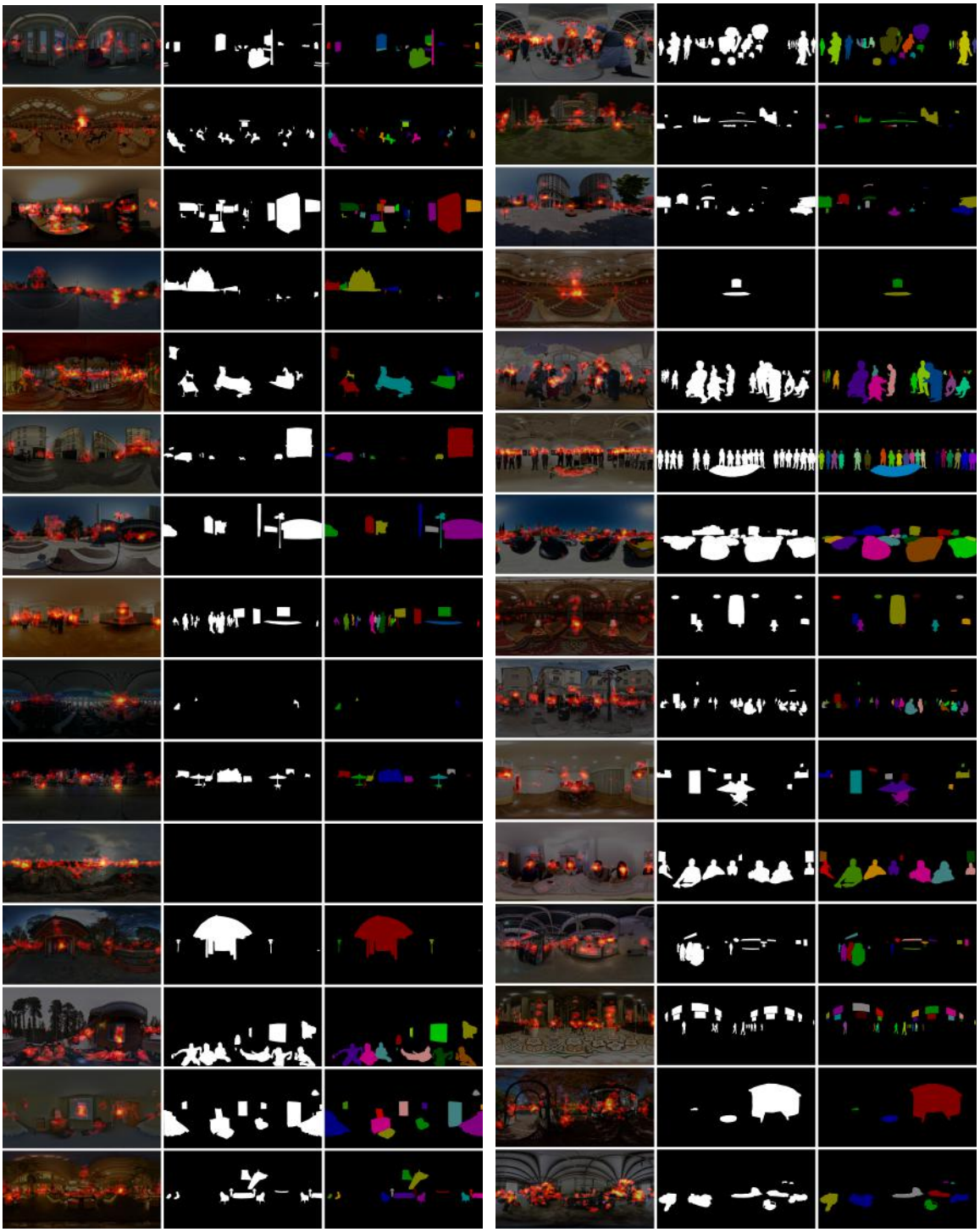


Fig. 3.5 Visualization of the proposed F-360iSOD (1/4).

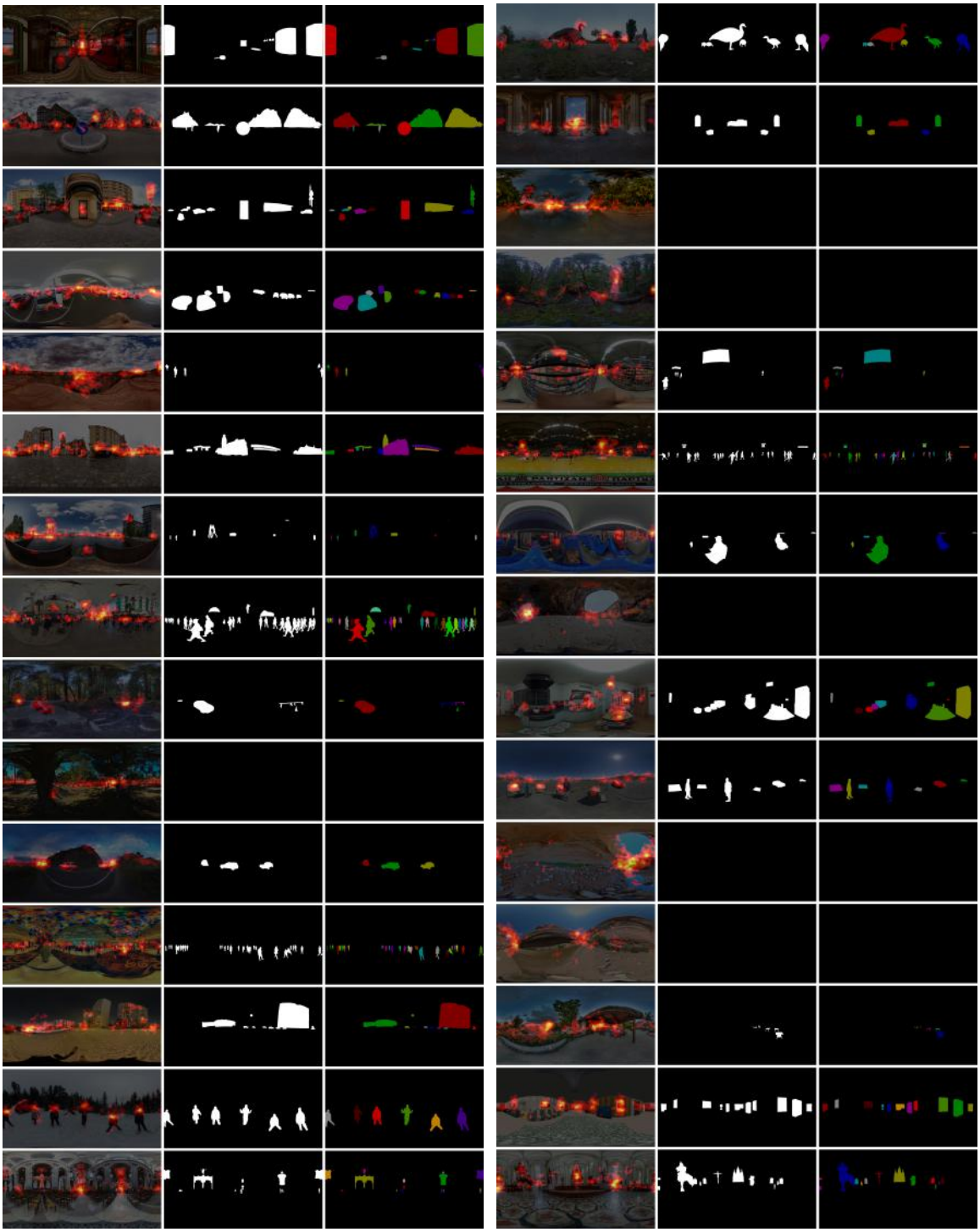


Fig. 3.6 Visualization of the proposed F-360iSOD (2/4).

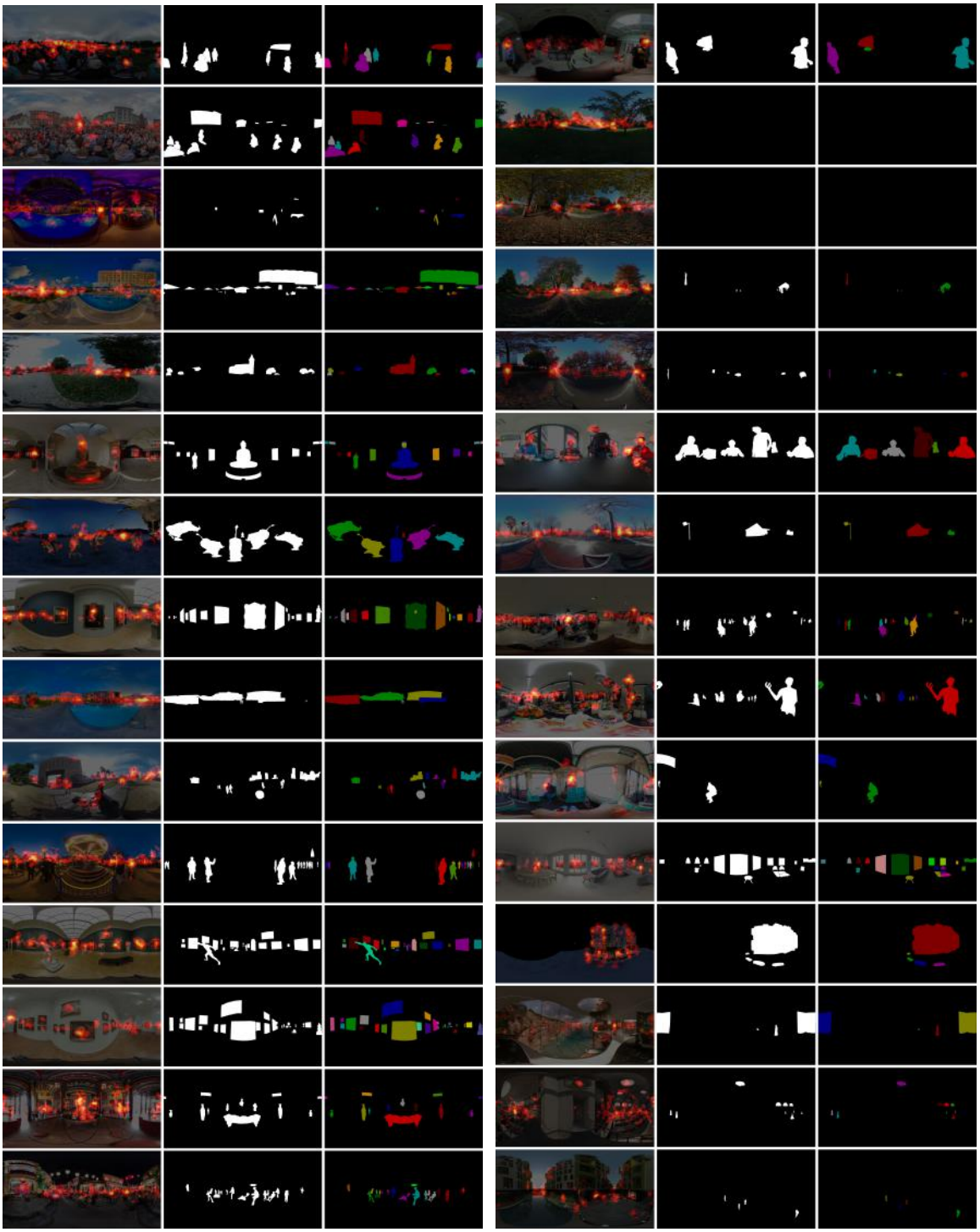


Fig. 3.7 Visualization of the proposed F-360iSOD (3/4).

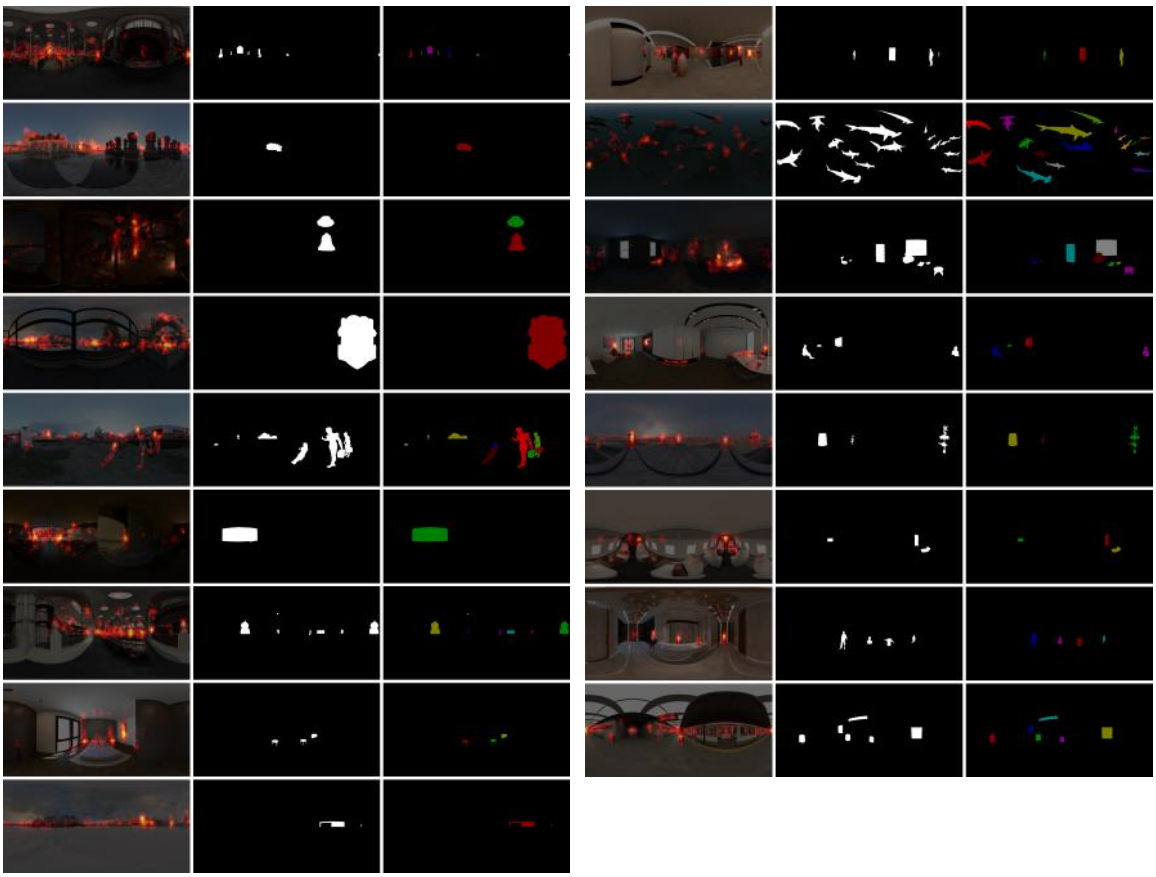


Fig. 3.8 Visualization of the proposed F-360iSOD (4/4).

Dataset statistics. In F-360iSOD, each salient object belongs to one specific class. Generally, there are 1,165 salient objects from 72 categories, thus reflecting 7 aspects (human, text, vehicle, architecture, artwork, animal and daily stuff) of the real-life common scenes (Fig. 3.9). The "person" category occupies the largest proportion with a number of instances of 386; other relative large object classes include "painting", "text", "building", "face" and "car", with a number of instances of 92, 89, 86, 75 and 72, respectively.

3.3.3 Benchmark studies

In this sub-section, we detail our works towards extensive benchmark studies based on our proposed F-360iSOD. A comprehensive benchmark study usually consists of a consistent protocol (which usually includes dataset split, traning/testing strategies for benchmark models and evaluation metrics), and systematical qualitative/quantitative analysis towards experimental results.

Dataset split. The F-360iSOD consists of one training set and two testing sets, which are denoted as F-360iSOD-train, F-360iSOD-testA and F-360iSOD-testB, respectively. The F-360iSOD-train contains 68 equirectangular images from the Salient!360, while the F-360iSOD-testA collects the remaining 17 (85 in total). Besides, the F-360iSOD-testB is established to enable the cross-testing forsalient object segmentation models, with 22 images from the other panoramic image dataset, *i.e.*,

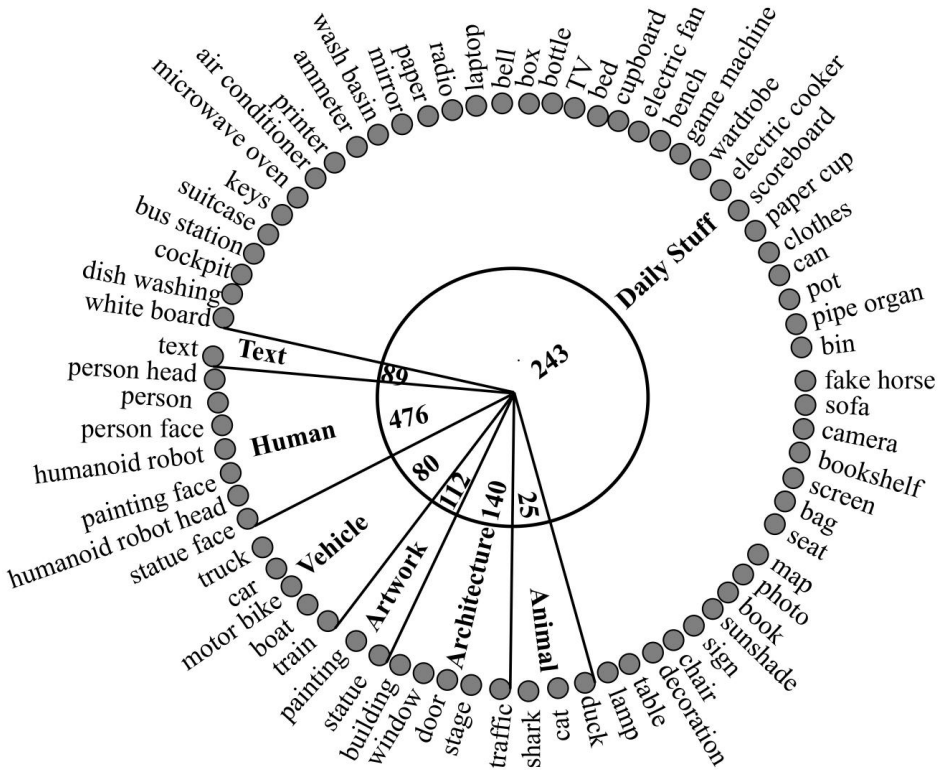


Fig. 3.9 Statistics of object categories of the proposed 360° image salient object segmentation dataset, *i.e.*, F-360iSOD.

Stanford360.

Projection methods. By wearing HMDs, people are able to freely rotate their head to make multiple viewports focusing on the attractive regions of the surrounding 360° content. Based on this prior knowledge, we apply cube map projection (where a 360° image is projected into 6 rectangular patches) to process 68 panoramic images (from F-360iSOD-train) with multiple rotation angles (0°, 30°, 60° both horizontally and vertically [81]). Thus, we gain 54 (6×3×3) patches representative of multiple fields of view for each of the 360° image. 3,672 (54×68) 2D patches (256×256) are therefore generated and used as inputs for the fine-tuning of 2D salient object segmentation models.

Evaluation metrics. To measure the agreement between manually labeled ground truth and model predictions, five widely used salient object segmentation metrics were adopted: F_β-measure [276], weighted F_β-measure (Fbw) [298], mean absolute error (MAE) [277], structural measure (S-measure) [278] and enhanced-alignment measure (E-measure) [279]. The details of salient object segmentation metrics are illustrated in Section 2.7. And it is worth noting that,

$$S = \alpha \times S_o + (1-\alpha) \times S_r, \tag{3.11}$$

where S_o and S_r denote the object-/region-aware structure similarities, respectively; α is empirically set to 0.7 ($\alpha = 0.5$ in 2D) to attach more importance on object structure, based on the observation that panoramic images are usually dominated by small salient objects distributed over the whole image

Table 3.4 A quantitative comparison between six state-of-the-art salient object segmentation models on F-360iSOD, where F_β^w means Fbw, S represents S-measure. Note that the top three results of each column are highlighted in red, green and blue, respectively.

Methods	F-360iSOD-testA			F-360iSOD-testB		
	$F_\beta^w \uparrow$	$S \uparrow$	$MAE \downarrow$	$F_\beta^w \uparrow$	$S \uparrow$	$MAE \downarrow$
SCRN [297]	.551	.809	.050	.124	.708	.034
BASNet [138]	.567	.825	.046	.118	.683	.048
CPD [139]	.521	.763	.052	.129	.695	.032
PoolNet [140]	.500	.834	.068	.136	.716	.058
GCPANet [121]	.630	.822	.045	.106	.693	.039
EGNet [141]	.715	.864	.045	.190	.714	.041

(e.g., Fig. 3.4), rather than one or multiple spatially connected foreground objects located at the center of the image.

Benchmark models. As stated in Chapter 2, convolutional networks (CNNs)-based models dominate the field of salient object segmentation. The CNN segmentation models differentiate themselves from other deep learning methods by predicting saliency maps as outputs, rather than classification scores.

EGNet [141] is one of the recently proposed state-of-the-art models. The method was motivated by the idea that simultaneously learning the salient edge and object information can help improving performance of salient object segmentation models. It modeled these two complementary information with an independent network outside the VGG-based backbone [2]. SCRN [297] is another newly proposed salient object segmentation model that considers the edge information. It also implements the salient object segmentation and salient edge detection in a synchronous manner, by stacking several so-called cross refinement units in an end-to-end manner. BASNet [138] proposed residual refinement module and hybrid loss to refine the salient objects boundaries in predicted saliency maps. PoolNet [140] improved the feature extraction efficiency of multiple layers of current U-shape architecture by adding two new modules, which were both designed based on simple pooling techniques. GCPANet [121] is a more recently proposed method which brought improvements to the traditional bottom-up/top-down networks by proposing four new modules. CPD [139] modified the traditional encoder-decoder framework to directly refine high-level features by generated saliency maps, without the consideration of low-level features. The idea here is different from PoolNet and GCPANet, which integrated both the low-/high-level features.

Benchmark results. In this study, each of the salient object segmentation models is fine-tuned on the F-360iSOD-train with an initial learning rate of one-10th of their default, and a batch size of 1. The training process will stop as the S-measure value on the F-360iSOD-testA starts to go down. As a result, it takes about 20 epochs for BASNet [138], EGNet [141], CPD [139] and SCRN [297] to converge, while 70 for PoolNet [140] and 15 for GCPANet [121]. The quantitative and qualitative comparison between the six state-of-the-art 2D salient object segmentation models on both the F-360iSOD-testA/B are illustrated in Table 3.4, Fig. 3.10 and Fig. 3.11, respectively.

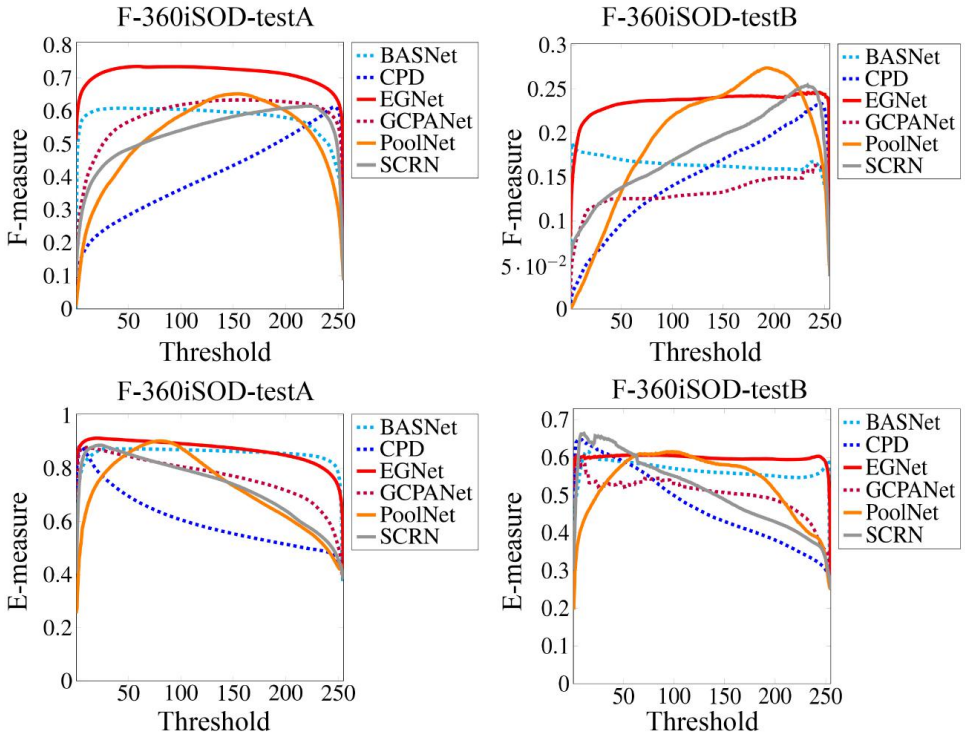


Fig. 3.10 F-measure curves and E-measure curves of six state-of-the-art salient object segmentation models on the proposed F-360iSOD.

3.3.4 Discussion

Challenging 360° salient object segmentation dataset. All benchmark models are constrained to some extent on the proposed F-360iSOD, even though achieving high performances in 2D domain. The limitation is mainly due to the challenges brought by the features of 360° dataset, such as equirectangular projection-induced distortions, small objects and clutter scenes, etc.

Fixation-based complexity analysis. Since the panoramic images tend to contain much more scenes and objects than 2D images, the ambiguity of saliency judgements in panoramas should also be considered, which can be quantified by inter observer congruency (IOC) [300] and entropy based on fixation maps, which are re-smoothed with a Gaussian with a standard deviation of 1° visual angle to reflect human foveal size [300]. As an image with high IOC and low entropy is usually considered to be simple, the F-360iSOD-testB should be easier to explore when compared with the F-360iSOD-testA (Fig. 3.12), from a perspective of human judgements.

Unseen object classes. All competing models fail on the F-360iSOD-testB, mainly due to the presence of unseen object classes in Stanford360, such as sharks, bells, robots, etc. People are capable of recognizing new object categories when provided with high-level descriptions. This strong generalization ability is still absent in current salient object segmentation models.

Instance-level ground truth. The proposed F-360iSOD is the first 360° dataset that provides instance-level semantic labels for salient objects. Future salient object segmentation models are capable of recognizing the individual instances from multiple classes, which is crucial for practical applications,

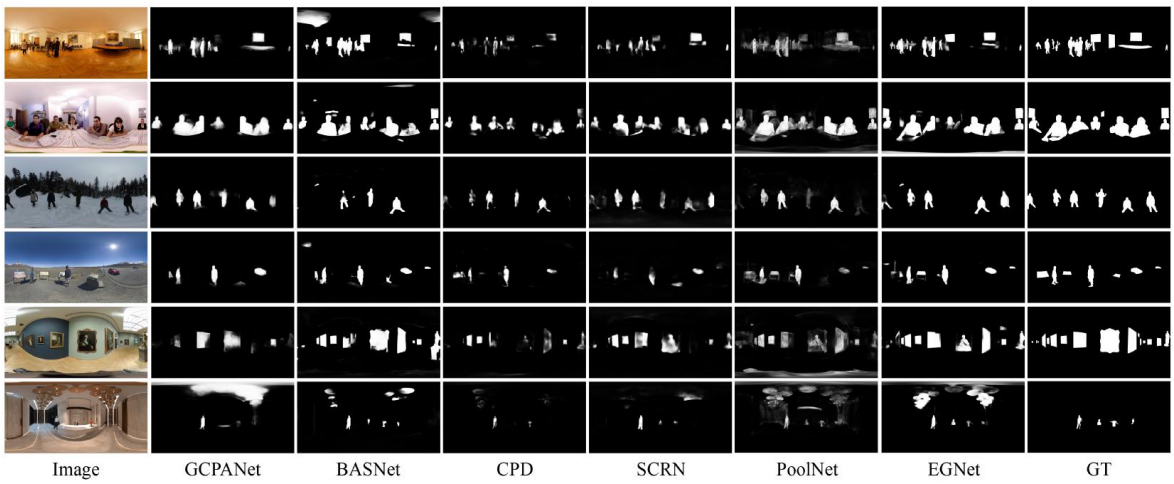


Fig. 3.11 A qualitative comparison between six state-of-the-art salient object segmentation models on F-360iSOD.

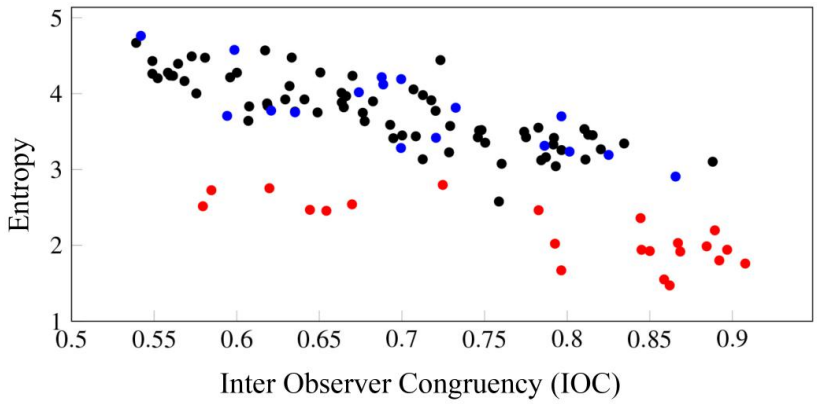


Fig. 3.12 A fixation-based complexity analysis of the proposed F-360iSOD. The F-360iSOD-train, F-360iSOD-testA/B are annotated in black, blue and red, respectively.

e.g., image captioning and scene understanding.

3.3.5 Conclusion

This section presents the proposed fixation-based 360° image dataset (F-360iSOD), with precisely annotated salient objects/instances from multiple classes representative of real-world daily scenes. Six recently proposed salient object segmentation methods are fine-tuned and tested on the F-360iSOD. Results show a limit of current 2D models when directly applied to the salient object segmentation in panoramas.

3.4 A dataset for salient object segmentation in 360° videos

This section introduces the details towards our PAVS10K, which is the first video-based 360° dataset proposed for salient object segmentation. Importantly, as PAVS10K uses both audio and visual cues for salient object judgment and annotations, this is also the first audio-visual dataset in the salient object segmentation community.

3.4.1 Introduction

Studying and modeling human attention in 360° panoramic real-life[10] environment has been an important issue in the fields of computer vision and multi-modal learning. Recently, the issue gains increasing attention from both communities as the booming development of virtual and augmented reality industries (*e.g.*, the recent boom of "metaverse"[11] which seeks to establish a new immersive digital extended world facilitating more efficient social network, education, entertainment, *etc.*).

Particularly, as the popularization of civil 360° cameras such as GoPro Max, Ricoh Theta Z1 and Insta360 ONE series, 360° panoramic images and videos are nowadays easily acquired. In this case, several 360° visual saliency prediction datasets [25–28, 76, 77, 80] have been proposed, to enable deep learning researches towards human attention modeling in 360° static and dynamic real-life daily scenes. Besides, more recent audio-visual dataset [79] investigates the influence of audio cues towards human perception in 360° videos. However, these datasets provide only head or eye movement data as ground truth, thus not being able to strictly reflect human attention to specific salient targets.

Besides, recent researches [301–303] have brought much attention to audio-visual object localization. Specifically, as the development of large-scale audio-visual datasets such as MUSIC [304], AudioSet [305], AVE [71], VGGSound [306] and ObjectFolder [307], the community has recently witnessed a booming trend of audio-visual researches, *e.g.*, [308–315]. Particularly, recent audio-visual object localization methods [301–303, 316–323] are closely related to salient object segmentation in terms of object-level attention modeling. It is worth noting that, these researches focus on the detection of sounding objects, rather than the salient objects. As a comparison, panoramic video salient object segmentation aims to finely segment the audio-visual salient objects, where manually labeled pixel-wise ground truth are necessary for the training and quantitative evaluation of models. In fact, mixed reality applications such as remote collaboration [324] and virtual object rendering [325] are closely related to object-level human attention modeling in dynamic 360° panoramas. However, so far there is no work focuses on object-level audio-visual saliency detection in challenging panoramic videos representing realistic scenes.

On the other hand, salient object segmentation, which mimics human attention by finely segmenting the visual salient objects in given images, has been constantly appealing attention from the computer vision community during the last decade [15]. As illustrated in Chapter 2, according to the types of training data, current salient object segmentation(*a.k.a.* SOD) methodologies can be classified into eight categories, *i.e.*, I (image)-SOD [16, 21, 326], V (video)-SOD [23, 162, 163], Co-

[10] "Real-life" targets indicate the objects/scenes captured by photographers in real life, thus distinguishing itself from virtual rendered ones.

[11] Meta: `https://about.facebook.com/meta/`

SOD [195, 198, 327], RGBD (depth)-SOD [225, 228, 270], RGBT (thermal infrared)-SOD [231–233], LF (light field)-SOD [238, 244, 251], HR (high resolution)-SOD [201, 328] and RS (remote sensing)-SOD [205, 206]. Despite a prosperous development of the salient object segmentation community, current state-of-the-art methods still suffer from two limitations that prevent them from modeling unbiased human attention as in real-world daily scenes. First, these models all rely on only visual data for the detection, thus hardly reflecting human attention in realistic circumstances where audio cues indeed play an important role (*e.g.*, audio-visual saliency network predicts more accurate results than visual-only ones [59]). Besides, these methods, with so far the only audio-visual salient object segmentation method [329] (not released), all focus on visual data with limited field-of-views (FoVs), *e.g.*, common 2-D images and videos, thus ignoring the rich visual cues as observed in real-life daily scenes, where people are able to explore an omnidirectional view with the FoV of 360°×180° by freely rotating their heads. Recent researches [256–258] shift attention to 360° panoramic image based salient object segmentation by proposing new datasets consisting of hundreds of equirectangular (ER) images[12] and corresponding pixel-wise ground truth. However, with only limited static visual cues provided by the datasets, current panoramic image-based salient object segmentation methods [257, 258] are far from representing real-life object-level human attentions, where the modeling of dynamic visual and audio information are essential.

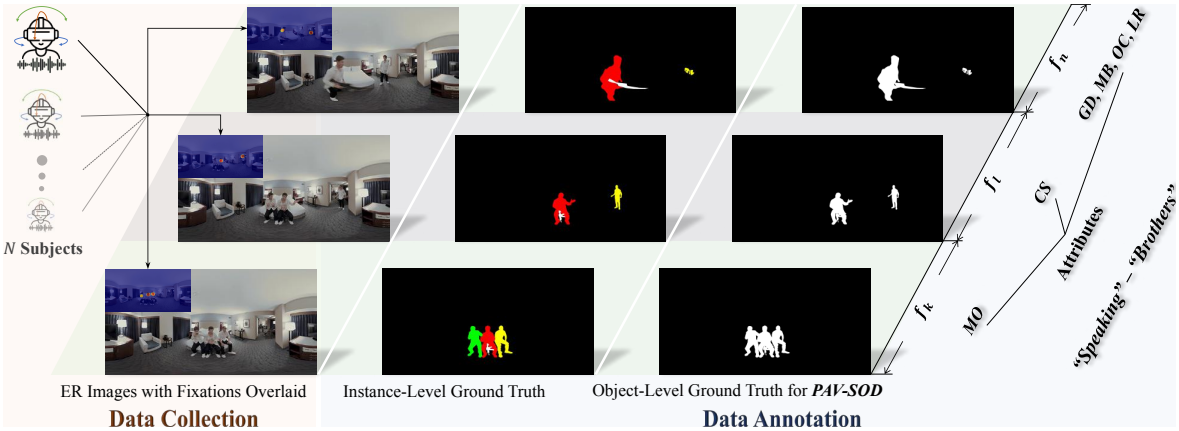


Fig. 3.13 An example of our **PAVS10K** where coarse-to-fine annotations are provided, based on a guidance of fixations acquired from subjective experiments conducted by multiple (N) subjects wearing Head-Mounted Displays (HMDs) and headphones. Each (*e.g.*, f_k, f_l and f_n, where random integral values $\{k,l,n\} \in [1,T]$) of the total equirectangular (ER) video frames (T) of the sequence "Speaking"(Super-class)-"Brothers"(sub-class) are manually labeled with both object-level and instance-level pixel-wise masks. According to the features of defined salient objects within each of the sequences, multiple attributes, *e.g.*, "*multiple objects*" (MO), "*competing sounds*" (CS), "*geometrical distortion*" (GD), "*motion blur*" (MB), "*occlusions*" (OC) and "*low resolution*" (LR) are further annotated to enable detailed analysis for **panoramic video salient object segmentation** modeling.

To model object-level audio-visual attention in realistic omnidirectional dynamic scenes, we conduct systematical researches, *i.e.*, establishing the first 360° video salient object segmentation dataset

[12] ER images are the most widely used lossless planar representation of 360° images.

with hierarchical annotations (*i.e.*, PAVS10K), building a new benchmark with state-of-the-art methods collected from multiple related fields including image-based salient object segmentation, video-based salient object segmentation and video object segmentation. Specifically, the main contributions of our proposed dataset and benchmark are:

- We propose a large-scale panoramic video-based salient object segmentation dataset, namely **PAVS10K**, which consists of uniformly sampled 10,465 4K-resolution ER video frames (from total 62,455 frames), with corresponding super-/sub-class labels and manually labeled object-level and instance-level pixel-wise masks[13] (Fig. 3.13). We further attach 360° salient object segmentation related challenging attributes to each of the 67 sequences in our PAVS10K (*e.g.*, Fig. 3.13). The coarse-to-fine labels enable comprehensive benchmark studies and detailed analysis regarding 360° salient object segmentation modeling.

- We establish so far the largest 360° video-based salient object segmentation benchmark which collects 13 state-of-the-art methods from the fields of 2D image salient object segmentation (7), 2D video salient object segmentation (2), video object segmentation (3) and panoramic image salient object segmentation (1). For fair comparison, we systematically evaluate all 13 models based on our PAVS10K, with four widely used salient object segmentation metrics.

3.4.2 Dataset statistics

Our PAVS10K dataset aims at segmenting the salient objects by taking advantage of both audio and visual cues in 360° dynamic scenes. A comparison between our PAVS10K and the current widely used salient object segmentation related datasets is shown in Table 3.5, in terms of scales, annotation types and diversities.

In this section, we elaborate our challenging large-scale PAVS10K, *i.e.*, the first panoramic video salient object segmentation dataset, in terms of stimuli collection, subjective experiments, annotation pipeline and dataset statistics.

Stimuli collection. The stimuli of PAVS10K were gained from *YouTube* with multiple searching keywords (*e.g.*, 360°/panoramic/omnidirectional video, spatial audio, ambisonics [311]). As a result, our collected stimuli cover various real-world dynamic scenes (*e.g.*, indoor/outdoor scenes), multiple occasions (*e.g.*, sports, travel, concerts, interviews, dramas), different motion patterns (*e.g.*, static/moving camera), and diverse object categories (*e.g.*, human, instruments, animals). They possess a wide range of major challenges for object detection in 360° content, *e.g.*, objects scattered far from the equirectangular image's equator thus suffering from serious geometrical distortions (*e.g.*, salient persons annotated with attribute "geometrical distortion – GD" as shown in Fig. 3.13).

The abundant on-line audio-visual sources provide us with a solid foundation to establish a challenging and representative benchmark dataset. As a result, we obtained 67 high-quality video sequences with a total of 62,455 frames recorded with 62,455×40 eye movement based fixations. Specifically, the 67 sequences are selected based on three criteria:

[13] Collecting the pixel-wise labels was a costly and time-consuming work, and it took us about one year to set up this large-scale dataset.

Table 3.5 A comparison between our proposed PAVS10K and the widely used salient object segmentation (*a.k.a.* SOD)/video object segmentation (VOS) datasets . #Img: The number of images/video frames. #GT: The number of object-level pixel-wise masks (ground truth for SOD). Pub. = Publication. Obj.-Level = Object-Level Labels. Ins.-Level = Instance-Level Labels. Fix. GT = Fixation Maps. † denotes equirectangular (ER) images.

Dataset	Task	Year	#Img	#GT	$\min(W,H)$	$\max(W,H)$	Obj.-Level	Ins.-Level	Attribute	Fix. GT	Audio
ECSSD [90]	I-SOD	CVPR'13	1,000	1,000	139	400	✓				
DUT-OMRON [17]	I-SOD	CVPR'13	5,168	5,168	139	401	✓			✓	
PASCAL-S [18]	I-SOD	CVPR'14	850	850	139	500	✓			✓	
HKU-IS [19]	I-SOD	CVPR'15	4,447	4,447	100	500	✓				
DUTS [20]	I-SOD	CVPR'17	15,572	15,572	100	500	✓				
ILSO [91]	I-SOD	CVPR'17	1,000	1,000	142	400	✓	✓			
SOC [21]	I-SOD	ECCV'18	6,000	6,000	161	849	✓	✓	✓		
SegTrack V2 [147]	VOS	ICCV'13	1,065	1,065	212	640	✓				
FBMS [148]	VOS	TPAMI'13	13,860	720	253	960	✓				
MCL [149]	V-SOD	TIP'15	3,689	463	270	480	✓				
ViSal [150]	V-SOD	TIP'15	963	193	240	512	✓				
DAVIS2016 [151]	VOS	CVPR'16	3,455	3,455	900	1,920	✓		✓		
UVSD [152]	V-SOD	TCSVT'16	3262	3262	240	877	✓				
VOS [22]	V-SOD	TIP'18	116,103	7,467	312	800	✓			✓	
DAVSOD [23]	V-SOD	CVPR'19	23,938	23,938	360	640	✓	✓	✓	✓	
F-360iSOD [256]	PI-SOD	ICIP'20	107†	107	1,024	2,048	✓	✓		✓	
360-SOD [257]	PI-SOD	JSTSP'20	500†	500	512	1,024	✓				
360SSOD [258]	PI-SOD	TVCG'20	1,105†	1,105	546	1,024	✓				
PAVS10K(Ours)	PAV-SOD	2022	62,455†	10,465	**1,920**	**3,840**	✓	✓	✓	✓	✓

- The collected video frames must be in good visual quality, *i.e.*, 4K resolution of each video frame.
- The collected videos must have corresponding audio files including both ambisonics and mono sound.
- The collected video scenes must include recognizable objects which constantly grasp subjects' attention.

Note that we manually trimmed the videos into small clips (29.6s on average) to avoid fatigue during the collection of human eye fixations. As a result, the total video duration is about 1983s (67×29.6s).
Subjective experiments. We detail the supportive subjective experiments from the following three aspects, *i.e.*, equipment, observers and experimental settings.

- *Equipment*. All the video clips were displayed using a HTC Vive HMD embedded with a Tobii eye tracker with 120Hz sample rate to collect eye fixations.
- *Observers*. We recruited 40 participants (8 females and 32 males) aging from 18 to 34 years old who reported normal or corrected-to-normal visual and audio acuity. Twenty participants were randomly selected to watch videos with mono sound (group #1), while the other participants watched videos without sound (group #2). Note that the two groups own the same gender and age distributions. Hence, each video with each audio modality (*i.e.*, with or without sound) was viewed by 20 participants, and each participant viewed (task-free) each video only once.
- *Settings*. All the participants seated in a swivel chair, wearing a HMD with headphones, and asked to explore the 360° panoramic videos without any specific intention. During the experi-

ments, the starting position was fixed to the center at the beginning of every video display. To avoid motion sickness and eye fatigue, we inserted a short rest of a five-second gray screen between two successive videos and a long break of 20 minutes after every 20 videos. We calibrated the system for each participant at the beginning and the end of every long break.

Coarse-to-Fine Annotations. Our annotations vary from scene/sequence level to fine pixel level, thus enabling detailed analysis towards panoramic dynamic audio-visual salient object segmentation modeling.

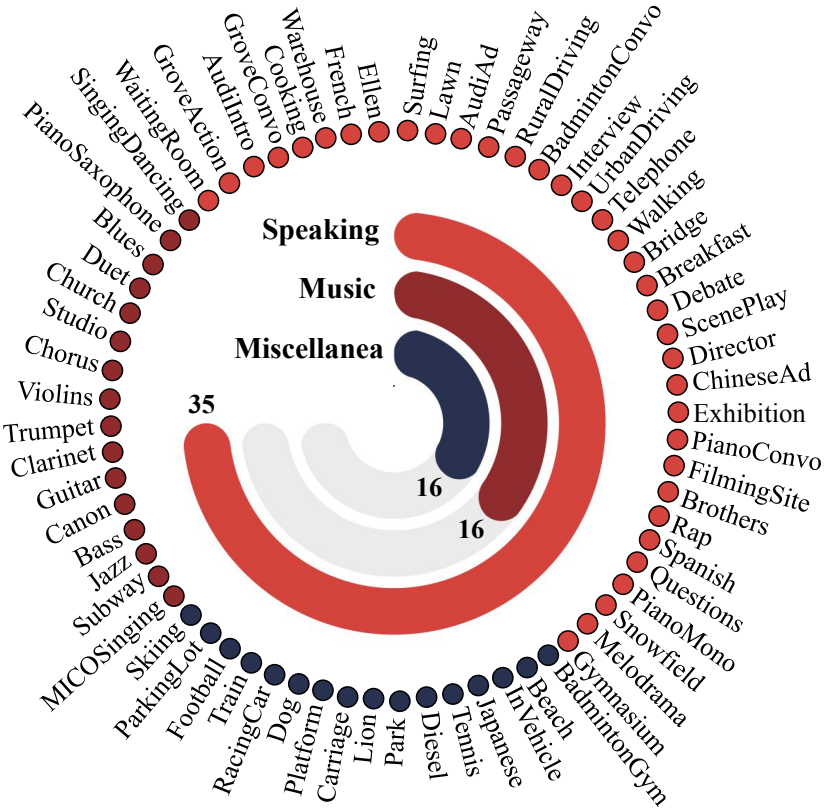


Fig. 3.14 Statistics of the proposed PAVS10K – super-/sub-category information.

Super-/Sub-Scene-Class Labeling. As shown in Fig. 3.14, our PAVS10K contains 67 videos representing 67 audio-visual scene classes. The 67 sub-classes can be categorized to three super classes with a cue of primary sound sources, *i.e.*, speaking (*e.g.*, conversation, monologue), music (*e.g.*, singing, instrument playing) and miscellanea (*e.g.*, the sound of vehicle engines and horns on the streets, crowd noise in the open air). The commonly seen sound sources are shown in Fig. 3.15.

Protocol of pixel-wise manual annotations. Our object-level and instance-level ground truth for conducting panoramic dynamic audio-visual salient object segmentation strictly follow the audio-visual eye fixations acquired from subjective experiments conducted by group #1 (please refer to details in "**subjective experiments**"). The annotation protocol is detailed as follows:

- Inspired by the widely used empirical IoU threshold AP50 (threshold set as 50%) in the field of object detection, we define the salient objects as the objects overlapped with top 50% saliency (*e.g.*, please refer to overlaid fixations as shown in Fig. 3.16).

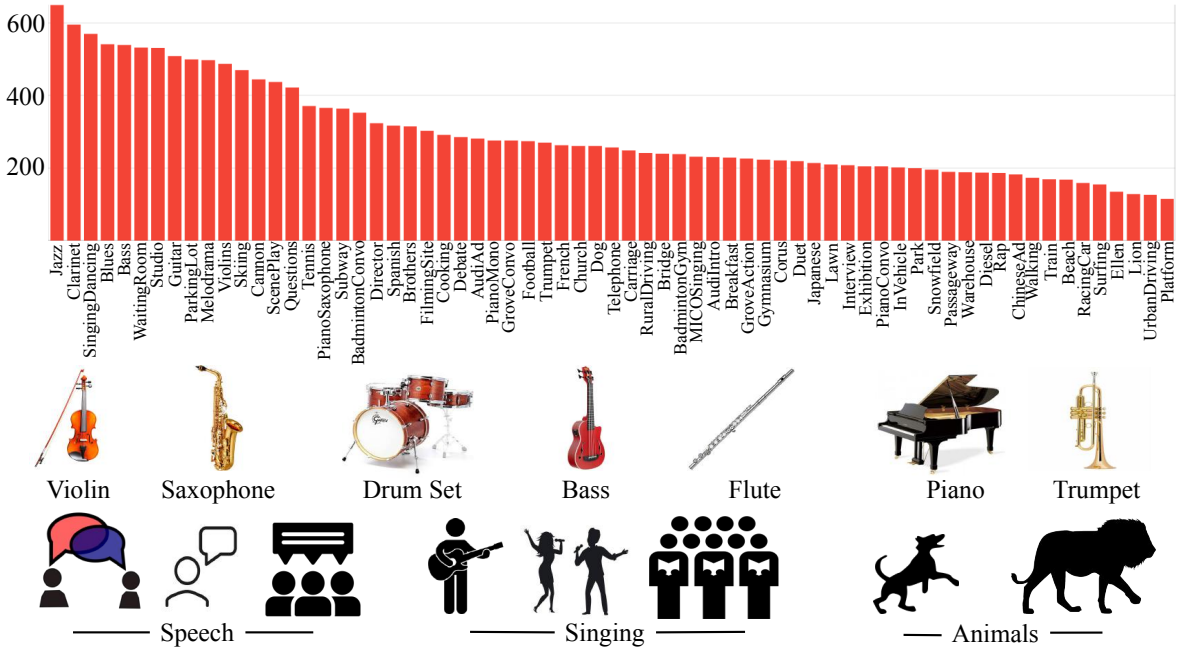


Fig. 3.15 Statistics of the proposed PAVS10K – Instance density (labeled frames per sequence) of each sub-class. Sound sources of PAVS10K scenes, such as musical instruments, human instances and animals.

- Following classical video datasets [22, 148–150], we uniformly extracted 10,465 video frames from the total 62,455 frames with a sampling rate of 1/6, for the pixel-wise annotation.
- We apply the commonly used CVAT toolbox[14] to conduct manual labeling.
- We conducted multiple annotation quality examinations to ensure high-quality pixel-wise labels (Examples are shown in Fig. 3.17).

Object-level masks. Object-level masks denote the pixel-wise binary masks (Fig. 3.13) representing object-level saliency. Researchers participated the manual annotation for the fixation-based salient objects in 10,465 frames. During the labeling process, the annotators were asked to correctly segment the salient objects by finely tracing objects' boundaries, rather than drawing rough polygons. Finally, we obtained 10,465 object-level masks corresponding to 10,465 uniformly extracted video frames, which were then used for panoramic dynamic audio-visual salient object segmentation model training and quantitative evaluation. Fig. 3.15 shows the number of object-level masks of each of the sequences of our PAVS10K.

Instance-level masks. As shown in Fig. 3.13 or Fig. 3.16, an important contribution of our PAVS10K is the instance-level pixel-wise masks, which are rarely seen in current salient object segmentation datasets (Table. 3.5). In fact, compared to conventional salient object segmentation, instance-level salient object segmentation is able to mimic more realistic human visual attention. As a result, we finally gained 19,904 instance-level salient object labels. Please refer to Fig. 3.18, Fig. 3.19, Fig. 3.20, Fig.

[14]CVAT Toolbox: `https://github.com/openvinotoolkit/cvat`

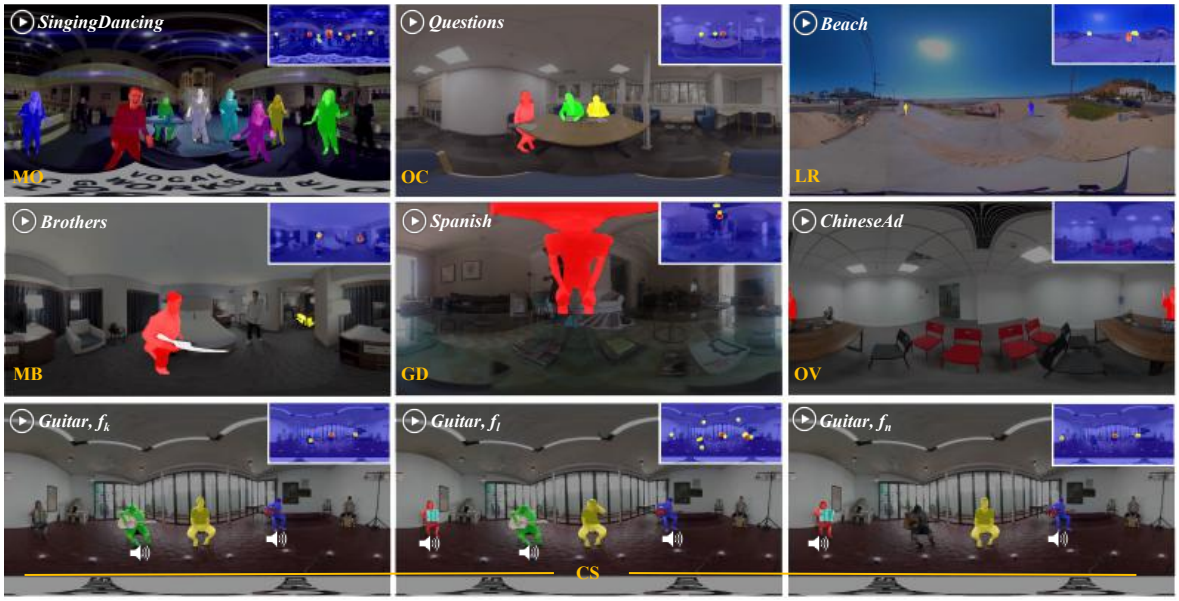


Fig. 3.16 Examples of challenging attributes (detailed description in Table. 3.6) on equirectangular images from our PAVS10K, with instance-level ground truth and fixations as annotation guidance. $\{f_k, f_l, f_n\}$ denote random frames of a given video. Please zoom-in for better view of overlaid fixations.

3.8, Fig. 3.22, Fig. 3.23 and Fig. 3.24 for randomly sampled video frames and their corresponding object-/instance-level masks, of each of the 67 sequences within our PAVS10K.

Attributes labeling. Following the recently proposed large-scale video object segmentation [151] and video-based salient object segmentation [23] datasets, we provide seven attributes to represent the challenges within our PAVS10K, *i.e.*, "*Multiple Objects*" (MO), "*Occlusions*" (OC), "*Low Resolution*" (LR), "*Motion Blur*" (MB), "*Out-of-View*" (OV), "*Geometrical Distortion*" (GD) and "*Competing Sounds*" (CS) (Table. 3.6).

Table 3.6 Description of each of the seven proposed attributes towards panoramic audio-visual salient object segmentation.

Attributes.	Description
MO	*Multiple Objects.* $\geqslant$ three objects occur simultaneously.
OC	*Occlusions.* Object is partially occluded.
LR	*Low Resolution.* Object occupies $\leqslant$ 0.5% of image area.
MB	*Motion Blur.* Moving object with fuzzy boundaries.
OV	*Out-of-View.* Object is cut in half in ER projection.
GD	*Geometrical Distortion.* Distorted object in ER projection.
CS	*Competing Sounds.* Sound objects compete for attention.

It is worth mentioning that, *OV* and *GD* (Fig. 3.13) are exclusive geometrical attributes of ER images, and *CS* is a novel attribute attached to sounding stimuli, thus representing challenging audio-visual scenes where multiple sounding objects compete for human attention. Detailed statistics of the proposed attributes towards each of the sequences of PAVS10K are shown in Table 3.7 and Table 3.8.

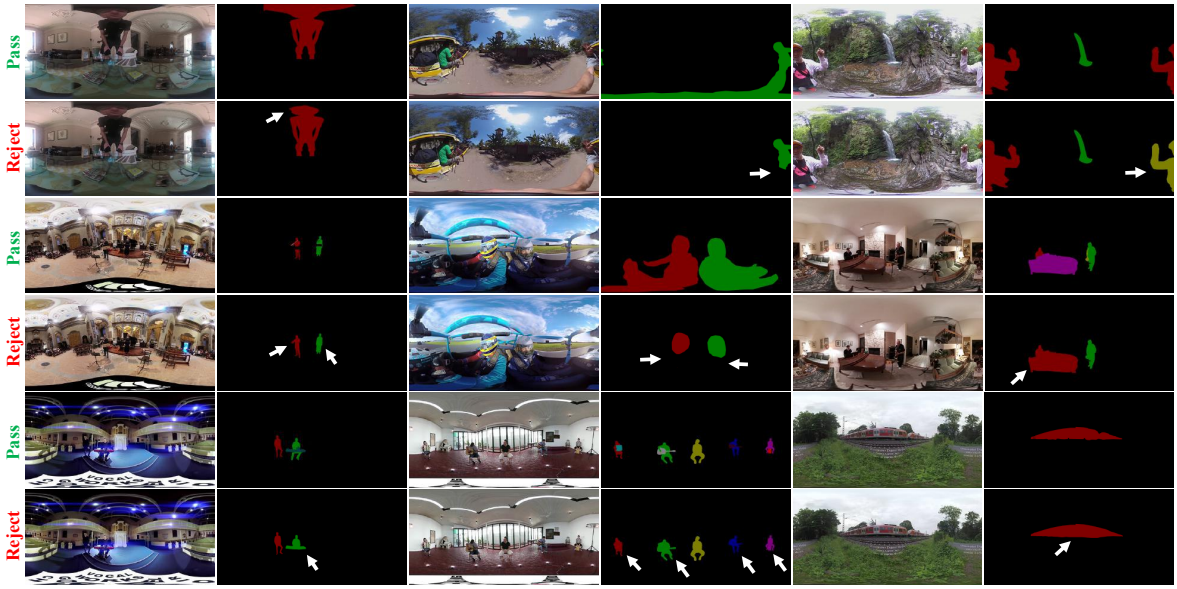


Fig. 3.17 Passed and rejected instance-level pixel-wise labels during quality examination processes.

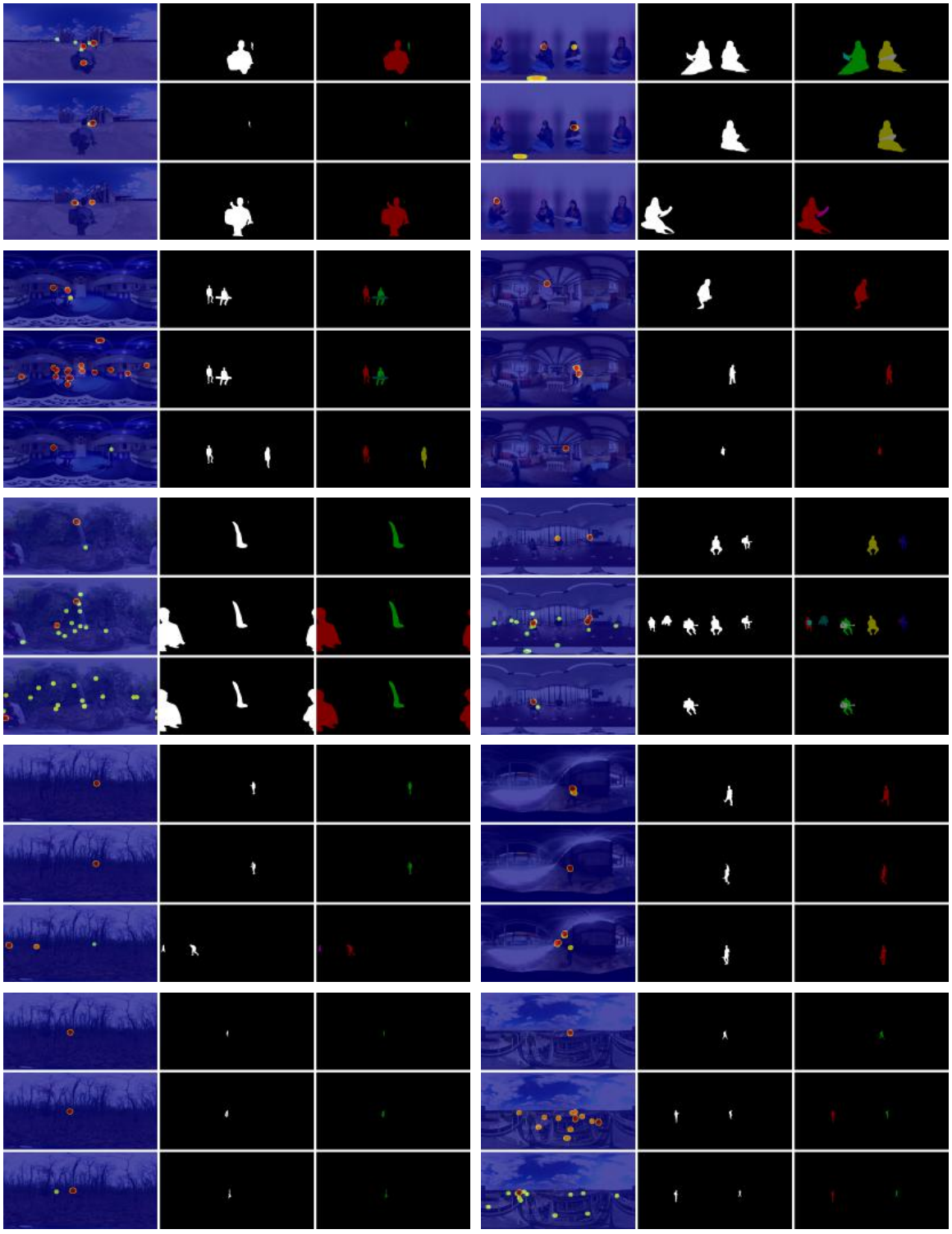


Fig. 3.18 Visualization of the proposed PAVS10K (1/7).

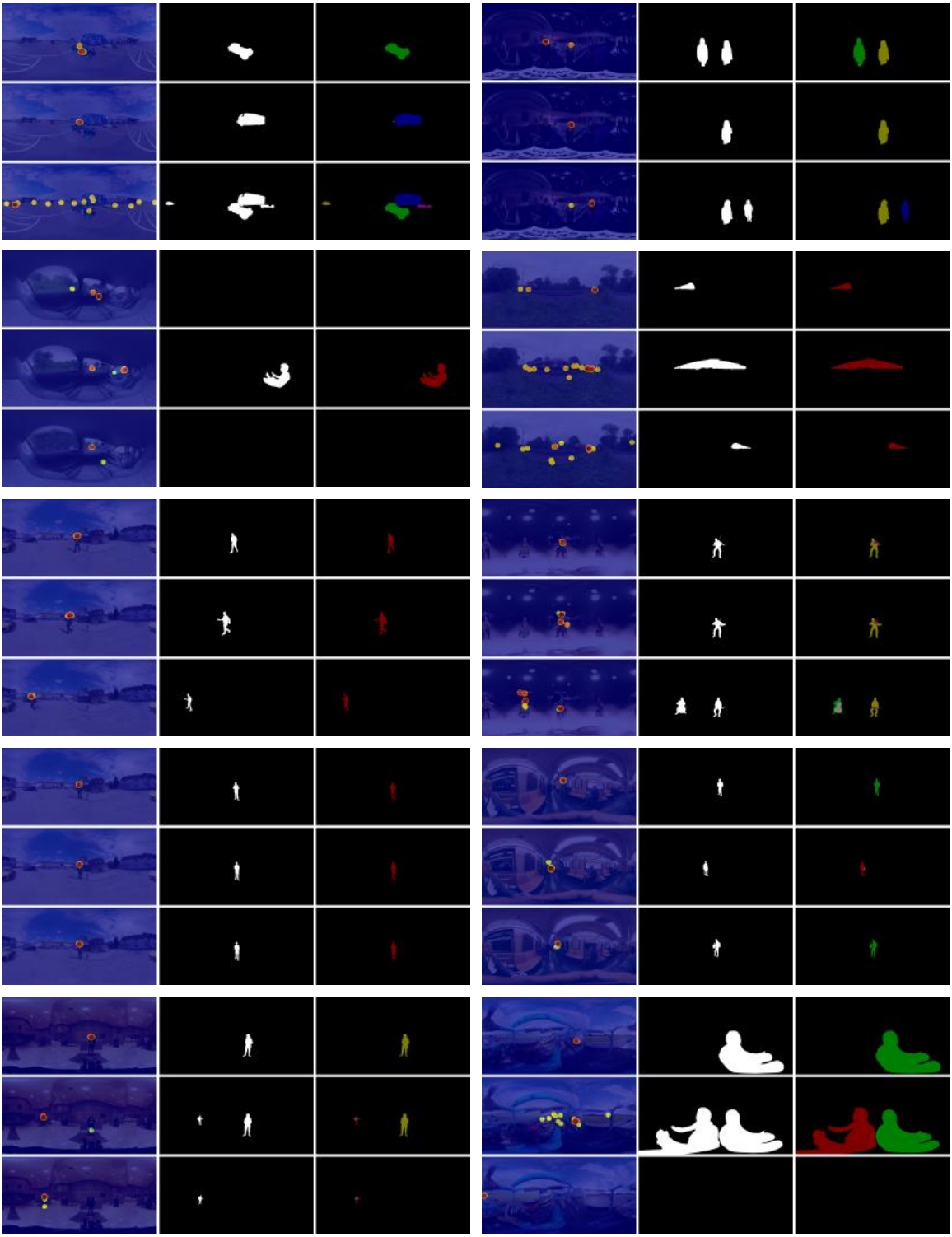


Fig. 3.19 Visualization of the proposed PAVS10K (2/7).

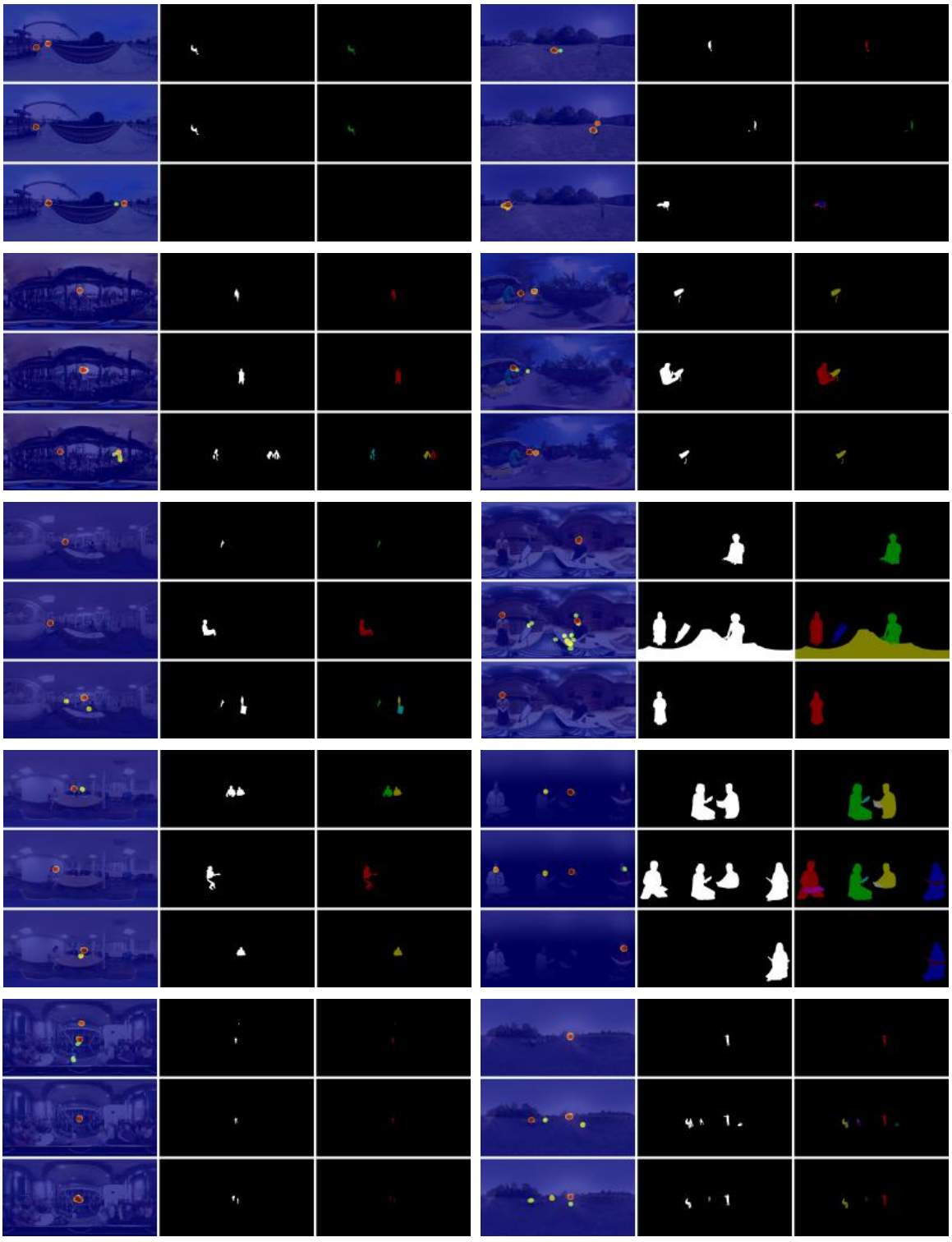


Fig. 3.20 Visualization of the proposed PAVS10K (3/7).

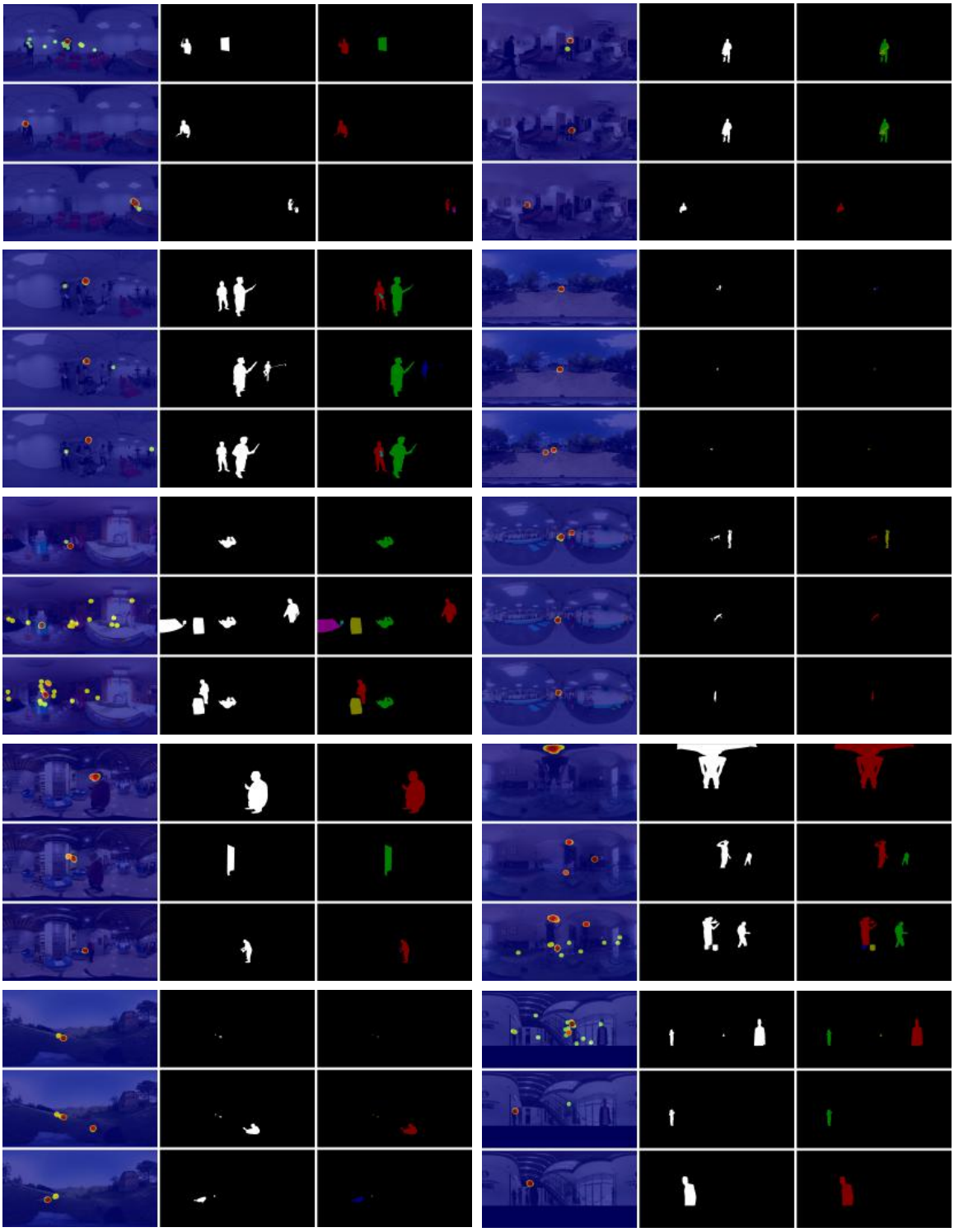


Fig. 3.21 Visualization of the proposed PAVS10K (4/7).

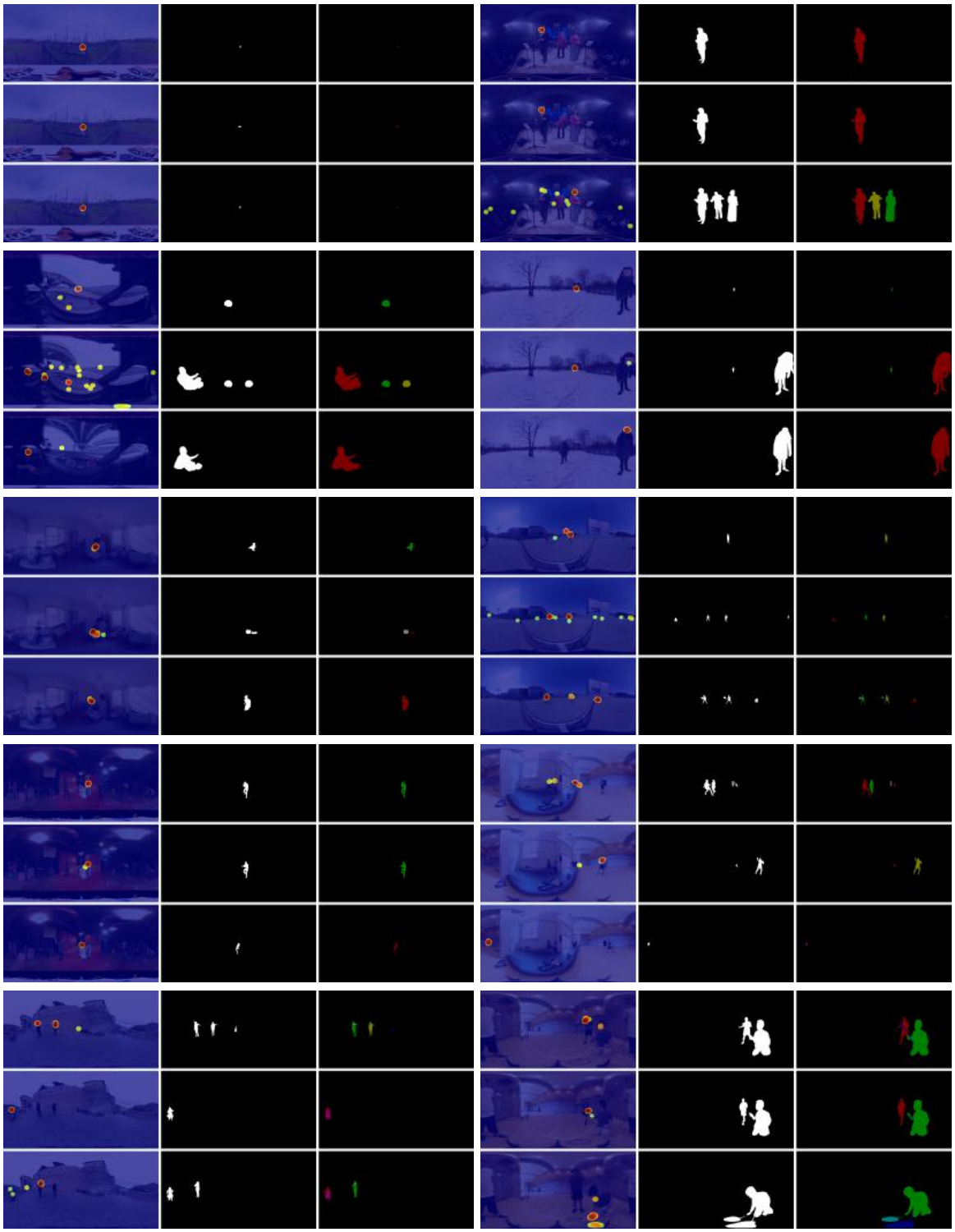


Fig. 3.22 Visualization of the proposed PAVS10K (5/7).

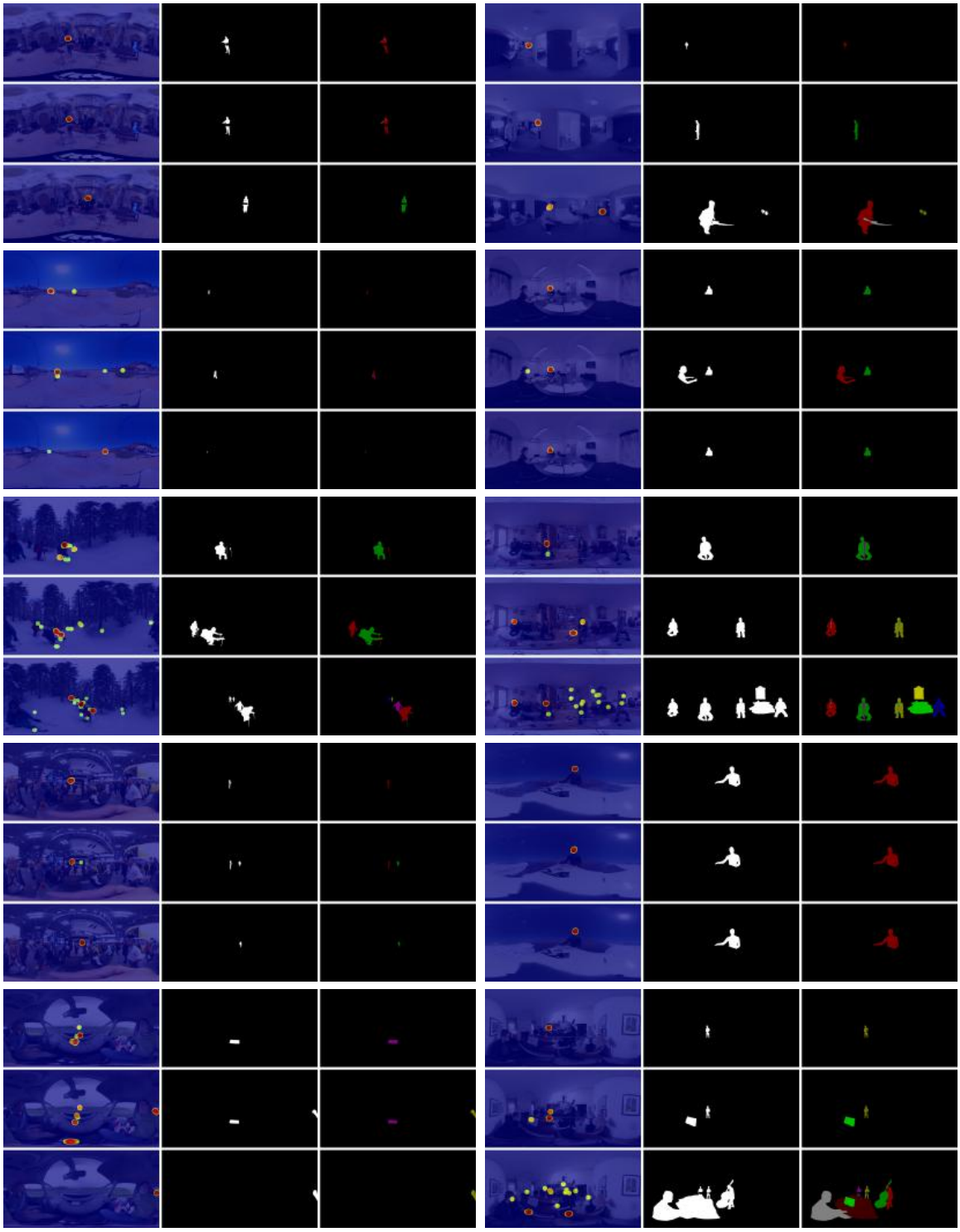


Fig. 3.23 Visualization of the proposed PAVS10K (6/7).

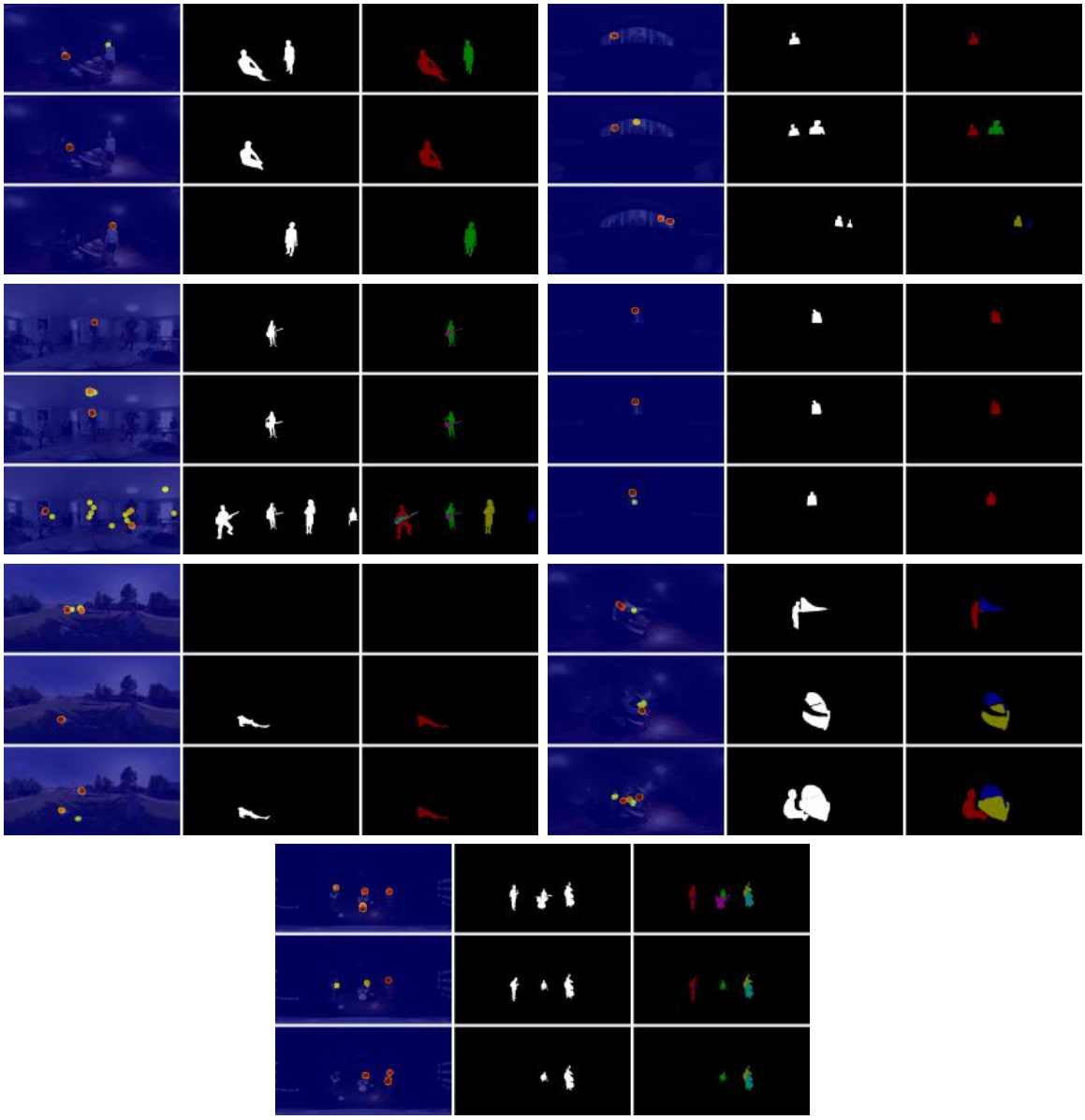


Fig. 3.24 Visualization of the proposed PAVS10K (7/7).

Table 3.7 Attribute details (1/2). General attributes: MO = Multiple Objects. OC = Occlusions. LR = Low Resolution. MB = Motion Blur. 360° geometrical attributes: OV = Out-of-View. GD = Geometrical Distortion. Audio attributes: CS = Competing Sounds.

	Sequence	General				360°		Audio	No.
		MO	OC	LR	MB	OV	GD	CS	
Speaking (35)	French	✓	✓	✓	✓		✓		5
	WaitingRoom	✓	✓	✓	✓			✓	5
	Cooking	✓	✓	✓	✓			✓	5
	AudiIntro	✓		✓		✓			3
	Ellen			✓					1
	GroveAction	✓	✓	✓	✓			✓	5
	Warehouse			✓	✓				2
	GroveConvo	✓	✓	✓	✓			✓	5
	Surfing		✓		✓		✓		3
	Passageway	✓		✓	✓	✓			4
	RuralDriving	✓	✓	✓			✓		4
	Lawn				✓		✓		2
	AudiAd	✓	✓	✓	✓	✓	✓		6
	ScenePlay	✓	✓	✓			✓	✓	5
	UrbanDriving	✓		✓			✓		3
	Interview	✓	✓	✓				✓	4
	Telephone	✓	✓	✓	✓		✓		5
	Walking			✓	✓		✓		3
	Bridge		✓	✓	✓			✓	4
	Breakfast	✓	✓	✓	✓		✓		5
	Debate	✓		✓				✓	3
	BadmintonConvo	✓	✓	✓	✓	✓	✓	✓	7
	Director	✓	✓	✓	✓		✓	✓	6
	ChineseAd	✓	✓	✓	✓	✓	✓		6
	Exhibition			✓					1
	PianoConvo	✓					✓	✓	3
	FilmingSite	✓	✓	✓	✓			✓	5
	Brothers	✓	✓	✓	✓		✓	✓	6
	Rap	✓	✓	✓	✓				4
	Spanish	✓	✓	✓	✓		✓		5
	Questions	✓	✓	✓				✓	4
	PianoMono	✓	✓	✓	✓		✓		5
	Snowfield				✓		✓	✓	3
	Melodrama	✓	✓	✓			✓	✓	5
	Gymnasium	✓	✓	✓	✓		✓		5

Table 3.8 Attribute details (2/2). General attributes: MO = Multiple Objects. OC = Occlusions. LR = Low Resolution. MB = Motion Blur. 360° geometrical attributes: OV = Out-of-View. GD = Geometrical Distortion. Audio attributes: CS = Competing Sounds.

Sequence		General				360°		Audio	No.
		MO	OC	LR	MB	OV	GD	CS	
Music (16)	Guitar	✓	✓	✓				✓	4
	Subway	✓	✓	✓	✓		✓		5
	Jazz	✓	✓	✓			✓	✓	5
	Bass	✓	✓	✓			✓	✓	5
	Canon	✓	✓	✓				✓	4
	MICOSinging	✓	✓				✓	✓	4
	Clarinet	✓	✓	✓			✓	✓	5
	Trumpet	✓	✓	✓					3
	PianoSaxophone	✓	✓	✓			✓	✓	5
	Chorus	✓	✓	✓				✓	4
	Studio	✓	✓	✓	✓			✓	5
	Church	✓	✓	✓				✓	4
	Duet	✓	✓				✓	✓	4
	Blues	✓	✓	✓				✓	4
	Violins	✓	✓	✓		✓		✓	5
	SingingDancing	✓	✓	✓	✓		✓	✓	6
Miscellanea (16)	Beach	✓	✓	✓	✓				4
	BadmintonGym	✓	✓	✓	✓				4
	InVehicle	✓	✓	✓			✓		4
	Japanese				✓	✓	✓	✓	4
	Tennis	✓	✓	✓	✓		✓		5
	Diesel	✓	✓	✓			✓		4
	Park	✓	✓	✓	✓				4
	Lion			✓	✓				2
	Carriage	✓	✓	✓	✓	✓	✓		6
	Platform	✓	✓	✓	✓		✓		5
	Dog	✓		✓	✓		✓		4
	RacingCar	✓		✓			✓	✓	4
	Train		✓		✓		✓	✓	4
	Football	✓	✓	✓	✓				4
	ParkingLot	✓	✓	✓	✓		✓	✓	6
	Skiing	✓	✓	✓	✓		✓	✓	6
No.		56	52	59	40	8	39	35	289

Dataset Features and Statistics. We analyze our proposed PAVS10K from three aspects, *i.e.*, dataset's attributes' distributions, dataset's ground truth distributions and salient objects' challenging features.

- *Attributes' distributions*. The attributes represent common challenges for conducting panoramic dynamic audio-visual salient object segmentation, thus facilitating detailed analysis regarding 2D image-/video-based salient object segmentation, 360° image-/video-based salient object segmentation and panoramic audio-visual salient object segmentation models. Specifically, as shown in Fig. 3.25 (a), the correlated attributes denote the attributes simultaneously appearing

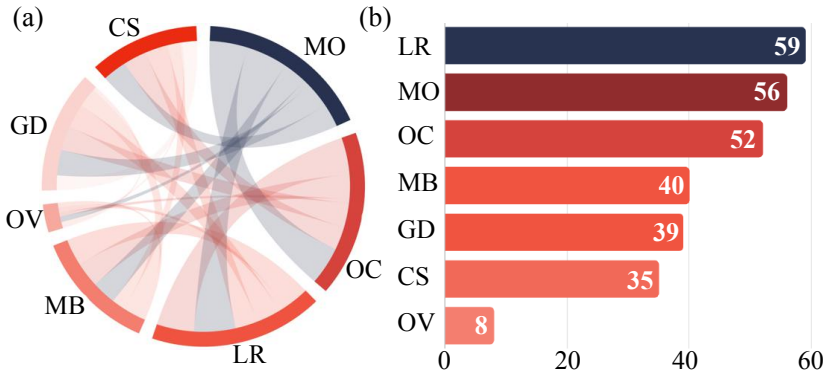


Fig. 3.25 Dataset features and statistics. (a) and (b) represent the correlation and frequency of PAVS10K's attributes, respectively.

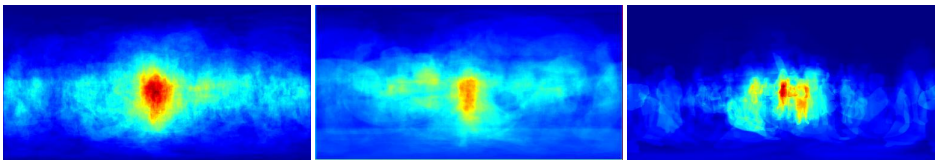


Fig. 3.26 A comparison of ground truth distribution of our PAVS10K and two recently proposed panoramic image salient object segmentation datasets, *i.e.*, 360-SOD [257] and 360-SSOD [258].

in the same sequence. *e.g.*, *LR* and *MO* show a strong correlation which indicates the two attributes tend to co-appear in most of the videos. Besides, as shown in Fig. 3.25 (b), *e.g.*, most of the videos (59) include small objects ($\leqslant 0.5\%$ of ER image area) and more than half of the videos (39) contain distorted objects, which illustrates that our PAVS10K is challenging.

- *Equator Center Bias*. As can be seen in Fig. 3.26, our PAVS10K, 360-SOD [257] and 360-SSOD [258] all show equator-center bias. The observation is consistent to the facts that photographers tend to frame the primary objects at the equator center of the 360° cameras, in addition, HMDs' users usually pay more attention to regions near the equator center during free-viewing [330, 331]. Besides inter-dataset comparison, we also show the ground truth distribution of our PAVS10K in terms of each of the three super-classes (Fig. 3.26). As a result, our PAVS10K clearly shows the equator-center biased pattern at both overall and super-class-based levels.

- *360° objects*. Following [21], we compute the normalized objects' size of our PAVS10K. The size distribution ranges from 0.03% to 23.00%, covering extremely small objects. In addition, we compare the situations of conducting object detection in 2D and 360° domains (Fig. 3.28). The appearances and sizes of 360° objects indicate the challenges for conducting salient object segmentation in panorama.

Fig. 3.27 Ground truth distribution over three super-classes of our proposed PAVS10K.

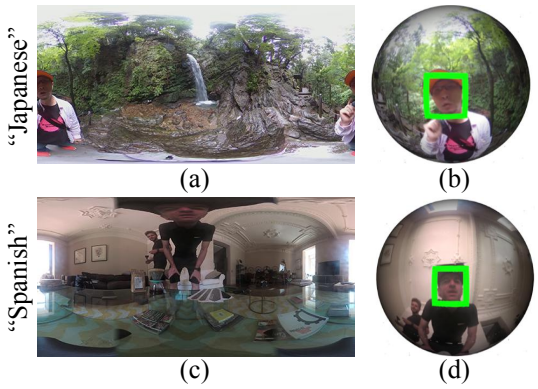


Fig. 3.28 An illustration of exclusive attributes regarding 360° salient object segmentation. (a) and (c) denote OV and GD salient objects in equirectangular images, (b) and (d) denote the ones in local viewports (*a.k.a.* common 2D cases).

3.4.3 Benchmark studies

In this sub-section, we introduce extensive benchmark experiments conducted based on the proposed PAVS10K. In the computer vision community, especially in the field of salient object segmentation, a comprehensive benchmark is essential to support a newly proposed task and its dataset, and is beneficial for future model development based on the new task/dataset. Following the same procedure as illustrated in Section 3.3.3, we introduce our new panoramic video salient object segmentation benchmark from the aspects of experimental settings and corresponding results.

Settings. Commonly, an integrated benchmark consists of consistent training/testing dataset split, benchmark models from multiple salient object segmentation related fields, and different metrics for quantitative evaluation of baseline models' predictions.

Data split. All 67 videos are split into separate training and testing sets by using a random selection strategy with a ratio of about 6:4. We thus reach a split of 40 training and 27 testing videos, with 5,796 and 4,669 video frames, respectively. Each of the 10,465 video frames are with per-pixel instance-/object-level ground truth. The testing set is further divided into "Miscellanea" (Test1), "Music" (Test2) and "Speaking" (Test3), consisting of 6, 6 and 15 videos respectively.

Benchmark models. To contribute a comprehensive benchmark to 360° video-based salient object segmentation, we collect 13 state-of-the-art methods from multiple related fields, including 2D image-base salient object segmentation methods (*i.e.*, CPD-R [139], SCRN [297], F3Net [122], MINet [126], LDF [128], CSFR2 [129] and GateNet [132]), 2D video-based salient object segmentation models (*i.e.*, RCRNet [178] and PCSA [175]), video object segmentation (*i.e.*, COSNet [180], 3DC-Seg [332]

and RTNet [333]) and panoramic image salient object segmentation baseline (FANet [334]). Note that all our collected benchmark models are able to be trained in an end-to-end manner, based on the most widely used open-source machine learning framework, *i.e.*, PyTorch. For fair comparison, we re-train the 13 baseline methods with the training set of our PAVS10K, based on their official publicly available codes and recommended parameter settings.

Evaluation metrics. Following the common settings in the field of salient object segmentation, we apply four widely used metrics, *i.e.*, mean F-measure (F_β, where β^2=0.3) [276], MAE ($\mathscr{M}$) [277], S-measure (S_α, where α=0.5) [278] and mean E-measure (E_ϕ) [279], to evaluate all benchmark models. Details about the four metrics are illustrated in Section 2.7.

Performance comparison. To contribute a comprehensive benchmark, we compare all baseline methods on our PAVS10K with and without PAVS10K training.

As a result, the quantitative results of the baseline models without/with PAVS10K training, are illustrated in Table 3.9 and Table 3.10, respectively. Besides overall performance, we also show the attributes-based performance of all baseline models in Table 3.11.

Table 3.9 Performance comparison of benchmark models without training on PAVS10K. I. = image-based salient object segmentation models. V. = video-based salient object segmentation or video object segmentation models. S_α = S-measure (α=0.5 [278]), F_β = mean F-measure (β^2=0.3) [276], E_ϕ = mean E-measure [279], $\mathscr{M}$ = mean absolute error [277]. Please note that FANet did not release its pre-trained model during the period when we conducted the benchmark studies.

Type	Year	Methods	Miscellanea (Test1)				Music (Test2)				Speaking (Test3)				PAVS10K-Test			
			$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$
I.	CVPR'19	CPD-R [139]	.261	.623	.604	.084	.151	.506	.483	.135	.190	.526	.488	.162	.195	.545	.515	.137
	ICCV'19	SCRN [297]	.271	.625	.606	.087	.206	.598	.594	.051	.218	.559	.518	.130	.226	.584	.558	.101
	AAAI'20	F3Net [122]	.236	.609	.573	.082	.152	.509	.524	.150	.215	.567	.505	.105	.204	.563	.526	.110
	CVPR'20	MINet [126]	.225	.606	.573	.093	.152	.542	.531	.073	.180	.523	.469	.151	.183	.548	.509	.118
	CVPR'20	LDF [128]	.268	.622	.606	.083	.204	.550	.557	.087	.227	.546	.503	.137	.230	.566	.541	.112
	ECCV'20	CSFR2 [129]	.305	.650	.624	.075	.139	.510	.471	.129	.189	.545	.511	.128	.202	.562	.529	.116
	ECCV'20	GateNet [132]	.243	.637	.588	.069	.206	.594	.611	.035	.206	.569	.554	.090	.214	.591	.576	.072
V.	CVPR'19	COSNet [180]	.280	.602	.581	.110	.181	.571	.614	.034	.232	.595	.587	.065	.230	.591	.592	.068
	ICCV'19	RCRNet [178]	.307	.666	.644	.062	.312	.630	.683	.040	.238	.591	.542	.065	.271	.619	.601	.058
	AAAI'20	PCSA [175]	.197	.629	.632	.042	.104	.543	.548	.030	.157	.565	.594	.037	.153	.575	.592	.036
	BMVC'20	3DC-Seg [332]	.231	.544	.523	.143	.268	.578	.663	.059	.193	.540	.584	.088	.220	.550	.588	.094
	CVPR'21	RTNet [333]	.331	.632	.602	.110	.436	.668	.769	.016	.338	.637	.639	.045	.361	.643	.661	.054

3.4.4 Discussion

The extensive benchmark studies based on our PAVS10K illustrate the challenges for conducting 360° video salient object segmentation.

Overall performance. According to the detailed quantitative results, we find that both salient object segmentation and video object segmentation state-of-the-art methods tend to show compromised performance (as for the image context) on the testing set of our PAVS10K, when compared to their performance on current salient object segmentation /video object segmentation benchmark datasets. For instance, as shown in Table 3.9 and Table. 3.10, the mean value of S_α of all competing methods on PAVS10K-Test are 0.534 and 0.626 without and with PAVS10K training, respectively. Besides, the maximum of S_α of these methods is 0.655. However, state-of-the-art video salient object

Table 3.10 Performance comparison of benchmark models with training on PAVS10K. (P)I. = (panoramic) image-based salient object segmentation models. V. = video-based salient object segmentation or video object segmentation models. S_α = S-measure (α=0.5 [278]), F_β = mean F-measure (β^2=0.3) [276], E_ϕ = mean E-measure [279], $\mathscr{M}$ = mean absolute error [277].

Type	Year	Methods	Miscellanea (Test1)				Music (Test2)				Speaking (Test3)				PAVS10K-Test			
			$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$
I.	CVPR'19	CPD-R [139]	.248	.654	.645	.035	.272	.608	.632	.018	.228	.588	.657	.026	.243	.609	.648	.026
	ICCV'19	SCRN [297]	.250	.665	.615	.046	.341	.683	.664	.023	.276	.636	.642	.034	.286	.655	.641	.034
	AAAI'20	F3Net [122]	.257	.655	.629	.040	.358	.662	.749	.021	.308	.626	.692	.027	.310	.642	.691	.029
	CVPR'20	MINet [126]	.238	.650	.625	.050	.380	.670	.716	.020	.261	.590	.635	.053	.286	.624	.652	.044
	CVPR'20	LDF [128]	.280	.663	.626	.044	.389	.671	.753	.023	.309	.625	.711	.037	.322	.645	.701	.035
	ECCV'20	CSFR2 [129]	.238	.652	.642	.033	.347	.665	.693	.018	.285	.636	.700	.026	.290	.646	.684	.026
	ECCV'20	GateNet [132]	.285	.677	.651	.044	.290	.673	.616	.018	.260	.633	.638	.034	.273	.653	.636	.033
V.	CVPR'19	COSNet [180]	.147	.610	.553	.031	.220	.577	.541	.016	.176	.572	.570	.023	.181	.582	.559	.023
	ICCV'19	RCRNet [178]	.272	.661	.640	.034	.403	.695	.738	.019	.282	.632	.687	.030	.310	.654	.688	.029
	AAAI'20	PCSA [175]	.123	.604	.574	.034	.310	.657	.645	.022	.150	.571	.534	.026	.184	.600	.570	.027
	BMVC'20	3DC-Seg [332]	.300	.668	.618	.062	.326	.635	.632	.046	.289	.629	.592	.056	.300	.640	.608	.055
	CVPR'21	RTNet [333]	.240	.622	.634	.038	.365	.638	.766	.020	.194	.555	.668	.028	.247	.591	.683	.029
PI.	SPL'20	FANet [334]	.164	.610	.529	.030	.380	.646	.758	.018	.207	.566	.663	.027	.241	.596	.654	.025

Table 3.11 Performance comparison of benchmark models based on each of the attributes. S_α = S-measure (α=0.5 [278]), F_β = mean F-measure (β^2=0.3) [276], E_ϕ = mean E-measure [279], $\mathscr{M}$ = mean absolute error [277].

Attr.	Metrics	Image-based salient object segmentation							Video-based salient object segmentation					-
		CPD-R [139]	SCRN [297]	F3Net [122]	MINet [126]	LDF [128]	CSFR2 [129]	GateNet [132]	COSNet [180]	RCRNet [178]	PCSA [175]	3DC-Seg [332]	RTNet [333]	FANet [334]
MO	$S_\alpha\uparrow$	.610	.657	.644	.624	.648	.649	.653	.588	.661	.607	.643	.595	.605
	$F_\beta\uparrow$	.244	.288	.315	.288	.324	.292	.270	.187	.319	.193	.302	.251	.258
	$E_\phi\uparrow$	.655	.649	.705	.665	.718	.694	.637	.571	.706	.580	.614	.703	.676
	$\mathscr{M}\downarrow$	.027	.034	.030	.045	.033	.027	.034	.024	.029	.027	.054	.028	.025
OC	$S_\alpha\uparrow$	.606	.655	.641	.619	.645	.645	.650	.577	.652	.600	.636	.586	.593
	$F_\beta\uparrow$	.260	.294	.329	.298	.335	.301	.276	.191	.316	.202	.308	.259	.258
	$E_\phi\uparrow$	.649	.639	.696	.651	.709	.682	.622	.554	.691	.570	.607	.694	.668
	$\mathscr{M}\downarrow$	.023	.029	.026	.043	.028	.023	.030	.020	.025	.024	.045	.024	.022
LR	$S_\alpha\uparrow$	.605	.649	.639	.618	.637	.644	.647	.585	.650	.609	.633	.590	.598
	$F_\beta\uparrow$	.229	.271	.301	.272	.303	.277	.255	.176	.294	.189	.286	.234	.238
	$E_\phi\uparrow$	.640	.636	.693	.642	.694	.683	.625	.565	.687	.586	.600	.688	.657
	$\mathscr{M}\downarrow$	.025	.034	.028	.045	.037	.025	.033	.022	.029	.026	.057	.029	.025
MB	$S_\alpha\uparrow$	.622	.651	.630	.620	.646	.638	.645	.582	.642	.586	.632	.595	.587
	$F_\beta\uparrow$	.281	.304	.299	.298	.330	.297	.281	.212	.307	.197	.302	.271	.247
	$E_\phi\uparrow$	.628	.630	.663	.637	.667	.668	.621	.563	.675	.563	.599	.676	.627
	$\mathscr{M}\downarrow$	.021	.029	.027	.047	.029	.021	.030	.019	.024	.022	.044	.023	.020
OV	$S_\alpha\uparrow$	.634	.661	.568	.633	.636	.636	.639	.582	.630	.599	.641	.573	.611
	$F_\beta\uparrow$	.311	.318	.167	.314	.309	.295	.258	.207	.276	.193	.362	.210	.310
	$E_\phi\uparrow$	.652	.638	.538	.691	.676	.697	.637	.633	.732	.536	.671	.703	.679
	$\mathscr{M}\downarrow$	.018	.021	.029	.038	.039	.021	.025	.021	.029	.021	.039	.022	.018
GD	$S_\alpha\uparrow$	.630	.662	.639	.633	.659	.646	.658	.588	.651	.578	.659	.587	.599
	$F_\beta\uparrow$	.285	.309	.299	.294	.341	.304	.300	.189	.311	.156	.320	.247	.245
	$E_\phi\uparrow$	.657	.653	.669	.676	.680	.674	.662	.564	.687	.538	.621	.666	.630
	$\mathscr{M}\downarrow$	.037	.042	.040	.045	.043	.035	.042	.032	.037	.036	.062	.038	.034
CS	$S_\alpha\uparrow$	.625	.680	.667	.654	.664	.670	.676	.592	.680	.621	.654	.602	.616
	$F_\beta\uparrow$	.277	.320	.357	.335	.361	.330	.304	.197	.354	.217	.324	.269	.279
	$E_\phi\uparrow$	.674	.664	.720	.691	.740	.696	.655	.550	.711	.590	.625	.711	.697
	$\mathscr{M}\downarrow$	.029	.035	.031	.035	.034	.028	.033	.026	.030	.029	.058	.031	.028

segmentation methods such as SSAV [23], PCSA [175] and DCFNet [163] show much better perfor-

mance on widely used video benchmark datasets, with S_α of 0.893/0.902/0.914 on DAVIS2016 [151], 0.819/0.827/0.846 on VOS [22] and 0.724/0.741/0.741 on DAVSOD [23].

Attribute-based performance. The proposed attributes (MO, LR, OC, MB, GD, OV and CS) of our PAVS10K enable detailed analysis towards panoramic video salient object segmentation modeling. Especially, compared to previous video datasets [23, 151], we propose extra attributes, *i.e.*, GD and OV, which reflect common challenges for modeling on 360° (Fig. 3.28). As a result, COSNet [180] acquires superior results on all attribute based testing sets only in terms of $\mathscr{M}$ [277], which is the mean value of per-pixel absolute error (Eq. 2.2). A superior $\mathscr{M}$ yet weak F_β (Table 3.11) indicate that video salient object segmentation method such as COSNet tends to be conservative for detecting salient objects in 360° panoramic videos.

3.4.5 Conclusion

In this section, we first propose a new task, panoramic dynamic audio-visual salient object segmentation, which aims at modeling both visual and audio cues to conduct salient object detection in 360° panoramic videos. To support the task, we establish a large-scale 360° video dataset, *i.e.*, PAVS10K, representing various real-life scenes with good visual quality (4K-resolution). Our PAVS10K provides multiple labels including three super-classes, 67 sub-classes, seven salient object segmentation attributes, 10,465 video frames with per-frame manually labeled object-level and instance-level masks. We further collect 13 state-of-the-art salient object segmentation /video object segmentation methods to establish so far the largest panoramic dynamic audio-visual salient object segmentation benchmark. We conduct extensive qualitative and quantitative experiments to achieve comprehensive benchmark studies.

3.5 Conclusion

This chapter introduces key issues towards large-scale salient object segmentation construction, and our newly proposed 360° image/video-based salient object segmentation datasets, *i.e.*, F-360iSOD and PAVS10K. Considering the blank of 360° salient object segmentation, the priority of this thesis work was to build both image and video datasets which observe common rules of current 2D datasets. To this end, we first summarized several key issues towards large-scale dataset construction, by detailing the statistics of current widely used 2D image/video salient object segmentation datasets. Based on the spotted key aspects of large-scale salient object segmentation datasets, we built the first 360° image-based salient object segmentation dataset, namely F-360iSOD, that provides both object-/instance-level pixel-wise ground truth. Inspired by the real-world scenes where human attention is affected by both audio and visual cues, we further proposed the first 360° dynamic audio-visual salient object segmentation dataset, namely PAVS10K, where the salient objects are annotated based on audio-visual eye fixations. To facilitate and inspire future works based on the newly proposed F-360iSOD and PAVS10K, we further conducted comprehensive benchmark studies.

Chapter 4

Salient object segmentation in light field

4.1 Introduction

Unlike traditional 2D RGB images, both light field and 360° based images contain extra visual cues reflecting real-life daily scenes. For instance, light field camera is able to capture visual details at different focus distances and thus generating a stack of images, namely focal stacks, with varying spatial texture across image depth. On the other hand, 360° camera is able to capture global context in a 360°×180° field-of-view. Therefore, modeling human attention in light field and 360° are both important for exploring human attention mechanism in real world. In this case, besides the main focus of our PhD work towards 360° vision, we have conducted multiple works in terms of light field salient object segmentation. In this chapter, we summarize our proposed new salient object segmentation methodologies in light field. Inspired by current state-of-the-art salient object segmentation methods with various attention modules as summarized in Chapter 2, we argue that recently proposed state-of-the-art attention models (*e.g.*, SENet [145], CBAM [146] and Non-local network [261]) can also be used as basic components for the development of light field salient object segmentation models. We hereby propose new methods (Section 4.2) consisting of multiple attention mechanisms to fuse and refine features extracted from multiple light field modalities (*i.e.*, all-in-focus, focal stack, depth).

Specifically, to explore multiple attention mechanisms for effective object-level attention modeling with multi-modal light field data, we first proposed SA-Net. Our SA-Net exploits the rich information of focal stacks via 3D convolutional neural networks, decodes the high-level features of multi-modal light field data with two cascaded synergistic attention modules, and predicts the saliency map using an effective feature fusion module in a progressive manner. As the development of large-scale vision transformers [143], we further explored the encoder of SA-Net and thus proposing SA-Net-V2, which replaces the ResNet blocks with hybrid-ViT based transformer blocks at the all-in-focus branch of the encoder. To improve the SA-Net from a perspective of model computational burden, we further proposed CMA-Net, which consists of two novel cascaded mutual attention modules aiming at fusing the high level features from the modalities of all-in-focus and depth.

In the following section, we detail the three proposed methods, *i.e.*, SA-Net [251], SA-Net-V2 and CMA-Net in a progressive manner.

4.2 Learning synergistic attention for light field salient object segmentation

4.2.1 Introduction

Recently, light field salient object segmentation [238] has attracted increasing attention owing to the introduction of various light field benchmark datasets, such as DUT-LF [240], LFSD [238], HFUT [239], DUT-MV [241], and Lytro Illum [242]. In addition to all-in-focus images, light field datasets [238, 239, 242] also provide focal stacks, multi-view images, and depth maps, where the focal stacks are usually known as a series of focal slices focusing at different depths of a given scene while the depth map contains holistic depth information. Unlike RGB-D salient object segmentation models, which utilize only two modalities, i.e., RGB images and depth maps, the light field salient object segmentation models also use multi-view images (e.g., [241, 248]), or focal stacks [240, 244–246] as auxiliary inputs to further improve the performance. It is worth noting that, most recent focal stack-based deep learning light field salient object segmentation models (e.g., ERNet [245]) have achieved state-of-the-art performances on three widely-used light field benchmark datasets [238–240].

Despite their advantages, existing works suffer from two major limitations. First, they explore little about the complementarities between all-in-focus images and the focal stacks. Existing focal stack-based methods [240, 244–246] applied only channel attention mechanisms to weight the key feature channels at the decoding stage, to aid the feature fusion between the modalities of all-in-focus and focal stack. Considering the fact that salient objects usually appear at specific depths of a given scene, all-in-focus image may include redundant texture details compared to focal stack, in which a focal slice focuses on a local region at specific depth and blurs the others. New cross-modal fusion strategy, which applies more sophisticated attention mechanisms learning robust cross-modal complementarities, may help solve the issue. Second, the methods [240, 244–246] all paid little attention to the inter-slice modeling during the encoding stage of focal stacks. In practice, the all-in-focus images are generated from focal stacks with a photo-montage technique [335], implying that the former simultaneously depict the spatial details of each local region, while the latter asynchronously focus on different local details along the sequential dimension. The relationship between focal slices reflects the context of given scenes as the changes of depth, which is appropriate to be encoded in a progressive manner.

SA-Net. To this end, we propose Synergistic Attention Network (SA-Net) to conduct light field salient object segmentation with rich information from all-in-focus images and focal stacks (Fig. 4.1). Specifically, we first employ 3D convolutional neural networks to progressively extract the sequential features from focal stacks. At the decoding stage, we propose a synergistic attention (SA) module, where the features from all-in-focus images and focal stacks are selectively fused and optimized to achieve a synergistic effect for salient object segmentation. Finally, the multi-modal features are fed to our progressive fusion (PF) module, which fuses multi-modal features and predicts the saliency map in a progressive manner.

SA-Net-V2. Furthermore, as the development of recent transformers (*e.g.*, ViT [143]), we replace the resnet50 with hybrid-ViT framework at the all-in-focus branch of the encoder to improve model

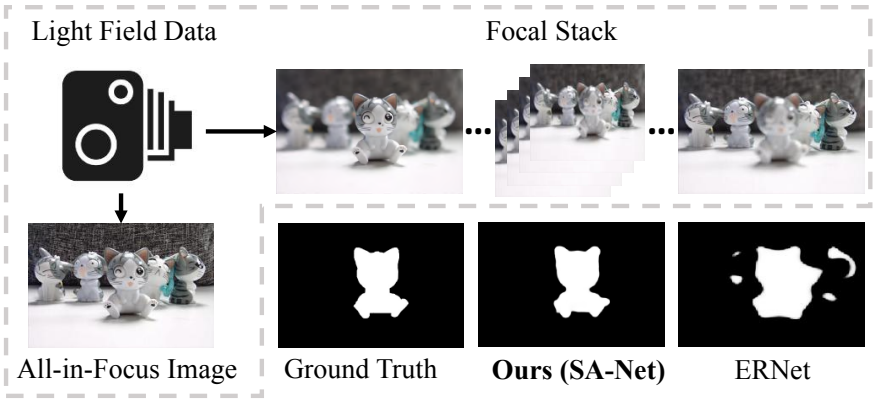


Fig. 4.1 An example of light field salient object segmentation using our SA-Net (**Ours**) and a state-of-the-art light field model, *i.e.*, ERNet [245].

performance. In addition, we take advantage of the inter-slice features in both the encoding and decoding processes, thus gaining an advanced version of SA-Net, namely SA-Net-V2.

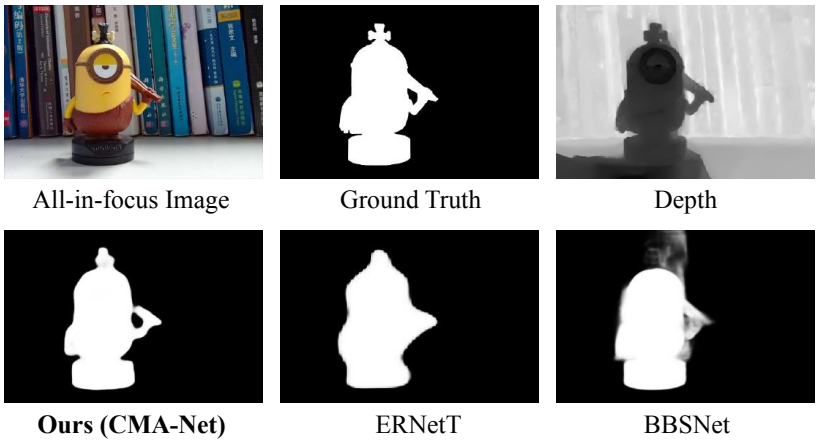


Fig. 4.2 An example of light field salient object segmentation using our CMA-Net (**Ours**) and a state-of-the-art RGB-D model, *i.e.*, BBSNet [270].

CMA-Net. As the computational burden is also an important issue for evaluating the effectiveness of deep learning models, we are inspired by RGB-D methods and thus using all-in-focus images and depth information to conduct light field salient object segmentation with a fine-tuned SA-Net, namely CMA-Net (Fig. 4.2).

In a nutshell, we provide several contributions as follows:

- In SA-Net, we propose the SA module to decode the high-level features from both all-in-focus images and focal stacks with a synergistic attention mechanism. Our SA module exploits the most meaningful information from the multi-modal multi-level features, allowing accurate salient object segmentation by taking advantage of light field data.
- In SA-Net, we introduce a dual-branch backbone to encode the all-in-focus and focal stack information, simultaneously. To the best of our knowledge, our work is the first attempt to

utilize 3D convolutional neural networks for the feature extraction of focal stacks in the field of light field salient object segmentation.

- In SA-Net, we design the PF module to gradually fuse the selective high-level features for the final saliency prediction.
- In SA-Net-V2, we discard the three 3D convolutional layers at the focal stack branch of the encoder, and replace the original 2D receptive field blocks [139] with our proposed 3D ones. To fuse the 3D focal stack-based features and 2D all-in-focus features at high-level, we further design a multi-head synergistic attention module (Multi-head SA).
- In CMA-Net, we propose a new cascaded mutual attention (CMA) mechanism to efficiently fuse the RGB-D high level features. Our CMA-Net does not apply focal stacks, also avoids processing low-level features from both the modalities, thus performing competitive inference speed.
- Extensive experiments demonstrate that all our proposed models (SA-Net,SA-Net-V2 and CMA-Net) outperforms dozens of state-of-the-art salient object segmentation models upon widely-used light field datasets.

4.2.2 Related works

Related datasets and most recent representative works towards RGB-D salient object segmentation and light field salient object segmentation are summarized in Chapter 2. In this sub-section, we further detail the related works towards light field salient object segmentation, several applications regarding mutual attention and 3D convolutional neural networks.

Light field salient object segmentation. By the time of the release of our proposed SA-Net, there are only 18 (11/7 traditional/deep learning-based, respectively) published methods. For traditional ones, the early method [238] conducted light field salient object segmentation by considering background and location related prior knowledge. In addition, [336] proposed a unified architecture based on weighted sparse coding. Later methods [239, 337–341] explored and further combined multiple visual cues (e.g., depth, color contrast, light field flows and boundary prior) to detect saliency. Most recent methods [215, 342] shifted more attention to depth information and employed cellular automata for the saliency detection in light field. With the development of public light field datasets, deep learning-based methods were proposed to conduct salient object segmentation task. Specifically, [241] developed a view synthesis network to detect salient objects by involving multi-views. With multi-views as inputs, [248] further established a unified structure to synchronously conduct salient object and edge detection. Besides, [242] applied DeepLab-v2 for salient object segmentation with multi-lens. As a mainstream, [240, 244–246] all employed ConvLSTM and channel attention mechanisms at the decoding stage to detect salient objects in all-in-focus images and focal stacks. [240] and [246] both modeled the all-in-focus and focal stack with separate encoder-decoder architectures. Specifically, [240] added the outputs of the decoders of both the focal stack branch and all-in-focus branch. [246] concatenated the outputs of all-in-focus branch and focal stack branch and used the ConvLSTM [14] to refine the concatenated features. In [244], focal stack and all-in-focus

share the same encoder. The extracted multi-modal features were further fused with memory-oriented attention modules. The most recent [245] designed teacher network and student network to encode the focal stack and all-in-focus respectively, and used ConvLSTM based attention module to facilitate the distillation process between the teacher network and student network.

Mutual attentions. Mutual(co)-attention, as a specific type of attention mechanism within the multi-branch attention category, has been used in the fields of video object segmentation (e.g., COSNet [180]), RGB-D salient object segmentation (e.g., S2MA [343]), *etc.*. Specifically, COSNet first proposed to establish mutual attention module for effective feature fusion and refinement between different video frames. S2MA [343] designed a self-mutual attention module to automatically select useful high-level features learned from both modalities. However, mutual attention mechanism has been seldom studied in the field of light field salient object segmentation. DLLF [240], LFNet [246], MoLF [244] and ERNet [245] all employed classical channel attention [145] to aid the feature selection and refinement from the modality of focal stacks. Recent large-scale light field salient object segmentation benchmark studies (e.g., [244, 245]) indicate that it remains an open question how to efficiently fuse the intrinsic features from multiple modalities for advanced detecting accuracy.

3D convolutional networks. 3D convolutional networks have proved great competence in modeling spatial-temporal information of video data, thus dominating the video-based detection fields, such as action recognition [344] and video object segmentation [332]. Recently, RD3D [228] was proposed to address the task of salient object segmentation by using a 3D convolutional network-based encoder-decoder structure, and achieved promising performance on widely-used RGB-D salient object segmentation benchmarks. As for light field salient object segmentation, MTCNet [248] applied 3D convolutional network-based encoder to extract the depth features from multi-view images. The rich high-level features gained from 3D convolutional networks were then used to infer depth maps and facilitate the salient object segmentation task, synchronously. Since focal stacks are sequences of focal slices focusing at different depths, learning focal stacks' features via 3D convolutional networks possesses great potential to boost the model performance for light field salient object segmentation, but so far lacks investigation.

4.2.3 Focal stack-based methodologies

In SA-Net, we exploit rich cross-modal complementary information with channel attention and co-attention mechanisms to achieve a synergistic effect between multi-level all-in-focus and focal stack features. In addition, to capture the inter-slice information of focal stack, we employ 3D convolutional neural networks to extract rich features from focal stacks. Fig. 4.3 shows an overview architecture of our SA-Net, which consists of three major components, including a multi-modal encoder consisting of 2D and 3D convolutional neural networks, two cascaded synergistic attention modules, and a progressive fusion module.

Multi-modal encoder. As shown in Fig. 4.3, the encoder of our network is a dual-branch architecture for synchronous feeding of the two modalities, i.e., all-in-focus images and focal stacks. For the 2D branch, we encode an input all-in-focus image with a group of convolutional blocks. On the other hand, focal stack is represented as a 4D tensor with the last dimension T denoting the number of focal slices. We encode the focal stack with a stack of 3D convolutional blocks, which are able to jointly

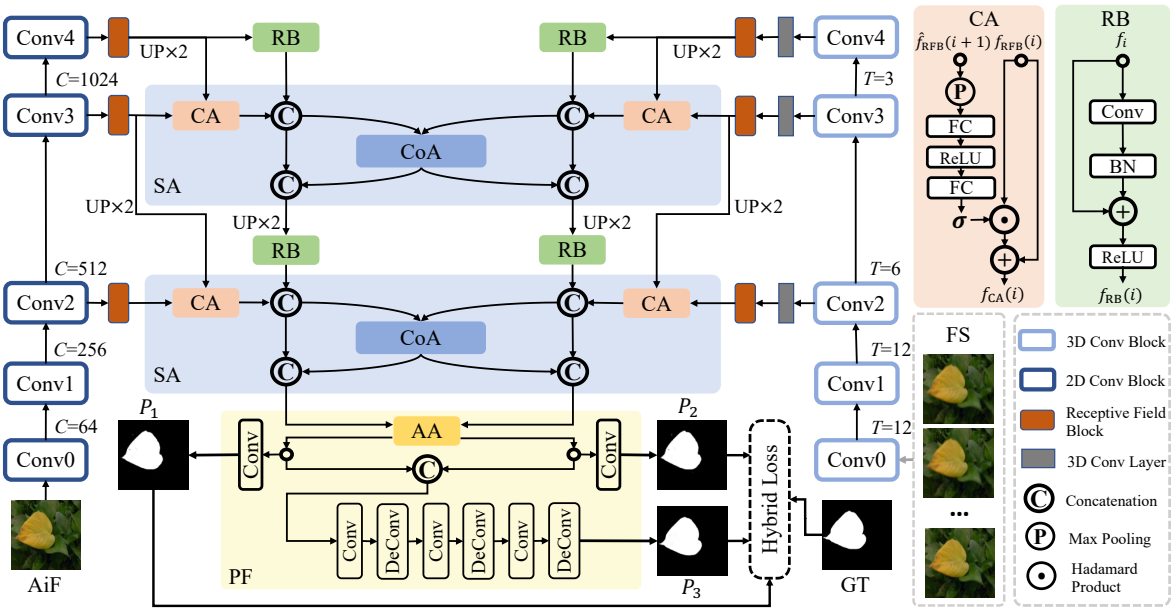


Fig. 4.3 An overview of our **SA-Net**. Multi-modal multi-level features extracted from our multi-modal encoder are fed to two cascaded synergistic attention (SA) modules followed by a progressive fusion (PF) module. The short names in the figure are detailed as follows: CoA = co-attention component. CA = channel attention component. AA = all-in-focus-induced attention component. RB = residual block. PF = progressive fusion module. P_n = the nth saliency prediction. (De)Conv = (de-)convolutional layer. BN = batch normalization layer. FC = fully connected layer.

capture the rich intra- and inter-slice information for accurate salient object segmentation. Note that the same setting ($T = 12$) as in [245] is adopted in our 3D branch, and a zero-padding strategy is applied to the focal stack with less than 12 focal slices.

Synergistic attention module. As high-level features tend to reserve the essential cues (e.g., location, shape) of salient objects while the low-level ones contain relative trivial information (e.g., edge) [139], our decoder only integrates high-level features to avoid redundant computational complexity. Specifically, we use $\{f_i^{2D}\}_{i=2}^4$ and $\{f_i^{3D}\}_{i=2}^4$ to denote the high-level all-in-focus and focal stack features extracted from the 2D and 3D convolutional networks of our dual-branch backbone network, respectively.

Multi-level attention. As shown in Fig. 4.3, a receptive field block (RFB) [139] is first employed to enrich the global context information for each convolution block. Taking the all-in-focus branch as an example, the adjacent high-level features from the encoder are then combined with a channel attention (CA) mechanism from [145], i.e.,

$$f_{\mathrm{CA}}^{2D}(i) = \sigma(FC(ReLU(FC(P(\hat{f}_{\mathrm{RFB}}^{2D}(i+1)))))) \odot f_{\mathrm{RFB}}^{2D}(i) + f_{\mathrm{RFB}}^{2D}(i), \tag{4.1}$$

where $f_{\mathrm{RFB}}^{2D}(i)$ represents the ith level features provided by RFB; $\hat{f}_{\mathrm{RFB}}^{2D}(i+1)$ is the up-sampled version of $f_{\mathrm{RFB}}^{2D}(i+1)$; $\sigma(\cdot)$, $FC(\cdot)$, $P(\cdot)$, and $\odot$ denotes the Sigmoid function, fully connected layer, max

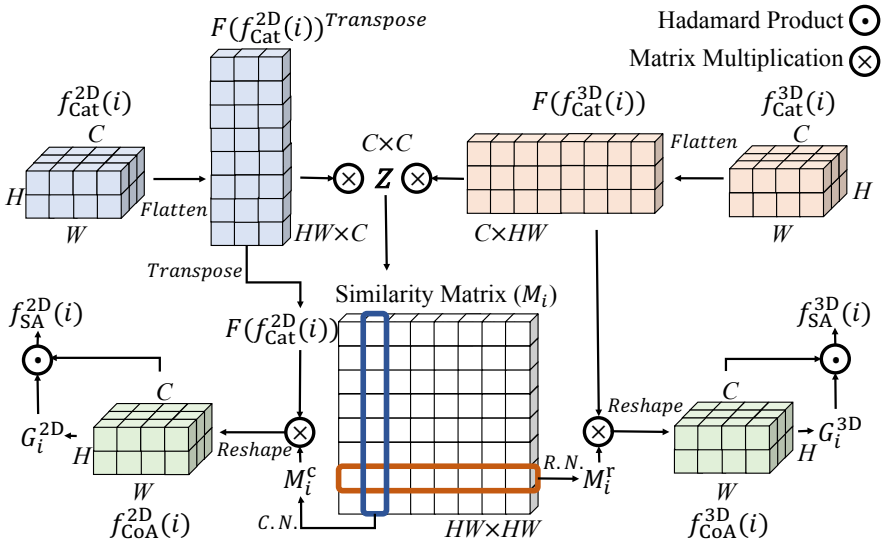


Fig. 4.4 The architecture of the co-attention (CoA) module. C.N. = column-wise normalization. R.N. = row-wise normalization.

pooling, and Hadamard product, respectively. The resulting feature $f_{\mathrm{CA}}^{2\mathrm{D}}(i)$ is further concatenated with the upper level feature $f_{\mathrm{RB}}^{2\mathrm{D}}(i+1)$ provided by a residual block (RB) for the feature $f_{\mathrm{Cat}}^{2\mathrm{D}}(i)$, which is one of the pair-wise inputs ($\{f_{\mathrm{Cat}}^{2\mathrm{D}}(i), f_{\mathrm{Cat}}^{3\mathrm{D}}(i)\}$) for the second stage of our SA module. Note that the focal stack branch follows the consistent procedure as in all-in-focus branch since the two branches are symmetric.

Multi-modal attention. Inspired by a mutual attention mechanism [180] proposed for cross-frame feature fusion in the field of video object segmentation, the high-level feature interaction between the two modalities is conducted with two cascaded co-attention (CoA) modules (Fig. 4.3). To be specific, as shown in Fig. 4.4, given the pair-wise features $\{f_{\mathrm{Cat}}^{2\mathrm{D}}(i), f_{\mathrm{Cat}}^{3\mathrm{D}}(i)\}$ at ith layer as inputs, a similarity matrix M_i can be computed as:

$$M_i = F(f_{\mathrm{Cat}}^{2\mathrm{D}}(i))^{\mathrm{T}} \otimes F(f_{\mathrm{Cat}}^{3\mathrm{D}}(i)), \tag{4.2}$$

where $F(\cdot)$ represents a flatten operation reshaping the 3D feature matrix $f_{\mathrm{Cat}}^{2\mathrm{D}}(i) \in \mathbb{R}^{H\times W\times C}$ to a 2D one with a dimension of $HW \times C$, $\otimes$ denotes matrix multiplication. Note that we do not apply extra weight matrix as in [180] to compute M_i, since the CoA module aims at fusing the cross-modal features with equally assigned attention. The M_i is then column-/row-wisely normalized via:

$$\begin{aligned} M_i^{\mathrm{c}} &= Softmax(M_i) \in [0,1]^{HW\times HW}, \\ M_i^{\mathrm{r}} &= Softmax(M_i^{\mathrm{T}}) \in [0,1]^{HW\times HW}, \end{aligned} \tag{4.3}$$

where $Softmax(\cdot)$ normalizes each column of the similarity matrix. Therefore, the co-attention-based

pair-wise features ($\{f^{2D}_{CoA}(i), f^{3D}_{CoA}(i)\}$) at ith layer are further defined as:

$$\begin{aligned} f^{2D}_{CoA}(i) &= R(f^{2D}_{Cat}(i) \otimes M^{c}_{i}) \in [0,1]^{H\times W\times C}, \\ f^{3D}_{CoA}(i) &= R(f^{3D}_{Cat}(i) \otimes M^{r}_{i}) \in [0,1]^{H\times W\times C}, \end{aligned} \tag{4.4}$$

where $R(\cdot)$ reshapes the given matrix from a dimension of $C \times HW$ to $H \times W \times C$. A self-gate mechanism [180] is further employed to automatically learn the co-attention confidences (G^{2D}_i, G^{3D}_i) for $f^{2D}_{CoA}(i)$ and $f^{3D}_{CoA}(i)$. Therefore the final outputs $\{f^{2D}_{SA}(i), f^{3D}_{SA}(i)\}$ of our SA module at ith layer are computed as:

$$f^{2D}_{SA}(i) = G^{2D}_i \odot f^{2D}_{CoA}(i) \text{ and } f^{3D}_{SA}(i) = G^{3D}_i \odot f^{3D}_{CoA}(i), \tag{4.5}$$

where the co-attention confidence $G^{2D}_i = \sigma(Conv(f^{2D}_{CoA}(i)))$ with $Conv(\cdot)$ denoting a convolutional layer.

By combining the channel attention (CA) and co-attenrion (CoA) module, our SA module is particularly effective in exploiting the multi-level and multi-modal complementary information, which, therefore, provides significantly improved performance, as demonstrated by our ablation studies in Section 4.2.5.

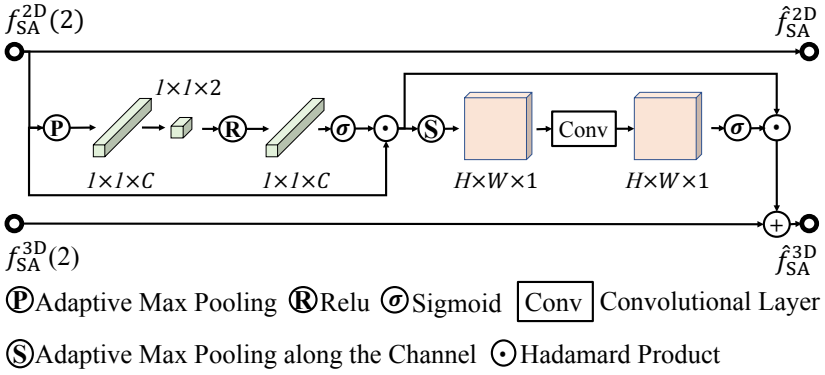


Fig. 4.5 The architecture of all-in-focus-induced attention (AA) component.

Progressive fusion module. To obtain the final prediction, we further add a progress fusion (PF) module to gradually up-sample the selective high-level features provided by our SA module (Fig. 4.3). Specifically, we first balance the focal stack and all-in-focus features with an all-in-focus-induced attention (AA) component (Fig. 4.5) before the final fusion of the two modalities. The AA component follows the same procedure applied in RGB-D fusion [270], i.e., unifying the channel and spatial attention by computing:

$$\hat{f}^{2D}_{SA} = f^{2D}_{SA}(2) \text{ and } \hat{f}^{3D}_{SA} = SA(CA(f^{2D}_{SA}(2))) + f^{3D}_{SA}(2), \tag{4.6}$$

where $CA(\cdot)$ and $SA(\cdot)$ denote spatial and channel attention components, respectively. We then concatenate the balanced cross-modal features and feed them to a deconvolutional block for the final prediction P_3, i.e.,

$$P_3 = DB(Cat(\hat{f}^{2D}_{SA}, \hat{f}^{3D}_{SA})), \tag{4.7}$$

where $Cat(\cdot)$ denotes the concatenation operation, and $DB(\cdot)$ represents a deconvolutional block consisting of three deconvolutional layers [270] and convolutional layers that are organized in a cascaded manner (Fig. 4.3).

Loss function. As shown in Fig. 4.3, our model predicts three saliency maps: $\{P_n\}_{n=1}^{3} \in [0,1]$. Let $G \in \{0,1\}$ denotes the ground-truth saliency map, we jointly optimize the three-way predictions by defining a hybrid loss ℓ:

$$\ell = \sum_{n=1}^{N} \ell_{\text{BCE}}(P_n, G) + \ell_{\text{IoU}}(P_n, G) + \ell_{\text{EM}}(P_n, G), \tag{4.8}$$

where ℓ_{BCE} and ℓ_{IoU} denote Binary Cross Entropy (BCE) and Intersection over Union (IoU) loss, respectively; the loss $\ell_{\text{EM}} = 1 - E_\phi$ with E_ϕ denoting E-Measure [279].

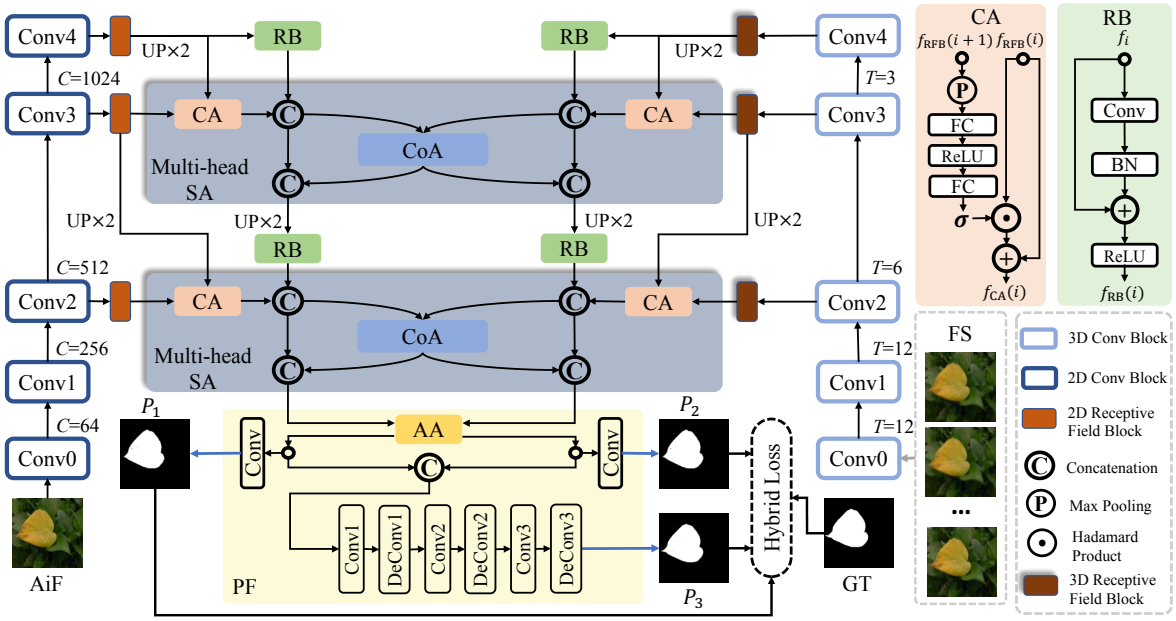


Fig. 4.6 The architecture of advanced version of SA-Net, *i.e.*, SA-Net-V2.

SA-Net-V2. In SA-Net-V2, as shown in Fig. 4.6, the difference between SA-Net and SA-Net-V2 mainly lies in all-in-focus encoder (as for SA-Net-V2, we took the hybrid-vit layers in DPT [345]), SA modules (as for SA-Net-V2, we designed multi-head SA module to fuse 2D features from all-in-focus branch and 3D futures from focal stack branch) and RFB modules (we fine-tuned RFB [139] and adopted it to 3D feature refinement).

Implementation details. Our SA-Net and SA-Net-V2 are implemented in PyTorch and optimized with Adam algorithm [3]. The backbone of SA-Net is based on a 2D standard ResNet50 for all-in-focus images and an inflated 3D ResNet50 [344] for focal stacks. The 2D convolution layers in our backbone are initialized with ImageNet-pretrained ResNet50, while the 3D convolutional layers are initialized with a 2D weight transfer strategy [344]. During the training stage, the batch size is set to 2, the learning rate is initialized as 1e-5 and decreased by 10% when training loss reaches a flat.

It takes about 14 hours to train the proposed model based on a platform consists of Intel® i9-7900X CPU@3.30GHz and one TITAN XP GPU.

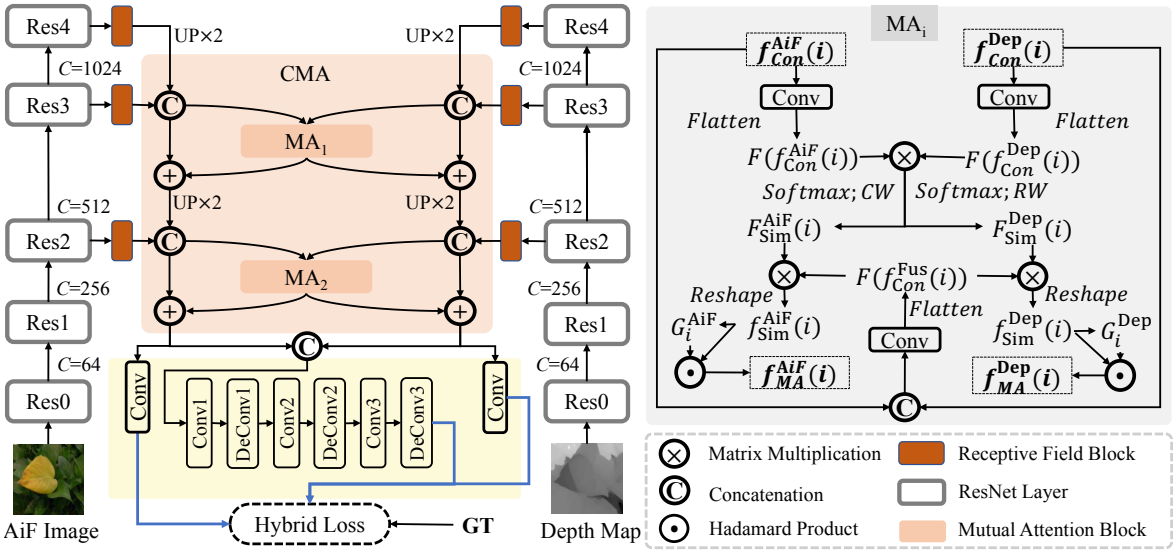


Fig. 4.7 An overview of our **CMA-Net**. RGB-D high level features extracted from duel-branch encoder are fed into two proposed cascaded mutual attention modules, followed by a group of (de-)convolutional layers from [270]. The abbreviations in the figure are detailed as follows: AiF Image = all-in-focus image. GT = ground truth. Res$i$ = the ith ResNet [142] layer. (De)Conv = (de-)convolutional layer. MA$_i$ = the ith mutual attention module. CMA = cascaded mutual attention module. CW = column-wise normalization. RW = row-wise normalization.

4.2.4 RGB-D-based methodologies

The CMA-Netconsists of a duel-branch ResNet50 [142]-based encoder and a cascaded mutual attention-based decoder.

RGB-D Encoder. Our encoder is a duel-branch architecture that consists of symmetrical convolutional layers transferred from ImageNet-pretrained ResNet50 [142]. In CMA-Net, we only process the high-level features, i.e., the features ($\{f_i^{\text{AiF}}\}_{i=2}^{4}$ and $\{f_i^{\text{Dep}}\}_{i=2}^{4}$) from the last three layers of ResNets, to focus on salient objects' shape and location cues [139] also to avoid extra computational cost. The $\{f_i^{\text{AiF}}\}_{i=2}^{4}$ and $\{f_i^{\text{Dep}}\}_{i=2}^{4}$ are then fed into a series of receptive field blocks [139] to enrich the global context information from each encoding level (Fig. 4.7).

Cascaded mutual attention. Similar to SA-Net, we illustrate the CMA-Net architecture from the aspects of multi-level and multi-modal processing.

Multi-level Concatenation. The refined high level features from adjacent encoding stages, e.g., $f_{\text{RFB}}^{\text{AiF}}(i)$ and $f_{\text{RFB}}^{\text{AiF}}(i+1)$ are further concatenated as $f_{\text{Con}}^{\text{AiF}}(i)$, where i ($i \in \{2,3\}$) denotes the ith decoding stage corresponding to the ith ResNet layer.

Mutual attention. Continually taking the ith decoding stage as an example, a similarity matrix (Sim_i)

between the features from two branches is computed as:

$$Sim_i = F(f_{\text{Con}}^{\text{Dep}}(i))^{\text{T}} \otimes F(f_{\text{Con}}^{\text{AiF}}(i)), \tag{4.9}$$

where $F(\cdot)$ represents a flatten operation reshaping the 3D feature matrix $f_{\text{Con}}^{\text{AiF}}(i) \in \mathbb{R}^{H\times W\times C}$ to a 2D one with a dimension of $C \times HW$, $\otimes$ denotes matrix multiplication. Inspired by [180], the Sim_i is then column-/row-wisely normalized via:

$$\begin{aligned} F_{\text{Sim}}^{\text{AiF}}(i) &= Softmax(Sim_i) \in [0,1]^{HW\times HW}, \\ F_{\text{Sim}}^{\text{Dep}}(i) &= Softmax(Sim_i^{\text{T}}) \in [0,1]^{HW\times HW}, \end{aligned} \tag{4.10}$$

where $Softmax(\cdot)$ normalizes each column of the $HW \times HW$ matrix. As shown in Fig. 4.7, the mutual attentions $(f_{\text{Sim}}^{\text{AiF}}(i), f_{\text{Sim}}^{\text{Dep}}(i))$ for each of the branches are computed as:

$$\begin{aligned} f_{\text{Sim}}^{\text{AiF}}(i) &= R(F_{\text{Sim}}^{\text{AiF}}(i) \otimes F(f_{\text{Con}}^{\text{Fus}}(i))^{\text{T}}) \in [0,1]^{H\times W\times C}, \\ f_{\text{Sim}}^{\text{Dep}}(i) &= R(F_{\text{Sim}}^{\text{Dep}}(i) \otimes F(f_{\text{Con}}^{\text{Fus}}(i))^{\text{T}}) \in [0,1]^{H\times W\times C}, \end{aligned} \tag{4.11}$$

where $R(\cdot)$ reshapes the given matrix from a dimension of $C \times HW$ to $H \times W \times C$, $F(f_{\text{Con}}^{\text{Fus}}(i))$ denotes fused features from both branches (Fig. 4.7), which is the main difference when compared to the counterpart in SA-Net. To further avoid unstable feature updating during the model training process, a pair of self-adapted gate functions (G_i^{AiF}, G_i^{Dep}) are computed to gain the final mutual attention matrix $(f_{\text{MA}}^{\text{AIF}}(i), f_{\text{MA}}^{\text{Dep}}(i))$. The process can be described as:

$$f_{\text{MA}}^{\text{AIF}}(i) = G_i^{AiF} \odot f_{\text{Sim}}^{\text{AiF}}(i) \text{ and } f_{\text{MA}}^{\text{Dep}}(i) = G_i^{Dep} \odot f_{\text{Sim}}^{\text{Dep}}(i), \tag{4.12}$$

where $\odot$ represents Hadamard product, the gate function $G_i^{AiF} = \sigma(Conv(f_{\text{Sim}}^{\text{AiF}}(i)))$ with $Conv(\cdot)$ and $\sigma(\cdot)$ denoting a convolutional layer and a Sigmoid function, respectively. In CMA-Net, we cascade two identical mutual attention modules to establish the decoder, thus acquiring the best performance (see detailed ablation studies in Section 4.2.5).

Co-supervision and hybrid loss. As shown in Fig. 4.7, to stabilize the multi-modal learning process, we apply a three-way strategy to co-supervise the training of our CMA-Net. Besides, inspired by a multi-loss function training setting applied in [122], we combine three loss functions including widely used binary cross entropy loss (ℓ_{BCE}), intersection over union loss (ℓ_{IoU}) and E-loss ($\ell_{\text{EM}} = 1 - E_\phi$), which is based on a recently proposed salient object segmentationmetric (E_ϕ [279]). Therefore, our hybrid loss function is denoted as:

$$\ell = \sum_{n=1}^{N} \ell_{\text{BCE}}(P_n, G) + \ell_{\text{IoU}}(P_n, G) + \ell_{\text{EM}}(P_n, G), \tag{4.13}$$

where $\{P_n\}_{n=1}^{3} \in [0,1]$ denotes the predicted three-way saliency maps, while $G \in \{0,1\}$ denotes the corresponding ground-truth binary mask.

Implementation details. Our CMA-Net is implemented in PyTorch 1.8 and optimized with Adam algorithm [3]. During the training stage, the batch size is set to 16, the learning rate is initialized as

1e-4 with a decay rate of 0.1 for every 50 epochs. It takes about one hour to finish the training of CMA-Net based on a platform consists of Intel® Xeon(R) W-2255 CPU @ 3.70GHz and one Quadro RTX 6000 GPU.

4.2.5 Experiments

Settings.

Datasets. We evaluate our SA-Net, SA-Net-V2, CMA-Net and 28 state-of-the-art salient object segmentation methods based on three widely-used light field datasets: DUT-LF, HFUT and LFSD, which all provide focal stack and semantic ground truth corresponding to each of the all-in-focus images (see detailed statistics of light field datasets in Section 2.4.3). For fair comparison, we simply follow the settings of a top-ranking method, i.e., ERNet [245]. To be specific, 1000/100 all-in-focus images of DUT-LF/HFUT are randomly selected as the training set, respectively, while the remains (462+155) and the whole LFSD are used for testing. Notably, as for competing methods, we report the results directly provided by authors or generated by officially released codes.

Metrics. We adopt the recently proposed S-measure (S_α) [278] and E-measure (E_ϕ) [279], also the generally agreed Mean Absolute Error (M) [277] and F-measure (F_β) [276] as evaluation metrics for the quantitative comparison between benchmark models and SA-Net and SA-Net-V2. Please note that, following the benchmark in [245], we report adaptive F/E-measure scores of each of the benchmark models.

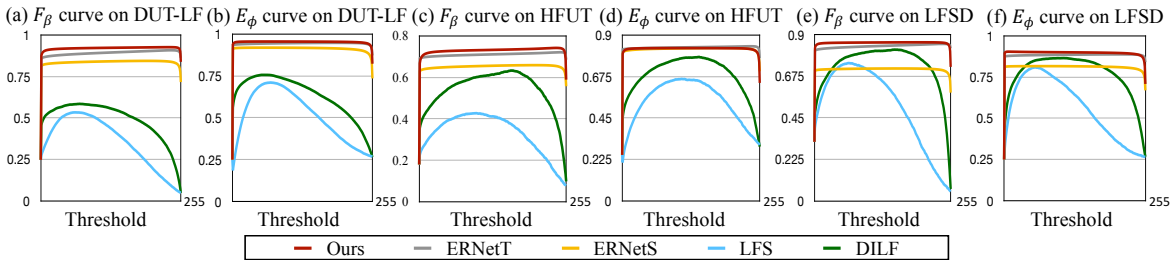


Fig. 4.8 F-measure (F_β) and E-measure (E_ϕ) curves of state-of-the-art light field salient object segmentation models and our SA-Net upon three datasets.

Comparison with state-of-the-art methods.

Quantitative results. We quantitatively compare our SA-Net, SA-Net-V2 and CMA-Net with 12/9/7 state-of-the-art RGB/RGB-D/light field salient object segmentation methods, respectively. As shown in Table 4.1, our SA-Net-V2 outperforms all state-of-the-art salient object segmentation models by a large margin in terms of all four evaluation metrics. We also perform a detailed comparison between our SA-Net, CMA-Net and the competing light field salient object segmentation methods by using F/E-measure curves. The results, shown in Fig. 4.8 and Fig. 4.10, indicate the F/E-measure curves of our SA-Net and CMA-Net are higher than those ones of competing models.

Qualitative results. Furthermore, we show some of the predicted saliency maps in Fig. 4.9 and Fig. 4.11. As can be observed, our SA-Net and CMA-Net provide saliency maps closest to the ground truth on various aspects, e.g., correct localization, intact object structure and clear details.

Table 4.1 Quantitative results for different models on three benchmark datasets. The best scores are in **boldface**. We train and test our SA-Net, SA-Net-V2 and CMA-Net with the setting that is consistent with [245], which is the state-of-the-art model at present. $\star$ indicates tradition methods. - denotes no available result. $\uparrow$ indicates the higher the score the better, and vice versa for $\downarrow$. LF = light field salient object segmentation methods. 3D = RGB-D salient object segmentation models. 2D = 2D image-based salient object segmentation methods.

Types	Models	DUT-LF [240]				HFUT [239]				LFSD [238]			
		$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$
LF	**SANetV2**	**0.941**	**0.940**	**0.964**	**0.023**	**0.803**	**0.853**	**0.887**	**0.055**	**0.865**	**0.873**	**0.900**	**0.061**
	SANet	0.920	0.918	0.954	0.032	0.736	0.784	0.849	0.078	0.844	0.841	0.889	0.074
	CMANet	0.917	0.918	0.949	0.033	0.744	0.807	0.865	0.069	0.823	0.830	0.864	0.083
	ERNetT [245]	0.889	0.899	0.943	0.040	0.705	0.777	0.831	0.082	0.842	0.838	0.889	0.080
	ERNetS [245]	0.838	0.848	0.916	0.061	0.651	0.736	0.824	0.085	0.721	0.726	0.820	0.137
	DLFS [241]	0.801	0.841	0.891	0.076	0.615	0.741	0.783	0.098	0.715	0.737	0.806	0.147
	LFS* [238]	0.484	0.563	0.728	0.240	0.430	0.579	0.686	0.205	0.740	0.680	0.771	0.208
	MCA* [239]	-	-	-	-	-	-	-	-	0.815	0.749	0.841	0.150
	WSC* [336]	-	-	-	-	-	-	-	-	0.706	0.706	0.794	0.156
	DILF* [337]	0.641	0.705	0.805	0.168	0.555	0.695	0.736	0.131	0.728	0.755	0.810	0.168
3D	S2MA [343]	0.754	0.787	0.841	0.103	0.647	0.761	0.787	0.100	0.819	0.837	0.863	0.095
	D3Net [216]	0.790	0.822	0.869	0.084	0.692	0.778	0.827	0.080	0.804	0.825	0.853	0.095
	CPFP [346]	0.730	0.741	0.808	0.101	0.594	0.701	0.768	0.096	0.524	0.599	0.669	0.186
	TANet [347]	0.771	0.803	0.861	0.096	0.638	0.744	0.789	0.096	0.804	0.803	0.849	0.112
	MMCI [348]	0.750	0.785	0.853	0.116	0.645	0.741	0.787	0.104	0.796	0.799	0.848	0.128
	PDNet [349]	0.763	0.803	0.864	0.111	0.629	0.770	0.786	0.105	0.780	0.786	0.849	0.116
	PCA [350]	0.762	0.800	0.857	0.100	0.644	0.748	0.782	0.095	0.801	0.807	0.846	0.112
	CTMF [351]	0.790	0.823	0.881	0.100	0.620	0.752	0.784	0.103	0.791	0.801	0.856	0.119
	DF [352]	0.733	0.716	0.838	0.151	0.562	0.670	0.742	0.138	0.756	0.751	0.816	0.162
2D	F3Net [122]	0.882	0.888	0.900	0.057	0.718	0.777	0.815	0.095	0.797	0.806	0.824	0.106
	GCPANet [121]	0.867	0.885	0.898	0.064	0.691	0.777	0.799	0.105	0.805	0.822	0.809	0.097
	EGNet [141]	0.870	0.886	0.914	0.053	0.672	0.772	0.794	0.094	0.762	0.784	0.776	0.118
	PoolNet [140]	0.868	0.889	0.919	0.051	0.683	0.776	0.802	0.092	0.769	0.800	0.786	0.118
	PAGRN [353]	0.828	0.822	0.878	0.084	0.635	0.717	0.773	0.114	0.725	0.727	0.805	0.147
	C2S [354]	0.791	0.844	0.874	0.084	0.650	0.763	0.786	0.111	0.749	0.806	0.820	0.113
	R^{3}Net [355]	0.783	0.819	0.833	0.113	0.625	0.727	0.728	0.151	0.781	0.789	0.838	0.128
	Amulet [356]	0.805	0.847	0.882	0.083	0.636	0.767	0.760	0.110	0.757	0.773	0.821	0.135
	UCF [357]	0.769	0.837	0.850	0.107	0.623	0.754	0.764	0.130	0.710	0.762	0.776	0.169
	SRM [358]	0.832	0.848	0.899	0.072	0.672	0.762	0.801	0.096	0.827	0.826	0.863	0.099
	NLDF [359]	0.778	0.786	0.862	0.103	0.636	0.729	0.807	0.091	0.748	0.745	0.810	0.138
	DSS [360]	0.728	0.764	0.827	0.128	0.626	0.715	0.778	0.133	0.644	0.677	0.749	0.190

Robustness of the proposed SA-Net. It is worth nothing that our SA-Net and SA-Net-V2 trained on DUT-LF and HFUT also achieves promising performance on the unseen dataset, i.e., LFSD, indicating its superior generalization ability and robustness. Theoretically, the robustness of SA-Net owes to the synergistic attention mechanism. In practice, attention mechanisms can improve network robustness [146, 353, 359] since they emphasize the most informative features and reduce the disturbance of noisy features. Our SA module employs both channel attention and co-attention for better feature representation, which can also improve the robustness of our model.

Efficiency of the propposed CMA-Net. It is worth mentioning that our CMA-Net is capable of running at 53 fps, being much more efficient than the top-ranked ERNetT [245] which reports an inference speed of only 14 fps. Besides, our proposed SA-Net and SA-Net-V2 have run-time of 47 fps and 26 fps during testing, respectively. Please note that the inference speed of all proposed models, *i.e.*,

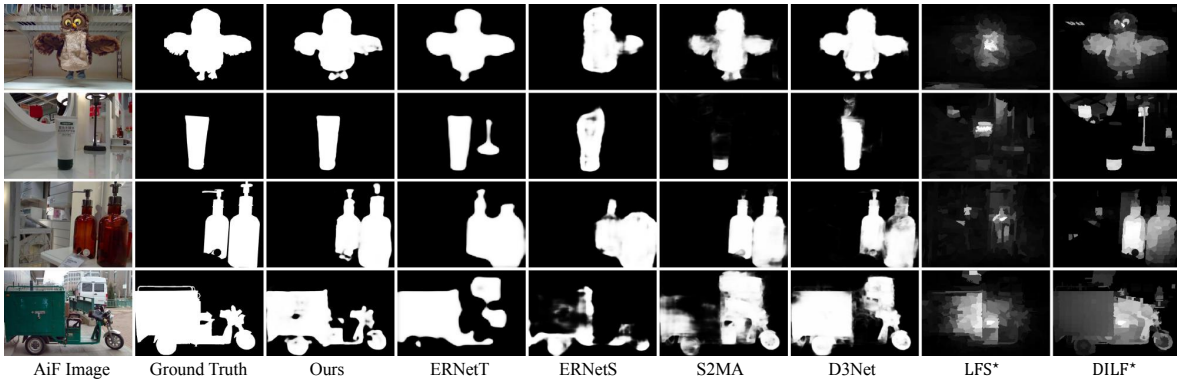


Fig. 4.9 Qualitative comparison between our SA-Net and state-of-the-art light field salient object segmentation models. ⋆ denotes traditional methods. Our SA-Net provides predictions closest to the ground truth on various aspects. More visual results are shown in Fig. 4.14, Fig. 4.15, Fig. 4.16 and Fig. 4.17.

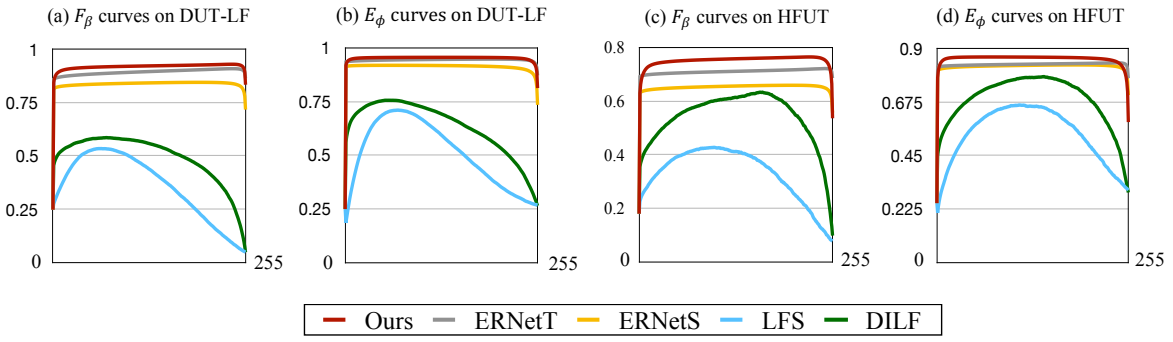


Fig. 4.10 F-measure (F_β) and E-measure (E_ϕ) curves of state-of-the-art light field salient object segmentation models and our CMA-Net upon two benchmark datasets, *i.e.*, DUT-LF and HFUT.

SA-Net, SA-Net-V2 and CMA-Net are computed based on one Quadro RTX 6000 GPU.

Ablation studies of SA-Net.

To verify the effectiveness of each proposed module of our SA-Net, we conduct thorough ablation studies by gradually adding key components. We first construct a baseline "B", which extracts all-in-focus and focal stack features with two 2D ResNet50 backbones, simply concatenates, and up-samples the pair-wise high-level features for salient object segmentation.

Effectiveness of multi-modal encoder. To investigate the effectiveness of our multi-modal encoder, we construct the second ablated version "ME", which is similar to "B", but using a 3D backbone to extract focal stack features, consistent with our multi-model encoder (Section 4.2.3). The results, shown in Table 4.2, indicate that "ME" outperforms "B" in terms of all evaluations, demonstrating the effectiveness of our 3D convolutional neural network-based encoder. Besides, to confirm the effectiveness of RFB for multi-level feature refinement, we also construct "ME0" without using the RFB, when compared to "ME". The result (Table 4.2) shows that RFB benefits significantly to the task.

Effectiveness of synergistic attention (SA) module. To investigate the effectiveness of our SA module,

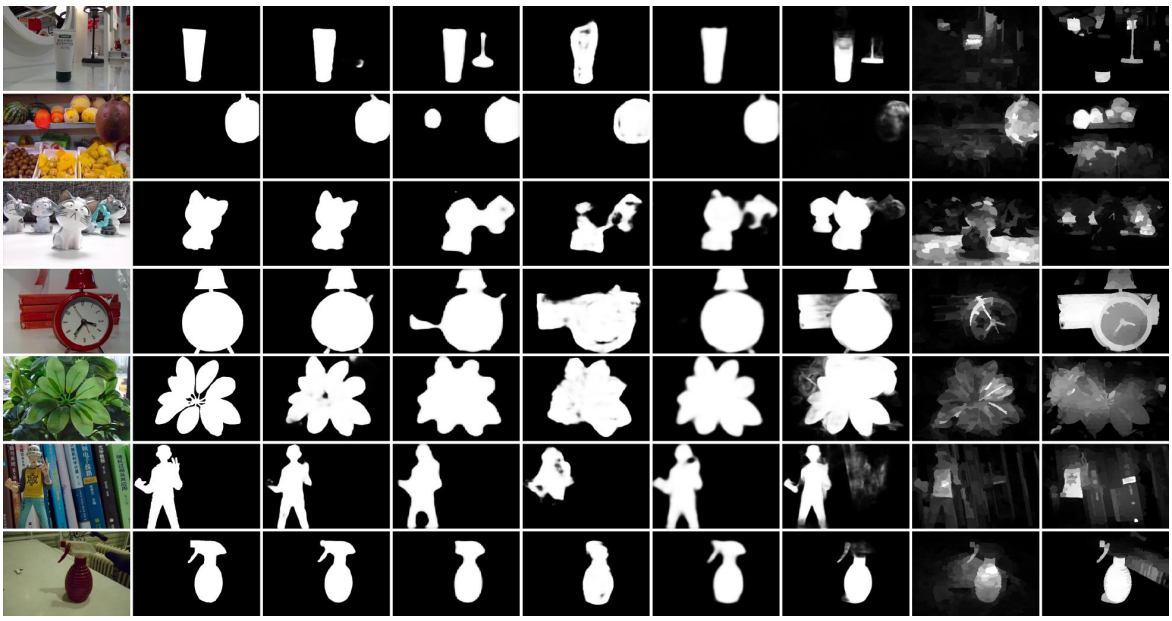


Fig. 4.11 Visual results our CMA-Net and state-of-the-art multi-modal salient object segmentation-models. ⋆ denotes traditional methods. AiF Image = all-in-focus image.

we further construct "SA1" and "SA2", which incorporate the SA into "ME" without and with CoA, respectively. As shown in Table 4.2, both "SA1" and "SA2" improve the performance in comparison with "ME". In particular, the full version of SA ("SA2") provides a significant improvement compared to "ME", indicating the importance of synergistic attention for learning the complementarities of multi-modal features. Besides, we compare SA-Net with F-SA (Figure 4.12), which consists of full four SA modules that fuse both the high and low level features for light field salient object segmentation. An interesting finding is that an increase of parameters (about 1.3 million of increment) focusing on low level features do not contribute to performance improvement (Table 4.2), which is also consistent with the conclusion in [139].

Effectiveness of progressive fusion (PF) module. Compared with "SA2", "PF1" uses the deconvolu-

Table 4.2 Quantitative results for the ablation studies of SA-Net on DUT-LF [240] and LFSD [238]. The best scores are in **boldface**. ↑ indicates the higher the score the better, and vice versa for ↓.

	Metric	B	ME0	ME	SA1	SA2	PF1	PF2	F-SA	SA-Net
DUT-LF	$F_\beta \uparrow$	0.871	0.874	0.881	0.890	0.899	0.912	0.919	0.913	**0.920**
	$M \downarrow$	0.051	0.051	0.048	0.041	0.037	0.035	0.034	0.037	**0.032**
LFSD	$F_\beta \uparrow$	0.811	0.825	0.835	0.836	0.835	0.838	0.839	**0.845**	0.844
	$M \downarrow$	0.095	0.089	0.080	0.079	0.077	0.075	0.075	0.078	**0.074**

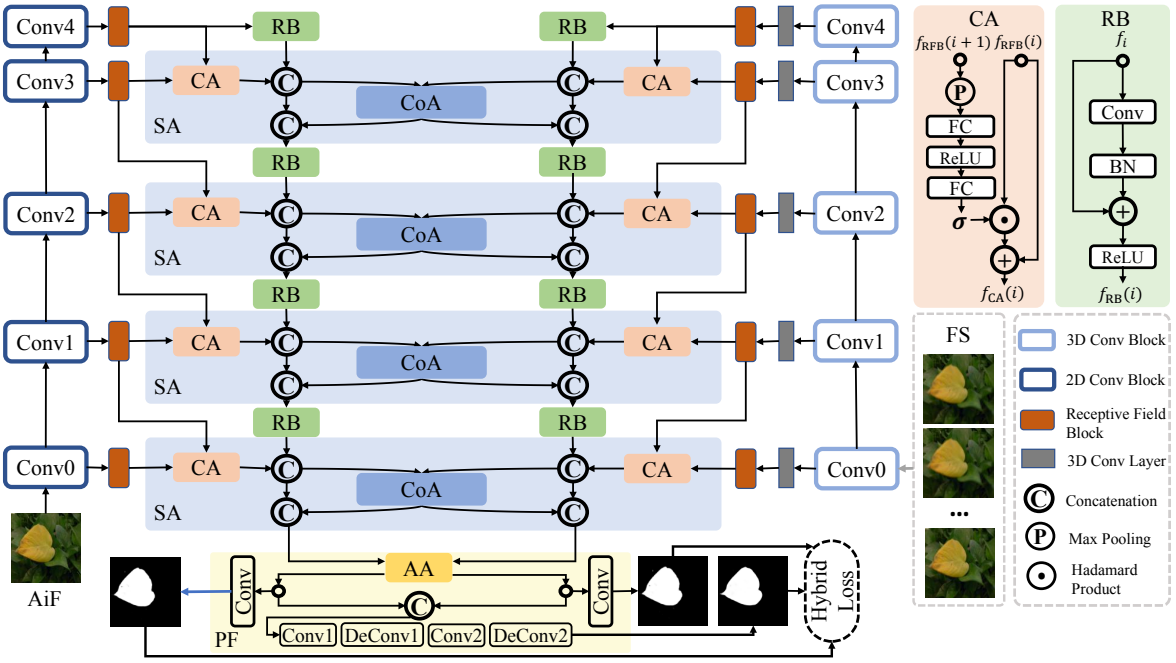


Fig. 4.12 An overview of F-SA with four SA modules. CoA = co-attention component. CA = channel attention component. AA = AiF-induced attention component. RB = residual block.

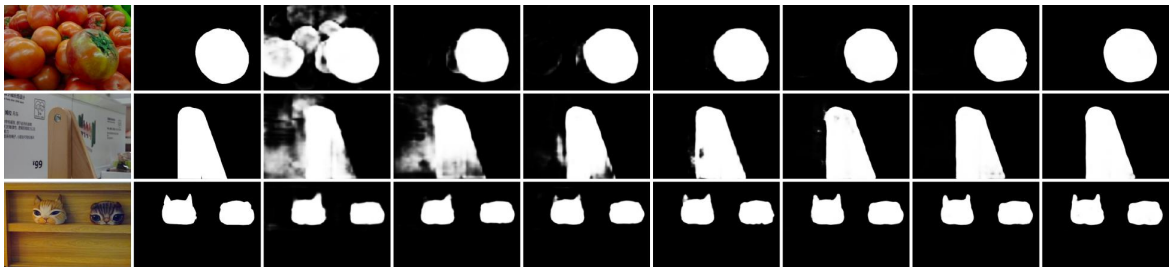


Fig. 4.13 Visual results of ablation models.

tional block (Figure 4.3) to gradually up-sample the features for predicting the saliency map. Besides, a three-way supervision ("PF2") is further employed to provide a deep supervision for the training. Finally, with the AA component (details in Section 4.2.3), our SA-Net achieves the best performance (Table 4.2), and provides the saliency maps closest to ground truth (Figure 4.13).

Ablation studies of CMA-Net.

We conduct thorough ablation studies to further verify the effectiveness of each module of the proposed method. We first construct basic "model1" which consists of single-branch ResNet layers and a group of (de-)convolutional layers, without the inputs of depth maps. Followed by "model2", which contains the duel-branch encoder with both all-in-focus images and depth maps as inputs. As a result, we find that depth information can be helpful for the salient object segmentation task (Tab. 4.3). We then carefully add mutual attention mechanisms to different decoding stages. The "model3" and "model4" are embedded with one mutual attention module at the 2^{nd} and 3^{rd} stages (Section 4.2.4),

Table 4.3 Quantitative results of the ablation studies of CMA-Net on DUT-LF [240] and HFUT [239]. The best scores are in **boldface**. ↑ indicates the higher the score the better, and vice versa for ↓.

Metric		Model1	Model2	Model3	Model4	CMA-Net
DUT-LF	$F_\beta \uparrow$	0.879	0.895	0.914	0.915	**0.917**
	$S_\alpha \uparrow$	0.893	0.911	0.916	0.916	**0.918**
	$E_\phi \uparrow$	0.931	0.943	0.949	**0.950**	0.949
	$M \downarrow$	0.047	0.039	0.034	0.034	**0.033**
HFUT	$F_\beta \uparrow$	0.697	0.704	0.727	0.729	**0.744**
	$S_\alpha \uparrow$	0.792	0.795	0.791	0.791	**0.807**
	$E_\phi \uparrow$	0.837	0.828	0.842	0.858	**0.865**
	$M \downarrow$	0.074	0.078	0.076	0.071	**0.069**

respectively. Finally, we cascade two mutual attention modules (*i.e.*, CMA-Net) and thus gaining the best performance compared to all ablation models (Tab. 4.3).

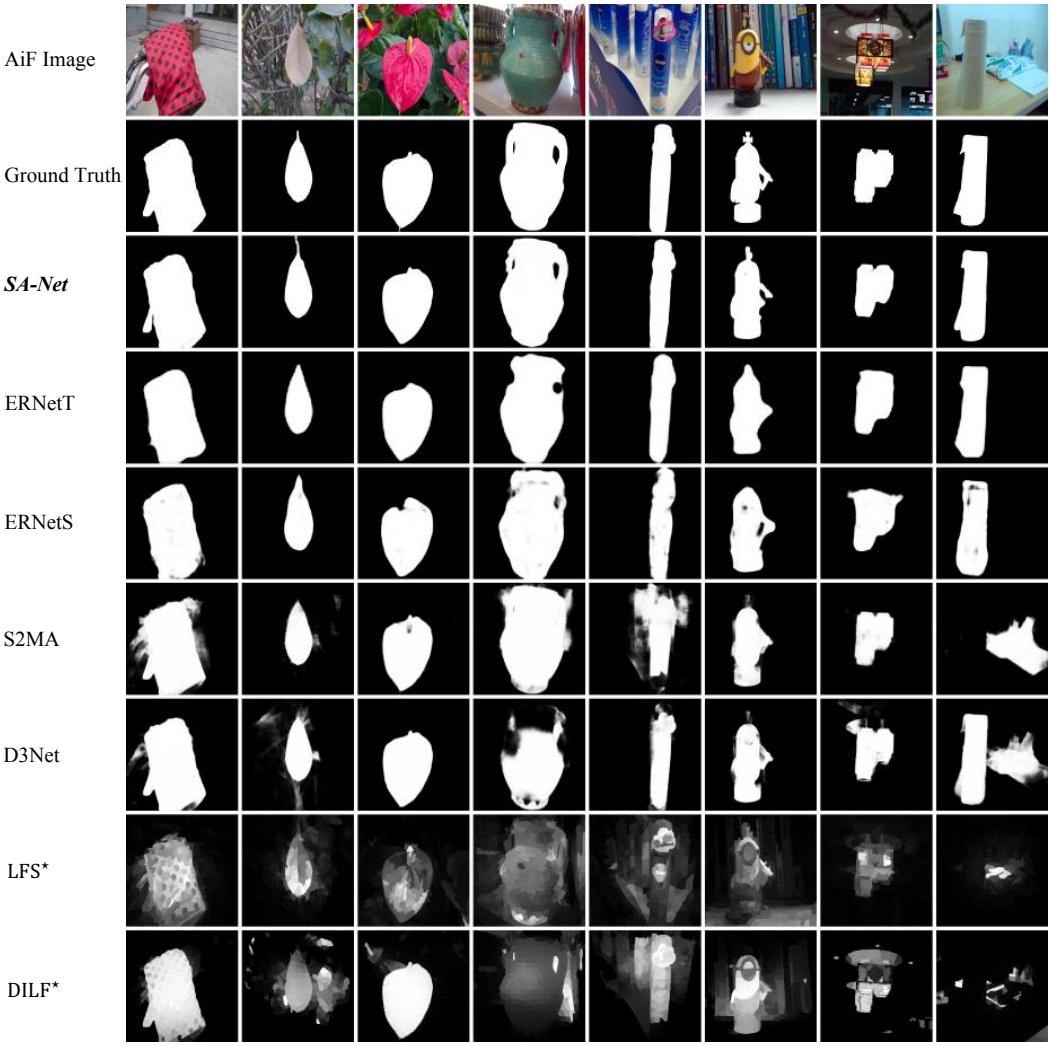


Fig. 4.14 Visual comparison of our SA-Net and state-of-the-art salient object segmentation models upon DUT-LF [240] (1/2). ⋆ indicates tradition methods.

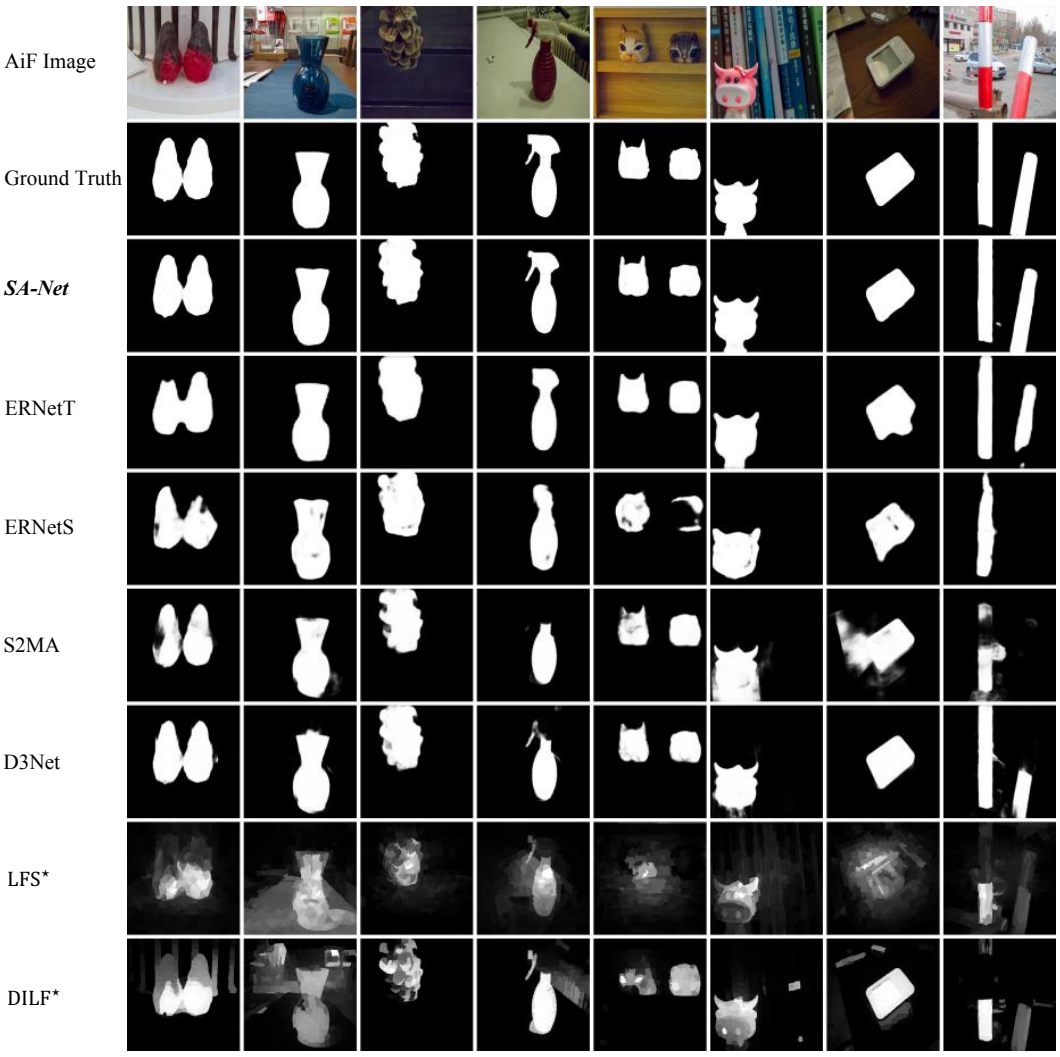


Fig. 4.15 Visual comparison of our SA-Net and state-of-the-art salient object segmentation models upon DUT-LF [240] (2/2). ⋆ indicates tradition methods.

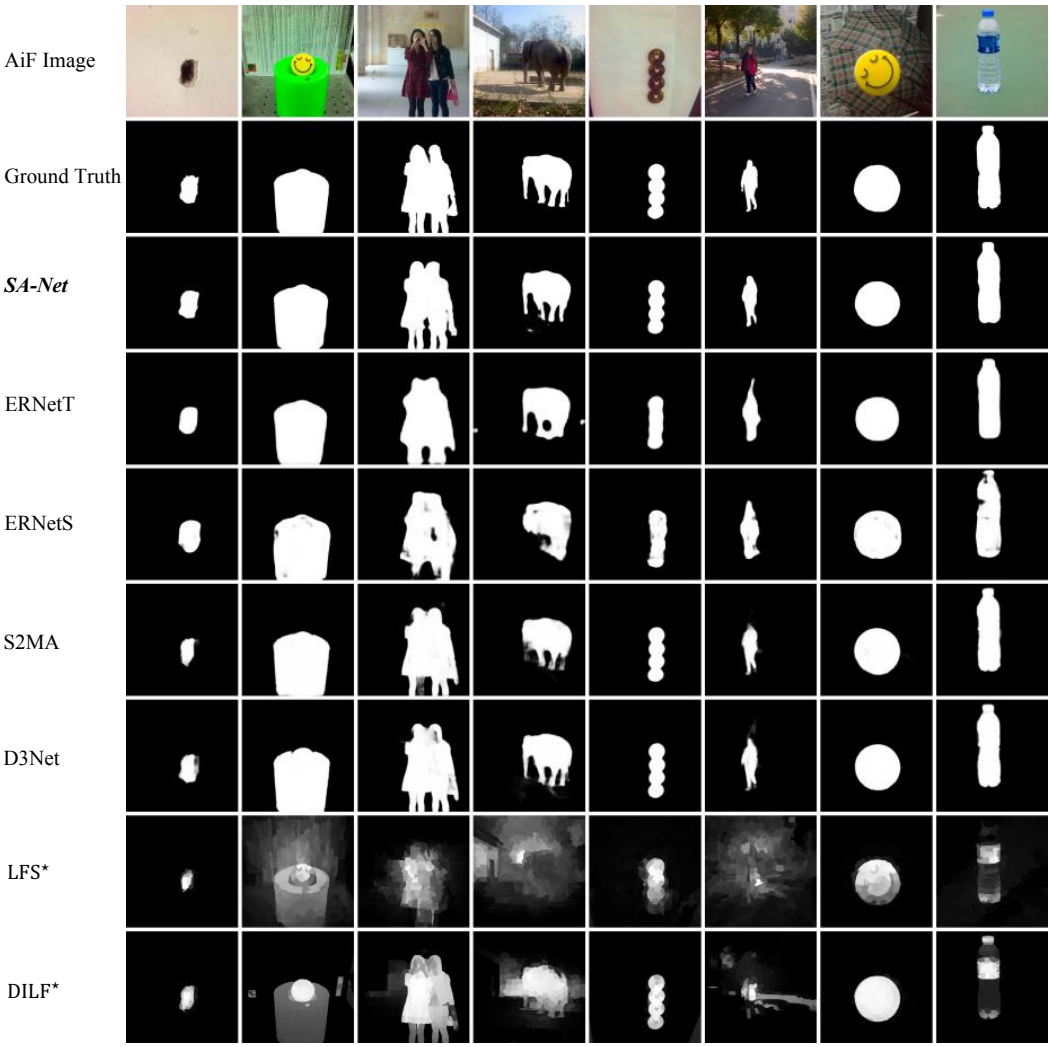


Fig. 4.16 Visual comparison of our SA-Net and state-of-the-art salient object segmentation models upon HFUT [239]. $\star$ indicates tradition methods.

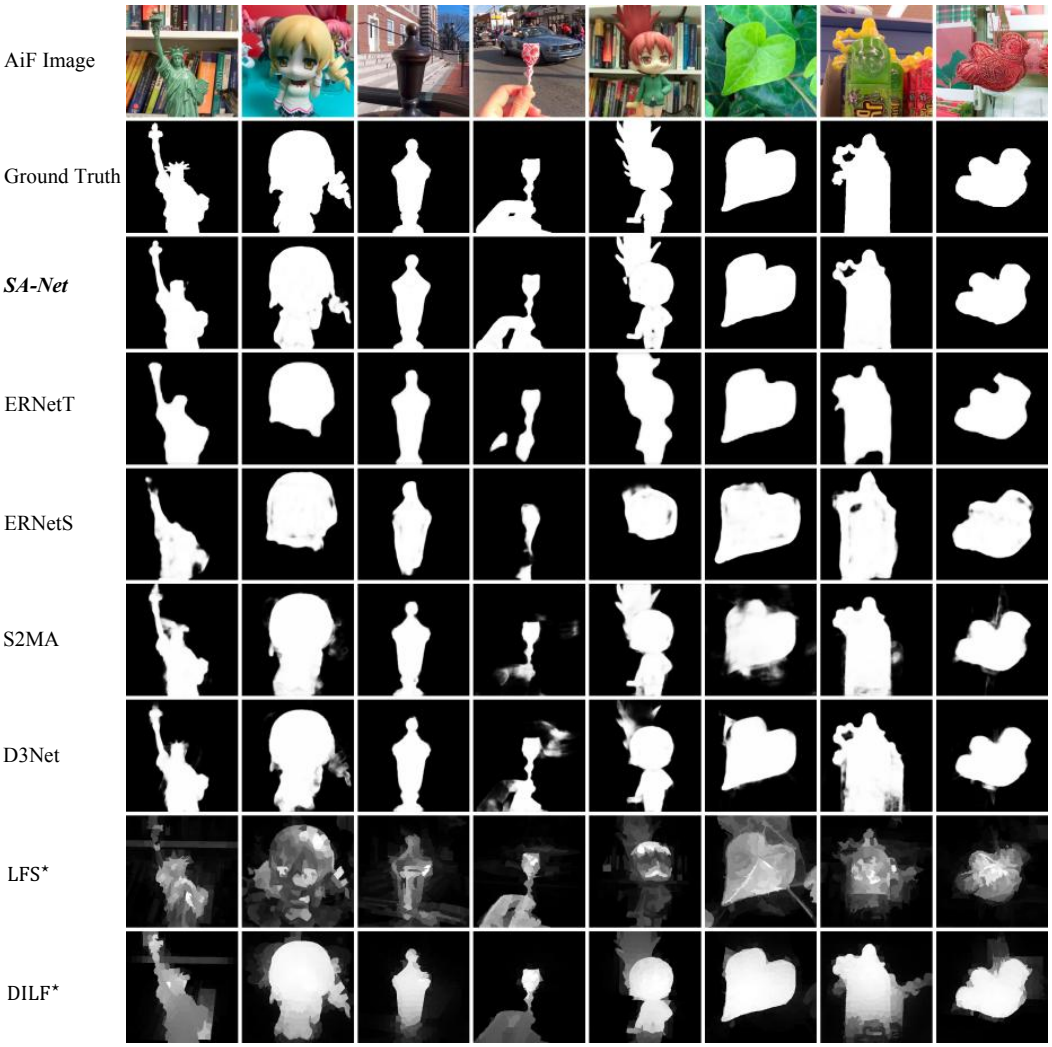


Fig. 4.17 Visual comparison of our SA-Net and state-of-the-art salient object segmentation models upon LFSD [238]. ⋆ indicates tradition methods.

4.3 Conclusion

In this chapter, we proposed new methodologies for light field salient object segmentation. The new methods can be summarized from the following two aspects:

Focal stack-based strategy. We proposed SA-Net and SA-Net-V2, which address the light field salient object segmentationby learning synergistic attention between two light field modalities, *i.e.*, all-in-focus images and focal stacks. The innovative attributes of our SA-Net are three-fold:

- It exploits the cross-modal complementary information by establishing a synergistic effect between multi-modal features.
- It is the first attempt to learn both the spatial and inter-slice features of focal stacks with 3D convolutional neural networks.
- It predicts the saliency map with an effective fusion model in a progressive manner.

RGB-D-based strategy. Our CMA-Net, which consists of two cascaded novel mutual attention modules for RGB-D cross modal high-level feature fusion. CMA-Net achieves comparable performance on widely used light field benchmark datasets based on four widey used salient object segmentation metrics, and a superior inference speed of 53 fps.

To varify the effectiveness of the proposed models, we conducted extensive experiments on three widely-used light field datasets where 28 state-of-the-art salient object segmentation models and four widely-adopted metrics are involved. Extensive qualitative and quantitative experimental results on three light field datasets demonstrate the superiority of our SA-Net [251], SA-Net-V2 and CMA-Net when compared to 28 competing models.

Our work towards light field salient object segmentation proved the ability of mutual attention mechanism in multi-modal feature fusion and refinement, also inspired global-local feature fusion in the 360° domain (Chapter 5).

Chapter 5

Salient object segmentation in 360° images&videos

5.1 Introduction

In this chapter, we first illustrate our work towards 360° image-based salient object segmentation (Section 5.2). Specifically, inspired by the synergistic attention proposed in our SA-Net [251], we further propose channel-spatial mutual attention to fuse global-local features for effective salient object segmentation in 360° images. As a result, our new method outperforms state-of-the-art segmentation methods based on a 360° image-based salient object segmentation benchmark where multiple fine-tuning and testing strategies are applied to the widely-used 360° datasets. Extensive experimental results illustrate the effectiveness and robustness of the proposed method.

To further approximate the scenario where persons depend on both audio and visual cues to locate and recognize the salient objects in dynamic immersive environments, we summarize our work (Section 5.3) towards 360° audio-visual salient object segmentation based on our newly proposed panoramic audio-visual dataset, *i.e.*, PAVS10K (Section 3.4). Specifically, we proposed a new audio-visual conditional variational auto-encoder combining both audio and visual cues for effective and interpretable 360° video-based salient object segmentation. As a result, our new 360° audio-visual model is able to outperform state-of-the-art salient object segmentation and video object segmentation methods and to estimate uncertainties towards model predictions.

5.2 Channel-spatial mutual attention for 360° image-based salient object segmentation

In this section, we introduce our work towards 360° image-based salient object segmentation. Specifically, we conduct 360° panoramic salient object segmentation by taking advantage of both global and local visual cues of 360° images, with a novel **c**hannel-**s**patial **m**utual **a**ttention **net**work (CSMA-Net). The key component of the CSMA-Net is the proposed CSMA module, which cascades channel-/spatial-weighting-based mutual attentions. And it is worth noting that, the new CSMA is inspired by the SA module in SA-Net as introduced in the last section.

The objective of our CSMA module is to refine and fuse the bottleneck features from two separate encoders with different planar representations of 360° panorama as inputs, *i.e.*, equirectangular image and cube map.

5.2.1 Introduction

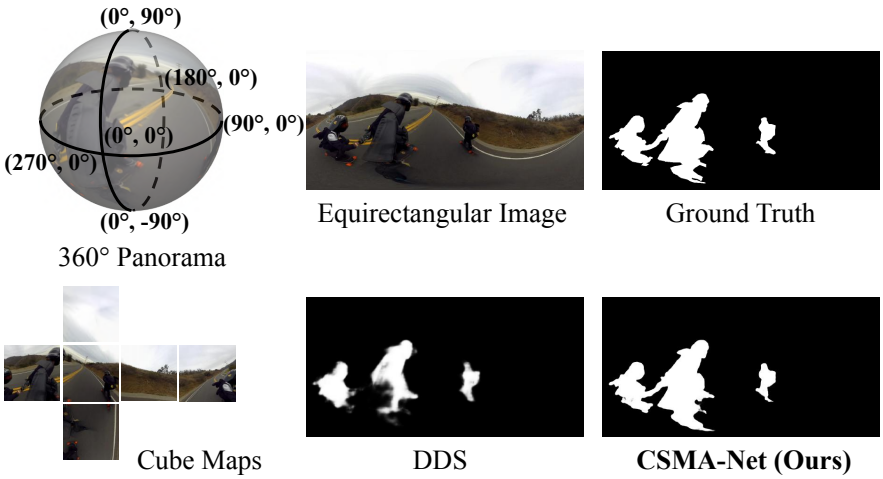


Fig. 5.1 An example of 360° panoramic salient object segmentation in terms of our CSMA-Net and DDS [257], which is a state-of-the-art model.

To recall, salient object segmentation is a task aiming to finely segment the objects that grasp most of the human attention within a given image, thus distinguishing itself from common visual saliency prediction [361] where only informative and salient regions are identified. During the past years, hundreds of deep learning methods have been proposed [16] to solve salient object segmentation in 2D images. Inspired by the benchmarks and methodologies of 2D salient object segmentation, 360° panoramic salient object segmentation [256] recently becomes a burgeoning field where the algorithms learn to segment the salient objects in images with a 360°×180° (Fig. 5.1) field-of-view depicting the real-life daily scenes. Owing to the importance of learning object-level human visual attention in immersive environments [257], 360° panoramic salient object segmentation is thus considered as one of the most essential joinpoints between salient object segmentation academia and augmented-/virtual-reality industries.

Compared to 2D salient object segmentation, 360° salient object segmentation is a burgeoning field where, as illustrated in Chapter 3, only three datasets have been established, *i.e.*, 360-SOD [257], F-360iSOD [256] and 360-SSOD [258]. To the best of our knowledge, DDS [257], FANet [334] and SW360 [258] are the only methods exclusively designed for 360° salient object segmentation. DDS [257] proposes a distortion-adaptive module which divides an inputting equirectangular image into four blocks, and tries to learn the features representing the specific level of geometrical distortions of each block locating at different regions of the given equirectangular image. The method is able to outperform contemporary 2D salient object segmentation methods, yet considers visual cues only from equirectangular blocks which focus less on local details compared to cube maps. FANet [334] uses channel attention mechanism [145] to weight and fuse the features of encoders for both the equirectangular image and cube maps. SW360 [258] designs a multi-stage framework to obtain the final predictions with the features based on 2D patches representing the content of local viewports. However, the proposed framework can not be trained in an end-to-end manner. Considering the large field-of-view, unique geometrical attributes and small objects, 360° salient object segmentation is still an opening issue that is far from being solved.

As there is no perfect planar representation for 360° images, equirectangular image contains the entire global context while brings extreme distortions to the region far from the equator (Fig. 5.1), cube maps alleviate the distortions of local content while give artificial boundaries between each of the cubes thus compromising the global continuity (Fig. 5.1). To acquire a balance from this trade-off, we follow the FANet [334] where the equirectangular image and cube map are considered for the global and local visual cues' modeling, respectively. Besides, to achieve more effective feature fusion, we argue that more sophisticated attention mechanisms (*e.g.*, mutual attention [180]) should be applied to explore the complementary information of visual cues of equirectangular image and cube map. To this end, we propose CSMA-Net, which consists of two separated encoder-decoder architectures and a novel channel-spatial mutual attention (CSMA) module fusing the bottleneck features of both branches of equirectangular image and cube maps. As a result, our CSMA-Net outperforms 10 state-of-the-art models, as being fine-tuned and tested on multiple 360° salient object segmentation datasets.

5.2.2 Methodologies

In this section, we introduce the framework of our CSMA-Net (Fig. 5.2), which consists of two encoder-decoder architectures and a new channel-spatial mutual attention (CSMA) module. The whole architecture of our CSMA-Net is trained in an end-to-end manner.

Uniqueness of the proposed CSMA-Net. Our CSMA-Net is inspired by both COSNet [180] and our SA-Net which designed mutual attention modules that operate only on spatial domain. Recently, [223] proposed channel-wise mutual attention mechanism to facilitate feature refinement for the task of scene segmentation. However, the idea of designing channel-spatial mutual attention in a cascaded manner still lacks discussion. In this work, we propose to build a cascaded channel-spatial mutual attention module to aid global-local feature fusion based on 360° images.

Encoder-decode architecture. Inspired by FANet [334], we use equirectangular image and cube maps as the inputs, which represent the global and local visual cues of 360° image, respectively. With

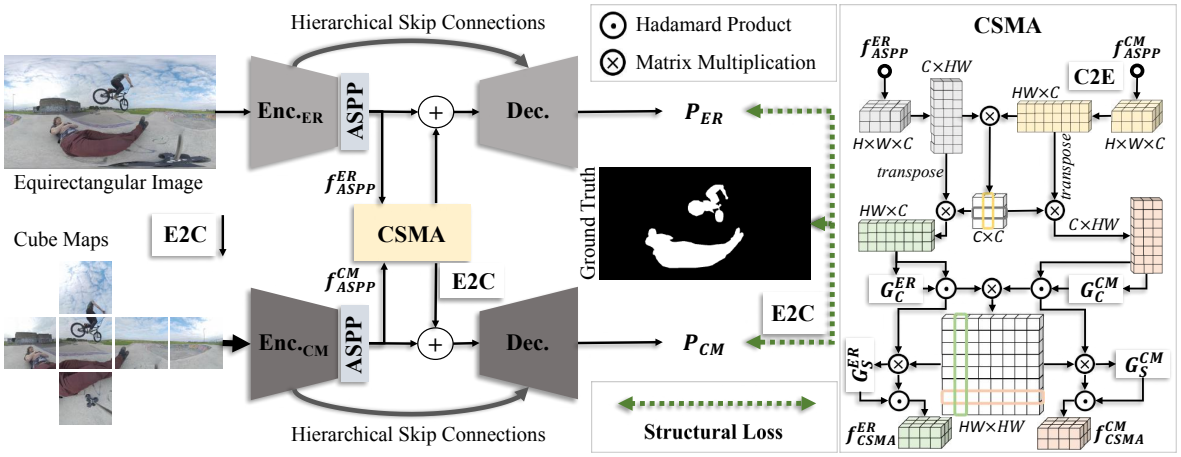


Fig. 5.2 The architecture of our CSMA-Net. The short names in the figure are detailed as follows: CSMA = the proposed channel-spatial mutual attention module. E2C/C2E = the projection interaction module which transforms the **e**quirectangular (ER) image/**c**ube maps to cube maps/ER image, respectively. ASPP = atrous spatial pyramid pooling module [181]. Enc.$_{\text{ER}}$ = the hybrid-ViT-based encoder [345] for equirectangular image. Enc.$_{\text{CM}}$ = the Res2Net-based encoder [362] for cube maps. Dec. = the decoder from RCRNet [178].

the given equirectangular image from 360-SOD [257] or 360-SSOD [258], we first apply the "E2C" module [363] to transform the equirectangular image to cube maps representing the local viewports observed from six orientations. We then feed the equirectangular image and corresponding cube maps to the separate encoders (Fig. 5.2), thus gaining hierarchical features $\{f_{\text{i}}^{\text{ER}}\}_{\text{i=1}}^{3}$ and $\{f_{\text{i}}^{\text{CM}}\}_{\text{i=1}}^{3}$, also bottleneck features, *i.e.*, $f_{\text{ASPP}}^{\text{ER}}$ and $f_{\text{ASPP}}^{\text{CM}}$. Further, we use the decoder structure from RCRNet [178] and add skip connections to link the $\{\{f_{\text{i}}^{\text{ER}}\}_{\text{i=1}}^{3}, \{f_{\text{i}}^{\text{CM}}\}_{\text{i=1}}^{3}\}$ to the corresponding decoding layers. The bottleneck features $\{f_{\text{ASPP}}^{\text{ER}}, f_{\text{ASPP}}^{\text{CM}}\}$ are then fed into the CSMA module for 360° spatial global-local feature fusion.

Channel-spatial mutual attention (CSMA). Inspired by the co-attention network, *i.e.*, COSNet [180], which is used for video object segmentation, we propose CSMA module establishing global connections between the bottleneck features $\{f_{\text{ASPP}}^{\text{ER}}, f_{\text{ASPP}}^{\text{CM}}\} \in \mathbb{R}^{H\times W\times C}$ respectively encoded from equirectangular image and cube maps. Being different to mutual-attention networks such as COSNet [180] and SA-Net [251], where only spatial-wise mutual attention is considered, we also introduce channel-wise operations [223] and thus establishing channel-spatial mutual attention in a cascaded manner, as shown in Fig. 5.2. Specifically, we first compute a similarity matrix M_c:

$$M_c = F(f_{\text{ASPP}}^{\text{ER}})^{\text{T}} \otimes F(f_{\text{ASPP}}^{\text{CM}}), \tag{5.1}$$

where $\otimes$ denotes matrix multiplication. $(\cdot)^{\text{T}}$ means the transpose of a matrix. $F(\cdot)$ denotes a flatten operation which reshapes $f_{\text{ASPP}}^{\text{ER}} \in \mathbb{R}^{H\times W\times C}$ to a 2D matrix with the dimension of $HW \times C$. $M_c \in \mathbb{R}^{C\times C}$. We thereby gain the channel-wise mutual attention-based outputs, *i.e.*, $f_{\text{CMA}}^{\text{ER}}$ and $f_{\text{CMA}}^{\text{CM}}$, by

computing:

$$\begin{aligned} f_{\mathrm{CMA}}^{\mathrm{ER}} &= reshape(f_{\mathrm{ASPP}}^{\mathrm{ER}} \otimes Softmax(M_c)), \\ f_{\mathrm{CMA}}^{\mathrm{CM}} &= reshape(f_{\mathrm{ASPP}}^{\mathrm{CM}} \otimes Softmax(M_c^{\mathrm{T}})). \end{aligned} \tag{5.2}$$

We are then able to gain the similarity matrix M_s for spatial mutual attention, by computing:

$$M_s = F(G_C^{ER}(f_{\mathrm{CMA}}^{\mathrm{ER}}) \odot f_{\mathrm{CMA}}^{\mathrm{ER}}) \otimes F(G_C^{CM}(f_{\mathrm{CMA}}^{\mathrm{CM}}) \odot f_{\mathrm{CMA}}^{\mathrm{CM}})^{\mathrm{T}}, \tag{5.3}$$

where $\odot$ denotes Hadamard product. $G_C^{ER}(\cdot)$ means a gate function [180] that learns confidence for the given features $f_{\mathrm{CMA}}^{\mathrm{ER}}$. With the M_s, the corresponding spatial mutual attention-based results are defined as:

$$\begin{aligned} \widehat{f}_{\mathrm{CSMA}}^{\mathrm{ER}} &= reshape(f_{\mathrm{CMA}}^{\mathrm{ER}} \otimes Softmax(M_s)), \\ \widehat{f}_{\mathrm{CSMA}}^{\mathrm{CM}} &= reshape(f_{\mathrm{CMA}}^{\mathrm{CM}} \otimes Softmax(M_s^{\mathrm{T}})). \end{aligned} \tag{5.4}$$

With the the other pair of gate functions $\{G_S^{ER}, G_S^{CM}\}$, we gain the final outputs of the CSMA module:

$$\begin{aligned} f_{\mathrm{CSMA}}^{\mathrm{ER}} &= G_S^{ER}(\widehat{f}_{\mathrm{CSMA}}^{\mathrm{ER}}) \odot \widehat{f}_{\mathrm{CSMA}}^{\mathrm{ER}}, \\ f_{\mathrm{CSMA}}^{\mathrm{CM}} &= G_S^{CM}(\widehat{f}_{\mathrm{CSMA}}^{\mathrm{CM}}) \odot \widehat{f}_{\mathrm{CSMA}}^{\mathrm{CM}}. \end{aligned} \tag{5.5}$$

Please refer to Fig. 5.2 for the model visualization. Besides, the effectiveness of the proposed CSMA module is tested in Section 5.2.3 where thorough ablation studies are presented.

Loss function. As shown in Fig. 5.2, our CSMA-Net is trained in an end-to-end manner by using structure loss (L^{S}) [122], which is the sum of the weighted BCE loss ($L_{\mathrm{wbce}}^{\mathrm{S}}$) and the weighted IoU loss ($L_{\mathrm{wiou}}^{\mathrm{S}}$). Therefore, the objective function (L) for our CSMA-Net is thus defined as:

$$L = L^{\mathrm{S}}(P_{\mathrm{ER}}, Y) + L^{\mathrm{S}}(P_{\mathrm{CM}}, E2C(Y)), \tag{5.6}$$

where P_{ER} and P_{CM} are the predictions of equirectangular image-based global and cube maps-based local branches, respectively. Y is the ground truth (Fig. 5.2). $E2C(Y)$ outputs cube maps corresponding to the given Y.

Implementation details. Our CSMA-Net is implemented in PyTorch, trained with Adam optimizer [3]. Following the common settings in 2D salient object segmentation, we initialize our dual-branch encoder-decoder framework with DUTS-tr [20]-based pre-training. On the other hand, the proposed CSMA is randomly initialized. For fair comparison, we simply follow FANet [334] and resize each input equirectangular image to 512×1,024, without using multi-scale or any data augmentation strategies. During fine-tuning, the batch size is set to 1, the default learning rate is fixed to 2.5-6. It takes about 2.5 hours to fine-tune the model on the training set of 360-SOD [257], based on the PC consisting of Intel® Xeon® W-2255 CPU@3.70GHz and one Quadro RTX-6000 GPU.

5.2.3 Experiments

Settings.

Benchmark datasets. To validate the effectiveness as well as the robustness of our CSMA-Net, we establish a new benchmark (Table 5.1) based on cross-validation strategy where two 360° salient object segmentation datasets, *i.e.*, 360-SOD [257] and 360-SSOD [258] are applied for the fine-tuning and testing of our CSMA-Net and the 10 competing methods. 360-SOD [257], as the first 360° salient object segmentation dataset, provides 400 and 100 equirectangular images for training (360-SOD-tr) and testing (360-SOD-te), respectively. 360-SSOD [257] consists of training set (360-SSOD-tr) of 850 equirectangular images and a testing set (360-SSOD-te) of 255 equirectangular images. Both datasets provide pixel-wise binary masks as ground truth.

Benchmark models. The 10 competing models include 7 state-of-the-art 2D salient object segmentation methods (CPD [139], BASNet [138], MINet [126], LDF [128], CSFR2 [129], GateNet [132] and JointCS [364]), two 360° salient object segmentation models (*i.e.*, DDS [257] and FANet [334]) and one transformer-based segmentation model, *i.e.*, TransUNet [365]. Following the settings of the premier 360° salient object segmentation benchmark [257], all seven 2D salient object segmentation benchmark models are based on DUTS-tr (2D salient object segmentation training set) [20] pre-training and fine-tuned with 360° datasets in an end-to-end manner. For fair comparison, we also train the TransUNet [365] with DUST-tr [20] before the fine-tuning process. As for FANet [334], we directly fine-tune the model with 360° datasets and without using DUTS-tr based pre-training, since it only accepts equirectangular images as inputs. We fine-tune each of the benchmark models with their recommended hyper-parameters. Please note that we do not include SW360 [258] in our benchmark since it can not be fine-tuned in an end-to-end manner.

Evaluation metrics. Following the common settings in 2D salient object segmentation, we apply four widely used metrics, *i.e.*, mean F-measure (F_β) [276], MAE ($\mathscr{M}$) [277], S-measure (S_α) [278] and mean E-measure (E_ϕ) [279], for the evaluation of all benchmark models and our CSMA-Net.

Table 5.1 Performance comparison of our CSMA-Net and 10 state-of-the-art methods. S_α = S-measure (α=0.5) [278], F_β = mean F-measure (β^2=0.3) [276], E_ϕ = mean E-measure [279], $\mathscr{M}$ = mean absolute error [277]. ↑/↓ denotes a larger/smaller value is better. ‡ denotes codes not released. The three best results of each column are in **red**, green and blue.

Methods	Year	Fine-tuning on 360-SOD-tr [257]								Fine-tuning on 360-SSOD-tr [258]							
		360-SOD-te [257]				360-SSOD-te [258]				360-SOD-te [257]				360-SSOD-te [258]			
		$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$
CPD [139]	CVPR'19	.765	.624	.798	.030	.666	.432	.698	.051	.721	.573	.736	.031	.748	.578	.785	.031
BASNet [138]	CVPR'19	.801	.689	.840	.024	.659	.451	.728	.050	.778	.661	.804	.027	.746	.597	.809	.031
MINet [126]	CVPR'20	.797	.708	.866	.022	.664	.456	.714	.052	.746	.632	.788	.027	.747	.602	.807	.029
LDF [128]	CVPR'20	.813	.706	.869	.021	.673	.471	.730	.050	.783	.671	.824	.028	.763	.622	.836	.031
CSFR2 [129]	ECCV'20	.847	.779	.879	.017	.664	.447	.691	.049	.752	.550	.696	.035	.700	.448	.681	.043
GateNet [132]	ECCV'20	.793	.639	.791	.028	.668	.419	.681	.056	.747	.560	.730	.038	.730	.504	.723	.041
JointCS [364]	CVPR'21	.829	.749	.889	.022	.679	.489	.741	.050	.791	.700	.845	.026	.764	.620	.835	.032
DDS [257]	JSTSP'19	.803	.696	.866	.023	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡
FANet [334]	SPL'20	.826	.749	.873	.021	.642	.420	.688	.053	.735	.566	.703	.034	.717	.523	.727	.039
TransUNet [365]	arXiv'21	.815	.719	.887	.023	.671	.474	.754	.057	.784	.693	.851	.028	.771	.642	.847	**.028**
CSMA-Net	ICPR'22	**.873**	**.833**	**.924**	**.016**	**.698**	**.531**	**.757**	**.048**	**.829**	**.777**	**.881**	**.020**	**.784**	**.661**	**.859**	**.028**

Performance comparison.

Quantitative results. As shown in Table 5.1, our CSMA-Net outperforms all the competing models by a large margin over two fine-tuning 360° datasets (360-SOD [257] and 360-SSOD [258]), in terms of four widely used salient object segmentation metrics, *i.e.*, F-measure, MAE, S-measure

and E-measure. Among all benchmark models, CSFR2 [129] and TransUNet [365] are able to predict competing results owing to the advanced backbones, *i.e.*, Res2Net [362] and hybrid-ViT [143]. JointCS [364], as the most recently proposed salient object segmentation model in our benchmark, also provides close results when compared to our CSMA-Net.

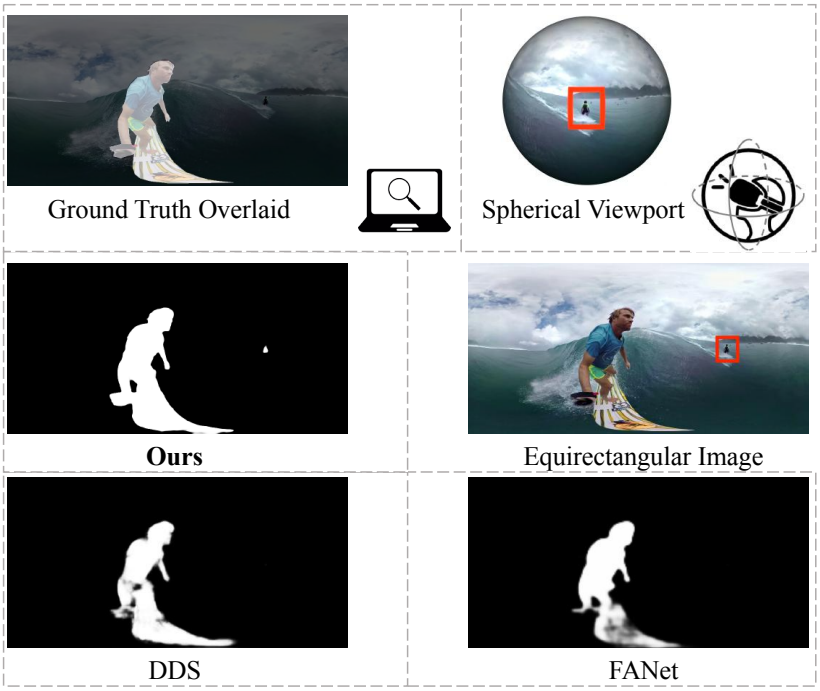


Fig. 5.3 An example illustrates the situation where our CSMA-Net(**Ours**) is able to detect small meaningful object, which is simply neglected by current 360° salient object segmentation datasets and models.

Qualitative results. Besides, as shown in Fig. 5.4, our CSMA-Net is able to predict results closest to the ground truth, from the respects of four cross-validation strategies. Specifically, our CSMA-Net provides saliency maps where salient objects are accurately spotted and finely depicted. On the other hand, the competing models sometimes fail to detect the 360° geometrical distortions (*e.g.*, the 2nd, 5th and 6th rows in Fig. 5.4 or small targets (*e.g.*, the 1st, 3rd and 11th rows in Fig. 5.4). Due to the limited space, we do not include the visualization results regrading all benchmark models in Fig. 5.4. Please refer to our supplementary materials for more qualitative results.

Ablation study.

First, as shown in Table 5.2, we verify the effectiveness of dual-branch encoder-decoder framework via "w/o ASPP", which is not loaded with ASPP [181] and any attention mechanisms. To further prove the effectiveness of the proposed CSMA module, we design another four ablated versions of our method, train and fine-tune them according to the same setting of our CSMA-Net. The quantitative results of fine-tuning and testing on two 360° salient object segmentation datasets respectively are shown in Table 5.2. Specifically, we first design a baseline version, denoted as "w/o Atten.", where no attention-based global-local feature interaction conducted at the bottleneck of the dual-branch encoder-decoder framework. We then focus on the validation of the proposed mutual-

Table 5.2 Ablation studies for our CSMA-Net. S_α = S-measure (α=0.5), F_β = mean F-measure (β^2=0.3), E_ϕ = mean E-measure, $\mathscr{M}$ = mean absolute error. ↑/↓ denotes a larger/smaller value is better. The two best results of each column are in **red** and blue.

Methods	360-SOD-te [257]				360-SSOD-te [258]			
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$\mathscr{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$\mathscr{M} \downarrow$
w/o ASPP	.864	.809	.920	.018	.777	.645	.849	.033
w/o Atten.	.869	.806	.916	.017	.776	.656	.837	.032
w/ SM-Atten.	.869	.820	.923	.017	.780	.666	.846	**.028**
w/ CM-Atten.	.870	.820	.918	**.016**	.781	**.670**	.854	**.028**
w/ CSMA(**Ours**)	**.873**	**.833**	**.924**	**.016**	**.784**	.661	**.859**	**.028**

attention mechanism, *i.e.*, CSMA. Inspired by SA-Net [251] and COSNet [180], we add spatial-wise mutual-attention module to the bottleneck of the framework, thus gaining the version "w/ SM-Atten.". Similarly, we replace the spatial-wise operation with channel-wise one and thus acquiring "w/ CM-Atten.". Further, we cascade both the channel-wise and spatial-wise mutual attentions and gain the final version of our CSMA module, *i.e.*, "w/ CSMA".

As a result, all mutual attention-based versions (*i.e.*, "w/ SM-Atten.", "w/ CM-Atten." and "w/ CSMA") outperform the baseline version ("w/o Atten."). More importantly, "w/ CSMA" provides the best results compared to the others, indicating the effectiveness and importance of our CSMA module.

5.2.4 Discussion

Mutual attention for 360°. Our CSMA module successfully fuse the global-local spatial information of 360° panoramic images based on equirectangular and cubemap projections. The ablation studies validate the effectiveness of the proposed CSMA module. Further, both the qualitative and quantitative experimental results illustrate the superiority of our CSMA-Net, which owes to the proposed mutual-attention-based global-local interactive architecture (Fig. 5.2).

Cross-validation strategy. As illustrated in Section 5.2.3, we establish so far the first 360° salient object segmentation benchmark involving multiple fine-tuning strategies, *i.e.*, fine-tuning on 360-SOD-tr [257] and testing on 360-SOD-te, fine-tuning on 360-SOD-tr and testing on 360-SSOD-te, fine-tuning on 360-SSOD-tr [258] and testing on 360-SOD-te, fine-tuning on 360-SSOD-tr and testing on 360-SSOD-te. Based on the sufficient quantitative results (Table 5.1), we observe a performance gap between different strategies, *e.g.*, the mean F-measure scores of all methods based on "fine-tuning on 360-SOD-tr and testing on 360-SOD-te" and "fine-tuning on 360-SOD-tr and testing on 360-SSOD-te" are about 0.717 and 0.459, respectively. The significant performance divergence indicates that current deep learning-based segmentation methods are strongly data-biased.

Small objects. Besides superior performance, an interesting finding is that our CSMA-Net is able to detect the small meaningful object in equirectangular images (Fig. 5.3), which tends to be ignored by current 360° salient object segmentation methods such as DDS [257] and FANet [334], also be easily regarded as non-salient objects in current datasets (*e.g.*, an example collected from 360-SOD [257] in Fig. 5.3) where the annotators conduct saliency judgements based on equirectangular image shown

on PC screen, rather than Head-Mounted Displays which captures a wide field-of-view (360°×180°) reflecting more realistic scenes (Fig. 5.3). Thus, our CSMA-Net probably provides supports for future augmented-/virtual-reality applications, where omnidirectional field-of-view is widely applied for human visual attention modeling and viewport-based meaningful objects may be considered important.

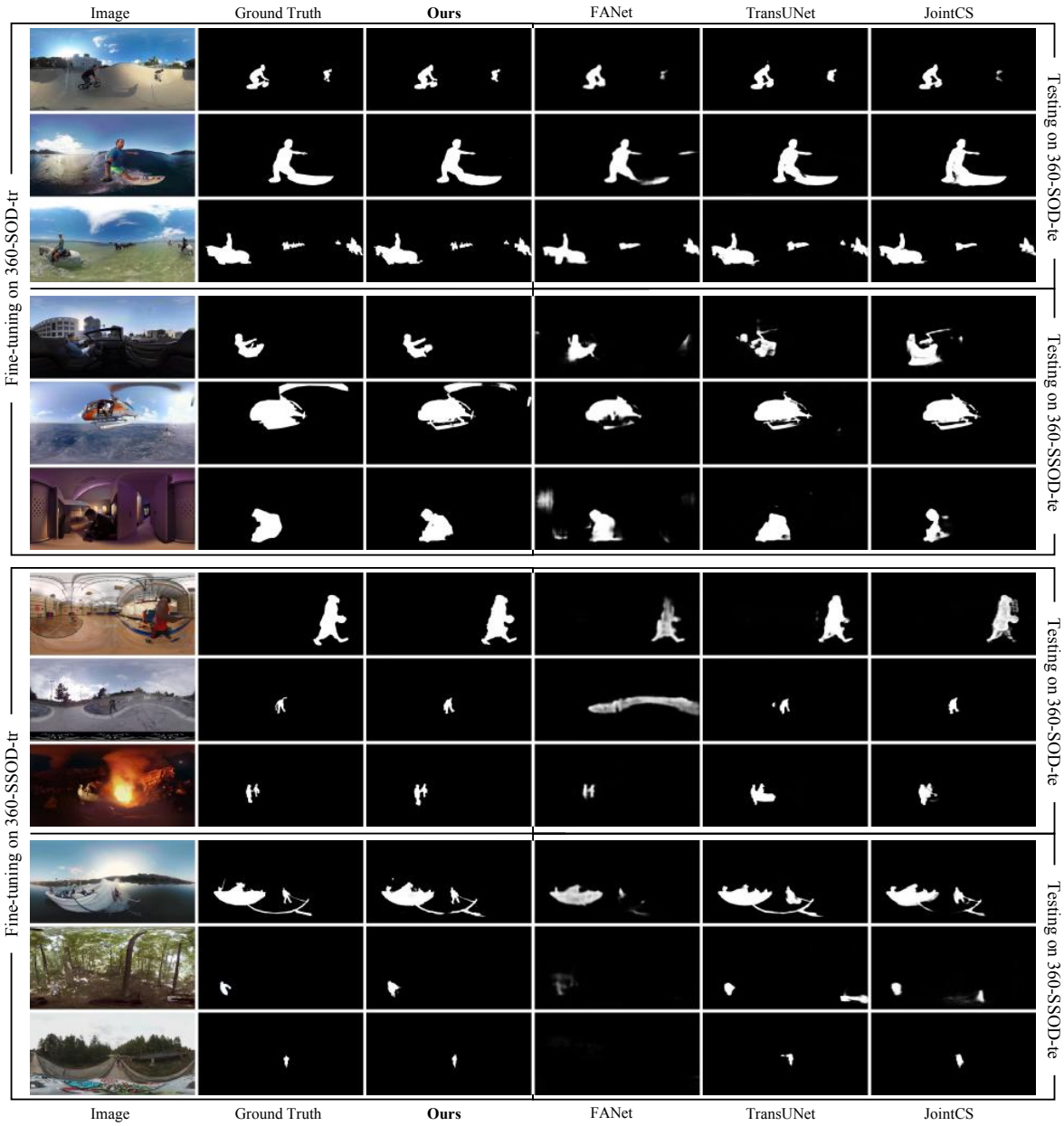


Fig. 5.4 Visualization results of our CSMA-Net(**Ours**) and the state-of-the-art methods. Our CSMA-Net is able to provide results closest to the ground truth. More visual results are presented in Fig. 5.5 and Fig. 5.6.

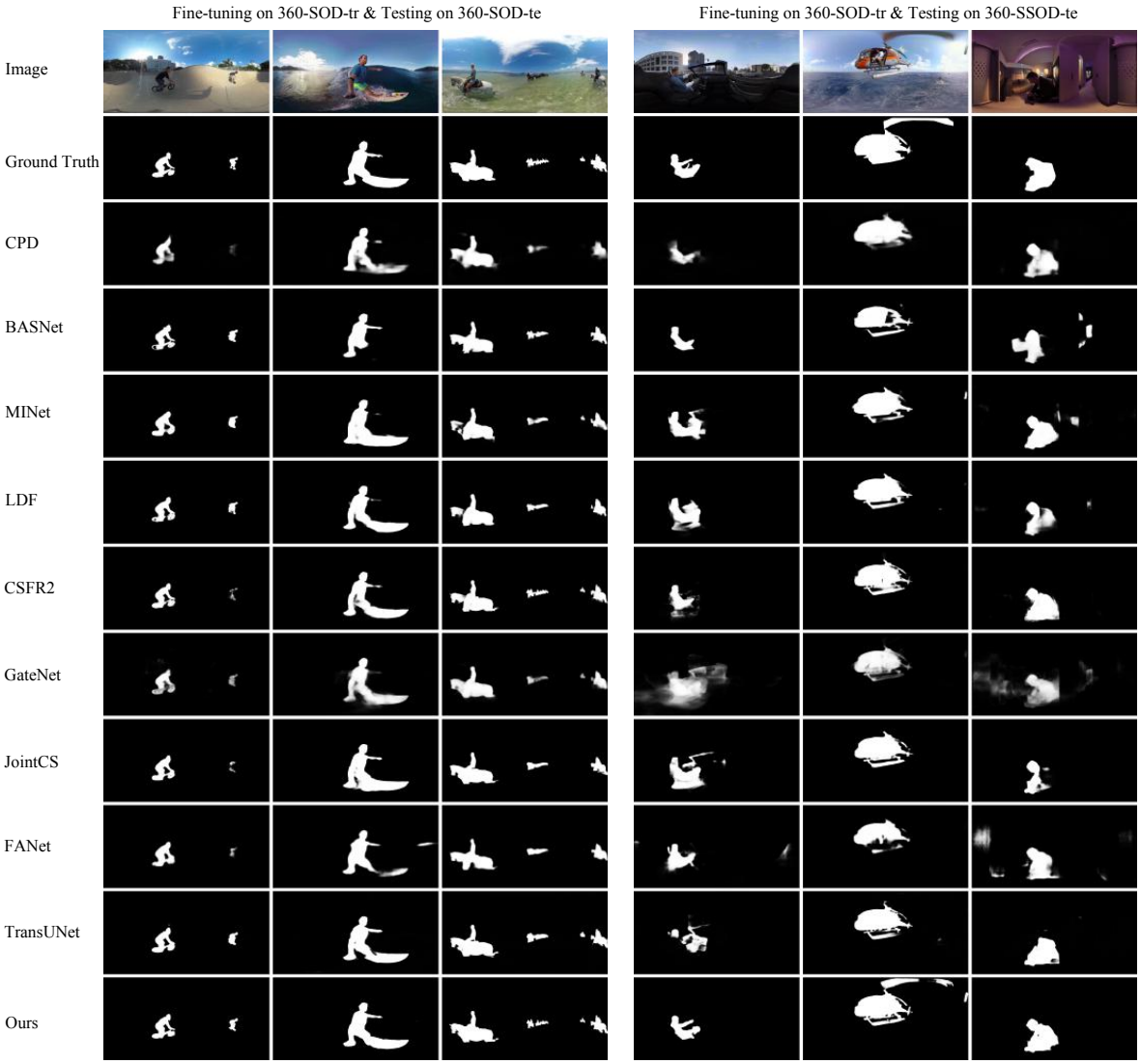


Fig. 5.5 More visualization results (part 1/2) of our CSMA-Net(**Ours**) and the state-of-the-art methods. Note that our CSMA-Net finely depicts small or distorted salient targets in equirectangular images.

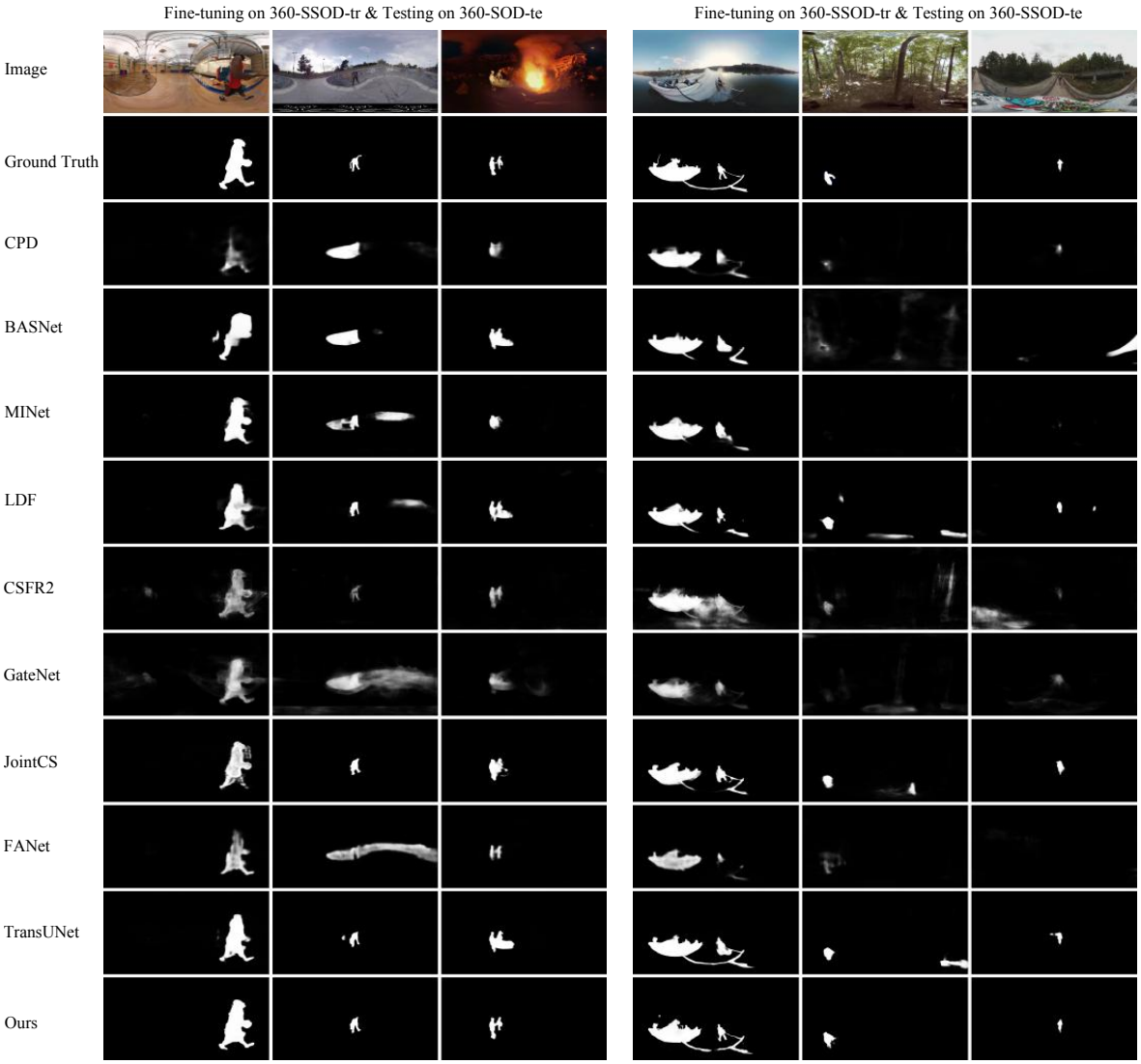


Fig. 5.6 More visualization results (part 2/2) of our CSMA-Net(**Ours**) and the state-of-the-art methods. Note that our CSMA-Net finely depicts small or distorted salient targets in equirectangular images.

5.2.5 Conclusion

In this section, we propose a new end-to-end deep learning method, *i.e.*, CSMA-Net, to conduct 360° salient object segmentation by combing the global-local priors based on multiple 360° projection techniques. To carefully explore the complementary information between equirectangular image and cube maps, we further design a channel-spatial mutual attention (CSMA) module which is able to effectively fuse the 360° multi-projection-based bottleneck features. Our CSMA-Net is able to outperform current 2D/360° state-of-the-art methods by a large margin, based on a new cross-validation-based 360° salient object segmentation benchmark.

5.3 Audio-visual salient object segmentation in 360° videos

5.3.1 Introduction

In the Section 3.4, we illustrate our newly proposed video dataset, *i.e.*, PAVS10K, which is the first 360° audio-visual salient object segmentation dataset reflecting various real-world scenes. Based on PAVS10K, in this chapter, we introduce a new baseline model, *i.e.*, the first conditional variational auto-encoder based audio-visual network (CAV-Net), which combines both audio and visual cues to conduct salient object segmentation in 360° immersive dynamic scenes. Generally, our CAV-Net consists of a spatial-temporal visual segmentation network, a convolutional audio-encoding network and audio-visual distribution estimation modules.

As a result, CAV-Net models both audio and visual cues for the segmentation of salient objects in 360° videos and outperforms all benchmark models. Besides, the conditional variational auto-encoder architecture of our CAV-Net enables aleatoric uncertainty estimation upon PAVS10K. Sufficient ablation studies and qualitative uncertainty estimation results indicate the effectiveness and explainability of our method. Besides, we also illustrate several findings based on extensive qualitative and quantitative experimental results, from the aspects of audio-visual modeling and uncertainty-aware object segmentation.

5.3.2 Methodologies

In this sub-section, we introduce a new **c**onditional variational auto-encoder based **a**udio**v**isual 360° salient object segmentation baseline model, *i.e.*, **CAV-Net**, from the aspects of its motivation, formulation, architecture and implementation details.

Motivation.

Audio-visual modeling. To the best of our knowledge, so far there is no released panoramic audio-visual salient object segmentation or video object segmentation method and the common issue existed among current state-of-the-art methods is the ignorance of audio cues. Since the salient objects in our PAVS10K are defined based on both audio and visual cues, a new baseline model which combines both audio and visual information, seeking to achieve better performance is worth attempting.

Unique aleatoric uncertainty estimation. As shown in Fig. 5.7, an interesting finding is that, neither sounding objects nor visual-only salient objects are necessarily regarded as salient targets from a perspective of audio-visual-based saliency judgments. In other words, the audio-visual saliency can not be regarded as a simple adding-up of the visual-only saliency and auditory stimuli. The other finding is that our PAVS10K reflects a realistic phenomenon where subjects tend to show different sensitivities towards similar audio-visual information, and thus making very different choices when deciding to which target to pay attention. The two findings indicate that subjects' personal preference introduce unavoidable ambiguities to audio-visual saliency judgments and thus bringing "unique aleatoric uncertainty" to the ground truth of audio-visual salient object segmentation dataset, especially when it comes to 360° panoramic scenes where multiple foreground objects and wide background context are included.

To estimate the "unique aleatoric uncertainty", we further add distribution estimation modules

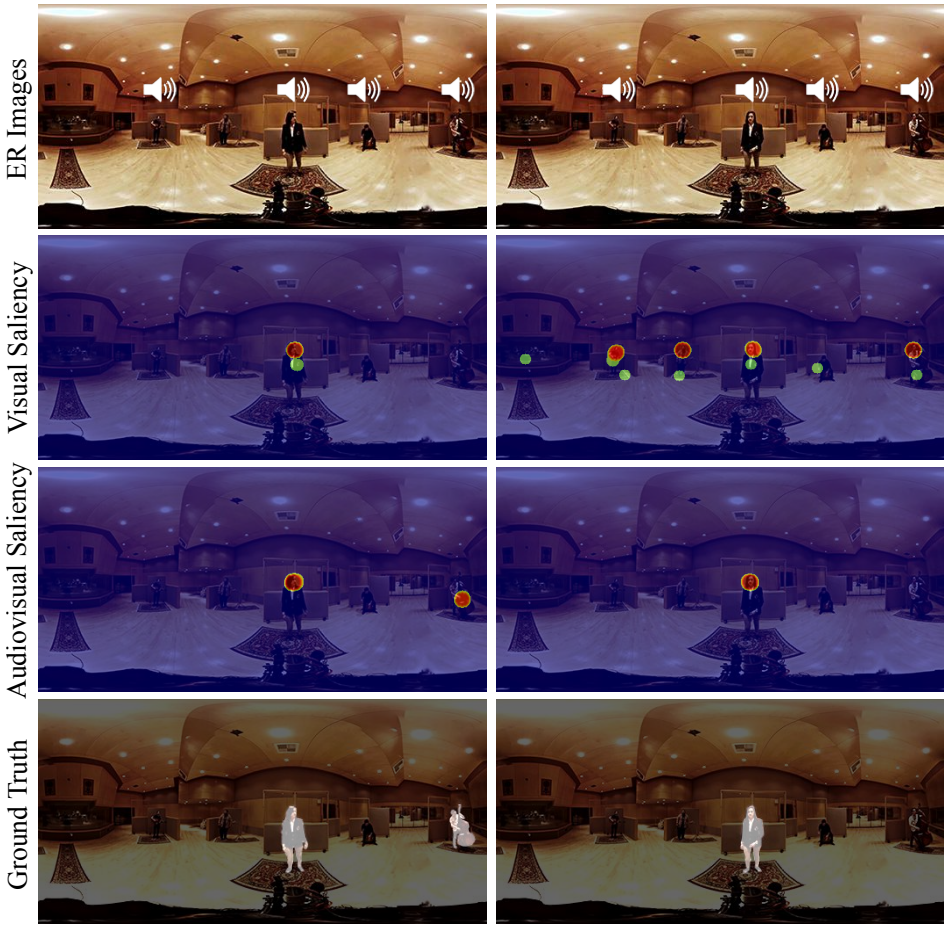


Fig. 5.7 An example ("Music"-"studio") that illustrates the divergence of saliency judgments based on visual-only and audio-visual cues. The annotations in PAVS10K are based on audio-visual saliency, which may vary in different frames with similar audio-visual cues.

and thus adapting our salient object segmentation deterministic model to a conditional variational auto-encoder [366], which is able to compute the distribution of model prediction. Therefore, our new **C**VAE-based **a**udio**v**isual panoramic salient object segmentation **net**work (**CAV-Net**) is capable of not only modeling dynamic audio-visual cues, but also estimating the uncertainty brought by the subjective stochasticity towards 360° audio-visual data.

Uniqueness of CAV-Net. First, our CAV-Net considers both audio and visual cues to segment the salient objects in 360 videos, thus distinguishing itself from current image/video-based salient object segmentation methods which consider visual-only static/dynamic cues. Besides, our CAV-Net is formulated as an end-to-end conditional variational auto-encoder which is able to conduct salient object segmentation and aleatoric uncertainty estimation simultaneously.

Audio-visual conditional variational auto-Encoder.

As shown in Fig. 5.8, we first design an end-to-end encoder-decoder framework learning audio-visual input data X (consisting of both visual sequence X^V and associated audio record X^A) via parameter set θ. Specifically, θ is the ensemble of parameters ($\{\theta^S, \theta^D, \theta^A\}$) modeling **s**tatic-/**d**ynamic-visual cues and corresponding **a**udio cues.

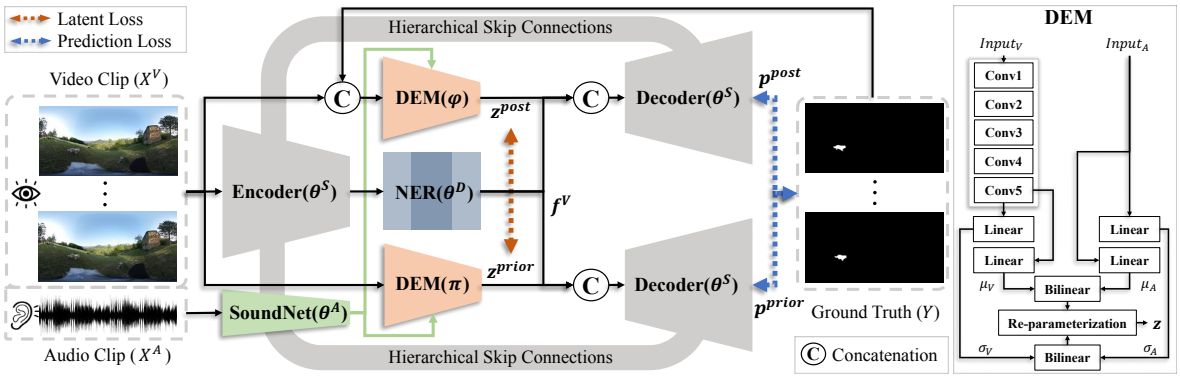


Fig. 5.8 The architecture of our **CAV-Net**, which consists of the proposed audio-visual posterior and prior distribution estimation modules (DEMs), a SoundNet-based audio encoder [70], a hybrid-ViT [345]-based visual encoder, fully convolutional decoders and one non-locally enhanced temporal module (NER) cited from RCRNet [178]. θ^S, θ^D, θ^A, π and φ are model parameter sets modeling static visual and dynamic visual cues, audio cues, prior and posterior distributions, respectively.

To further model the "unique aleatoric uncertainty" within the audio-visual salient object segmentation dataset, we add extra inference modules to adapt the original audio-visual deterministic model (θ) to a new conditional variational auto-encoder, namely CAV-Net, which enables the modeling of distribution of model prediction, *i.e.*, $P(Y|X;\theta)$. Specifically, following the common implementation of conditional variational auto-encoder [367], we apply two convolutional encoders for the inference of latent variable z which is capable of generating stochastic predictions and thus enabling the estimation of uncertainty of model prediction. It is worth mentioning that, this model prediction based uncertainty reflects intrinsic noises from training data [368], thus representing the aleatoric uncertainty within salient object segmentation dataset. The two inference modules are thus named as the prior ($P_\pi(z|X)$) and posterior ($P_\varphi(z|X,Y)$) distribution estimation modules, where π and φ indicate the parameter sets of the prior-/posterior-based encoders, respectively.

To train the proposed audio-visual conditional variational auto-encoder framework (CAV-Net), we use the posterior distribution estimation module to approximate the true posterior distribution of latent variable z. To this end, we further apply the Stochastic Gradient Variational Bayes [369] framework to estimate the parameter sets of our CAV-Net, by maximizing the evidence lower bound:

$$L(\theta,\varphi,\pi;X) = \mathbb{E}_{z\sim P_\varphi(z|X,Y)}[\log(P_\theta(Y|X,z))] - D_{KL}(P_\varphi(z|X,Y)||P_\pi(z|X)), \tag{5.7}$$

where $D_{KL}(P_\varphi(z|X,Y)||P_\pi(z|X))$ denotes the Kullback–Leibler divergence loss, regarded as a regularization closing the gap between the prior $P_\pi(z|X)$ and the posterior $P_\varphi(z|X,Y)$. With the conditional variational auto-encoder framework, the aleatoric uncertainty (σ^2) can then be computed as the mean entropy of multiple model predictions:

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T}\mathbb{H}[P(Y|X;\theta,\varphi)], \tag{5.8}$$

where T is the number of iterations of sampling. $\mathbb{H}[\cdot]$ denotes entropy operation.

Table 5.3 Components (Comp.) of each of the modules of our CAV-Net. ‡ denotes component does not exist.

Comp.	Modules of *CAV-Net*				
	π	φ	θ^A	θ^D	θ^S
#Conv2D	5	5	7	7	98
#Conv3D	‡	‡	‡	8	‡
#Identity	‡	‡	‡	‡	56
#ReLU	‡	‡	7	‡	69
#GELU	‡	‡	‡	‡	14
#Linear	4	4	‡	‡	51
#Bilinear	2	2	‡	‡	‡

Network architectures.

We further illustrate the structural details of our CAV-Net, consisting of a static visual encoder-decoder (θ^S), a dynamic visual module (θ^D), an audio encoder (θ^A) and audio-visual prior-/posterior-distribution estimation modules (π, φ). The detailed statistics of each module of our CAV-Net are shown in Table 5.3.

Visual encoder-decoder. As shown in Fig. 5.8, our visual encoder-decoder framework consists of two parts, *i.e.*, θ^S and θ^D, thus modeling the static and dynamic visual cues respectively. Specifically, we resort to the strong encoding ability of vision transformers [143] and thus using the hybrid-ViT based encoder [345] to extract the abundant visual information of 360° images. The static visual bottleneck features f^S are then fed into a non-locally enhanced temporal module (NER) [178] to seek inter-frame connections and thus aiding the dynamic visual cues modeling. As for the decoder, we simply follow the state-of-the-art video-based salient object segmentation method, RCRNet [178], and use its U-Net like skip connections to gradually refine the final visual bottleneck features f^V with an aid of hierarchical features $\{f_i^S\}_{i=1}^3$ gained from the first three vision transformer layers [345] of the encoder.

Audio encoder. To encode the mono sound X^A extracted from the given video clip, we apply the first seven 1-D convolutional layers of the state-of-the-art network, *i.e.*, SoundNet [70]. The output audio feature vector f^A is then used to synchronously model the prior and posterior distributions of model predictions (Fig. 5.8).

Audio-visual distribution estimation module. Following [367], our prior and posterior distribution estimation modules both use five convolutional layers to extract the latent features from visual input ($Input_V$ in Fig. 5.8). Importantly, to fit the task of audio-visual salient object segmentation, our distribution estimation modules take advantage of not only visual cues but also audio feature vector f^A ($Input_A$) to estimate the distributions. Specifically, the audio-visual posterior distribution estimation module takes the concatenation of video clip X^V and ground truth Y as visual input and thus modeling the visual latent space distribution with mean and standard deviation pair $\{\mu_V^{post}, \sigma_V^{post}\}$. Similarly, the audio counterpart with mean and standard deviation pair $\{\mu_A^{post}, \sigma_A^{post}\}$ can be easily gained with f^A as the input. To effectively use the audio-visual latent features, we are inspired by STAViS [59] which uses bilinear operations to combine the multi-modal features, thus adding an extra bilinear layer (Fig.

5.8) to the distribution estimation module to gain an audio-visual latent space distribution with mean and standard deviation pair $\{\mu^{post}, \sigma^{post}\}$. Following the same procedure, the prior distribution with mean and standard deviation pair $\{\mu^{prior}, \sigma^{prior}\}$ is acquired by using audio-visual prior distribution estimation module. The only difference is that the prior distribution estimation module dose not need ground truth Y as the input (Fig. 5.8).

The latent variables of both distributions are then obtained with the re-parameterization trick as:

$$z^{post} = \mu^{post} + \sigma^{post} \odot \varepsilon, \;\; z^{prior} = \mu^{prior} + \sigma^{prior} \odot \varepsilon, \tag{5.9}$$

where $\odot$ is the dot product operation, and $\varepsilon \sim \mathcal{N}(0,1)$. The z^{prior} and z^{post} are then tiled to a 3-D feature map with the same spatial size of bottleneck features f^V to enable the feature concatenation.

Implementation details.

Loss function. Following the conditional variational auto-encoder formulation (Eq. 5.7), the total loss L of our CAV-Net (Fig. 5.8) is defined as the sum of a prediction loss L^P and a latent loss L^L. The L^P is the widely used structure loss [122] consisting of a weighted binary cross entropy loss L^P_{wbce} and a weighted IoU loss L^P_{wiou}, while the L^L denotes the Kullback–Leibler divergence loss (the D_{KL} in Eq. 5.7). Thus, the total loss of CAV-Net is formulated as:

$$L = L^P(p^{prior}, Y) + L^P(p^{post}, Y) + L^L(z^{prior} || z^{post}), \tag{5.10}$$

where p^{prior}/p^{post} are model predictions sampled from prior/posterior distributions respectively. Y is ground truth.

Algorithms. The training and testing procedures of our CAV-Net are shown in Algorithm. 1 and Algorithm. 2 respectively, to facilitate the re-implementation of our method.

Algorithm 1 Training CAV-Net.

Input: (1) Training video clips $\{\mathbf{X}^{\mathbf{V}}_i\}^n_i$, associated audio clips $\{\mathbf{X}^{\mathbf{A}}_i\}^n_i$ and ground truth $\{\mathbf{Y}_i\}^n_i$; (2) Maximum of learning iterations M.

Output: Parameters θ^S, θ^D and θ^A for the static visual, dynamic visual and audio feature extraction modules respectively, π and φ for the audio-visual prior and posterior distribution estimation modules respectively (please refer to Fig. 5.8 for structural details).

1: Initialize θ^S, θ^D, θ^A, π and φ
2: **for** $t \leftarrow 1$ to M **do**
3: Sample video clip, corresponding audio clip and ground truth, $\{\mathbf{X}^{\mathbf{V}}_i, \mathbf{X}^{\mathbf{A}}_i, \mathbf{Y}_i\}^b_i$ where b is the batch size.
4: For each $\{\mathbf{X}^{\mathbf{V}}_i, \mathbf{X}^{\mathbf{A}}_i\}$, sample the prior $z^{prior}_i \sim P_\pi(z|\mathbf{X}^{\mathbf{V}}_i, \mathbf{X}^{\mathbf{A}}_i)$ for T times, compute the prior-based mean prediction p^{prior}_i.
5: For each $\{\mathbf{X}^{\mathbf{V}}_i, \mathbf{X}^{\mathbf{A}}_i, \mathbf{Y}_i\}$, sample the posterior $z^{post}_i \sim P_\varphi(z|\mathbf{X}^{\mathbf{V}}_i, \mathbf{X}^{\mathbf{A}}_i, \mathbf{Y}_i)$ for T times, compute the posterior-based mean prediction p^{post}_i.
6: Synchronously update all parameters (θ^S, θ^D, θ^A, π and φ) via the sum of prediction loss and latent loss (Eq. 5.10).
7: **end for**

Hyper-Parameters. CAV-Net is implemented with PyTorch, optimized with Adam algorithm [3].

Algorithm 2 Testing CAV-Net.

Input: Testing video clips $\{\mathbf{X}_i^{\mathbf{V}}\}_i^n$ and audio clips $\{\mathbf{X}_i^{\mathbf{A}}\}_i^n$.
Output: Prediction p_i and uncertainty σ_i^2.

1: **for** $i \leftarrow 1$ to n **do**
2: For each $\{\mathbf{X}_i^{\mathbf{V}}, \mathbf{X}_i^{\mathbf{A}}\}$, sample $z_i^{prior} \sim P_\pi(z|\mathbf{X}_i^{\mathbf{V}}, \mathbf{X}_i^{\mathbf{A}})$ for T times, compute the mean prediction p_i and the mean entropy of multiple predictions, *i.e.*, σ_i^2.
3: **end for**

Following the common settings of salient object segmentation methods, the static visual cues modeling parts (θ^S) of our CAV-Net are initialized with DUTS-tr [20] pre-training, while parameter sets θ^D, π and φ are randomly initialized. For fair comparison, we resize the input equirectangular video frames to 416×832 (smaller than 512×1,024 applied in state-of-the-art 360° salient object segmentation method, FANet [334]), without using multi-scale or any other data augmentation tricks. During training, the batch size is set as 1, default video clip length is 3, learning rate initialized as 2.5×10^{-6}. It takes about 9.5 hours to train the whole framework with the training set of our PAVS10K, based on a PC consisting of Intel® Xeon® W-2255 CPU@3.70GHz and one Quadro RTX-6000 GPU.

5.3.3 Experiments

The detailed experimental settings are illustrated in Section 3.4.3. Following the same settings of our proposed PAVS10K benchmark, we conduct thorough quantitative and qualitative experiments to varify the effectiveness and superiority of the proposed new baseline model, *i.e.*, CAV-Net.

Performance comparison.

Table 5.4 Performance comparison of our panoramic audio-visual network, *i.e.*, CAV-Net and 12 state-of-the-art salient object segmentation/video object segmentation methods without training on PAVS10K. I. = image-based salient object segmentation. V. = video-based salient object segmentation or video object segmentation. Best result of each column is **bolded**.

Type	Year	Methods	Miscellanea (Test1)				Music (Test2)				Speaking (Test3)				PAVS10K-Test			
			$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$
I.	CVPR'19	CPD-R [139]	.261	.623	.604	.084	.151	.506	.483	.135	.190	.526	.488	.162	.195	.545	.515	.137
	ICCV'19	SCRN [297]	.271	.625	.606	.087	.206	.598	.594	.051	.218	.559	.518	.130	.226	.584	.558	.101
	AAAI'20	F3Net [122]	.236	.609	.573	.082	.152	.509	.524	.150	.215	.567	.505	.105	.204	.563	.526	.110
	CVPR'20	MINet [126]	.225	.606	.573	.093	.152	.542	.531	.073	.180	.523	.469	.151	.183	.548	.509	.118
	CVPR'20	LDF [128]	.268	.622	.606	.083	.204	.550	.557	.087	.227	.546	.503	.137	.230	.566	.541	.112
	ECCV'20	CSFR2 [129]	.305	.650	.624	.075	.139	.510	.471	.129	.189	.545	.511	.128	.202	.562	.529	.116
	ECCV'20	GateNet [132]	.243	.637	.588	.069	.206	.594	.611	.035	.206	.569	.554	.090	.214	.591	.576	.072
V.	CVPR'19	COSNet [180]	.280	.602	.581	.110	.181	.571	.614	.034	.232	.595	.587	.065	.230	.591	.592	.068
	ICCV'19	RCRNet [178]	.307	.666	.644	.062	.312	.630	.683	.040	.238	.591	.542	.065	.271	.619	.601	.058
	AAAI'20	PCSA [175]	.197	.629	.632	.042	.104	.543	.548	.030	.157	.565	.594	.037	.153	.575	.592	.036
	BMVC'20	3DC-Seg [332]	.231	.544	.523	.143	.268	.578	.663	.059	.193	.540	.584	.088	.220	.550	.588	.094
	CVPR'21	RTNet [333]	.331	.632	.602	.110	.436	.668	.769	**.016**	.338	.637	.639	.045	.361	.643	.661	.054
PAV.	‡	**CAV-Net**	**.410**	**.704**	**.705**	**.040**	**.466**	**.675**	**.801**	.018	**.391**	**.659**	**.742**	**.024**	**.414**	**.674**	**.747**	**.027**

General Performance. To conduct thorough benchmark studies, we compare our new baseline model CAV-Net with the competing salient object segmentation/video object segmentation models based on two settings, *i.e.*, with and without PAVS10K training. Specifically, we first download the officially

Table 5.5 Performance comparison between our CAV-Net and 13 state-of-the-art methods (including seven image-based salient object segmentation (I.), five video-based salient object segmentation or video object segmentation (V.) and one 360° panoramic image-based salient object segmentation (PI.) methods) with PAVS10K training.

Type	Year	Methods	Miscellanea (Test1)				Music (Test2)				Speaking (Test3)				PAVS10K-Test			
			$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$\mathscr{M}\downarrow$
I.	CVPR'19	CPD-R [139]	.248	.654	.645	.035	.272	.608	.632	.018	.228	.588	.657	.026	.243	.609	.648	.026
	ICCV'19	SCRN [297]	.250	.665	.615	.046	.341	.683	.664	.023	.276	.636	.642	.034	.286	.655	.641	.034
	AAAI'20	F3Net [122]	.257	.655	.629	.040	.358	.662	.749	.021	.308	.626	.692	.027	.310	.642	.691	.029
	CVPR'20	MINet [126]	.238	.650	.625	.050	.380	.670	.716	.020	.261	.590	.635	.053	.286	.624	.652	.044
	CVPR'20	LDF [128]	.280	.663	.626	.044	.389	.671	.753	.023	.309	.625	.711	.037	.322	.645	.701	.035
	ECCV'20	CSFR2 [129]	.238	.652	.642	.033	.347	.665	.693	.018	.285	.636	.700	.026	.290	.646	.684	.026
	ECCV'20	GateNet [132]	.285	.677	.651	.044	.290	.673	.616	.018	.260	.633	.638	.034	.273	.653	.636	.033
V.	CVPR'19	COSNet [180]	.147	.610	.553	.031	.220	.577	.541	**.016**	.176	.572	.570	**.023**	.181	.582	.559	**.023**
	ICCV'19	RCRNet [178]	.272	.661	.640	.034	.403	**.695**	.738	.019	.282	.632	.687	.030	.310	.654	.688	.029
	AAAI'20	PCSA [175]	.123	.604	.574	.034	.310	.657	.645	.022	.150	.571	.534	.026	.184	.600	.570	.027
	BMVC'20	3DC-Seg [332]	.300	.668	.618	.062	.326	.635	.632	.046	.289	.629	.592	.056	.300	.640	.608	.055
	CVPR'21	RTNet [333]	.240	.622	.634	.038	.365	.638	.766	.020	.194	.555	.668	.028	.247	.591	.683	.029
PI.	SPL'20	FANet [334]	.164	.610	.529	**.030**	.380	.646	.758	.018	.207	.566	.663	.027	.241	.596	.654	.025
PAV.	‡	**CAV-Net**	**.410**	**.704**	**.705**	.040	**.466**	.675	**.801**	.018	**.391**	**.659**	**.742**	.024	**.414**	**.674**	**.747**	.027

released best models of each of the state-of-the-art salient object segmentation/video object segmentation methods and directly test these models on the testing set of our dataset (PAVS10K-Test). As a result, our CAV-Net outperforms all 12 state-of-the-art baselines (with publicly available pre-trained models) based on all four metrics (Table 5.4). Further, we re-train the 13 competing methods with the training set of our PAVS10K and test them on PAVS10K-Test (Table 5.5). Finally, our CAV-Net still outperforms all baselines in terms of F-/S-/E-measure.

Super-Class-wise Performance. The models' performance based on each of the super-classes of our PAVS10K are shown in Table 5.4 and Table 5.5. As a result, our audio-visual method, CAV-Net, is able to outperform the 13 benchmark models based on both settings. Note that following quantitative results are all based on PAVS10K training.

Attribute-wise Performance. As shown in Table 5.6, our CAV-Net outperforms all 13 competing baselines on all seven PAVS10K's attributes-based testing sets, in terms of F-/S-/E-measure. The superior performance of CAV-Net upon all attribute-based testing sets indicate that our new baseline model successfully considers all spotted challenges for salient object segmentation modeling.

Sub-Class-wise Performance. As shown in Table 5.7, our CAV-Net ranks first on 8 and 15 testing sequences in terms of S-measure and E-measure respectively, thus being the most robust model when compared to all competing baselines.

Qualitative Results. As shown in Fig. 5.9, Fig. 5.10 and Fig. 5.11, our CAV-Net is able to correctly detect the audio-visual salient objects labeled with multiple attributes. For instance, in Fig. 5.9, the train is finely depicted even though it is seriously distorted and blurred. In Fig. 5.10, the small person is accurately segmented. In "Spanish" (Fig. 5.11), the occluded and distorted people are correctly detected.

Table 5.6 Performance comparison of 13 competing models and our CAV-Net based on each of the attributes.

		I.							V.						PI.	PAV.
Attr.	Metrics	CPD-R [139]	SCRN [297]	F3Net [122]	MINet [126]	LDF [128]	CSFR2 [129]	GateNet [132]	COSNet [180]	RCRNet [178]	PCSA [175]	3DC-Seg [332]	RTNet [333]	FANet [334]	**CAV-Net Ours**
MO	$S_\alpha \uparrow$	.610	.657	.644	.624	.648	.649	.653	.588	.661	.607	.643	.595	.605	**.672**
	$F_\beta \uparrow$	.244	.288	.315	.288	.324	.292	.270	.187	.319	.193	.302	.251	.258	**.414**
	$E_\phi \uparrow$	.655	.649	.705	.665	.718	.694	.637	.571	.706	.580	.614	.703	.676	**.751**
	$\mathscr{M} \downarrow$	.027	.034	.030	.045	.033	.027	.034	**.024**	.029	.027	.054	.028	.025	.028
OC	$S_\alpha \uparrow$	.606	.655	.641	.619	.645	.645	.650	.577	.652	.600	.636	.586	.593	**.667**
	$F_\beta \uparrow$	.260	.294	.329	.298	.335	.301	.276	.191	.316	.202	.308	.259	.258	**.422**
	$E_\phi \uparrow$	.649	.639	.696	.651	.709	.682	.622	.554	.691	.570	.607	.694	.668	**.751**
	$\mathscr{M} \downarrow$	.023	.029	.026	.043	.028	.023	.030	**.020**	.025	.024	.045	.024	.022	.022
LR	$S_\alpha \uparrow$	.605	.649	.639	.618	.637	.644	.647	.585	.650	.609	.633	.590	.598	**.663**
	$F_\beta \uparrow$	.229	.271	.301	.272	.303	.277	.255	.176	.294	.189	.286	.234	.238	**.392**
	$E_\phi \uparrow$	.640	.636	.693	.642	.694	.683	.625	.565	.687	.586	.600	.688	.657	**.738**
	$\mathscr{M} \downarrow$	.025	.034	.028	.045	.037	.025	.033	**.022**	.029	.026	.057	.029	.025	.028
MB	$S_\alpha \uparrow$	.622	.651	.630	.620	.646	.638	.645	.582	.642	.586	.632	.595	.587	**.666**
	$F_\beta \uparrow$	.281	.304	.299	.298	.330	.297	.281	.212	.307	.197	.302	.271	.247	**.419**
	$E_\phi \uparrow$	.628	.630	.663	.637	.667	.668	.621	.563	.675	.563	.599	.676	.627	**.736**
	$\mathscr{M} \downarrow$	.021	.029	.027	.047	.029	.021	.030	**.019**	.024	.022	.044	.023	.020	.020
OV	$S_\alpha \uparrow$	.634	.661	.568	.633	.636	.636	.639	.582	.630	.599	.641	.573	.611	**.662**
	$F_\beta \uparrow$	.311	.318	.167	.314	.309	.295	.258	.207	.276	.193	.362	.210	.310	**.442**
	$E_\phi \uparrow$	.652	.638	.538	.691	.676	.697	.637	.633	.732	.536	.671	.703	.679	**.771**
	$\mathscr{M} \downarrow$	**.018**	.021	.029	.038	.039	.021	.025	.021	.029	.021	.039	.022	**.018**	.019
GD	$S_\alpha \uparrow$	.630	.662	.639	.633	.659	.646	.658	.588	.651	.578	.659	.587	.599	**.680**
	$F_\beta \uparrow$	.285	.309	.299	.294	.341	.304	.300	.189	.311	.156	.320	.247	.245	**.425**
	$E_\phi \uparrow$	.657	.653	.669	.676	.680	.674	.662	.564	.687	.538	.621	.666	.630	**.739**
	$\mathscr{M} \downarrow$	.037	.042	.040	.045	.043	.035	.042	**.032**	.037	.036	.062	.038	.034	.038
CS	$S_\alpha \uparrow$	.625	.680	.667	.654	.664	.670	.676	.592	.680	.621	.654	.602	.616	**.693**
	$F_\beta \uparrow$	.277	.320	.357	.335	.361	.330	.304	.197	.354	.217	.324	.269	.279	**.449**
	$E_\phi \uparrow$	.674	.664	.720	.691	.740	.696	.655	.550	.711	.590	.625	.711	.697	**.762**
	$\mathscr{M} \downarrow$	.029	.035	.031	.035	.034	.028	.033	**.026**	.030	.029	.058	.031	.028	.031

Table 5.7 $S_\alpha(\alpha{=}0.5)$ and E_ϕ performance comparison of 13 competing models and our CAV-Net based on each of the sequences. Sp. = Speaking. Mu. = Music. Mi. = Miscellanea. M. = Metrics.

		I.							V.					PI.	PAV.
Super-class/Sequence	M.	CPD-R [139]	SCRN [297]	F3Net [122]	MINet [126]	LDF [128]	CSFR2 [129]	GateNet [132]	COSNet [180]	RCRNet [178]	PCSA [175]	3DC-Seg [332]	RTNet [333]	FANet [334]	**CAV-Net Ours**
Sp./Debate	$S_\alpha\uparrow$	.547	.620	.605	.553	.566	.576	.628	.514	.559	.569	.601	.488	.557	**.669**
	$E_\phi\uparrow$	.600	.752	.818	.592	**.829**	.802	.768	.410	.809	.607	.640	.670	.702	.713
Sp./BadmintonConvo	$S_\alpha\uparrow$	**.712**	.669	.617	**.712**	.613	.647	.652	.613	.668	.551	.627	.556	.635	.654
	$E_\phi\uparrow$	.759	.672	.564	.824	.667	.671	.746	.612	**.782**	.449	.647	.682	.713	.762
Sp./Director	$S_\alpha\uparrow$	.679	.753	.701	.677	.756	**.772**	.726	.716	.755	.724	.726	.628	.672	.759
	$E_\phi\uparrow$	.735	.729	.852	.773	.849	.810	.681	.744	.774	.715	.690	.814	.768	**.859**
Sp./ChineseAd	$S_\alpha\uparrow$	.601	.645	.551	.477	.631	.630	.605	.553	.542	.567	.642	.534	.595	**.695**
	$E_\phi\uparrow$	.504	.544	.548	.410	.597	.656	.564	.662	.652	.490	.689	.754	.523	**.696**
Sp./Exhibition	$S_\alpha\uparrow$	.487	.469	.480	.469	.428	.492	.486	.487	.473	.510	.444	.460	.475	**.514**
	$E_\phi\uparrow$	.486	.365	.460	.350	.270	.508	.459	.514	.349	.510	.328	.309	.329	**.578**
Sp./PianoConvo	$S_\alpha\uparrow$	.577	.652	.579	.607	.639	.636	.586	.603	**.718**	.508	.685	.593	.632	.686
	$E_\phi\uparrow$	.774	.673	.745	.833	**.847**	.807	.693	.706	.803	.418	.641	.694	.804	.843
Sp./FilmingSite	$S_\alpha\uparrow$	.578	.633	.603	.610	.637	**.645**	.636	.578	.640	.631	.472	.551	.522	.631
	$E_\phi\uparrow$	.562	.627	.626	.636	.707	.654	.613	.540	.628	.652	.441	.720	.727	**.775**
Sp./Brothers	$S_\alpha\uparrow$	.673	.686	.638	.655	.652	.697	.685	.662	.664	.669	.663	.571	.623	**.715**
	$E_\phi\uparrow$	.702	.690	.726	.718	.719	.729	.643	.676	.743	.666	.628	.671	.695	**.783**
Sp./Rap	$S_\alpha\uparrow$	.498	.477	.521	.343	.507	.525	.463	.482	.506	.495	.578	**.633**	.532	.615
	$E_\phi\uparrow$	.530	.387	.548	.260	.484	.678	.400	.513	.590	.566	.557	.685	.733	**.756**
Sp./Spanish	$S_\alpha\uparrow$	.606	.765	.746	.679	**.793**	.713	.701	.724	.700	.541	.773	.580	.602	.766
	$E_\phi\uparrow$	.662	.817	.845	.737	**.876**	.808	.833	.795	.811	.495	.735	.577	.514	.874
Sp./Questions	$S_\alpha\uparrow$	.505	.640	**.740**	.563	.605	.691	.671	.576	.676	.595	.690	.530	.549	.702
	$E_\phi\uparrow$	.763	.609	**.870**	.576	.855	.700	.574	.569	.667	.540	.603	.747	.703	.724
Sp./PianoMono	$S_\alpha\uparrow$	.598	.555	.573	.572	.629	.522	**.637**	.506	.611	.503	**.637**	.570	.502	.512
	$E_\phi\uparrow$	.682	.736	.688	.739	.746	**.758**	.696	.500	.736	.399	.573	.693	.633	.748
Sp./Snowfield	$S_\alpha\uparrow$	.729	.811	.778	.800	.819	.779	**.823**	.601	.794	.584	.754	.704	.578	.819
	$E_\phi\uparrow$	.682	.783	.725	.769	.797	.721	.793	.490	.758	.523	.722	.777	.623	**.814**
Sp./Melodrama	$S_\alpha\uparrow$	.609	**.685**	.655	.673	.667	.664	.617	.467	.608	.605	.626	.472	.568	.613
	$E_\phi\uparrow$	.699	.744	.732	.773	**.784**	.717	.710	.296	.730	.523	.623	.624	.770	.657
Sp./Gymnasium	$S_\alpha\uparrow$	**.551**	.514	.492	.501	.501	.507	.537	.520	.520	.502	.501	.511	.505	.533
	$E_\phi\uparrow$	.584	.545	.461	.593	.469	.512	.487	.584	.518	.469	.420	.504	**.642**	.591
Mu./Studio	$S_\alpha\uparrow$	.741	.770	.753	**.788**	.758	.739	.724	.637	.778	.758	.665	.743	.760	.721
	$E_\phi\uparrow$	.745	.731	.832	.826	.847	.756	.601	.629	.800	.730	.677	.837	**.859**	.779
Mu./Church	$S_\alpha\uparrow$	.527	.589	.621	.566	.518	.624	.651	.562	.676	.627	.535	.546	.679	**.720**
	$E_\phi\uparrow$	.451	.575	.731	.576	.715	.601	.657	.487	.635	.579	.536	.687	.774	**.850**
Mu./Duet	$S_\alpha\uparrow$	.662	.704	.698	.653	.751	.648	.730	.553	.731	.540	.672	.577	.643	**.764**
	$E_\phi\uparrow$	.810	.705	.792	.821	.808	.693	.735	.542	.776	.508	.702	.769	.765	**.865**
Mu./Blues	$S_\alpha\uparrow$	.580	.742	**.776**	.722	.771	.734	.740	.595	.765	.747	.732	.730	.600	.639
	$E_\phi\uparrow$	.598	.688	.830	.698	.789	.766	.640	.473	.834	.716	.687	**.875**	.612	.768
Mu./Violins	$S_\alpha\uparrow$	.589	.668	.537	**.692**	.661	.631	.656	.578	.669	.670	.653	.621	.604	.642
	$E_\phi\uparrow$	.679	.685	.507	.805	.751	.754	.599	.627	.754	.655	.679	.679	.779	**.842**
Mu./SingingDancing	$S_\alpha\uparrow$	.506	**.601**	.582	.560	.561	.594	.568	.521	.569	.558	.566	.560	.557	.597
	$E_\phi\uparrow$	.500	.587	**.758**	.565	.618	.589	.547	.452	.637	.608	.532	.720	.705	.756
Mi./Dog	$S_\alpha\uparrow$	.497	.516	.571	.560	.569	.557	.562	.523	.562	.540	**.605**	.548	.520	.603
	$E_\phi\uparrow$	.470	.467	.503	.521	.434	.543	.525	.504	.558	.550	.530	.556	.335	**.565**
Mi./RacingCar	$S_\alpha\uparrow$	.770	.769	.763	.770	.772	.771	.791	.760	.772	.759	.749	.753	.762	**.809**
	$E_\phi\uparrow$	.760	.752	.726	.771	.757	.729	.788	.719	.733	.709	.715	.760	.708	**.811**
Mi./Train	$S_\alpha\uparrow$	.604	.616	.614	.607	.629	.594	.663	.501	.524	.515	.638	.527	.489	**.725**
	$E_\phi\uparrow$	.581	.553	.486	.493	.554	.558	.634	.351	.462	.418	.606	.423	.386	**.735**
Mi./Football	$S_\alpha\uparrow$	.653	.696	.618	.656	.668	.658	.676	.648	**.710**	.640	.604	.632	.556	.708
	$E_\phi\uparrow$	.634	.676	.755	.633	.770	.721	.663	.649	.732	.631	.637	.701	.477	**.811**
Mi./ParkingLot	$S_\alpha\uparrow$	.635	.627	.624	.564	.640	.562	.625	.548	.624	.501	**.688**	.598	.627	.656
	$E_\phi\uparrow$	.641	.551	.600	.597	.625	.602	.610	.482	.612	.501	.599	.639	.593	**.661**
Mi./Skiing	$S_\alpha\uparrow$	.697	.728	.689	.727	.632	**.757**	.695	.624	.745	.641	.647	.599	.590	.672
	$E_\phi\uparrow$	.705	.645	.669	.661	.517	.675	.605	.573	**.716**	.614	.554	.650	.500	.586

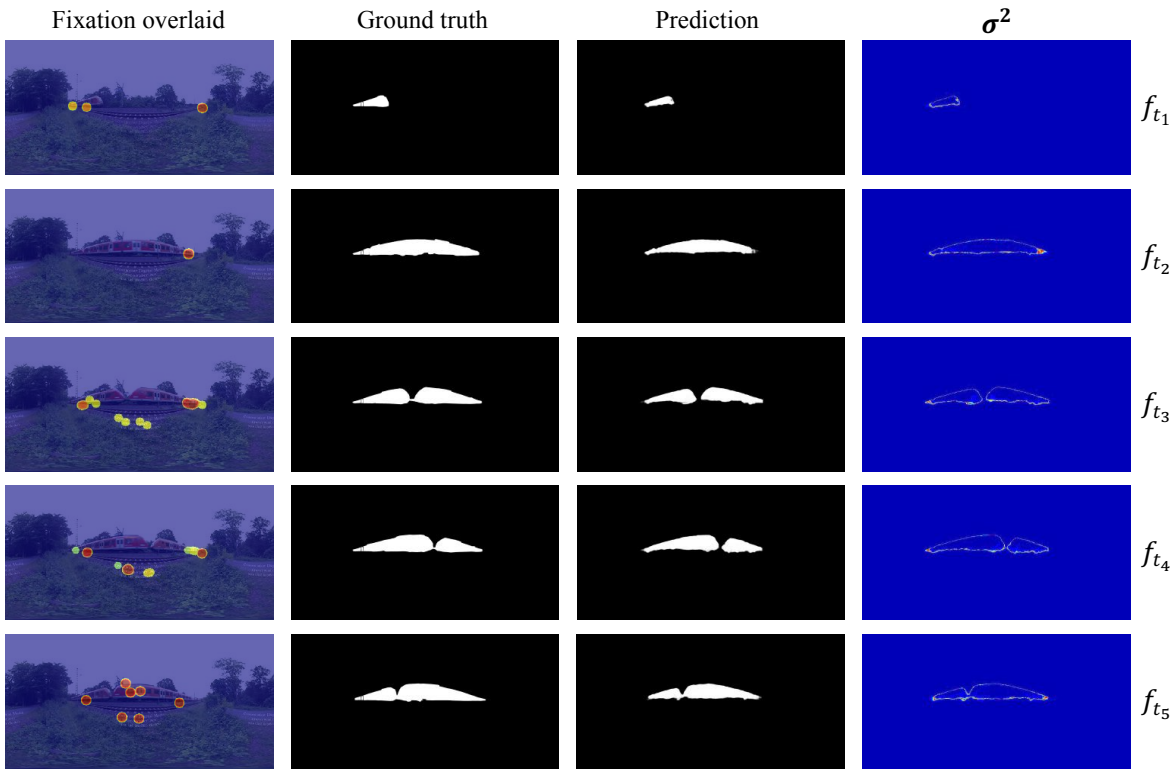


Fig. 5.9 Visualization results of our CAV-Net on sub-class "train". "σ^2" denotes uncertainty map corresponding to the prediction.

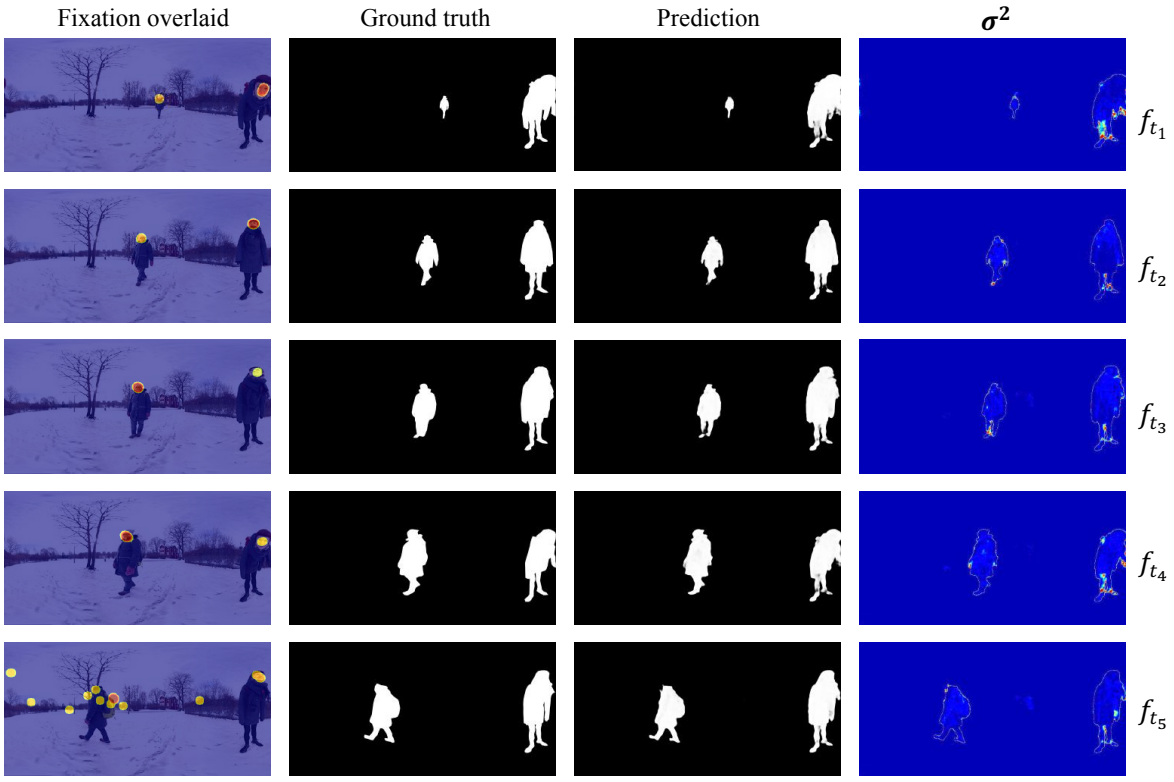


Fig. 5.10 Visualization results of our CAV-Net on sub-class "snowfield". "σ^2" denotes uncertainty map corresponding to the prediction.

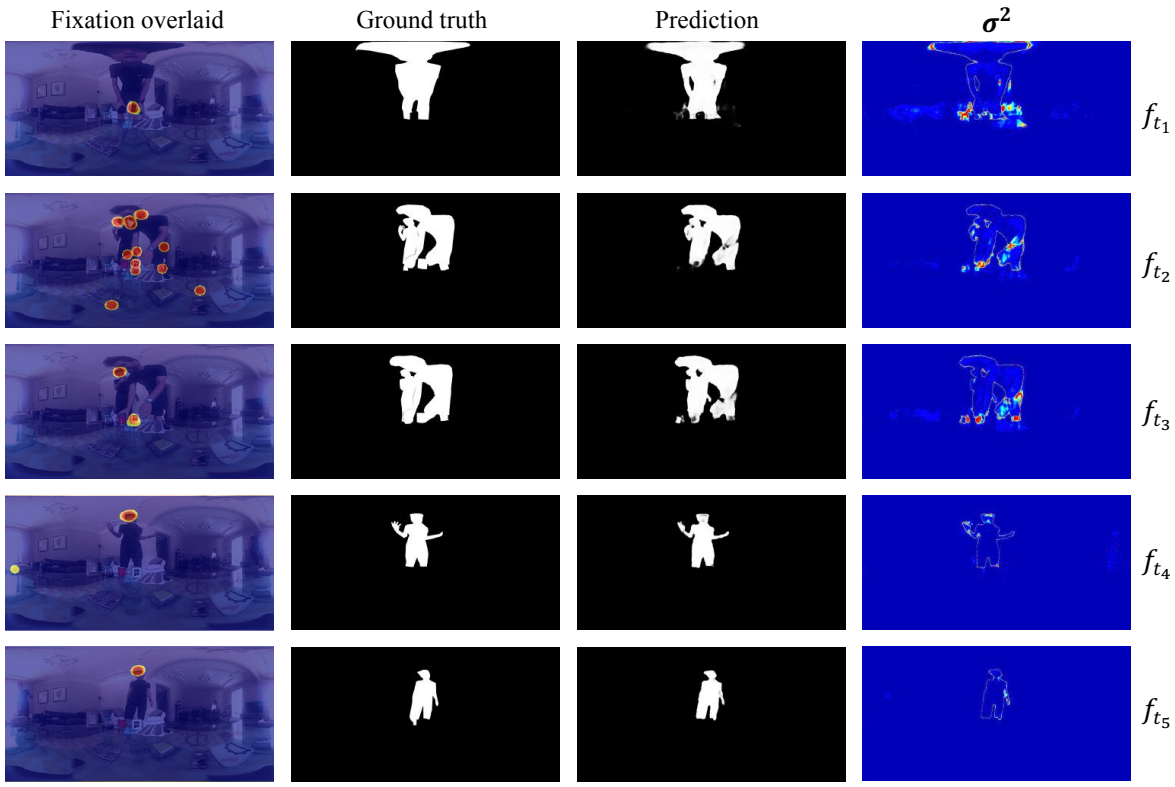


Fig. 5.11 Visualization results of our CAV-Net on sub-class "Spanish". "σ^2" denotes uncertainty map corresponding to the prediction.

Ablation studies.

To verify the effectiveness of the proposed audio-visual distribution estimation module in our CAV-Net, we conduct thorough ablation studies with multiple backbone strategies. Specifically, we directly use an off-the-shelf distribution estimation module [367] as the ablation version of our audio-visual distribution estimation module, thus gaining "Visual" and "Audio-visual" versions of our method. As shown in Table 5.8, our "Audio-visual" models are able to outperform "Visual" models based on each of the widely used backbones, *i.e.*, ResNet50 [142], Res2Net50 [362] and Hybrid-ViT [345]. As a result, the "Audio-visual" Hybrid-ViT version, which is exactly our CAV-Net, ranks first among all ablation models in terms of all four metrics.

Besides segmentation performance, we show two more statistics of each ablation model in Table 5.8, *i.e.*, the number of parameters (#Params) and frame-per-second (#FPS) during test-time (please note that all measurement are based on one Quadro RTX-6000 GPU with an input resolution of 416×832). As a result, the incremental computational burden of our CAV-Net mainly comes from the Transformer-based backbone (*e.g.*, 137.7 millions > 63.0 millions, 15 fps < 54 fps). As a comparison, the proposed audio-visual DEM only brings about 3 millions of extra model parameters and slight compromise to model inference speed (*i.e.*, 54 fps < 59 fps, 48 fps < 51 fps, 15 fps < 16 fps).

Table 5.8 Ablation studies of CAV-Net on our PAVS10K. S_α = S-measure (α=0.5), F_β = mean F-measure (β^2=0.3), E_ϕ = mean E-measure, $\mathscr{M}$ = mean absolute error.

Backbone	Modality	PAVS10K-Test					
		$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$\mathscr{M} \downarrow$	#Params	#FPS
ResNet50 [142]	Visual	.307	.628	.700	.028	59.8M	59
	Audio-visual	.325	.633	.698	**.027**	63.0M	54
Res2Net50 [362]	Visual	.283	.603	.659	.042	62.1M	51
	Audio-visual	.341	.630	.720	.036	65.2M	48
Hybrid-ViT [345]	Visual	.383	.656	.716	.028	134.6M	16
	Audio-visual	**.414**	**.674**	**.747**	**.027**	137.7M	15

5.3.4 Discussion

In this sub-section, we discuss about several new findings towards 360° audio-visual salient object segmentation, based on above extensive experimental results. Generally, we find that our new task, *i.e.*, 360° panoramic audio-visual salient object segmentation, is challenging for current salient object segmentation/video object segmentation state-of-the-art methods. Besides, we gain the conclusion that the modeling of both audio and visual cues help 360° audio-visual salient object segmentation. Finally, we obtain the evidence proving uncertainty-aware method helps exploring the intrinsic noises within audiovisual saliency detection dataset, thus inspiring new insights towards advanced and more reliable 360° audio-visual salient object segmentation modeling.

Audio-visual Modeling

Based on comprehensive benchmark studies (Table 5.4, Table 5.5 and Table 5.6), our new baseline CAV-Net proves its ability for 360° audio-visual salient object segmentation modeling. Specifically, our CAV-Net acquires better results on PAVS10K-Test and its three super-class-based testing sets (Table 5.4 and Table 5.5). Importantly, the consistent better results of CAV-Net on seven attribute testing sets (Table 5.6) shows that our new model gains significant improvement in terms of multiple aspects, including detecting accuracy (via F-measure F_β), the quality of object structure (via S-measure S_α) and the integrality of global context (E-measure E_ϕ). Besides superior segmentation performance, our CAV-Net also shows better computational efficiency than the current state-of-the-art 360° image-based salient object segmentation methods such as FANet [334] and SW360 [258]. It takes about 0.067s (which equals to 1/15s according to Table 5.8) for our CAV-Net while 0.26s/0.392s for FANet/SW360, to process one 360° image during test-time.

In addition, the multiple backbone based ablation studies regarding the proposed audio-visual distribution estimation module further verify the effectiveness and necessity of modeling both visual and audio cues when conducting 360° audio-visual salient object segmentation. The conclusion is consistent with human attention in real-world scenes where both visual and audio cues are regarded as inputs and share different weights for influencing human judgments towards visual saliency.

Uncertainty-aware Segmentation

As illustrated in Section 5.3.2, we observe an "unique aleatoric uncertainty", possibly introduced by two sources of subjective stochasticity, within the 360° audio-visual salient object segmentation dataset. Thus, being different to current mainstream salient object segmentation/video object

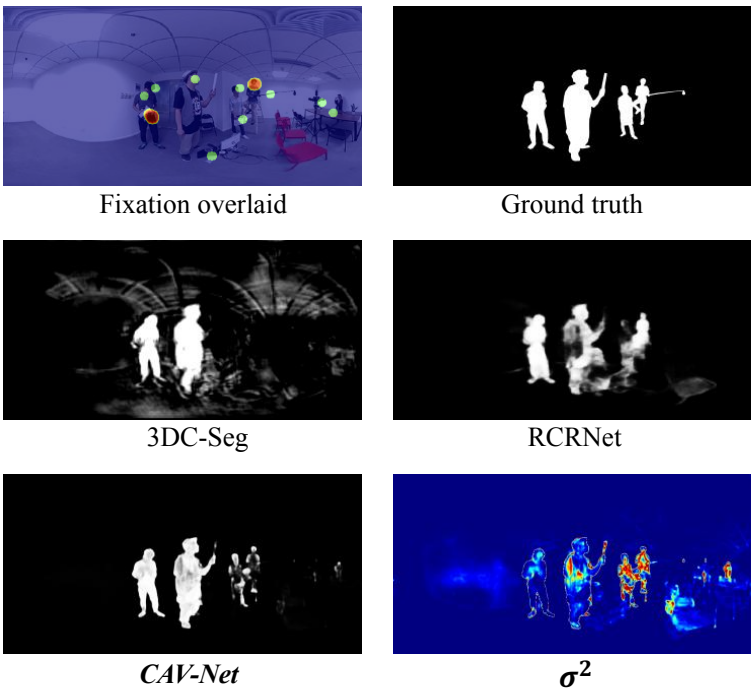


Fig. 5.12 An example that illustrates the interpretability of our CAV-Net via uncertainty estimation. σ^2 is the uncertainty map based on multiple sampling during model testing.

segmentation methods, we resort to uncertainty estimation methodology and propose to model this "unique aleatoric uncertainty" via conditional variational auto-encoder (Section 5.3.2). As a result, our CAV-Net successfully reflects this "unique aleatoric uncertainty" via predicting uncertainty map. Specifically, as shown in Fig. 5.12, our CAV-Net is able to explain the failure by highlighting the uncertain regions, while the competing video-based salient object segmentation and video object segmentation methods do not possess such an ability when making errors. Besides, the extreme uncertain regions estimated by our CAV-Net are exactly the regions with serious subjective stochasticity indicated by scattered fixations (Fig. 5.12).

Limitation and Future Work

Model performance. Although the proposed CAV-Net shows better overall results than all competing baselines, we are still limited by the challenges of PAVS10K (*e.g.*, our CAV-Net is unable to outperform all competing methods based on each of the sub-classes as shown in Table 5.7), thus failing to build a strong baseline model that outperforms current state-of-the-art methods by a large margin. Future works may explore deeper towards the low-level features (*e.g.*, contrast, sharpness, brightness) of 360° data of specific challenging sub-classes via image quality assessment techniques [370], to improve the generalization ability of 360° audio-visual salient object segmentation models.

Aleatoric uncertainty estimation. Although our CAV-Net successfully estimates the general aleatoric uncertainty [368] focusing on objects' boundaries (Fig. 5.9, Fig. 5.10 and Fig. 5.11), and the unique aleatoric uncertainty reflecting subjective stochasticity, we still have not explored deeply towards the details of subjective stochasticity that introduce such uncertainties. Besides, future works may consider to further improve the model performance by designing new uncertainty-aware frameworks.

Audio modality. Both the ground truth of our PAVS10K and CAV-Net are based on mono sound. Future works may consider to conduct panoramic audio-visual salient object segmentation via spatial audio or ambisonics.

5.3.5 Conclusion

In this section, we illustrate the details of our proposed new baseline model, namely CAV-Net, which is able to outperform all benchmark models and represent data uncertainty. Our CAV-Net verifies the superiority of modeling audio-visual cues for conducting 360° audio-visual salient object segmentation, and provides explanation for 360° audio-visual salient object segmentation modeling.

5.4 Conclusion

In this chapter, we first illustrated the details of our proposed CSMA-Net, which aims at accurately segmenting the salient objects in 360° images. To the best of our knowledge, we are the first to cascade channel-based and spatial-based mutual attentions, to effectively fuse and refine the high-level features extracted from global and local context of given 360° images. To further mimic the real-world scenes where human subjects tend to use both auditory and visual sensors to explore the surrounding world, we further introduced CAV-Net, which takes advantage of both audio and visual cues for salient object segmentation in 360° videos and reflects the aleatoric uncertainty within PAVS10K to some extent. To the best of our knowledge, our CAV-Net is the first publicly released audio-visual salient object segmentation model, also the first 360° video-based salient object segmentation method.

Chapter 6

Conclusion

6.1 Summary

To wrap this dissertation, we have successfully built new datasets and proposed new methodologies to address salient object segmentation in 360° panoramic images and videos, which we hope could serve as a starting point for object-level human visual attention modeling in immersive multi-media.

State-of-the-art methods for salient object segmentation. In Chapter 2, we have thoroughly summarized the state-of-the-art methods in the field of salient object segmentation academia. Based on the observation, we found that a lack of large-scale image/video 360° datasets seriously limited the development of 360° panoramic salient object segmentation, which is of great importance for mimicking real human visual attention in real-world. In addition, from a perspective of methodology, we concluded that current attention-based deep learning models have been widely applied in not only general computer vision tasks but also multiple types of salient object segmentation tasks. Besides empirical findings, modeling human visual attention with attention models is theoretically reasonable.

F-360iSOD&PAVS10K. In Chapter 3, we have detailed our works towards new dataset establishment in the field of 360° salient object segmentation. We first proposed a 360° image-based salient object segmentation dataset, namely F-360iSOD, which contains 1,165 pixel-wisely annotated salient instances belonging to 72 object/scene classes. Considering the real-world scenes where subjects depend on both audio and visual cues to locate and recognize the salient objects in 360° panoramic field-of-view, we further established so far the first 360° audio-visual dataset, *i.e.*, PAVS10K, which provides 19,904 manually labeled salient instances within 10,465 360° video frames.

Salient object segmentation in light field. Compared to 2D RGB salient object segmentation, light field salient object segmentation is relatively a new area to explore. In Chapter 4, following the mainstream of salient object segmentation researches where attention mechanisms have been widely applied to improve model performance, we proposed a synergistic attention network, *i.e.*, SA-Net, to segment salient objects by taking advantage of two light field modalities, *i.e.*, focal stacks and all-in-focus images. Besides, we have improved our SA-Net from both perspectives of computation burden and segmenting accuracy, via further proposing CMA-Net and SA-Net-V2, respectively.

Salient object segmentation in 360° images&videos. In Chapter 5, we have illustrated the details of our proposed CSMA-Net and CAV-Net, which address 360 image-based salient object segmenta-

tion and 360° audio-visual dynamic salient object segmentation, respectively. The key component of our CSMA-Net is a new mutual-attention module inspired by SA module in SA-Net. The CAV-Net is a new audio-visual conditional variational auto-encoder which is not only capable of segmenting salient objects but also estimating predictions' uncertainty.

6.2 Future works and perspectives

Though the objective of the thesis is successfully fulfilled, a gap between salient object segmentation and immersive vision still exists. In the following parts, we imagine future works towards immersive saliency detection from two perspectives, *i.e.*, an application of multi-modal visual cues and an involvement of multi-dimensional auditory information.

Modeling multi-modal visual cues for 360° saliency detection. This thesis explored both omnidirectional vision and light field, however, the real-world scenario is more similar to immersive light field vision [371] which combines both. A correct modeling of real-world lights via light field techniques may advance future datasets/models' development to a new level, thus enabling salient object segmentation to further fit real-world applications.

Specifically, future works may consider to establish new datasets collecting 360° images/videos with depth information, thus further mimicking the real-world scenes where subjects are able to observe and recognize salient objects with 6 degree-of-freedom (DoF) (an example illustrates 6 DoF shown in Fig. 6.1).

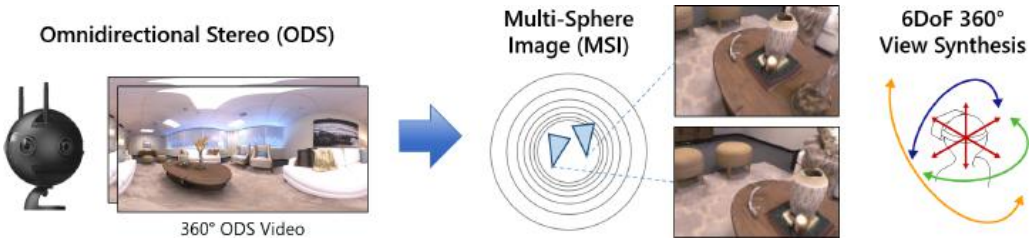


Fig. 6.1 An illustration of RGB-Depth 360° visual data. This figure is taken from [372].

Fig. 6.2 Ambisonics is able to be visualized as spatial-audio-based attention maps overlaid with 360° images. This figure is taken from [311].

Ambisonics for realistic audio-visual modeling in 360°. This thesis used mono sound to facilitate the establishment of both large-scale video dataset (details in Chapter 3) and new baseline model (details in Chapter 5). Future works may consider to establish datasets&models by taking advantage of ambisonics [311], which provides abundant auditory cues of multiple channels. The involvement of

realistic multi-channel audio information may improve the effectiveness of 360° audio-visual modeling via introducing realistic audio-based priors (Fig. 6.2).

Chapter 7

Appendix

7.1 A Predictive uncertainty estimation network for camouflaged object segmentation

7.1.1 Introduction

In this section, we briefly summarize our work towards the reverse task of salient object segmentation, *i.e.*, camouflaged object segmentation (Fig. 7.1). Being different to salient object segmentation which mimics the function of human attention mechanism towards visually discriminative targets, segmenting the targets concealed in natural scenes is always counter-intuitive and thus being difficult for human subjects.

Current state-of-the-art deep learning methods are able to learn the mapping between random inputting domain and target domain to solve challenging task such as camouflaged object segmentation, however the robustness and interpretability of the models are hardly guaranteed.

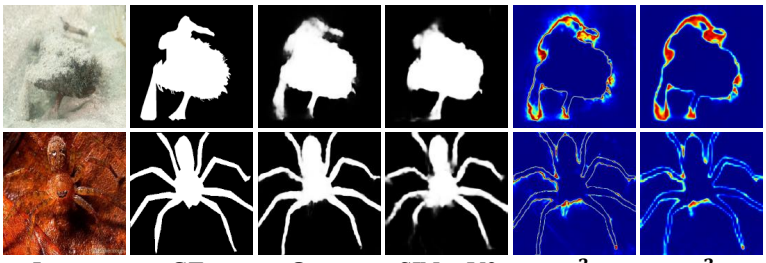


Fig. 7.1 An example illustrating camouflaged object segmentation and uncertainty estimation. "σ_e^2" is the sampling-based uncertainty of "Bayesian conditional variational auto-encoder". "σ_a^2" is the output of "predictive uncertainty approximation" module. SINet-V2 [373] is a state-of-the-art method.

Specifically, uncertainty is inherent in deep learning methods, especially those for camouflaged object segmentation aiming to finely segment the objects concealed in background. The strong "center bias" of the training dataset leads to models of poor generalization ability as the models learn to find camouflaged objects around image center, which we define as "model bias". Further, due to the similar appearance of camouflaged object and its surroundings, it is difficult to label the accurate

scope of the camouflaged object, especially along object boundaries, which we term as "data bias". To effectively model the two types of biases, we resort to uncertainty estimation and introduce predictive uncertainty estimation technique, which is the sum of model uncertainty and data uncertainty, to estimate the two types of biases simultaneously. Specifically, we present a predictive uncertainty estimation network (PUENet) that consists of a Bayesian conditional variational auto-encoder to achieve predictive uncertainty estimation, and a predictive uncertainty approximation module to avoid the expensive sampling process at test-time. Experimental results show that our PUENet achieves both highly accurate prediction, and reliable uncertainty estimation representing the biases within both model parameters and the datasets.

7.1.2 Methodology

In our PUENet, we design a Bayesian neural network to capture the distribution of model parameters. Further, we add extra inference model and adapt our network to a conditional variational auto-encoder [366], which is used to model the distribution of model prediction. In this way, our framework can estimate both model uncertainty (with the Bayesian neural network) and the data uncertainty (with the conditional variational auto-encoder). Further, we present predictive uncertainty approximation module to approximate the sampling-based predictive uncertainty of the proposed Bayesian conditional variational auto-encoder. The pipeline of our proposed PUENet is shown in Fig. 7.2

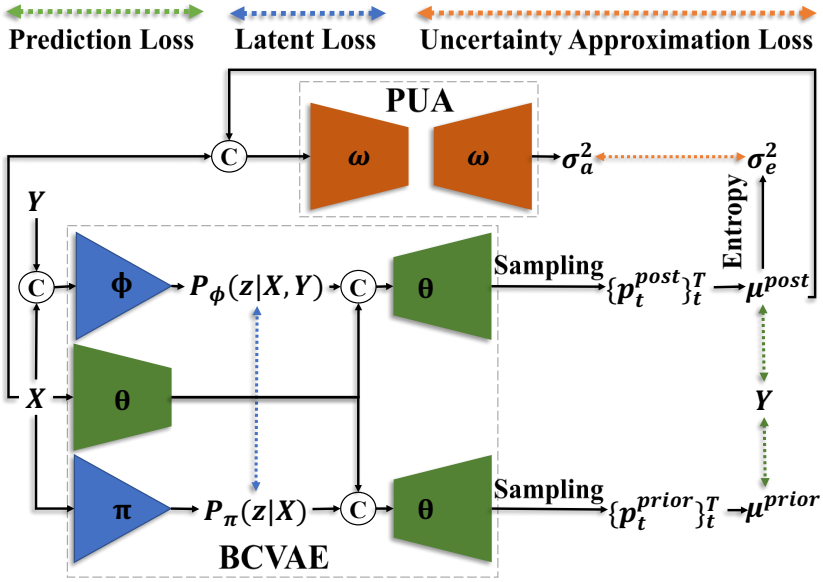


Fig. 7.2 The pipeline of our PUENet, which consists of a "Bayesian conditional variational auto-encoder" (BCVAE), and a "predictive uncertainty approximation" (PUA) module. "σ_e^2" and "σ_a^2" denote the sampling based uncertainty and approximated uncertainty, respectively.

7.1.3 Experiments

As a result, our PUENet is able to outperform the competing models by a large margin (Table 7.1), and gain uncertainty maps explaining the model predictions (Fig.).

Table 7.1 Performance comparison with state-of-the-art camouflaged object segmentation models on benchmark testing datasets. ↑ indicates the higher the score the better, and vice versa for ↓. The two best results of each column are in **red** and **blue**.

Method	Backbone	Year	CAMO [374]				CHAMELEON [375]				COD10K [373]				NC4K [376]			
			$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$\mathcal{M}\downarrow$
SINet [373]	ResNet50	CVPR'20	0.745	0.702	0.804	0.092	0.872	0.827	0.936	0.034	0.776	0.679	0.864	0.043	0.810	0.772	0.873	0.057
LSR [376]	ResNet50	CVPR'21	0.793	0.725	0.826	0.085	0.893	0.839	0.938	0.033	0.793	0.685	0.868	0.041	0.839	0.779	0.883	0.053
UJSC [364]	ResNet50	CVPR'21	0.803	0.759	0.853	0.076	0.894	0.848	**0.943**	0.030	0.817	0.726	0.892	0.035	0.842	0.806	0.898	0.047
MGL [377]	ResNet50	CVPR'21	0.775	0.726	0.812	0.088	0.893	0.834	0.918	0.030	0.814	0.711	0.852	0.035	0.833	0.782	0.867	0.052
PFNet [378]	ResNet50	CVPR'21	0.782	0.744	0.840	0.085	0.882	0.826	0.922	0.033	0.800	0.700	0.875	0.040	0.829	0.782	0.886	0.053
SINet-V2 [373]	Res2Net50	TPAMI'21	0.820	0.782	0.882	0.070	0.888	0.835	0.942	0.030	0.815	0.718	0.887	0.037	0.847	0.805	0.903	0.048
UJTR [379]	ResNet50	ICCV'21	0.785	0.686	0.859	0.086	0.888	0.796	0.918	0.031	0.818	0.667	0.850	0.035	0.839	0.786	0.873	0.052
PUENet (Ours)	ResNet50	2022	0.794	0.762	0.857	0.080	0.888	0.844	**0.943**	0.030	0.813	0.727	0.887	0.035	0.836	0.798	0.892	0.050
	Res2Net50	2022	**0.834**	**0.806**	**0.889**	**0.067**	**0.897**	**0.858**	0.940	**0.027**	**0.844**	**0.774**	**0.910**	**0.029**	**0.862**	**0.830**	**0.913**	**0.042**
	Hybrid-ViT	2022	**0.877**	**0.860**	**0.930**	**0.045**	**0.910**	**0.869**	**0.957**	**0.022**	**0.873**	**0.812**	**0.938**	**0.022**	**0.898**	**0.874**	**0.945**	**0.028**

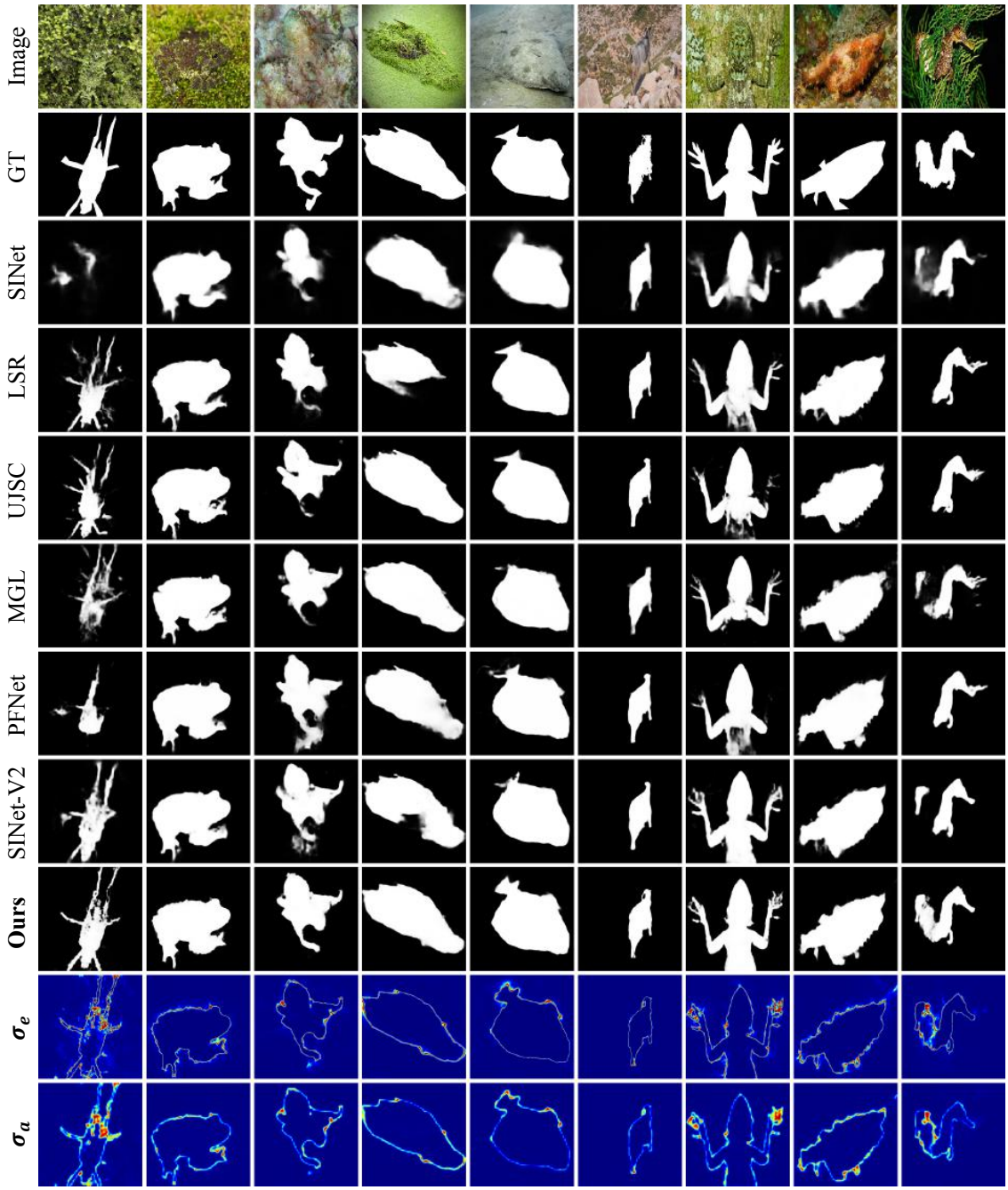


Fig. 7.3 Visual results of our method on CAMO [374].

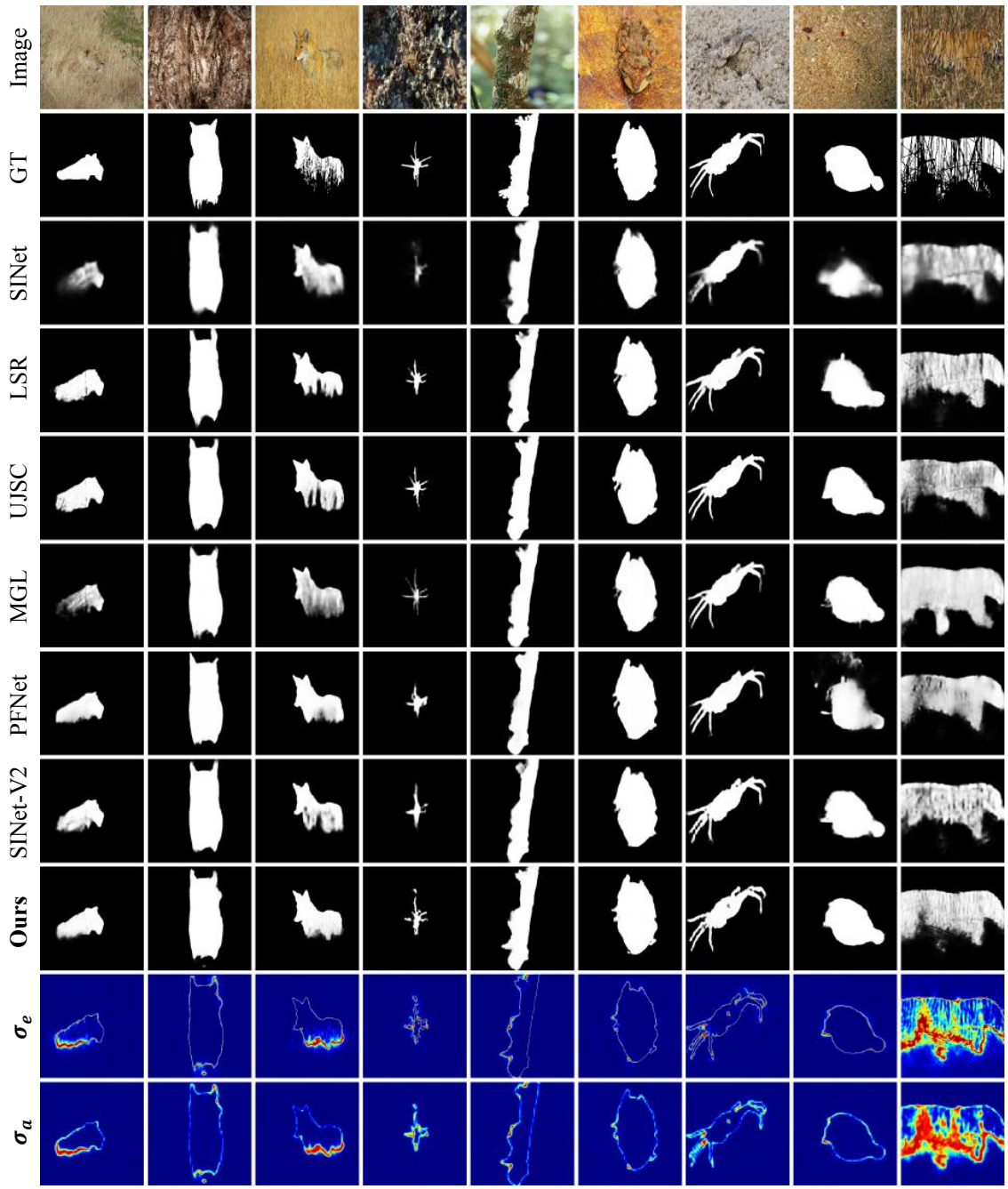


Fig. 7.4 Visual results of our method on CHAMELEMON [375].

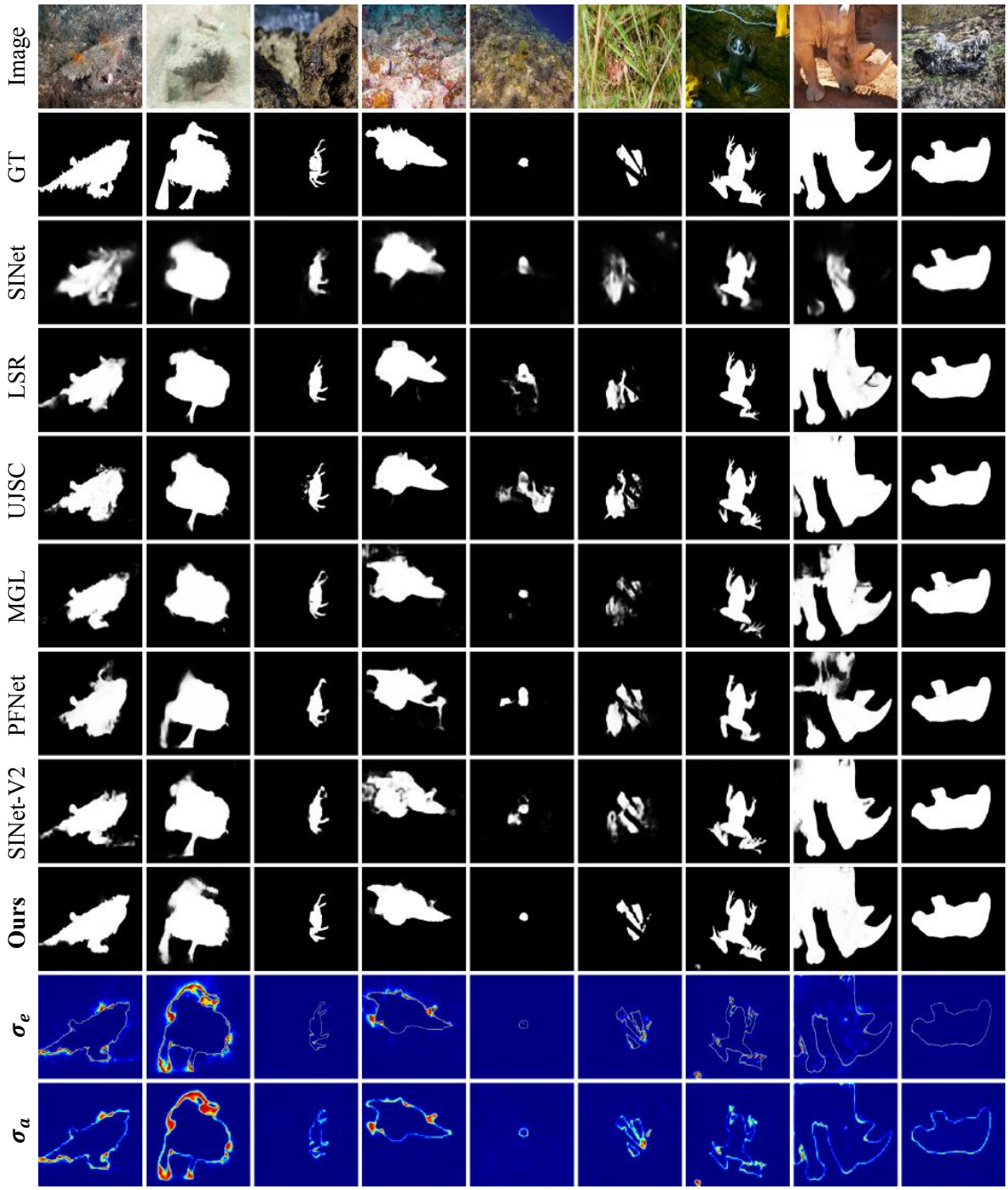


Fig. 7.5 Visual results of our method on COD10K [373].

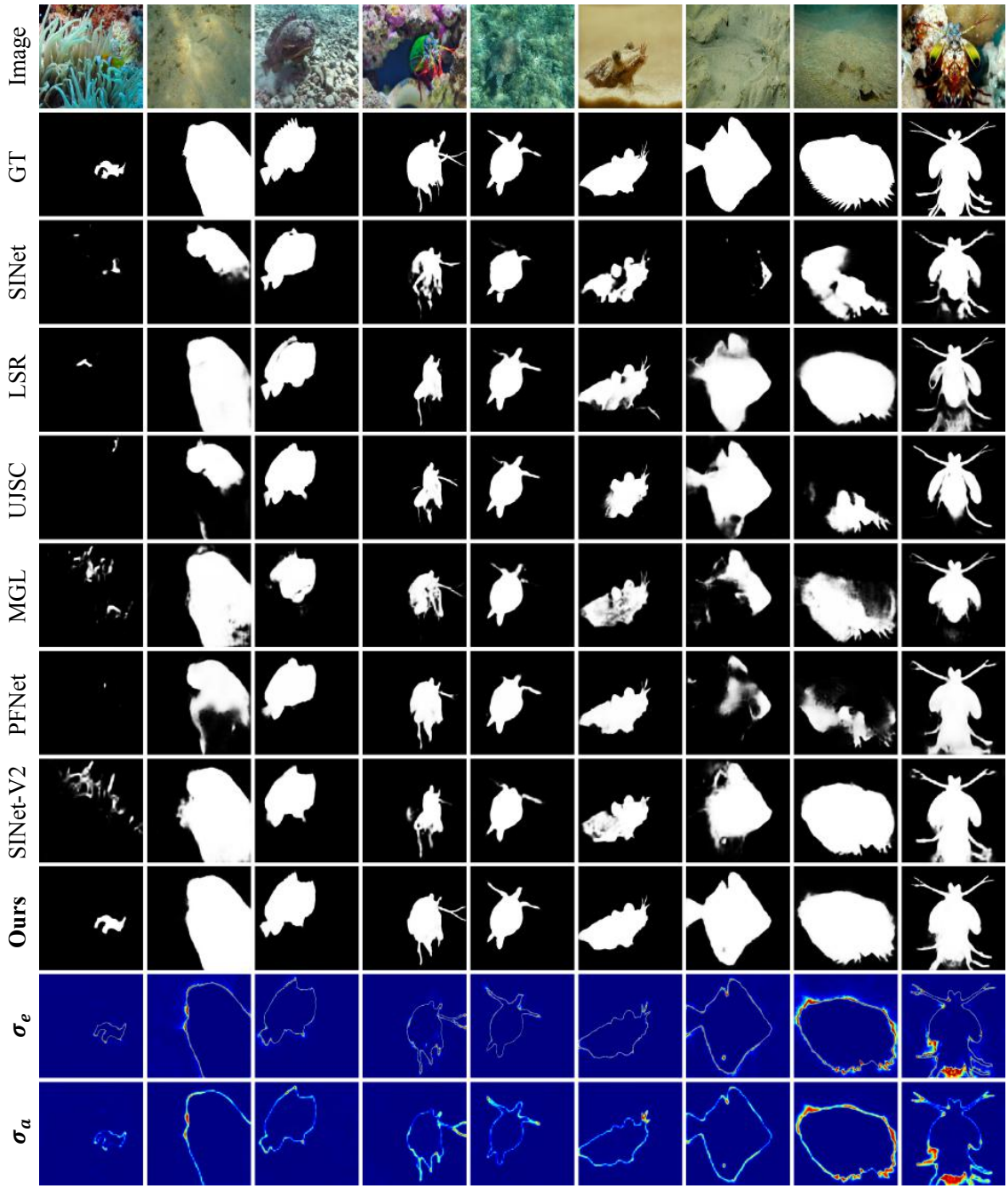


Fig. 7.6 Visual results of our method on NC4K [376].

7.1.4 Conclusion

Considering the inherent "model bias" and "data bias" of camouflaged object segmentation, we propose PUENet to achieve both accurate camouflaged object segmentation model and reliable uncertainty estimation. To reduce the sampling effort, we introduce PUA module to approximate the sampling based predictive uncertainty and achieve sampling-free uncertainty estimation during test-time. Further, Experimental results validate our solution. Importantly, the produced uncertainty map can represent our limited knowledge about this task, *i.e.*, center bias, data bias, and category bias. Although reliable uncertainty can be achieved with the proposed strategy, further investigation on uncertainty quantification and out-of-distribution sample estimation can lead to more advanced explainable camouflaged object segmentation model.

List of publications

[1] **Yi Zhang**, Lu Zhang, Wassim Hamidouche, and Olivier Deforges. "Key issues for the construction of salient object datasets with large-scale annotation." In 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 117-122. IEEE, 2020.
[2] **Yi Zhang**, Lu Zhang, Wassim Hamidouche, and Olivier Deforges. "A fixation-based 360 benchmark dataset for salient object detection." In 2020 IEEE International Conference on Image Processing (ICIP), pp. 3458-3462. IEEE, 2020.
[3] **Yi Zhang**, Geng Chen, Qian Chen, Yujia Sun, Yong Xia, Olivier Deforges, Wassim Hamidouche, and Lu Zhang. "Learning Synergistic Attention for Light Field Salient Object Detection." In 32nd British Machine Vision Conference (BMVC), 2021.
[4] **Yi Zhang**, Wassim Hamidouche, and Olivier Deforges. "Channel-Spatial Mutual Attention Network for 360° Salient Object Detection." In 26th International Conference on Pattern Recognition (ICPR), 2022.
[5] **Yi Zhang***, Fang-Yi Chao*, Wassim Hamidouche, and Olivier Deforges. "PAV-SOD: A New Task Towards Panoramic Audiovisual Saliency Detection." In Transactions on Multimedia Computing Communications and Applications (TOMM), 2022.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255. 3, 6, 9, 13, 23

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 3, 9, 18, 40, 67

[3] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 3, 9, 101, 103, 119, 131

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755. 3, 6, 9, 13, 15

[5] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833. 3, 9

[6] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," *MIT Laboratories*, 2012. 5, 11

[7] A. Borji and L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," *arXiv preprint arXiv:1505.03581*, 2015. 5, 11, 17

[8] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet," *arXiv preprint arXiv:1411.1045*, 2014. 5, 11

[9] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 262–270. 5, 11

[10] S. S. Kruthiventi, K. Ayush, and R. V. Babu, "Deepfix: A fully convolutional neural network for predicting human eye fixations," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4446–4456, 2017. 5, 11

[11] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4894–4903. 5, 6, 11, 15, 18, 52

[12] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "Deepvs: A deep learning based video saliency prediction approach," in *Proceedings of the european conference on computer vision (eccv)*, 2018, pp. 602–617. 5, 11, 18

[13] L. Jiang, M. Xu, and Z. Wang, "Predicting video saliency with object-to-motion cnn and two-layer convolutional lstm," *arXiv preprint arXiv:1709.06316*, 2017. 6, 11

[14] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015. 6, 11, 18, 28, 96

[15] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Computational visual media*, vol. 5, no. 2, pp. 117–150, 2019. 6, 11, 15, 24, 54, 55, 56, 70

[16] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 6, 11, 12, 15, 24, 46, 70, 116

[17] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3166–3173. 6, 11, 23, 24, 48, 52, 55, 60, 73

[18] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 280–287. 6, 11, 23, 24, 48, 49, 50, 51, 52, 55, 60, 73

[19] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5455–5463. 6, 11, 23, 48, 50, 52, 56, 60, 73

[20] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 136–145. 6, 11, 23, 52, 56, 60, 73, 119, 120, 132

[21] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 186–202. 6, 11, 23, 24, 48, 50, 52, 56, 57, 58, 70, 73, 87

[22] J. Li, C. Xia, and X. Chen, "A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 349–364, 2017. 6, 12, 26, 27, 48, 49, 51, 52, 56, 57, 60, 73, 75, 91

[23] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8554–8564. 6, 12, 26, 27, 28, 48, 49, 51, 52, 56, 57, 60, 70, 73, 76, 90, 91

[24] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, pp. 1–38, 2022. 6, 12

[25] Y. Rai, J. Gutiérrez, and P. Le Callet, "A dataset of head and eye movements for 360 degree images," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017, pp. 205–210. 6, 12, 21, 59, 60, 61, 70

[26] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, "Saliency in vr: How do people explore virtual environments?" *IEEE transactions on visualization and computer graphics*, vol. 24, no. 4, pp. 1633–1642, 2018. 6, 12, 21, 59, 60, 70

[27] X. Corbillon, F. De Simone, and G. Simon, "360-degree video head movement dataset," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017, pp. 199–204. 6, 12, 60, 70

[28] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang, "Predicting head movement in panoramic video: A deep reinforcement learning approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2693–2708, 2018. 6, 12, 20, 21, 70

[29] H.-N. Hu, Y.-C. Lin, M.-Y. Liu, H.-T. Cheng, Y.-J. Chang, and M. Sun, "Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 1396–1405. 6, 12, 21

[30] K.-H. Wang and S.-H. Lai, "Object detection in curved space for 360-degree camera," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3642–3646. 6, 12

[31] Y. Zhang, X. Xiao, and X. Yang, "Real-time object detection for 360-degree panoramic image using cnn," in *2017 International Conference on Virtual Reality and Visualization (ICVRV)*. IEEE, 2017, pp. 18–23. 6, 12

[32] P. Zhao, A. You, Y. Zhang, J. Liu, K. Bian, and Y. Tong, "Spherical criteria for fast and accurate 360 object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 959–12 966. 6, 12

[33] R. Benenson, S. Popov, and V. Ferrari, "Large-scale interactive object segmentation with human annotators," in *CVPR*, 2019. 7, 13

[34] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 302–321, 2019. 7, 13, 15

[35] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3485–3492. 7, 13

[36] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 15

[37] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 15

[38] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015. 15

[39] A. Borji, "What is a salient object? a dataset and a baseline model for salient object detection," *IEEE Transactions on Image Processing*, vol. 24, no. 2, pp. 742–756, 2014. 16, 51, 52, 55, 60

[40] ——, "Saliency prediction in the deep learning era: Successes and limitations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 679–700, 2019. 17, 18

[41] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark," http://saliency.mit.edu/. 17

[42] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1072–1080. 17

[43] P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and J. Xiao, "Turkergaze: Crowdsourcing saliency with webcam based eye tracking," *arXiv preprint arXiv:1504.06755*, 2015. 17

[44] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson, "Clustering of gaze during dynamic scene viewing is predicted by motion," *Cognitive computation*, vol. 3, no. 1, pp. 5–24, 2011. 18

[45] S. Mathe and C. Sminchisescu, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 7, pp. 1408–1424, 2014. 18

[46] A. Coutrot and N. Guyader, "How saliency, faces, and sound influence gaze in dynamic social scenes," *Journal of vision*, vol. 14, no. 8, pp. 5–5, 2014. 18

[47] Y. Liu, S. Zhang, M. Xu, and X. He, "Predicting salient face in multiple-face videos," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2017, pp. 4420–4428. 18

[48] K. Wang, S. Ma, F. Ren, and J. Lu, "Sbas: Salient bundle adjustment for visual slam," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–9, 2021. 18

[49] P. Linardos, E. Mohedano, J. J. Nieto, N. E. O'Connor, X. Giro-i Nieto, and K. McGuinness, "Simple vs complex temporal recurrences for video saliency prediction," *arXiv preprint arXiv:1907.01869*, 2019. 18, 19

[50] S. Yang, G. Lin, Q. Jiang, and W. Lin, "A dilated inception network for visual saliency prediction," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 2163–2176, 2019. 18, 19

[51] Q. Lai, W. Wang, H. Sun, and J. Shen, "Video saliency prediction using spatiotemporal residual attentive networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1113–1126, 2019. 18, 19

[52] T. Nguyen, M. Dax, C. K. Mummadi, N. Ngo, T. H. P. Nguyen, Z. Lou, and T. Brox, "Deepusps: Deep robust unsupervised saliency prediction via self-supervision," *Advances in Neural Information Processing Systems*, vol. 32, 2019. 18, 19

[53] L. Jiang, Z. Wang, M. Xu, and Z. Wang, "Image saliency prediction in transformed domain: A deep complex neural network method," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8521–8528. 18, 19

[54] S. He, H. R. Tavakoli, A. Borji, Y. Mi, and N. Pugeault, "Understanding and visualizing deep visual saliency models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 206–10 215. 18, 19

[55] H. R. Tavakoli, A. Borji, E. Rahtu, and J. Kannala, "Dave: A deep audio-visual embedding for dynamic saliency prediction," *arXiv preprint arXiv:1905.10693*, 2019. 19, 20

[56] K. Zhang, Z. Chen, and S. Liu, "A spatial-temporal recurrent neural network for video saliency prediction," *IEEE Transactions on Image Processing*, vol. 30, pp. 572–587, 2020. 18, 19

[57] X. Wu, Z. Wu, J. Zhang, L. Ju, and S. Wang, "Salsac: A video saliency prediction model with shuffled attentions and correlation-based convlstm," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 410–12 417. 19

[58] F. Hu and K. McGuinness, "Fastsal: a computationally efficient network for visual saliency prediction," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 9054–9061. 19

[59] A. Tsiami, P. Koutras, and P. Maragos, "Stavis: Spatio-temporal audiovisual saliency network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4766–4776. 19, 20, 22, 71, 130

[60] Y. Liu, M. Qiao, M. Xu, B. Li, W. Hu, and A. Borji, "Learning to predict salient faces: A novel visual-audio saliency model," in *European Conference on Computer Vision*. Springer, 2020, pp. 413–429. 19, 20

[61] X. Min, G. Zhai, J. Zhou, X.-P. Zhang, X. Yang, and X. Guan, "A multimodal saliency model for videos with high audio-visual correspondence," *IEEE Transactions on Image Processing*, vol. 29, pp. 3805–3819, 2020. 19, 20

[62] A. Linardos, M. Kümmerer, O. Press, and M. Bethge, "Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 919–12 928. 19

[63] G. Wang, C. Chen, D.-P. Fan, A. Hao, and H. Qin, "From semantic categories to fixations: A novel weakly-supervised visual-auditory saliency detection approach," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 119–15 128. 19, 20

[64] S. Yao, X. Min, and G. Zhai, "Deep audio-visual fusion neural network for saliency estimation," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 1604–1608. 19, 20

[65] S. Jain, P. Yarlagadda, S. Jyoti, S. Karthik, R. Subramanian, and V. Gandhi, "Vinet: Pushing the limits of visual modality for audio-visual saliency prediction," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 3520–3527. 19, 20

[66] F. Abawi, T. Weber, and S. Wermter, "Gasp: Gated attention for saliency prediction," *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, 2021. 19, 20

[67] G. Bellitto, F. Proietto Salanitri, S. Palazzo, F. Rundo, D. Giordano, and C. Spampinato, "Hierarchical domain-adapted feature learning for video saliency prediction," *International Journal of Computer Vision*, vol. 129, no. 12, pp. 3216–3232, 2021. 19

[68] Q. Lai, T. Zhou, S. Khan, H. Sun, J. Shen, and L. Shao, "Weakly supervised visual saliency prediction," *IEEE Transactions on Image Processing*, 2022. 19

[69] A. Umer, C. Termritthikun, T. Qiu, P. H. W. Leong, and I. Lee, "On-device saliency prediction based on pseudo knowledge distillation," *IEEE Transactions on Industrial Informatics*, 2022. 19

[70] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," *Advances in neural information processing systems*, vol. 29, 2016. 19, 20, 129, 130

[71] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 247–263. 20, 70

[72] S. Ramenahalli, "A biologically motivated, proto-object-based audiovisual saliency model," *AI*, vol. 1, no. 4, pp. 487–509, 2020. 20, 21, 22

[73] F.-Y. Chao, C. Ozcinar, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic, "Towards audio-visual saliency prediction for omnidirectional video with spatial audio," in *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2020, pp. 355–358. 20, 21, 22

[74] M. Cokelek, N. Imamoglu, C. Ozcinar, E. Erdem, and A. Erdem, "Leveraging frequency based salient spatial sound localization to improve 360° video saliency prediction," in *2021 17th International Conference on Machine Vision and Applications (MVA)*, 2021, pp. 1–5. 20, 21, 22

[75] E. J. David, J. Gutiérrez, A. Coutrot, M. P. Da Silva, and P. L. Callet, "A dataset of head and eye movements for 360 videos," in *Proceedings of the 9th ACM Multimedia Systems Conference*, 2018, pp. 432–437. 21

[76] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao, "Gaze prediction in dynamic 360 immersive videos," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5333–5342. 20, 21, 59, 60, 70

[77] Z. Zhang, Y. Xu, J. Yu, and S. Gao, "Saliency detection in 360 videos," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 488–503. 21, 59, 60, 70

[78] C. Li, M. Xu, X. Du, and Z. Wang, "Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 932–940. 21, 59, 60

[79] F.-Y. Chao, C. Ozcinar, C. Wang, E. Zerman, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic, "Audio-visual perception of omnidirectional video for virtual reality applications," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2020, pp. 1–6. 21, 70

[80] H.-T. Cheng, C.-H. Chao, J.-D. Dong, H.-K. Wen, T.-L. Liu, and M. Sun, "Cube padding for weakly-supervised saliency prediction in 360 videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1420–1429. 21, 22, 70

[81] F.-Y. Chao, L. Zhang, W. Hamidouche, and O. Deforges, "Salgan360: Visual saliency prediction on 360 degree images with generative adversarial networks," in *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2018, pp. 01–04. 21, 22, 66

[82] M. Qiao, M. Xu, Z. Wang, and A. Borji, "Viewport-dependent saliency prediction in 360° video," *IEEE Transactions on Multimedia*, vol. 23, pp. 748–760, 2020. 21, 22

[83] H. Lv, Q. Yang, C. Li, W. Dai, J. Zou, and H. Xiong, "Salgcn: Saliency prediction for 360-degree images based on spherical graph convolutional networks," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 682–690. 22

[84] Y. A. D. Djilali, K. McGuinness, and N. E. O'Connor, "Simple baselines can fool 360deg saliency metrics," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3750–3756. 21, 22

[85] M. Xu, L. Yang, X. Tao, Y. Duan, and Z. Wang, "Saliency prediction on omnidirectional image with generative adversarial imitation learning," *IEEE Transactions on Image Processing*, vol. 30, pp. 2087–2102, 2021. 21, 22

[86] Y. Dahou, M. Tliba, K. McGuinness, and N. O'Connor, "Atsal: An attention based architecture for saliency prediction in 360° videos," in *International Conference on Pattern Recognition*. Springer, 2021, pp. 305–320. 21, 22

[87] Y. A. D. Djilali, T. Krishna, K. McGuinness, and N. E. O'Connor, "Rethinking 360deg image visual attention modelling with unsupervised learning." in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 414–15 424. 21, 22

[88] D. Martin, A. Serrano, A. W. Bergman, G. Wetzstein, and B. Masia, "Scangan360: A generative model of realistic scanpaths for 360 images," *IEEE Transactions on Visualization & Computer Graphics*, vol. 28, no. 05, pp. 2003–2013, 2022. 21, 22

[89] J. Li, L. Han, C. Zhang, Q. Li, and Z. Liu, "Spherical convolution empowered viewport prediction in 360 video multicast with limited fov feedback," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2022. 21, 22

[90] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1155–1162. 23, 48, 49, 52, 55, 60, 73

[91] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2386–2395. 23, 24, 48, 49, 52, 56, 73

[92] P. Yan, Z. Wu, M. Liu, K. Zeng, L. Lin, and G. Li, "Unsupervised domain adaptive salient object detection through uncertainty-aware pseudo-label learning," *arXiv preprint arXiv:2202.13170*, 2022. 24, 25

[93] X. Lin, Z. Wu, G. Chen, G. Li, and Y. Yu, "A causal debiasing framework for unsupervised salient object detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 24, 25

[94] J. Zhang, J. Xie, Z. Zheng, and N. Barnes, "Energy-based generative cooperative saliency prediction," *arXiv preprint arXiv:2106.13389*, 2021. 24, 25

[95] S. Gao, W. Zhang, Y. Wang, Q. Guo, C. Zhang, Y. He, and W. Zhang, "Weakly-supervised salient object detection using point supervison," *arXiv preprint arXiv:2203.11652*, 2022. 25

[96] M. S. Lee, W. Shin, and S. W. Han, "Tracer: Extreme attention guided salient object tracing network," *arXiv preprint arXiv:2112.07380*, 2021. 25

[97] Y. Y. Ke and T. Tsubono, "Recursive contour-saliency blending network for accurate salient object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2940–2950. 25

[98] Y. Song, H. Tang, N. Sebe, and W. Wang, "Disentangle saliency detection into cascaded detail modeling and body filling," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2022. 25

[99] Y. Piao, W. Wu, M. Zhang, Y. Jiang, and H. Lu, "Noise-sensitive adversarial learning for weakly supervised salient object detection," *IEEE Transactions on Multimedia*, 2022. 25

[100] L. Zhang, Q. Zhang, and R. Zhao, "Progressive dual-attention residual network for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 25

[101] S. Yu, B. Zhang, J. Xiao, and E. G. Lim, "Structure-consistent weakly supervised salient object detection with local saliency coherence," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. AAAI Palo Alto, CA, USA, 2021. 25

[102] M. Ma, C. Xia, and J. Li, "Pyramidal feature shrinking for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2311–2318. 25

[103] B. Xu, H. Liang, R. Liang, and P. Chen, "Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection," in *Proceedings. of the AAAI Conference On Artificial Intelligence*, 2021, pp. 1–9. 25

[104] W. Wang, P. Li, and H.-T. Zheng, "Generating diversified comments via reader-aware topic modeling and saliency detection," *arXiv preprint arXiv:2102.06856*, 2021. 25

[105] Y. Liu, X.-Y. Zhang, J.-W. Bian, L. Zhang, and M.-M. Cheng, "Samnet: Stereoscopically attentive multi-scale network for lightweight salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3804–3814, 2021. 24, 25

[106] R. Song, W. Zhang, Y. Zhao, Y. Liu, and P. L. Rosin, "Mesh saliency: An independent perceptual measure or a derivative of image saliency?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8853–8862. 25

[107] M. Zhang, T. Liu, Y. Piao, S. Yao, and H. Lu, "Auto-msfnet: Search multi-scale fusion network for salient object detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 667–676. 25

[108] Z. Wu, L. Su, and Q. Huang, "Decomposition and completion network for salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 6226–6239, 2021. 25

[109] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4722–4732. 25

[110] L. Tang, B. Li, Y. Zhong, S. Ding, and M. Song, "Disentangled high quality salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3580–3590. 25

[111] Y.-C. Gu, S.-H. Gao, X.-S. Cao, P. Du, S.-P. Lu, and M.-M. Cheng, "inas: Integral nas for device-aware salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4934–4944. 25

[112] A. Siris, J. Jiao, G. K. Tam, X. Xie, and R. W. Lau, "Scene context-aware salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4156–4166. 25

[113] Y. Piao, J. Wang, M. Zhang, and H. Lu, "Mfnet: Multi-filter directive network for weakly supervised salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4136–4145. 25

[114] J. Li, J. Su, C. Xia, M. Ma, and Y. Tian, "Salient object detection with purificatory mechanism and structural similarity loss," *IEEE Transactions on Image Processing*, vol. 30, pp. 6855–6868, 2021. 25

[115] Z. Zhao, C. Xia, C. Xie, and J. Li, "Complementary trilateral decoder for fast and accurate salient object detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4967–4975. 25

[116] J. Zhang, J. Xie, N. Barnes, and P. Li, "Learning generative vision transformer with energy-based latent space for saliency prediction," *Advances in Neural Information Processing Systems*, vol. 34, 2021. 25

[117] I. Duta, A. Nicolicioiu, and M. Leordeanu, "Discovering dynamic salient regions for spatio-temporal graph neural networks," *Advances in Neural Information Processing Systems*, vol. 34, 2021. 25

[118] S. Yang, W. Lin, G. Lin, Q. Jiang, and Z. Liu, "Progressive self-guided loss for salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 8426–8438, 2021. 25

[119] X. Zhou, H. Fang, Z. Liu, B. Zheng, Y. Sun, J. Zhang, and C. Yan, "Dense attention-guided cascaded network for salient object detection of strip steel surface defects," *IEEE Transactions on Instrumentation and Measurement*, 2021. 25

[120] B. Wang, Q. Chen, M. Zhou, Z. Zhang, X. Jin, and K. Gai, "Progressive feature polishing network for salient object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 128–12 135. 26

[121] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 10 599–10 606. 26, 60, 67, 105

[122] J. Wei, S. Wang, and Q. Huang, "F^3net: fusion, feedback and focus for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 321–12 328. 24, 26, 88, 89, 90, 103, 105, 119, 131, 132, 133, 134, 135

[123] S. Song, H. Yu, Z. Miao, J. Fang, K. Zheng, C. Ma, and S. Wang, "Multi-spectral salient object detection by adversarial domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 023–12 030. 26

[124] Z. Zhou, Z. Wang, H. Lu, S. Wang, and M. Sun, "Multi-type self-attention guided degraded saliency detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13 082–13 089. 25, 26

[125] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, "Weakly-supervised salient object detection via scribble annotations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 546–12 555. 26

[126] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9413–9422. 24, 26, 88, 89, 90, 120, 132, 133, 134, 135

[127] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9141–9150. 26

[128] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 025–13 034. 26, 88, 89, 90, 120, 132, 133, 134, 135

[129] S.-H. Gao, Y.-Q. Tan, M.-M. Cheng, C. Lu, Y. Chen, and S. Yan, "Highly efficient salient object detection with 100k parameters," in *European Conference on Computer Vision*. Springer, 2020, pp. 702–721. 24, 26, 88, 89, 90, 120, 121, 132, 133, 134, 135

[130] Y. Luo, Y. Wong, M. S. Kankanhalli, and Q. Zhao, "n-reference transfer learning for saliency prediction," in *European Conference on Computer Vision*. Springer, 2020, pp. 502–519. 26

[131] J. Zhang, J. Xie, and N. Barnes, "Learning noise-aware encoder-decoder from noisy labels by alternating back-propagation for saliency detection," in *European conference on computer vision*. Springer, 2020, pp. 349–366. 26

[132] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *European conference on computer vision*. Springer, 2020, pp. 35–51. 25, 26, 88, 89, 90, 120, 132, 133, 134, 135

[133] J.-J. Liu, Q. Hou, and M.-M. Cheng, "Dynamic feature integration for simultaneous detection of salient object, edge, and skeleton," *IEEE Transactions on Image Processing*, vol. 29, pp. 8652–8667, 2020. 26

[134] S. Mohammadi, M. Noori, A. Bahri, S. G. Majelan, and M. Havaei, "Cagnet: Content-aware guidance for salient object detection," *Pattern Recognition*, vol. 103, p. 107303, 2020. 26

[135] Y. Liu, Y.-C. Gu, X.-Y. Zhang, W. Wang, and M.-M. Cheng, "Lightweight salient object detection via hierarchical visual perception learning," *IEEE Transactions on Cybernetics*, vol. 51, no. 9, pp. 4439–4449, 2020. 26

[136] D. Zhang, H. Tian, and J. Han, "Few-cost salient object detection with adversarial-paced learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 236–12 247, 2020. 26

[137] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1623–1632. 25, 26

[138] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7479–7489. 26, 27, 60, 67, 120

[139] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3907–3916. 26, 60, 67, 88, 89, 90, 96, 98, 101, 102, 107, 120, 132, 133, 134, 135

[140] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3917–3926. 24, 26, 60, 67, 105

[141] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8779–8788. 24, 26, 60, 67, 105

[142] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 24, 27, 28, 39, 102, 138, 139

[143] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 24, 25, 39, 40, 41, 42, 93, 94, 121, 130

[144] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597 24, 27

[145] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141. 25, 28, 39, 40, 41, 93, 97, 98, 117

[146] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19. 25, 28, 39, 40, 41, 42, 93, 105

[147] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2192–2199. 26, 27, 73

[148] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 6, pp. 1187–1200, 2013. 26, 27, 73, 75

[149] H. Kim, Y. Kim, J.-Y. Sim, and C.-S. Kim, "Spatiotemporal saliency detection for video sequences based on random walk with restart," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2552–2564, 2015. 26, 27, 73, 75

[150] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4185–4196, 2015. 26, 27, 48, 49, 51, 52, 56, 73, 75

[151] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 724–732. 26, 27, 48, 57, 58, 73, 76, 91

[152] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE transactions on circuits and systems for video technology*, vol. 27, no. 12, pp. 2527–2542, 2016. 26, 27, 48, 49, 51, 52, 56, 73

[153] Z. Yang, J. Miao, X. Wang, Y. Wei, and Y. Yang, "Associating objects with scalable transformers for video object segmentation," *arXiv preprint arXiv:2203.11442*, 2022. 27, 28

[154] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022. 27, 28

[155] D. Li, R. Li, L. Wang, Y. Wang, J. Qi, L. Zhang, T. Liu, Q. Xu, and H. Lu, "You only infer once: Cross-modal meta-transfer for referring video object segmentation," in *AAAI Conference on Artificial Intelligence*, 2022. 27, 28

[156] M. Lan, J. Zhang, F. He, and L. Zhang, "Siamese network with interactive transformer for video object segmentation," *arXiv preprint arXiv:2112.13983*, 2021. 27, 28

[157] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9226–9235. 27

[158] Y. Lee, H. Seong, and E. Kim, "Iteratively selecting an easy reference frame makes unsupervised video object segmentation easier," *arXiv preprint arXiv:2112.12402*, 2021. 27

[159] X. Xu, J. Wang, X. Li, and Y. Lu, "Reliable propagation-correction modulation for video object segmentation," *arXiv preprint arXiv:2112.02853*, 2021. 27

[160] Y.-W. Chen, X. Jin, X. Shen, and M.-H. Yang, "Video salient object detection via contrastive features and attention modules," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1320–1329. 27

[161] X. Jin, H. Xiao, X. Shen, J. Yang, Z. Lin, Y. Chen, Z. Jie, J. Feng, and S. Yan, "Predicting scene parsing and motion dynamics in the future," *Advances in neural information processing systems*, vol. 30, 2017. 27

[162] W. Zhao, J. Zhang, L. Li, N. Barnes, N. Liu, and J. Han, "Weakly supervised video salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 16 826–16 835. 27, 28, 70

[163] M. Zhang, J. Liu, Y. Wang, Y. Piao, S. Yao, W. Ji, J. Li, H. Lu, and Z. Luo, "Dynamic context-sensitive filtering network for video salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 1553–1563. 27, 28, 70, 90

[164] G.-P. Ji, K. Fu, Z. Wu, D.-P. Fan, J. Shen, and L. Shao, "Full-duplex strategy for video object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 4922–4933. 27, 28

[165] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 27

[166] K. Zhang, Z. Zhao, D. Liu, Q. Liu, and B. Liu, "Deep transport network for unsupervised video object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 8781–8790. 27

[167] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, vol. 26, 2013. 27

[168] C. Chen, G. Wang, C. Peng, Y. Fang, D. Zhang, and H. Qin, "Exploring rich and efficient spatial temporal interactions for real-time video salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3995–4007, 2021. 27

[169] Z. Yang, Y. Wei, and Y. Yang, "Associating objects with transformers for video object segmentation," *Advances in Neural Information Processing Systems*, vol. 34, 2021. 27

[170] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520. 27

[171] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125. 27

[172] H. K. Cheng, Y.-W. Tai, and C.-K. Tang, "Rethinking space-time networks with improved memory coverage for efficient video object segmentation," *Advances in Neural Information Processing Systems*, vol. 34, 2021. 27

[173] S. Ren, C. Han, X. Yang, G. Han, and S. He, "Tenet: Triple excitation network for video salient object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 212–228. 27

[174] B. Wang, W. Liu, G. Han, and S. He, "Learning long-term structural dependencies for video salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 9017–9031, 2020. 27

[175] Y. Gu, L. Wang, Z. Wang, Y. Liu, M.-M. Cheng, and S.-P. Lu, "Pyramid constrained self-attention network for fast video salient object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 10 869–10 876. 27, 28, 88, 89, 90, 132, 133, 134, 135

[176] S. Liu, D. Huang *et al.*, "Receptive field block net for accurate and fast object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 385–400. 27

[177] H. Li, G. Chen, G. Li, and Y. Yu, "Motion guided attention for video salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7274–7283. 27

[178] P. Yan, G. Li, Y. Xie, Z. Li, C. Wang, T. Chen, and L. Lin, "Semi-supervised video salient object detection using pseudo-labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7284–7293. 27, 28, 88, 89, 90, 118, 129, 130, 132, 133, 134, 135

[179] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8554–8564. 27

[180] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3623–3632. 27, 28, 44, 88, 89, 90, 91, 97, 99, 100, 103, 117, 118, 119, 122, 132, 133, 134, 135

[181] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017. 27, 118, 121

[182] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2. IEEE, 2005, pp. 1800–1807. 29

[183] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "icoseg: Interactive co-segmentation with intelligent scribble guidance," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3169–3176. 29, 52, 54

[184] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3365–3375, 2011. 29

[185] D. Zhang, J. Han, C. Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2994–3002. 29

[186] H. Yu, K. Zheng, J. Fang, H. Guo, W. Feng, and S. Wang, "Co-saliency detection within a single image," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 29

[187] D.-P. Fan, Z. Lin, G.-P. Ji, D. Zhang, H. Fu, and M.-M. Cheng, "Taking a deeper look at co-salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2919–2929. 29

[188] Z. Zhang, W. Jin, J. Xu, and M.-M. Cheng, "Gradient-induced co-saliency detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 455–472. 29

[189] D.-P. Fan, T. Li, Z. Lin, G.-P. Ji, D. Zhang, M.-M. Cheng, H. Fu, and J. Shen, "Re-thinking co-salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 29

[190] Y. Su, J. Deng, R. Sun, G. Lin, and Q. Wu, "A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection," *arXiv preprint arXiv:2203.04708*, 2022. 29

[191] S. Yu, J. Xiao, B. Zhang, and E. G. Lim, "Democracy does matter: Comprehensive feature mining for co-salient object detection," *arXiv preprint arXiv:2203.05787*, 2022. 29, 30

[192] R. Cong, N. Yang, C. Li, H. Fu, Y. Zhao, Q. Huang, and S. Kwong, "Global-and-local collaborative learning for co-salient object detection," *arXiv preprint arXiv:2204.08917*, 2022. 29

[193] R. Hu, Z. Deng, and X. Zhu, "Multi-scale graph fusion for co-saliency detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 7789–7796. 29

[194] K. Zhang, M. Dong, B. Liu, X.-T. Yuan, and Q. Liu, "Deepacg: Co-saliency detection via semantic-aware contrast gromov-wasserstein distance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 703–13 712. 29

[195] Q. Fan, D.-P. Fan, H. Fu, C.-K. Tang, L. Shao, and Y.-W. Tai, "Group collaborative learning for co-salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 288–12 298. 29, 30, 71

[196] N. Zhang, J. Han, N. Liu, and L. Shao, "Summarize and search: Learning consensus-aware dynamic convolution for co-saliency detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4167–4176. 29, 30

[197] W.-D. Jin, J. Xu, M.-M. Cheng, Y. Zhang, and W. Guo, "Icnet: Intra-saliency correlation network for co-saliency detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 749–18 759, 2020. 29, 30

[198] Q. Zhang, R. Cong, J. Hou, C. Li, and Y. Zhao, "Coadnet: Collaborative aggregation-and-distribution networks for co-salient object detection," *Advances in neural information processing systems*, vol. 33, pp. 6959–6970, 2020. 29, 30, 71

[199] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. 30, 42

[200] C. Xie, C. Xia, M. Ma, Z. Zhao, X. Chen, and J. Li, "Pyramid grafting network for one-stage high resolution saliency detection," *arXiv preprint arXiv:2204.05041*, 2022. 30

[201] Y. Zeng, P. Zhang, J. Zhang, Z. Lin, and H. Lu, "Towards high-resolution salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7234–7243. 30, 71

[202] G. Li, Z. Liu, Z. Bai, W. Lin, and H. Ling, "Lightweight salient object detection in optical remote sensing images via feature correlation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022. 31

[203] G. Li, Z. Liu, D. Zeng, W. Lin, and H. Ling, "Adjacent context coordination network for salient object detection in optical remote sensing images," *IEEE Transactions on Cybernetics*, 2022. 31

[204] X. Zhou, K. Shen, L. Weng, R. Cong, B. Zheng, J. Zhang, and C. Yan, "Edge-guided recurrent positioning network for salient object detection in optical remote sensing images," *IEEE Transactions on Cybernetics*, 2022. 31

[205] Q. Zhang, R. Cong, C. Li, M.-M. Cheng, Y. Fang, X. Cao, Y. Zhao, and S. Kwong, "Dense attention fluid network for salient object detection in optical remote sensing images," *IEEE Transactions on Image Processing*, vol. 30, pp. 1305–1317, 2020. 31, 71

[206] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, and S. Kwong, "Nested network with two-stream pyramid for salient object detection in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 9156–9166, 2019. 31, 71

[207] K. Fu, Y. Jiang, G.-P. Ji, T. Zhou, Q. Zhao, and D.-P. Fan, "Light field salient object detection: A review and benchmark," *arXiv preprint arXiv:2010.04968*, 2020. 32

[208] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "Rgb-d salient object detection: A survey," *Computational Visual Media*, vol. 7, no. 1, pp. 37–69, 2021. 32, 33

[209] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 454–461. 32

[210] A. Ciptadi, T. Hermans, and J. M. Rehg, "An in depth view of saliency," in *British Machine Vision Conference*. Georgia Institute of Technology, 2013. 32

[211] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proceedings of international conference on internet multimedia computing and service*, 2014, pp. 23–27. 32

[212] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: A benchmark and algorithms," in *European conference on computer vision*. Springer, 2014, pp. 92–109. 32

[213] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *2014 IEEE international conference on image processing (ICIP)*. IEEE, 2014, pp. 1115–1119. 32

[214] C. Zhu and G. Li, "A three-pathway psychobiological framework of salient object detection using stereoscopic technology," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 3008–3014. 32

[215] Y. Piao, X. Li, M. Zhang, J. Yu, and H. Lu, "Saliency detection via depth-induced cellular automata on light field," *IEEE Transactions on Image Processing*, vol. 29, pp. 1879–1889, 2019. 32, 96

[216] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Transactions on neural networks and learning systems*, vol. 32, no. 5, pp. 2075–2089, 2020. 32, 52, 56, 105

[217] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, and L. Shao, "Specificity-preserving rgb-d saliency detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4681–4691. 32, 34

[218] J. Zhang, D.-P. Fan, Y. Dai, X. Yu, Y. Zhong, N. Barnes, and L. Shao, "Rgb-d saliency detection via cascaded mutual information minimization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4338–4347. 32, 34

[219] W.-D. Jin, J. Xu, Q. Han, Y. Zhang, and M.-M. Cheng, "Cdnet: Complementary depth network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3376–3390, 2021. 32, 34

[220] P. Sun, W. Zhang, H. Wang, S. Li, and X. Li, "Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1407–1417. 32, 34

[221] X. Wang, L. Zhu, S. Tang, H. Fu, P. Li, F. Wu, Y. Yang, and Y. Zhuang, "Boosting rgb-d saliency detection by leveraging unlabeled rgb images," *IEEE Transactions on Image Processing*, 2022. 32, 34

[222] F. Wang, J. Pan, S. Xu, and J. Tang, "Learning discriminative cross-modality features for rgb-d saliency detection," *IEEE Transactions on Image Processing*, 2022. 32, 34

[223] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154. 32, 117, 118

[224] Y. Xu, X. Yu, J. Zhang, L. Zhu, and D. Wang, "Weakly supervised rgb-d salient object detection with prediction consistency training and active scribble boosting," *IEEE Transactions on Image Processing*, vol. 31, pp. 2148–2161, 2022. 34

[225] Y.-H. Wu, Y. Liu, J. Xu, J.-W. Bian, Y.-C. Gu, and M.-M. Cheng, "Mobilesal: Extremely efficient rgb-d salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 34, 71

[226] W. Ji, J. Li, Q. Bi, J. Liu, L. Cheng *et al.*, "Promoting saliency from depth: Deep unsupervised rgb-d saliency detection," in *International Conference on Learning Representations*, 2021. 34

[227] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu *et al.*, "Calibrated rgb-d salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9471–9481. 34

[228] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, "Rgb-d salient object detection via 3d convolutional neural networks," *arXiv preprint arXiv:2101.10241*, 2021. 34, 71, 97

[229] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3528–3542, 2021. 34

[230] Y. Zhao, J. Zhao, J. Li, and X. Chen, "Rgb-d salient object detection with ubiquitous target awareness," *IEEE Transactions on Image Processing*, vol. 30, pp. 7717–7731, 2021. 34

[231] J. Tang, D. Fan, X. Wang, Z. Tu, and C. Li, "Rgbt salient object detection: Benchmark and a novel cooperative ranking approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4421–4433, 2019. 33, 34, 35, 71

[232] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "Rgb-t image saliency detection via collaborative graph learning," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 160–173, 2019. 33, 34, 71

[233] Q. Zhang, T. Xiao, N. Huang, D. Zhang, and J. Han, "Revisiting feature fusion for rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1804–1818, 2020. 33, 34, 71

[234] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, "Apnet: Adversarial learning assistance and perceived importance fusion network for all-day rgb-t salient object detection," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021. 33, 34, 35

[235] C. Chen, S. Li, H. Qin, and A. Hao, "Real-time and robust object tracking in video via low-rank coherency analysis in feature space," *Pattern Recognition*, vol. 48, no. 9, pp. 2885–2905, 2015. 33

[236] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "Rgbt salient object detection: A large-scale dataset and benchmark," *arXiv preprint arXiv:2007.03262*, 2020. 33

[237] Y. Piao, Z. Rong, S. Xu, M. Zhang, and H. Lu, "Dut-lfsaliency: Versatile dataset and light field-to-rgb saliency detection," *arXiv preprint arXiv:2012.15124*, 2020. 35

[238] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2806–2813. 35, 71, 94, 96, 105, 107, 113

[239] J. Zhang, M. Wang, L. Lin, X. Yang, J. Gao, and Y. Rui, "Saliency detection on light field: A multi-cue approach," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 3, pp. 1–22, 2017. 35, 94, 96, 105, 109, 112

[240] T. Wang, Y. Piao, X. Li, L. Zhang, and H. Lu, "Deep learning for light field saliency detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8838–8848. 35, 36, 94, 96, 97, 105, 107, 109, 110, 111

[241] Y. Piao, Z. Rong, M. Zhang, X. Li, and H. Lu, "Deep light-field-driven saliency detection from a single view." in *International Joint Conference on Artificial Intelligence*, 2019, pp. 904–911. 35, 36, 94, 96, 105

[242] J. Zhang, Y. Liu, S. Zhang, R. Poppe, and M. Wang, "Light field saliency detection with deep convolutional networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 4421–4434, 2020. 35, 36, 94, 96

[243] Q. Zhang, S. Wang, X. Wang, Z. Sun, S. Kwong, and J. Jiang, "Geometry auxiliary salient object detection for light fields via graph neural networks," *IEEE Transactions on Image Processing*, vol. 30, pp. 7578–7592, 2021. 35, 36

[244] M. Zhang, J. Li, J. Wei, Y. Piao, and H. Lu, "Memory-oriented decoder for light field salient object detection," *Advances in neural information processing systems*, vol. 32, 2019. 36, 71, 94, 96, 97

[245] Y. Piao, Z. Rong, M. Zhang, and H. Lu, "Exploit and replace: An asymmetrical two-stream architecture for versatile light field saliency detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 865–11 873. 36, 94, 95, 96, 97, 98, 104, 105

[246] M. Zhang, W. Ji, Y. Piao, J. Li, Y. Zhang, S. Xu, and H. Lu, "Lfnet: Light field fusion network for salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 6276–6287, 2020. 36, 94, 96, 97

[247] Y. Jiang, W. Zhang, K. Fu, and Q. Zhao, "Meanet: Multi-modal edge-aware network for light field salient object detection," *Neurocomputing*, 2022. 36

[248] Q. Zhang, S. Wang, X. Wang, Z. Sun, S. Kwong, and J. Jiang, "A multi-task collaborative network for light field salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1849–1861, 2020. 36, 94, 96, 97

[249] D. Jing, S. Zhang, R. Cong, and Y. Lin, "Occlusion-aware bi-directional guided network for light field salient object detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1692–1701. 36

[250] N. Liu, W. Zhao, D. Zhang, J. Han, and L. Shao, "Light field saliency detection with dual local graph learning and reciprocative guidance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4712–4721. 36

[251] Y. Zhang, G. Chen, Q. Chen, Y. Sun, O. Deforges, W. Hamidouche, and L. Zhang, "Learning synergistic attention for light field salient object detection," *arXiv preprint arXiv:2104.13916*, 2021. 36, 71, 93, 114, 115, 118, 122

[252] A. Wang, "Three-stream cross-modal feature aggregation network for light field salient object detection," *IEEE Signal Processing Letters*, vol. 28, pp. 46–50, 2020. 36

[253] Y. Piao, Y. Jiang, M. Zhang, J. Wang, and H. Lu, "Panet: Patch-aware network for light field salient object detection," *IEEE Transactions on Cybernetics*, 2021. 36

[254] H. Cai, X. Zhang, R. Sun, J. Zhang *et al.*, "Multi-generator adversarial networks for light field saliency detection," in *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2021, pp. 1–6. 36

[255] Y. Liang, G. Qin, M. Sun, J. Qin, J. Yan, and Z. Zhang, "Dual guidance enhanced network for light field salient object detection," *Image and Vision Computing*, vol. 118, p. 104352, 2022. 36

[256] Y. Zhang, L. Zhang, W. Hamidouche, and O. Deforges, "A fixation-based 360 benchmark dataset for salient object detection," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 3458–3462. 37, 45, 59, 71, 73, 116, 117

[257] J. Li, J. Su, C. Xia, and Y. Tian, "Distortion-adaptive salient object detection in 360° omnidirectional images," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 38–48, 2019. 37, 45, 59, 60, 71, 73, 87, 116, 117, 118, 119, 120, 122

[258] G. Ma, S. Li, C. Chen, A. Hao, and H. Qin, "Stage-wise salient object detection in 360° omnidirectional image via object-level semantical saliency ranking," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 12, pp. 3535–3545, 2020. 37, 45, 71, 73, 87, 117, 118, 120, 122, 139

[259] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 39, 40, 42

[260] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," *Advances in neural information processing systems*, vol. 31, 2018. 39, 40

[261] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803. 39, 40, 41, 42, 93

[262] J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang, "Global-local temporal representations for video person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3958–3967. 39, 40

[263] Z. Liu, L. Wang, W. Wu, C. Qian, and T. Lu, "Tam: Temporal adaptive module for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 708–13 718. 39, 40

[264] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," *Advances in neural information processing systems*, vol. 28, 2015. 39, 40, 41

[265] B. Yang, G. Bender, Q. V. Le, and J. Ngiam, "Condconv: Conditionally parameterized convolutions for efficient inference," *Advances in Neural Information Processing Systems*, vol. 32, 2019. 39, 40

[266] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3139–3148. 39, 40

[267] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1347–1360, 2017. 39, 40

[268] Y. Fu, X. Wang, Y. Wei, and T. Huang, "Sta: Spatial-temporal attention for large-scale video-based person re-identification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8287–8294. 39, 40

[269] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global second-order pooling convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3024–3033. 40, 41

[270] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network," in *European Conference on Computer Vision*. Springer, 2020, pp. 275–292. 40, 71, 95, 100, 101, 102

[271] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," *Advances in neural information processing systems*, vol. 27, 2014. 40, 41

[272] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015. 41

[273] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 40

[274] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 510–519. 40, 41

[275] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018. 42

[276] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 1597–1604. 43, 49, 52, 54, 60, 66, 89, 90, 104, 120

[277] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 733–740. 43, 60, 66, 89, 90, 91, 104, 120

[278] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4548–4557. 43, 60, 66, 89, 90, 104, 120

[279] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," *arXiv preprint arXiv:1805.10421*, 2018. 43, 60, 66, 89, 90, 101, 103, 104, 120

[280] H. Jiang, M.-M. Cheng, S.-J. Li, A. Borji, and J. Wang, "Joint salient object detection and existence prediction," *Frontiers of Computer Science*, vol. 13, no. 4, pp. 778–788, 2019. 48

[281] Y. Zhang, L. Zhang, W. Hamidouche, and O. Deforges, "Key issues for the construction of salient object datasets with large-scale annotation," in *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2020, pp. 117–122. 49, 51, 53, 54, 55, 56, 57

[282] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8. 48, 49, 50, 52, 54

[283] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2007, pp. 1–8. 48, 49, 52, 54

[284] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 898–916, 2010. 48, 49, 52, 54

[285] E. L. Kaufman, M. W. Lord, T. W. Reese, and J. Volkmann, "The discrimination of visual number," *The American journal of psychology*, vol. 62, no. 4, pp. 498–525, 1949. 49

[286] A. Borji, D. N. Sihite, and L. Itti, "What stands out in a scene? a study of human explicit saliency judgment," *Vision research*, vol. 91, pp. 62–77, 2013. 50, 51, 52, 55, 57

[287] J. Li, Y. Tian, T. Huang, and W. Gao, "A dataset and evaluation methodology for visual saliency in video," in *2009 IEEE International Conference on Multimedia and Expo*. IEEE, 2009, pp. 442–445. 51, 52, 54

[288] M. Brown and S. Süsstrunk, "Multi-spectral sift for scene category recognition," in *CVPR 2011*. IEEE, 2011, pp. 177–184. 51, 52, 54

[289] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 4, pp. 996–1010, 2012. 51, 52, 55

[290] K. Koehler, F. Guo, S. Zhang, and M. P. Eckstein, "What do saliency models predict?" *Journal of vision*, vol. 14, no. 3, pp. 14–14, 2014. 51, 52, 55

[291] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *Journal of vision*, vol. 14, no. 1, pp. 28–28, 2014. 51, 52, 55

[292] Y. Wu, N. Zheng, Z. Yuan, H. Jiang, and T. Liu, "Detection of salient objects with focused attention based on spatial and temporal coherence," *Chinese Science Bulletin*, vol. 56, no. 10, pp. 1055–1062, 2011. 52, 55

[293] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior." in *BMVC*, vol. 6, no. 7, 2011, p. 9. 52, 54

[294] M.-M. Cheng, N. J. Mitra, X. Huang, and S.-M. Hu, "Salientshape: group saliency in image collections," *The visual computer*, vol. 30, no. 4, pp. 443–453, 2014. 52, 55

[295] C. Xia, J. Li, X. Chen, A. Zheng, and Y. Zhang, "What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4142–4150. 52, 56

[296] A. G. Huth, S. Nishimoto, A. T. Vu, and J. L. Gallant, "A continuous semantic space describes the representation of thousands of object and action categories across the human brain," *Neuron*, vol. 76, no. 6, pp. 1210–1224, 2012. 57

[297] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7264–7273. 60, 67, 88, 89, 90, 132, 133, 134, 135

[298] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 248–255. 60, 66

[299] M. Xu, C. Li, Y. Liu, X. Deng, and J. Lu, "A subjective visual quality assessment method of panoramic videos," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 517–522. 60

[300] O. Le Meur, T. Baccino, and A. Roumy, "Prediction of the inter-observer visual congruency (iovc) and application to image ranking," in *Proceedings of the 19th ACM international conference on Multimedia*, 2011, pp. 373–382. 68

[301] R. Arandjelovic and A. Zisserman, "Objects that sound," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 435–451. 70

[302] A. B. Vasudevan, D. Dai, and L. V. Gool, "Semantic object prediction and spatial sound super-resolution with binaural sounds," in *European conference on computer vision*. Springer, 2020, pp. 638–655. 70

[303] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, "Self-supervised learning of audio-visual objects from video," in *European Conference on Computer Vision*. Springer, 2020, pp. 208–224. 70

[304] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 570–586. 70

[305] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780. 70

[306] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725. 70

[307] R. Gao, Y.-Y. Chang, S. Mall, L. Fei-Fei, and J. Wu, "Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations," *arXiv preprint arXiv:2109.07991*, 2021. 70

[308] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman, "Audio-visual synchronisation in the wild," *arXiv preprint arXiv:2112.04432*, 2021. 70

[309] C. Gan, Y. Zhang, J. Wu, B. Gong, and J. B. Tenenbaum, "Look, listen, and act: Towards audio-visual embodied navigation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9701–9707. 70

[310] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, "Soundspaces: Audio-visual navigation in 3d environments," in *European Conference on Computer Vision*. Springer, 2020, pp. 17–36. 70

[311] P. Morgado, N. Nvasconcelos, T. Langlois, and O. Wang, "Self-supervised generation of spatial audio for 360 video," *Advances in Neural Information Processing Systems*, vol. 31, 2018. 70, 72, 144

[312] R. Garg, R. Gao, and K. Grauman, "Geometry-aware multi-task learning for binaural audio generation from video," *arXiv preprint arXiv:2111.10882*, 2021. 70

[313] M. Narasimhan, S. Ginosar, A. Owens, A. A. Efros, and T. Darrell, "Strumming to the beat: Audio-conditioned contrastive video textures," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3761–3770. 70

[314] C. Chen, Z. Al-Halah, and K. Grauman, "Semantic audio-visual navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 516–15 525. 70

[315] S. Majumder, Z. Al-Halah, and K. Grauman, "Move2hear: Active audio-visual source separation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 275–285. 70

[316] A. Rouditchenko, H. Zhao, C. Gan, J. McDermott, and A. Torralba, "Self-supervised audio-visual co-segmentation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2357–2361. 70

[317] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman, "Localizing visual sounds the hard way," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 867–16 876. 70

[318] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3879–3888. 70

[319] X. Liu, R. Qian, H. Zhou, D. Hu, W. Lin, Z. Liu, B. Zhou, and X. Zhou, "Visual sound localization in the wild by cross-modal interference erasing," *arXiv preprint arXiv:2202.06406*, 2022. 70

[320] Y. Tian, D. Hu, and C. Xu, "Cyclic co-learning of sounding object visual grounding and sound separation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2745–2754. 70

[321] D. Hu, R. Qian, M. Jiang, X. Tan, S. Wen, E. Ding, W. Lin, and D. Dou, "Discriminative sounding objects localization via self-supervised audiovisual matching," *Advances in Neural Information Processing Systems*, vol. 33, pp. 10 077–10 087, 2020. 70

[322] D. Hu, Y. Wei, R. Qian, W. Lin, R. Song, and J.-R. Wen, "Class-aware sounding objects localization via audiovisual correspondence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 70

[323] T. Afouras, Y. M. Asano, F. Fagan, A. Vedaldi, and F. Metze, "Self-supervised object detection from audio-visual correspondence," *arXiv preprint arXiv:2104.06401*, 2021. 70

[324] J. Yang, P. Sasikumar, H. Bai, A. Barde, G. Sörös, and M. Billinghurst, "The effects of spatial auditory and visual cues on mixed reality remote collaboration," *Journal on Multimodal User Interfaces*, vol. 14, no. 4, pp. 337–352, 2020. 70

[325] T. Rhee, L. Petikam, B. Allen, and A. Chalmers, "Mr360: Mixed reality rendering for 360 panoramic videos," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 4, pp. 1379–1388, 2017. 70

[326] M.-M. Cheng, S. Gao, A. Borji, Y.-Q. Tan, Z. Lin, and M. Wang, "A highly efficient model to study the semantics of salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 70

[327] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Deepco3: Deep instance co-segmentation by co-peak search and co-saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8846–8855. 71

[328] P. Zhang, W. Liu, Y. Zeng, Y. Lei, and H. Lu, "Looking for the detail and context devils: High-resolution salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3204–3216, 2021. 71

[329] S. Cheng, L. Song, J. Tang, and S. Guo, "Audio-visual salient object detection," in *International Conference on Intelligent Computing*. Springer, 2021, pp. 510–521. 71

[330] M. Xu, C. Li, S. Zhang, and P. Le Callet, "State-of-the-art in 360 video/image processing: Perception, assessment and compression," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 5–26, 2020. 87

[331] C.-L. Fan, W.-C. Lo, Y.-T. Pai, and C.-H. Hsu, "A survey on 360 video streaming: Acquisition, transmission, and display," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–36, 2019. 87

[332] S. Mahadevan, A. Athar, A. Ošep, S. Hennen, L. Leal-Taixé, and B. Leibe, "Making a case for 3d convolutions for object segmentation in videos," *arXiv preprint arXiv:2008.11516*, 2020. 88, 89, 90, 97, 132, 133, 134, 135

[333] S. Ren, W. Liu, Y. Liu, H. Chen, G. Han, and S. He, "Reciprocal transformations for unsupervised video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 455–15 464. 89, 90, 132, 133, 134, 135

[334] M. Huang, Z. Liu, G. Li, X. Zhou, and O. Le Meur, "Fanet: Features adaptation network for 360° omnidirectional salient object detection," *IEEE Signal Processing Letters*, vol. 27, pp. 1819–1823, 2020. 89, 90, 117, 119, 120, 122, 132, 133, 134, 135, 139

[335] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen, "Interactive digital photomontage," in *ACM SIGGRAPH*, 2004, pp. 294–302. 94

[336] N. Li, B. Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5216–5223. 96, 105

[337] J. Zhang, M. Wang, J. Gao, Y. Wang, X. Zhang, and X. Wu, "Saliency detection with a deeper investigation of light field," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015. 96, 105

[338] H. Sheng, S. Zhang, X. Liu, and Z. Xiong, "Relative location for light field saliency detection," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 1631–1635. 96

[339] A. Wang, M. Wang, X. Li, Z. Mi, and H. Zhou, "A two-stage bayesian integration framework for salient object detection on light field," *Neural Processing Letters*, vol. 46, no. 3, pp. 1083–1094, 2017. 96

[340] H. Wang, B. Yan, X. Wang, Y. Zhang, and Y. Yang, "Accurate saliency detection based on depth feature of 3d images," *Multimedia Tools and Applications*, vol. 77, no. 12, pp. 14 655–14 672, 2018. 96

[341] S. Wang, W. Liao, P. Surman, Z. Tu, Y. Zheng, and J. Yuan, "Salience guided depth calibration for perceptually optimized compressive light field 3d display," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2031–2040. 96

[342] X. Wang, Y. Dong, Q. Zhang, and Q. Wang, "Region-based depth feature descriptor for saliency detection on light field," *Multimedia Tools and Applications*, vol. 80, no. 11, pp. 16 329–16 346, 2021. 96

[343] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 756–13 765. 97, 105

[344] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308. 97, 101

[345] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 179–12 188. 101, 118, 129, 130, 138, 139

[346] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for rgbd salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3927–3936. 105

[347] H. Chen and Y. Li, "Three-stream attention-aware network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2825–2835, 2019. 105

[348] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection," *Pattern Recognition*, vol. 86, pp. 376–385, 2019. 105

[349] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "Pdnet: Prior-model guided depth-enhanced network for salient object detection," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 199–204. 105

[350] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for rgb-d salient object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3051–3060. 105

[351] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion," *IEEE transactions on cybernetics*, vol. 48, no. 11, pp. 3171–3183, 2017. 105

[352] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "Rgbd salient object detection via deep fusion," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2274–2285, 2017. 105

[353] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 714–722. 105

[354] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 355–370. 105

[355] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R3net: Recurrent residual refinement network for saliency detection," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press Menlo Park, CA, USA, 2018, pp. 684–690. 105

[356] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 202–211. 105

[357] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proceedings of the IEEE International Conference on computer vision*, 2017, pp. 212–221. 105

[358] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4019–4028. 105

[359] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2017, pp. 6609–6617. 105

[360] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3203–3212. 105

[361] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2798–2805. 116

[362] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, 2019. 118, 121, 138, 139

[363] F.-E. Wang, Y.-H. Yeh, M. Sun, W.-C. Chiu, and Y.-H. Tsai, "Bifuse: Monocular 360 depth estimation via bi-projection fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 462–471. 118

[364] A. Li, J. Zhang, Y. Lv, B. Liu, T. Zhang, and Y. Dai, "Uncertainty-aware joint salient object and camouflaged object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 071–10 081. 120, 121, 149

[365] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021. 120, 121

[366] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015. 128, 148

[367] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, "Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 129, 130, 138

[368] J. Zhang, Y. Dai, M. Xiang, D.-P. Fan, P. Moghadam, M. He, C. Walder, K. Zhang, M. Harandi, and N. Barnes, "Dense uncertainty estimation," *arXiv preprint arXiv:2110.06427*, 2021. 129, 140

[369] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013. 129

[370] K. Gu, G. Zhai, W. Lin, X. Yang, and W. Zhang, "No-reference image sharpness assessment in autoregressive parameter space," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 3218–3231, 2015. 140

[371] M. Broxton, J. Flynn, R. Overbeck, D. Erickson, P. Hedman, M. Duvall, J. Dourgarian, J. Busch, M. Whalen, and P. Debevec, "Immersive light field video with a layered mesh representation," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 86–1, 2020. 144

[372] B. Attal, S. Ling, A. Gokaslan, C. Richardt, and J. Tompkin, "Matryodshka: Real-time 6dof video view synthesis using multi-sphere images," in *European Conference on Computer Vision*. Springer, 2020, pp. 441–459. 144

[373] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 147, 149, 152

[374] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," *Computer Vision and Image Understanding*, vol. 184, pp. 45–56, 2019. 149, 150

[375] P. Skurowski, H. Abdulameer, J. Baszczyk, T. Depta, A. Kornacki, and P. Kozie, "Animal camouflage analysis: Chameleon database." in *Unpublished Manuscript*, 2018. 149, 151

[376] Y. Lv, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D.-P. Fan, "Simultaneously localize, segment and rank the camouflaged objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 591–11 601. 149, 153

[377] Q. Zhai, X. Li, F. Yang, C. Chen, H. Cheng, and D.-P. Fan, "Mutual graph learning for camouflaged object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 997–13 007. 149

[378] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan, "Camouflaged object segmentation with distraction mining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8772–8781. 149

[379] F. Yang, Q. Zhai, X. Li, R. Huang, A. Luo, H. Cheng, and D.-P. Fan, "Uncertainty-guided transformer reasoning for camouflaged object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4146–4155. 149

DOCTORAT BRETAGNE LOIRE / MATHSTIC

INSA RENNES

Titre: Segmentation d'objets saillants dans des images/videos 360° et champ de lumière

Mot clés : Segmentation d'objets saillants, 360°, champ de lumière, attention, audio-visuel, estimation de l'incertitude.

Résumé : La segmentation d'objets saillants est une tâche imitant l'attention visuelle humaine, et a constamment attiré l'attention de la communauté de la vision par ordinateur en raison de son énorme potentiel pour le développement de futures applications de réalité augmentée. Cependant, les méthodes de segmentation d'objets saillants sont principalement formées et testées avec des images et des vidéos 2D où des stimuli visuels sont collectés en fonction de rayons lumineux et d'un champ de vision limités, échouant ainsi à s'adapter au scénario du monde réel où les sujets humains reconnaissent les objets saillants en (i) capturant des informations sur le champ lumineux, (ii) en observant des scènes dans un champ de vision panoramique à 360°. Dans cette thèse, nous avons mené des études systématiques sur la segmentation d'objets saillants sur des images/vidéos à 360°, et proposé de nouvelles méthodologies pour la segmentation d'objets saillants en champ lumineux. Nous avons d'abord proposé respectivement des jeux de données image et vidéo pour permettre la segmentation des objets saillants à 360°. Nos ensembles de données proposés fournissent des données visuelles couvrant diverses scènes quotidiennes du monde réel, avec des objets saillants garantis annotés avec des masques pixel par pixel au niveau de l'objet et de l'instance, des étiquettes de classe d'objet/scène grossières à fines, et des attributs indiquant le commun défis pour mener le segmentation d'objets saillants dans les images/vidéos 360°. Pour contribuer davantage à la segmentation d'objets saillants à base d'images/vidéos à 360°, nous suivons les procédures courantes de segmentation d'objets saillants 2D et établissons ainsi des études de référence complètes basées sur nos jeux de données d'images et de vidéos à 360° proposés, obtenant de nouvelles découvertes qui facilitent le développement de nouveaux modèles 360°. Pour imiter l'attention visuelle humaine dans des scènes du monde réel, nous avons donc proposé de nouvelles méthodologies basées respectivement sur le champ lumineux 2D, et les images/vidéos 360°. Pour être précis, nos nouveaux modèles basés sur le champ lumineux ont appris une attention synergique multimodale pour une segmentation efficace des objets saillants. Notre méthode proposée basée sur l'image à 360° a permis d'obtenir une amélioration significative sur plusieurs références à 360°. Notre méthode basée sur la vidéo à 360 ° a eu recours à une technique d'estimation aléatoire de l'incertitude et a tiré parti des signaux visuels et audio pour segmenter les objets saillants de manière explicable. Nous espérons que cette thèse pourra servir de point de départ pour un développement futur vers une modélisation immersive de l'attention visuelle humaine au niveau de l'objet basée sur le multimédia.

Title: Salient object segmentation in 360° images/videos and light field

Keywords : Salient object segmentation, 360°, light field, attention, audio-visual, uncertainty estimation.

Abstract: Salient object segmentation is a task mimicking human visual attention, and has been constantly appealing attention from the computer vision community owing to its huge potential for the development of future augmented reality applications. However, state-of-the-art salient object segmentation methods are mostly trained and tested with 2D images and videos where visual cues are collected based on limited light rays and field-of-view, thus failing to adapt to the real-world scenario where human subjects recognize the salient objects by (i) capturing light field information, (ii) observing scenes in a 360° panoramic field-of-view. To close the gap between salient object segmentation academia and real-world applications, in this thesis, we conducted systematical studies towards 360° image-/video-based salient object segmentation, and proposed new methodologies for light field salient object segmentation. As current top-ranked salient object segmentation methods are mostly fully-supervised deep learning models, a lack of large-scale 360° image and video datasets surely limits the development of 360° models based on the same learning paradigm. To this end, we first respectively proposed image and video datasets to enable salient object segmentation in 360°. Our proposed datasets provide visual data covering various real-world daily scenes, with guaranteed salient objects annotated with both object-level and instance-level pixel-wise masks, coarse-to-fine object-/scene-class labels and attributes indicating the common challenges for conducting salient object segmentation in both 360° images and 360° videos. To further contribute to 360° image-/video-based salient object segmentation, we follow the common procedures in 2D salient object segmentation and thus establishing comprehensive benchmark studies based on our proposed 360° image and video datasets, gaining new findings that facilitate the development of new 360° models. To mimic the human visual attention in real-world scenes, we thus proposed new methodologies based on 2D light field, 360° images and 360° videos, respectively. To be specific, our new light field-based models learned multi-modal synergistic attention for effective salient object segmentation. Our proposed 360° image-based method achieved significant improvement on multiple 360° benchmarks. Our 360° video-based method resorted to aleatoric uncertainty estimation technique and took advantage of both visual and audio cues to segment salient objects in an explainable manner. We hope this thesis could serve as a starting point for future development towards immersive multi-media-based object-level human visual attention modeling.