https://doi.org/10.26599/AIR.2024.9150044 CAAI Artificial Intelligence Research

A Survey of Camouflaged Object Detection and Beyond

Fengyang Xiao[1, 2#], Sujie Hu[1#], Yuqi Shen[1], Chengyu Fang[1], Jinfa Huang[3], Longxiang Tang[1], Ziyun Yang[2], Xiu Li[1] ✉, and Chunming He[1, 2] ✉

ABSTRACT

Camouflaged object detection (COD) refers to the task of identifying and segmenting objects that blend seamlessly into their surroundings, posing a significant challenge for computer vision systems. In recent years, COD has garnered widespread attention due to its potential applications in surveillance, wildlife conservation, autonomous systems, and more. While several surveys on COD exist, they often have limitations in terms of the number and scope of papers covered, particularly regarding the rapid advancements made in the field since mid-2023. To address this void, we present the most comprehensive review of COD to date, encompassing both theoretical frameworks and practical contributions to the field. This paper explores various COD methods across four domains, including both image-level and video-level solutions, from the perspectives of traditional and deep learning approaches. We thoroughly investigate the correlations between COD and other camouflaged scenario methods, thereby laying the theoretical foundation for subsequent analyses. Furthermore, we delve into novel tasks such as referring-based COD and collaborative COD, which have not been fully addressed in previous works. Beyond object-level detection, we also summarize extended methods for instance-level tasks, including camouflaged instance segmentation, counting, and ranking. Additionally, we provide an overview of commonly used benchmarks and evaluation metrics in COD tasks, conducting a comprehensive evaluation of deep learning-based techniques in both image and video domains, considering both qualitative and quantitative performance. Finally, we discuss the limitations of current COD models and propose 9 promising directions for future research, focusing on addressing inherent challenges and exploring novel, meaningful technologies. This comprehensive examination aims to deepen the understanding of COD models and related methods in camouflaged scenarios. For those interested, a curated list of COD-related techniques, datasets, and additional resources can be found at https://github.com/ChunmingHe/awesome-concealed-object-segmentation.

KEYWORDS

camouflaged object detection; camouflaged scenario understanding; deep learning; artificial intelligence

Object detection, a fundamental task in computer vision, involves identifying and locating objects within images or videos. It comprises various fine-grained subfields: generic object detection (GOD)[1–4], salient object detection (SOD)[5, 6], and camouflaged object detection (COD)[7–9]. GOD aims to detect general objects, while SOD identifies prominent objects that stand out from the background. In contrast, COD targets those objects that blend into their surroundings, making COD an extremely challenging task. Figure 1 illustrates the relationship between the target dog and its background across various tasks, sourced from the classical datasets of GOD[10], SOD[11], and COD[12].

Fig. 1 Differences in inputs for GOD, SOD, and COD tasks. The three dogs, from left to right, are from the GOD dataset COCO[10], the SOD dataset DUTS[11], and the COD dataset COD10K[12], respectively. Notably, the rightmost dog is deeply camouflaged within its surroundings. Thus, it is evident that the COD task is significantly more challenging compared to the GOD and SOD tasks.

COD has recently garnered increasing attention and rapid development for its advantages in facilitating the development of visual perception for nuance discrimination and promoting various valuable real-life applications, ranging from concealed defect detection[13] in industry and pest monitoring[14, 15] in agriculture to lesion segmentation[16] in medical diagnosis and art, such as recreational art[17] and photo-realistic blending[18].

However, unlike GOD and SOD, COD involves detecting objects that are purposefully designed to be inconspicuous, like the Dalmatian hidden in the forest on the far right of Fig. 1 which is difficult to detect due to its camouflage with the surroundings, thus requiring more sophisticated detection strategies. COD can be further classified into image and video tasks[19, 20]. Normal COD, i.e., image-level, involves detecting the camouflaged objects in static images, whereas video-level COD (VCOD), deals with detecting these objects within video sequences. The latter one introduces additional complexity due to temporal continuity and dynamic changes, which necessitates models capable of extracting both spatial and temporal features effectively.

Traditional methods for COD and VCOD, including texture[21], intensity[22], color[23], motion[24], optical flow[25], and multi-modal analysis[26], have demonstrated their strengths in specific scenarios

1 Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China
2 Department of Biomedical Engineering, Duke University, Durham, NC 27708, USA
3 School of Electrical and Computer Engineering, Peking University, Shenzhen 518055, China
Fengyang Xiao and Sujie Hu contribute equally to this work.
Address correspondence to Xiu Li, li.xiu@sz.tsinghua.edu.cn; Chunming He, chunming.he@duke.edu

but also exhibit notable shortcomings. These approaches, relying on manually designed operators, suffer from limited feature extraction capacity, thus struggling with complex backgrounds and varying object appearances. This can lead to limited detection accuracy and robustness.

In contrast, deep learning-based COD method, e.g., convolutional neural network (CNN), transformer, and diffusion model, offer significant advantages by automatically learning rich feature representations[7, 9]. In addition, these methods utilize various strategies to address such challenging task, e.g., aggregating multi-scale features[27–31], bio-inspired mechanism-simulation[12, 19, 32–36], fusing multi-source information[9, 14, 37–39], learning multi-task[8, 40–43], jointing SOD[39, 44–46], and setting novel-task[47–51]. Despite their advantages, these methods also face intractable challenges, including high computational demands[52] and the requirement for large annotated, clean, and paired datasets[28].

Several surveys have been conducted on COD, with three seminal works[53–55] providing valuable overviews of the field. However, these surveys have limitations due to the narrow scope and limited number of papers they cover. For example, most of the methods discussed in these surveys are from before the first half of 2023, resulting in insufficient historical depth and domain breadth. As illustrated in Fig. 2, the COD field has seen rapid development in 2023. To address these gaps, we propose a more comprehensive survey that not only covers traditional and deep learning COD methods across both image and video domains but also benchmarks deep learning models in these areas. Furthermore, to the best of our knowledge, this survey is the first to deeply explore novel tasks such as referring-based COD[48] and collaborative COD[50]. We also provide a broader review of commonly used COD datasets and comprehensively cover recent advancements, challenges, and future trends.

The motivation for this paper stems from the critical importance of COD and the inadequacies of existing surveys. Our survey aims to provide a more thorough and detailed examination of COD, address gaps in the current literature, and highlight recent developments. We systematically categorize and analyze existing cutting-edge techniques, identify critical challenges, and suggest future research directions to advance the field.

Our contributions are summarized as follows:

- We provide a comprehensive review of existing COD methods and related tasks in camouflaged scenario understanding (CSU), along with commonly used datasets and evaluation metrics. To the best of our knowledge, this work represents the most extensive investigation to date, encompassing approximately 150 CSU-relevant cutting-edge studies.
- We methodically benchmark 40 representative image-level models and 8 representative video-level models based deep features on 6 characteristic datasets and 6 typical evaluation metrics, providing the quantitative and qualitative analysis of them.
- We systematically identify the limitations of existing COD methods and propose potential directions for future research. By shedding light on these challenges and opportunities, our work serves to guide and inspire further research efforts to advance the state-of-the-art in COD technology.
- We create a repository, at https://github.com/ChunmingHe/awesome-concealed-object-segmentation, that houses a carefully curated collection of COD methods, datasets, and relevant resources, which will be consistently updated to ensure the latest information is accessible.

We hope that this survey on COD will not only enhance understanding of the field but also stimulate greater interest within the computer vision community, fostering further research initiatives in related areas.

Note. In developing our search strategy, we conducted a thorough investigation across a variety of databases, including DBLP, Google Scholar, and arXiv Sanity Preserver. Our focus is particularly directed toward reputable sources, such as TPAMI and IJCV, as well as prominent conferences like CVPR, ICCV, and ECCV. We prioritized studies that provided official codes to enhance reproducibility, as well as those with higher citations and GitHub stars, indicative of significant recognition and adoption within the academic community. Following this initial screening, our literature selection process involved a rigorous evaluation of each paper's novelty, contribution, and significance, and an assessment of its status as seminal work in the field. While we acknowledge the possibility of omitting some noteworthy papers, our aim is to present a comprehensive overview of the most influential and impactful research, promoting research advancement and suggesting potential future trends and directions.

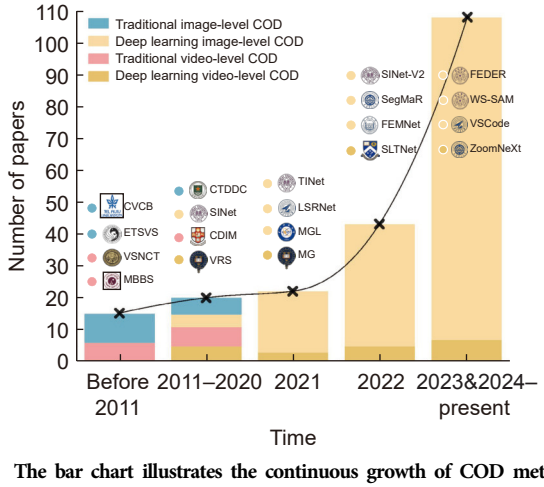

Fig. 2 The bar chart illustrates the continuous growth of COD methods across four scenarios. Representative works from various periods are categorized and marked on the line graph, with different colors corresponding to each scenario.

1 Image-Level COD Model

Image-level COD refers to the process of identifying and distinguishing objects that are designed to blend into their surroundings within static images, garnering significant attention currently. In this section, we categorize these methods into two primary approaches based on the features they utilize: traditional COD methods and deep learning COD methods. Traditional approaches typically rely on handcrafted features, whereas deep learning methods leverage neural networks to automatically learn and extract discriminative features from data. Given the rapid advancements in technology, we will focus primarily on deep learning COD methods, which have recently become the predominant approach.

1.1 Traditional COD methods

COD is initially rooted in traditional image-level methods, leveraging hand-crafted low-level features tailored to capture nuances in textures, intensities, and colors. These approaches constituted the foundation of early efforts in this domain. Table 1 summarizes the relevant models and characteristics of these methods.

Table 1 14 representative traditional methods of image-level COD. For more details, please refer to Section 1.1.

Index	Method	Year	Feature	Core component
1	CBDCE[22]	1998	Intensity	Distinguishing the convex and concave, D-Arg operator
2	CVCB[56]	2001	Intensity	strengthened D-Arg operator
3	TSMA[21]	2003	Texture	Bottom-up integration of structural features, filtered response
4	ETSVS[57]	2006	Texture	Four volunteer experiments
5	CDIA[58]	2006	Texture	Co-occurrence matrix, watershed technology, cluster analysis
6	DTEIA[59]	2008	Texture	Gray-level co-occurrence matrix, dendrogram plot
7	TTSCT[60]	2009	Color	Volunteer experiment based on the color difference
8	CYEM[61]	2010	Texture	Weight structural similarity, evaluation method
9	ROHS[62]	2010	Color & intensity	Background subtraction, color statistics, edge information
10	IRCT[63]	2011	Color & texture	Hue, saturation, and value (HSV) color, gray level co-occurrence matrix
11	CTDDC[64]	2011	Intensity	Gray-level image, 3D convexity
12	CTESM[65]	2015	Texture	Saliency map, evaluation method
13	TGWV[66]	2017	Texture & intensity	Texture-guided voting, stationary wavelet transform
14	CODMVA[23]	2020	Texture & intensity	Characterizing entity texture, statistical modeling

Texture feature-based approaches. Texture features capture the surface properties and distinctive patterns of images, usually manifested through grayscale distributions across pixels and their spatial neighborhoods. These features are utilized to distinguish camouflaged objects from their backgrounds based on texture differences.

Galun et al.[21] introduced a bottom-up approach, TSMA, for texture segmentation. This method leverages the adaptive identification and characterization of texture elements–such as size, aspect ratio, orientation, and brightness–combined with filter response statistics for enhanced differentiation and noise reduction. Unlike TSMA, which extracts global texture features, Bhajantri et al.[58] employed a co-occurrence matrix for block-based local texture analysis. Their method uses watershed segmentation for each defective block, processed through a dendrogram to distinguish the target, and concludes with cluster analysis to further confirm detection results.

Building upon earlier works, Sengottuvelan et al.[59] proposed an unsupervised technique for decamouflaging that converts images to grayscale and then splits them into blocks for gray-level co-occurrence matrix analysis, capturing pixel-neighbor relationships. They used a dendrogram plot to identify camouflaged objects without requiring prior background knowledge. Traditional evaluation methods often rely on subjective assessments, which can be cumbersome and ambiguous processes. To address this, Song et al.[61] leveraged a weighted structural similarity index and intrinsic image feature analysis to assess camouflage textures. To be more objective, Feng et al.[65] utilized human visual-based saliency maps to quantitatively assess texture differences.

Intensity feature-based approaches. These methods progressively advance from basic intensity-based techniques to more sophisticated exploitation of 3D convexity. CBDCE[22] detects regions of interest by processing intensity images directly, distinguishing between 3D convex and concave regions. This approach demonstrates robustness against variations in illumination, orientation, and scale. To enhance the detection of 3D objects camouflaged in complex scenes, CVCB[56] enhances the D-Arg operator in CBDCE. It maximizes the response to curved 3D objects against flat backgrounds, effectively mitigating visual camouflage in both natural and artificial environments. CTDDC[64] employes another 3D convexity-based operator that exploits image gray levels and median filtering to eliminate background noise, enabling effective detection of camouflaged objects.

Color feature-based approaches. In certain scenarios, color contrast and distribution can provide significant distinctiveness for separating camouflaged objects from their surroundings. Siricharoen et al.[62] optimized a statistical background subtraction and shadow detection algorithm by integrating color, edge, and intensity features for outdoor human segmentation, where strong shadows and low contrast are common. They employed vector median filtering to remove outlier pixels and combine color statistics with edge to generate initial coarse results, which are then refined using intensity features for accuracy. Kavitha et al.[63] proposed an image retrieval technique for COD, which involves segmenting images into blocks and extracting hue, saturation, value (HSV) color, and gray-level co-occurrence matrix texture features from each block. They use the principle of matching images based on similarity and Euclidean distance to improve detection accuracy.

Summary of the traditional COD methods. Handcrafted low-level features are engineered to be highly discriminative, making them effective for object detection and segmentation by emphasizing differences in texture and color. These features facilitate the isolation of concealed objects. Nevertheless, camouflage aims to minimize such distinctions, reducing visibility by blending objects with the background. Consequently, traditional methods struggle with camouflaged scenarios, performing well only in simple scenes with uniform backgrounds. These approaches often underperform when dealing with low-resolution images or when the foreground and background regions exhibit significant visual similarities.

1.2 Deep learning COD methods

While traditional methods rely on hand-crafted low-level features to capture key attributes of an image, deep learning methods extract complex and deep features directly from the data through automatic learning representations, demonstrating superior performance across various computer vision tasks. According to the existing surveys[54, 55], deep learning methods for COD can be broadly categorized based on three fundamental criteria: network architecture, learning paradigm, and supervision level. A detailed description of the three criteria is shown in Fig. 3. What's more,

Tables 2 and 3 outline the key characteristics of a total of 104 representative methods for image-level COD, published in 2019–2022 and 2023 & 2024, respectively.

Network architecture delineates how the input and output configurations are structured within the models. The linear[7, 12, 27, 28, 30, 31, 34, 35, 51, 73–75, 77, 78, 81–83, 86, 88, 89, 91, 92, 97, 99, 100, 104, 105, 107, 110–113, 115, 116, 120, 122, 124, 127, 131, 135–140, 142–146] employs bottom-up/top-down network, where the data flows in a single feed-forward pass. Aggregative architecture[14, 36, 37, 47, 48, 68, 72, 94–96, 101, 103, 123, 132, 147–150] combines features from multiple input streams, while branched architecture[8, 9, 32, 40–43, 46, 52, 67, 69–71, 76, 79, 80, 84, 85, 87, 90, 93, 98, 102, 106, 108, 114, 117, 119, 121, 125, 133, 134, 151–155] is characterized by multiple pathways for multiple outputs, associated with the multi-task learning paradigm. The hybrid[19, 39, 44, 45, 49, 50, 118, 129, 130, 156–160] integrates elements and strategies from the aforementioned ones to leverage their strengths.

The learning paradigm pertains to the approach adopted by the models to learn and adapt. This includes single-task and multi-task learning, specifically the former only involving COD, while the latter often importing auxiliary tasks, e.g., localization/ranking[32, 117], reconstruction[32, 117] and predicting associated cues, such as boundary[9, 19, 41, 43, 45, 52, 67, 71, 79, 84, 85, 87, 98, 102, 106, 114, 118, 121, 125, 133, 139], texture[69, 87, 108, 139], and uncertainty[8, 42, 44, 70, 76, 80, 90, 121, 129, 161], leading to improved accuracy and performance.

The supervision level describes the degree and nature of supervision provided during the training phase. This can be classified into four categories: fully supervised, with complete ground truth data; weakly-supervised, which utilizes limited or imprecise annotations[28, 92]; semi-supervised, which combines labeled and unlabeled data, generating pseudo labels for the latter[67]; and unsupervised, where no explicit labels are provided[51]. Notably, the "train-free" model[68, 132, 162] refers to models that do not require a traditional training process.

Beyond these categories, existing works also employ several strategies, such as aggregating multi-scale features, simulating bio-inspired mechanisms, fusing multi-source information, learning multiple tasks, jointing SOD, and establishing novel tasks, all aimed at enhancing COD performance. These strategies underscore the diverse and innovative approaches that researchers adopt to address the challenges in COD, revealing the underlying principles and techniques driving advancements. By focusing on these strategies, we can better understand the strengths and limitations of different methods, identify emerging trends, and provide a clearer roadmap for future research. However, there is currently a lack of detailed categorization and analysis of the various strategies employed. Hence, we aim to provide an elaborate introduction to this field. Besides, recent COD methods commonly utilize encoder-decoder architectures with the above strategies. The encoder is responsible for capturing rich semantic and structural information from input images, while the decoder reconstructs the image with the desired changes with the discriminative features extracted by the encoder. As shown in Tables 2 and 3, they often extract discriminative features via convolution-based encoders, such as VGG[163], ResNet[164], Res2Net[165], and EfficientNet[166], or transformer-based encoders, such as ViT[167], CLIP[168], BLIP2[169], PVT[170], and Swin[171]. Finally, Fig. 4 presents a comprehensive breakdown of the various methods discussed in this paper, along with their respective contributions to the overall analysis as detailed in the following subsections.

1.2.1 Multi-scale-context based strategy

This strategy captures the diverse appearances and varying scales

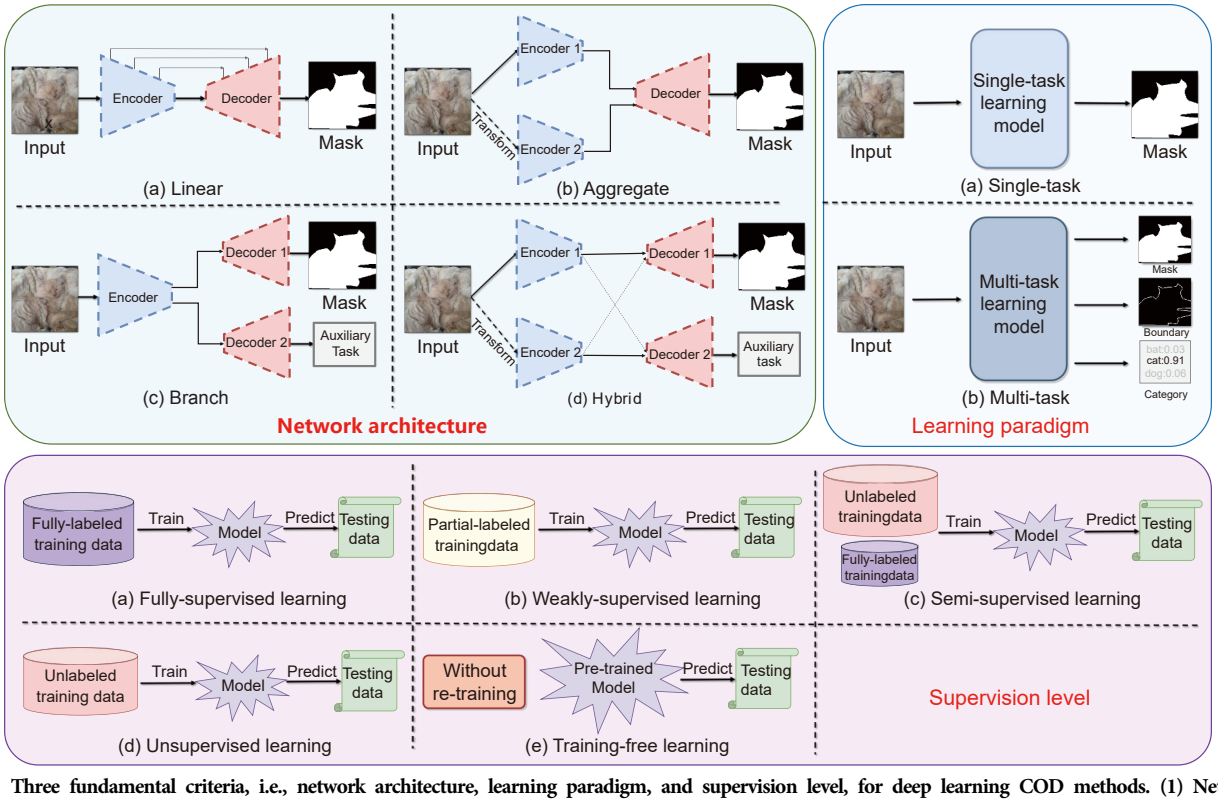

Fig. 3 Three fundamental criteria, i.e., network architecture, learning paradigm, and supervision level, for deep learning COD methods. (1) Network architecture: the input and output configurations of models, including linear[30], aggregative[14], branch[9], and hybrid[39] architectures. (2) Learning paradigm: the approach adopted by the models to learn, including single-task[19] and multi-task[49] paradigms. (3) Supervision level: the degree and nature of supervision provided during the training phase, including full-supervision[36], weak-supervision[28], semi-supervision[67], unsupervision[51], and training-free[68] levels. For more details, please refer to Section 1.2.

Table 2 36 representative deep learning methods of image-level COD published from 2019 to 2022. N.A.: network architecture according to input-output configuration, including L (Linear), A (Aggregative), B (Branched), and H (Hybrid). L.P.: learning paradigm, including S (Single-task) and M (Multi-task). S.L.: supervision level, including F (Full-supervision), W (Weak-supervision), S (Semi-supervision), U (Unsupervision) and TF (Training-free). Clicking on some method names may redirect you to their open-source codes or projects. For more details, please refer to Section 1.2.

Index	Method	Year	Backbone	N.A.	L.P.	S.L.	Core component
1	ANet[40]	2019	DHS, DSS, SRM, WSS	B	M	F	Classification & segmentation streams
2	SINet[7]	2020	ResNet50	L	S	F	Search & identification
3	TINet[69]	2021	ResNet50	B	M	F	Feature interaction, texture & holistic perception
4	UR-COD[70]	2021	ResNet50, ResNet34	B	M	F	Pseudo-edge & pseudo-map generators, uncertainty-aware refinement
5	UJSCOD[44]	2021	ResNet50	H	M	F	Uncertainty-aware adversarial learning, joint SOD & COD, confidence estimation
6	LSR[32]	2021	ResNet50	B	M	F	Simultaneously localization, segmentation & ranking
7	PFNet[34]	2021	ResNet50	L	S	F	Distraction mining strategy, positioning & focus
8	MGL[71]	2021	ResNet50	B	M	F	Mutual Graph Learning
9	UGTR[8]	2021	ResNet50	B	M	F	Uncertainty quantification network with Bayesian learning, prototyping & uncertainty-guided transformers
10	MirrorNet[72]	2021	ResNet50, ResNet101	H	M	F	Main & mirror streams embedded image flipping
11	C^2F-Net[27]	2021	Res2Net50	L	S	F	Attention-induced cross-level fusion & dual-branch global context modules
12	TANet[73]	2021	ResNeXt50	L	S	F	Residual refinement, texture-aware refinement, boundary-consistency loss
13	D^2C-Net[74]	2021	Res2Net50	L	S	F	Dual-branch features extraction & gradually refined cross fusion
14	POCINet[75]	2021	VGG16	L	S	F	Search & identification, part-object relationship & contrast integrated
15	DCE[76]	2021	ResNet50	B	M	F	Depth contribution exploration, multi-modal confidence-aware loss
16	BAS[77]	2021	ResNet34	L	S	F	Residual refinement, hybrid loss
17	MCIF-Net[78]	2021	ResNet50	L	S	F	Dual-branch mixture convolution & multi-level interactive fusion
18	BSA-Net[79]	2022	Res2Net	B	M	F	Boundary guider, reverse and normal attention streams
19	PreyNet[80]	2022	ResNet50	B	M	F	Initial detection and predator learning, bidirectional bridging interaction, uncertainty estimation
20	NCHIT[81]	2022	ResNet50	L	S	F	Neighbor connection mode, hierarchical information transfer
21	FEMNet[37]	2022	Res2Net, ResNet50	A	S	F	Frequency enhancement & high-order relation modules, frequency perception loss
22	SegMaR[35]	2022	ResNet50	L	S	F	Iteratively segmenting & magnifying, discriminative mask
23	ZoomNet[36]	2022	ResNet50	A	S	F	Scale integration & hierarchical mixed-scale units, uncertainty-aware loss
24	PINet[82]	2022	ResNet50	L	S	F	Cascaded decamouflage, classification-based label reweighting
25	TCU-Net[83]	2022	SwinT	L	S	F	Transformer and CNN-based U-Net, weighted hybrid loss, multi-dilated residual blocks, attention inception decoder
26	BGNet[84]	2022	Res2Net50	B	M	F	Edge-aware, edge-guidance feature & context aggregation modules
27	CubeNet[85]	2022	ResNet50	B	M	F	Square fusion & sub edge decoders, X-connection
28	ERRNet[52]	2022	ResNet50	B	M	F	Selective edge aggregation, reversible re-calibration, NGES Priors
29	C^2F-Net-V2[86]	2022	Res2Net50	L	S	F	Attention-induced cross-level fusion & dual-branch global context modules[27], extention
30	FAP-Net[41]	2022	Res2Net50	B	M	F	Boundary guidance, multi-scale feature aggregation, cross-level fusion and propagation
31	FindNet[87]	2022	Res2Net	B	M	F	Boundary & texture enhancement modules
32	DTC-Net[88]	2022	ResNet50	L	S	F	Local bilinear, spatial coherence organization
33	FBNet[89]	2022	ResNet50	L	S	F	Frequency-aware context aggregation and adaptive frequency attention
34	SINet-V2[12]	2022	Res2Net50	L	S	F	Search & identification, neighbor connection decoder, group-reversal attention[7], extention
35	OCENet[90]	2022	ResNet50	B	M	F	Dynamic confidence supervision, explicitly aleatoric uncertainty estimation
36	DQnet[91]	2022	ResNet50, ViTB	L	S	F	Cross-model detail querying, relation-based querying

Table 3 68 representative deep learning methods of image-level COD published in 2023 & 2024. As there exist some plug-and-play methods, "–" signifies no relevant information, e.g., backbone, of them. The same caption as Table 2.

Index	Method	Year	Backbone	N.A.	L.P.	S.L.	Core component
1	HitNet[31]	2023	PVTv2	L	S	F	High-resolution iterative feedback, iterative feedback loss
2	CRNet[92]	2023	ResNet50	L	S	W	consistency loss & feature-guided loss
3	FRINet[93]	2023	ViT	B	M	F	Frequency representation integration & reasoning
4	DaCOD[94]	2023	SwinL, ResNet50	A	S	F	Multi-modal collaborative learning, cross-modal asymmetric fusion
5	FPNet[95]	2023	PVT	A	S	F	Frequency-guided positioning, detail-preserving fine localization
6	XMSNet[96]	2023	PVT	A	S	F	All-round attentive fusion, coarse-to-fine decoder, cross-layer self-supervision
7	HCM[97]	2023	ResNet50	L	S	F	Hierarchical coherence modeling, reversible re-calibration decoder
8	ASBI[98]	2023	ResNet50, Res2Net50, PVTv2, EfficientNet	B	M	F	Attention-induced semantic & boundary interaction
9	FSPNet[30]	2023	ViT	L	S	F	Nonlocal token enhancement, a feature shrinkage decoder
10	FEDER[9]	2023	ResNet50, Res2Net50	B	M	F	Deep wavelet-like decomposition, ODE-inspired edge reconstruction
11	Semi-SINet[67]	2023	ResNet50	B	M	S	Search & identification, edge attention, semi-supervised
12	diffCOD[99]	2023	ViT	L	S	F	Denoising diffusion, injection attention module
13	TinyCOD[100]	2023	TinyNet-a	L	S	F	Adjacent scale features fusion, edge area focus
14	PopNet[101]	2023	–	A	S	F	Source-free depth, object popping, segmentating contact surface, object separation
15	UCOS-DA[51]	2023	DINO	L	S	U	Source-free unsupervised domain adaptation task, foreground-background contrastive self-adversarial
16	EAMNet[102]	2023	Res2Net	B	M	F	Segmentation-induced edge aggregation, edge-induced integrity aggregation, guided-residual channel attention
17	FDNet[103]	2023	PVT	A	S	F	Feature grafting & distractor aware
18	CFANet[104]	2023	Res2Net50	L	S	F	Cross-layer feature fusion, uniqueness enhancement strategy
19	OAFormer[105]	2023	PVTv2	L	S	F	Hierarchical location guidance, neighborhood searching
20	PENet[106]	2023	Res2Net50	B	M	F	Locatint, refining & restoring
21	OPNet[107]	2023	Conformer-B	L	S	F	Pyramid positioning, dual focus
22	DGNet[108]	2023	EfficientNet	B	M	F	Deep gradient network, gradient-induced transition
23	PrObeD[109]	2023	–	–	–	–	Wrapper based on proactive schemes
24	WS-SAM[28]	2023	ResNet50	L	S	W	Pseudo labeling with SAM, multi-scale feature grouping
25	Bi-RRNet[110]	2023	VAN-samll	L	S	F	Multi-scale scene perception, region-consistency enhancement
26	CPNet[111]	2023	ResNet50, ResNet101, SwinT	L	S	F	Ternary cascade perception
27	DBFN[112]	2023	Res2Net50	L	S	F	Double-branch fusion, parallel attention selection mechanism
28	PRNet[113]	2023	SMT-T	L	S	F	Cascaded attention perceptron, guided refinement decoder
29	DCNet[114]	2023	PVTv2	B	M	F	Area-boundary decoder, search & refinement
30	CMNet[45]	2023	Res2Net50	H	M	F	De-camouflaging manner, modeling task-conflicting&consistent attribute
31	SARNet[115]	2023	PVT	L	S	F	Searching, amplifying & recognizing
32	MSCAF-Net[116]	2023	PVTv2	L	S	F	Enhanced receptive field, cross-scale feature fusion, dense interactive
33	LSR-V2[117]	2023	ResNet50	B	M	F	Triple-task learning[32], extention
34	JCNet[118]	2023	SwinT	H	M	F	Contrastive learning with salient object
35	ZSCOD[119]	2023	ResNet50	B	M	F	Dynamic graph searching, camouflaged visual reasoning generator
36	PUENet[42]	2023	ResNet50, Res2Net50, ViT	B	M	F	Bayesian conditional variational auto-encoder, predictive uncertainty approximation
37	FSNet[120]	2023	SwinT	L	S	F	Two-stage focus & scanning, dynamic difficulty aware loss

(to be continued)

Table 3 68 representative deep learning methods of image-level COD published in 2023 & 2024. As there exist some plug-and-play methods, "–" signifies no relevant information, e.g., backbone, of them. The same caption as Table 2. (Continued)

Index	Method	Year	Backbone	N.A.	L.P.	S.L.	Core component
38	MGL-V2[43]	2023	ResNet50	B	M	F	Mutual graph learning, multi-source attention contextual recovery[71], extention
39	UEDG[121]	2023	PVTv2	B	M	F	Uncertainty reasoning & edge inference
40	BBNet[50]	2023	Res2Net	H	S	F	Inter-image collaborative feature exploration, intra-image object feature search, local-global refinement
41	MRR-Net[122]	2023	ResNet50, Res2Net50	L	S	F	Matching, recognition & refinement
42	MFFN[123]	2023	ResNet50	A	S	F	Co-attention of multi-view, channel fusion unit
43	OWinCANet[124]	2023	PVTv2	L	S	F	Overlapped window cross-level attention
44	BTSNet[125]	2023	ResNet50	B	M	F	Three-stage coarse-to-fine, boundary enhancement, mask-guided fusion
45	MLKG[47]	2023	SAM, CLIP	A	S	F	Multi-level knowledge descriptions from multimodal LLM
46	R2CNet[48]	2023	ResNet50	A	S	F	Reference & segmentation branches
47	SAM-Adapter[126]	2023	SAM	–	–	–	Incorporating domain-specific information or visual prompts
48	PFRNet[127]	2023	Res2Net50	L	S	F	Adaptive feature aggregation, feature refinement, context-aware feature decoding
49	CamoFourier[128]	2023	PatchGAN	–	–	F	Learnable Fourier-based augmentation, adaptive hybrid swapping
50	UJSCOD-V2[129]	2023	ResNet50	H	M	F	Joint-task contrastive & uncertainty-aware learning,[44] extention
51	PAD[130]	2023	ViT	H	M	F	Parameter efficient adapter, pre-train-adapt-detect
52	CamoDiffusion[131]	2024	PVT	L	S	F	Adaptive transformer conditional network, denoising network
53	GenSAM[68]	2024	CLIP, BLIP2	A	S	TF	Cross-modal chains of thought prompting, progressive mask generation
54	CoVP[132]	2024	Shikra, SAMHQ	A	S	TF	Chain of Visual Perception
55	RISNet[13]	2024	PVT	A	S	F	Depth-guided feature decoder, iterative feature refinement
56	VSCode[39]	2024	Swin	H	M	F	2D domain-specific and task-specific prompt learning, prompt discrimination loss
57	OVCoser[49]	2024	CLIP	H	M	F	Semantic guidance, structure enhancement, iterative refinement
58	EANet[133]	2024	EfficientNetB4	B	M	F	Edge-attention guidance, progressive recognition
59	Camouflageator[19]	2024	ResUNet, ResNet50	H	S	F	Adversarial training, internal coherence and edge guidance
60	FS-CDIS[134]	2024	ResNet101	B	M	F	Instance memory storage, instance triplet loss
61	DINet[135]	2024	Res2Net50	L	S	F	Body & detail blocks, global context unit
62	CamoFormer[136]	2024	PVTv2, ResNet50, Swin, ConvNeXt-B	L	S	F	Masked separable attention
63	CamoFocus[137]	2024	PVTv2	L	S	F	Feature split and modulation, context refinement
64	GreenCOD[138]	2024	EfficientNetB4	L	S	F	Gradient boosting without back propagation
65	AGLNet[139]	2024	EfficientNetB4	L	M	F	Hierarchical feature combination, recalibration decoder
66	SENet[46]	2024	ViT	B	M	F	Local information capture module, dynamic weighted loss
67	CoFiNet[140]	2024	SwinV2	L	S	F	Multi-scale feature integration, multi-activation selective kernel, dual-mask
68	SCLoss[141]	2024	–	–	–	–	Spatial coherence loss

of camouflaged objects with rich context information, and then aggregate cross-level features[52, 93, 104] while gradually refining features[113, 127, 137, 172], specifically in a hierarchical[36, 81, 105, 139], residual[73, 77], dual-branch[74, 78, 86, 107, 112, 140], X-connection[85] or iterative[14, 31, 35] manner, to enhance the representation. Some methods are further guided by edge[41, 84, 102, 133], frequency[89, 95], querying[91] or coarse prediction maps[125] in fusion process.

ERRNet[52] and HCM[97] propose a reversible re-calibration mechanism that leverages prior prediction maps, specifically targeting low-confidence regions to detect previously missed parts. This approach refines detection by focusing on regions that are initially overlooked. To improve efficiency, TinyCOD[100] introduces an adjacent scale feature fusion strategy using the lightweight TinyNet[100] as the backbone. Inspired by the Transformer architecture[173], CamoFormer[136] adopts masked separable attention, where multi-head self-attention is divided into three components. This approach allows for the simultaneous refinement of features at different levels in a top-down manner. While vision Transformers excel in global context modeling, they often struggle with locality modeling and feature fusion. To

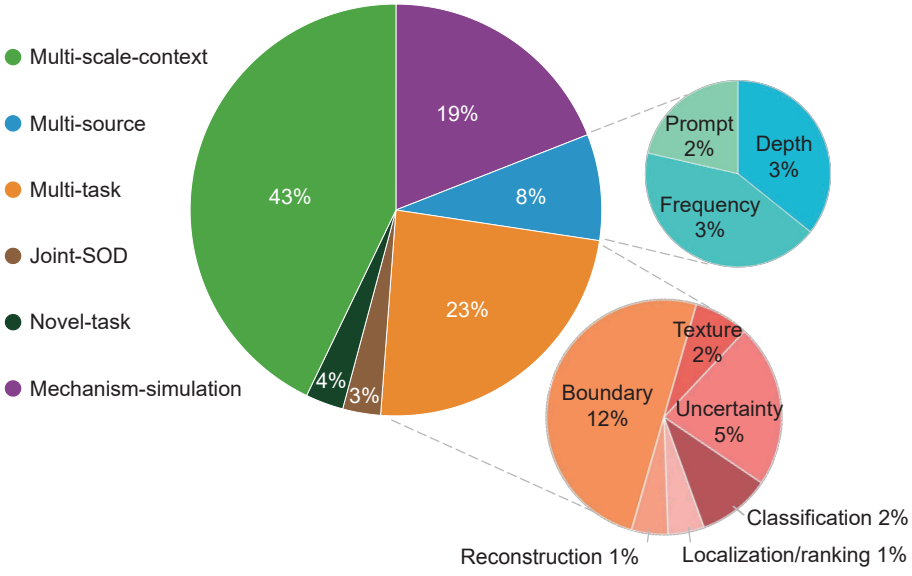

Fig. 4 Ratio of deep learning COD papers analyzed in this survey to each strategy.

address these issues, FSPNet[30] introduces a non-local token enhancement mechanism for improved feature interaction and a feature shrinkage decoder. Building on the Swin Transformer[171] and its shifted window strategy, OWinCANet[124] employs overlapped window cross-level attention. This method enhances low-level features with high-level guidance by sliding aligned window pairs across feature maps, ensuring a balance between local and global for superior performance.

Pixel-wise annotation of camouflaged objects is time-consuming and labor-intensive. Weakly supervised methods, using limited or scribble annotations, aim to reduce labeling effort and address boundary ambiguities. CRNet[92] designs a local-context contrasted module to enhance image contrast and a logical semantic relation module to analyze semantic relations, combined with feature-guided and consistency losses to impose stability on the predictions. He et al.[28] utilized the visual foundation model SAM[174] with sparse annotations as prompts to achieve initial coarse segmentation. Enhanced with multi-scale feature grouping[175], they generate reliable pseudo-labels for training off-the-shelf methods, addressing intrinsic similarity issues in coherent segmentation.

Summary. The multi-scale context-based strategy, a fundamental and widely adopted approach in COD, enhances feature representation through various architectures, such as hierarchical, residual, and iterative structures. This strategy effectively refines feature fusion, excelling at capturing diverse object appearances across varying scales and leveraging both local and global contextual information for robust feature extraction.

1.2.2 Mechanism-simulation based strategy

This bio-inspired strategy simulates the behavior of predators in nature or the visual detection mechanisms of humans. This is a multi-stage, coarse-to-fine strategy to incrementally enhance the accuracy of detection outcomes. SINet[7] simulates the search and identification process of predators, utilizing densely connected features and receptive fields, further enhanced by SINet-V2[12] with group-reversal attention. Semi-SINet[67] addresses training data scarcity with semi-supervised learning, generating pseudo labels for unlabeled data and introducing edge attention to improve SINet. Such localization-segmentation two-stage process has also inspired many approaches[34, 75, 95, 105, 107, 114, 120, 176, 177]. Besides, extra stages, e.g., restoring[106], matching[122], and amplifying[115], have also been introduced for more precise detection and enhanced adaptability.

MirrorNet[72] employs a mirror stream with embedded image flipping as a bio-inspired strategy to disrupt camouflage. ZoomNet[36] emulates human visual patterns when observing ambiguous images, specifically through zooming in and out, and leverages scale integration and hierarchical units to capture mixed-scale semantics. In contrast to multi-scale input strategies, MFFN[123] acquires complementary information through multi-view inputs from various angles, distances, and perspectives. This approach is designed to address the visually indistinguishable characteristics of camouflaged objects that arise under complex conditions, such as tiny object size, fuzzy objects, and blurred boundaries.

Moreover, incorporating auxiliary tasks could contribute to a better understanding of camouflage. PreyNet[80] mimics the sensory and cognitive mechanism of predation combined with uncertainty estimation, through a bidirectional interaction module for feature aggregation, as well as a policy-and-calibration paradigm for feature calibration. Camouflageator[19] inspired by the prey-predator dynamics presents an adversarial training framework with an auxiliary generator creating challenging camouflaged objects on the prey side. As for the predator side, it employs a camouflaged feature coherence module and edge-guided calibration to enhance complete segmentation and boundary clarity.

Summary. The mechanism-simulation-based strategy emulates predator behavior and human visual processes, adopting a multi-stage, coarse-to-fine detection approach. Its primary advantage lies in the incremental refinement of features across stages, progressively enhancing performance by concentrating on increasingly finer details.

1.2.3 Multi-source information fusion strategy

This strategy integrates diverse supplementary information sources to enhance COD, improving robustness and accuracy by leveraging complementary data from different domains, mainly frequency, depth, or text.

Frequency-domain integrated approaches. FEMNet[37] is the first to extend COD into the spatial domain, incorporating frequency enhancement and high-order relation for effective digging and fusion of the frequency clues with RGB features. FEDER[9] further decomposes the features into different frequency

bands via learnable wavelets, using frequency attention and guidance-based aggregation to differentiate foreground-background, complemented by an ordinary differential equation-inspired edge reconstruction for precise boundaries.

Depth-perceptual integrated approaches. Xiang et al.[76] introduced the first depth-guided COD network, DCE, which incorporates an auxiliary depth estimation branch and a multi-modal confidence-aware loss function via GAN[178] for effective depth integration. Building on DCE, DaCOD[94] also utilizes existing monocular depth estimation methods to generate depth maps and proposes a novel cross-modal asymmetric fusion strategy to blend these modalities asymmetrically. XMSNet[96] implements all-around attentive fusion to facilitate explicit cross-modal semantic mining and consistency constraints across decoding layers, thereby mitigating bias from inherent noise in depth estimation. PopNet[101] employs source-free depth for object pop-out, identifying contact surfaces under weak supervision and leveraging 3D priors for segmentation without source data, which enables efficient depth-to-semantics transfer. RISNet[14] is designed for extreme agricultural application scenarios, using multi-scale receptive fields and depth feature fusion to enhance dense and small COD in multiple stages.

Prompt-learning integrated approaches. These approaches utilize textual or visual prompts to adaptively guide and enhance COD models. CoVP[132] introduces a chain of visual perception and language text prompts, which linguistically and visually enhance the camouflaged scene perception of large vision-language models (LVLM), thereby reducing hallucinations. In contrast, GenSAM[68] employs cross-modal chains of thought prompting to generate visual prompts and progressively produce masks, iteratively refining the detection results. Both CoVP and GenSAM operate in a training-free manner, leveraging pre-trained models to avoid the need for extensive retraining, save computational resources, and enable rapid adaptation to COD. Rather than relying on text prompts, VSCode[39] introduces 2D domain-specific and task-specific prompt learning, effectively disentangling domain and task peculiarities. Through joint training, VSCode emerges as the first generalist model capable of addressing multimodal SOD and COD tasks with remarkable zero-shot generalization ability.

Summary. The multi-source information fusion strategy enhances COD by integrating diverse data modalities, e.g., frequency, depth, and prompt. However, challenges remain in effectively utilizing translated multi-modal information for more accurate segmentation and avoiding errors caused by incorrect translation, which may mislead the model's decision-making process.

1.2.4 Multi-task learning strategy

The fundamental premise of multi-task learning is that different tasks share common information for data processing. By leveraging these additional cues, multi-task learning is extensively employed to extract complementary insights from positively correlated tasks—such as reconstruction, classification, and localization—to enhance the detection accuracy of camouflaged objects.

Boundary-supervision integrated approaches. MGL[71] integrates camouflaged object-aware edge extraction (COEE) with graph-based mutual learning, incorporating typed functions to enhance feature representation by capturing both semantic and spatial information, which benefits COD and boundary detail accuracy. To address interpolation accuracy loss and computational redundancy in MGL, MGL-V2[43] incorporates multi-source attention contextual recovery, iteratively leveraging pixel feature information. Inspired by human COD processes, BSA-Net[79] employs a two-stream separated attention mechanism—reverse and normal attention streams—followed by a boundary guider to enhance performance by focusing on object boundaries. Beyond boundary enhancement, the extension of BSA-Net, FindNet[87], embeds texture information into feature representation, focusing on local patterns to perform effectively under complex COD conditions. To better integrate multiple features, ASBI[98] introduces an attention-induced semantic and boundary interaction network, utilizing attention-induced interaction to completely fuse multivariate and heterogeneous information.

Category-prediction integrated approaches. ANet[40] is the very classic method for COD, which includes a classification stream to predict whether they contain camouflaged objects and a segmentation stream to recognize them. To leverage the strong generalization ability and rich semantic knowledge of large-scale pre-trained foundation models, PAD[130] utilizes the pre-trained ViT[167] with lightweight parallel adapters by only tuning a small number of parameters for multi-task learning in a "pre-train, adapt, and detect" paradigm, achieving zero-shot task transferability, multi-task adaption, and cross-task generalization. As unseen classes are more general in real-world scenarios, Li et al. proposed ZSCOD[119] for zero-shot learning, which employs dynamic graph searching to adaptively capture edge details and a camouflaged visual reasoning generator for generating pseudo-features with the object-wise graph learning strategy to dynamically sample nodes to reduce background interference. Considering limited data on COD, FS-CDIS[134] leverages few-shot learning, simultaneously implementing instance segmentation task, with proposed instance triplet loss and instance memory storage, enhancing distinguishable features between background and foreground areas.

Localization/ranking integrated approaches. LSR[32] introduces the first ranking-based COD network, which infers the detectability of different camouflaged objects through instance segmentation and classification branches. LSR-V2[117] builds on the pioneering work of LSR by further exploring the interdependencies among these tasks. It not only establishes a new baseline and benchmark for both individual and joint tasks within a triple-task learning framework but also revisits the role of the camouflaged object ranking (COR) task.

Reconstruction integrated approaches. FRINet[93] uses Laplacian pyramid-like decomposition and transformer-CNN hybrid encoders with a reasoning module to capture and integrate high and low-frequency components, guided by an auxiliary image reconstruction task. Considering the strong performance of the transformer, Hao et al.[46] used ViT and regards both image reconstruction and binary segmentation tasks as training targets, with a local information capture module and dynamic weighted loss to enhance local modeling and handle complex cases.

Texture-detection integrated approaches. Motivated by the need to leverage complementary texture and camouflaged object cues, Zhu et al. proposed TINet[69], which utilizes texture labels and an interactive guidance framework. This framework consists of feature interaction guidance as well as texture and holistic perception decoders, aimed at refining segmentation with a particular focus on indefinite boundaries and texture differences. To better exploit the "discriminative patterns" within the object, DGNet[108] decouples COD into context and texture branches based on gradient generation. It employs a gradient-induced transition to softly group features from both branches, resulting in

superior efficiency.

Uncertainty-estimation based approaches. By approximating the uncertainty across different areas, these methods enhance the model's ability to focus on less identifiable regions, ultimately achieving high-confidence detection. UR-COD[70] utilizes uncertainty-aware refinement to reduce the noise of pseudo-edge and pseudo-map labels. Yang et al.[8] introduced Bayesian learning into transformer-based reasoning which revolutionizes the traditional deterministic mapping process employed in conventional COD by transitioning it into an uncertainty-guided context reasoning procedure. OCENet[90] employs aleatoric uncertainty estimation for confidence-aware COD, using dynamic supervision for accurate maps and assessing pixel-wise accuracy without ground truth. PUENet[42] integrates model and data uncertainty, implementing predictive uncertainty estimation and predictive uncertainty approximation for efficient test-time alongside SAM refining hierarchical features.

Summary. The multi-task learning strategy leverages shared information across tasks to enhance feature extraction, incorporating techniques such as boundary supervision, category prediction, localization, texture detection, and uncertainty estimation. While these approaches improve accuracy, they also introduce challenges, including increased computational complexity and data limitations. Researchers have sought to mitigate these issues by developing lightweight models and employing zero-shot learning techniques.

1.2.5 Joint-SOD based strategy

SOD and COD seem opposing but share common ground in the necessity to discern objects from a background based on contrast and contextual features. This strategy combines SOD with COD, leveraging the contradictory information or shared characteristics to gain a more thorough comprehension of COD. Many works are built in a multi-task manner, e.g., boundary detection[45, 118, 179] and image reconstruction[46].

UJSCOD[44] uses uncertainty-aware adversarial learning with a similarity measure module for modeling contradicting attributes and a data interaction strategy by defining simple COD samples as hard SOD samples for SOD data augmentation and higher robustness. Additionally, UJSCOD-V2[129] introduces contrastive learning to further investigate the cross-task correlations and random sampling-based foreground-cropping for COD data augmentation. CMNet[45] proposes a new perspective, de-camouflaging, modeling task-conflicting and task-consistent attributes to destroy the camouflage.

Summary. The joint-SOD strategy integrates SOD and COD tasks by leveraging their complementary and conflicting attributes. Techniques such as adversarial learning and contrastive learning are employed to enhance cross-task understanding. However, despite its effectiveness in modeling task similarities, the complexity of task-conflicting features presents optimization challenges. A major drawback of this strategy is the difficulty in distinguishing between salient and camouflaged objects. While both are anomalies, only camouflaged objects blend into their surroundings. This distinction is often overlooked, leading to misclassification when COD methods are applied to SOD tasks, where the camouflaged aspect is not adequately addressed.

1.2.6 Novel-task setting strategy

As COD continues to gain attention, its application scenarios have become increasingly diverse. Consequently, researchers have proposed various novel task settings to address COD challenges in different contexts, offering significant advantages, such as improved generalization to unseen data, better handling of complex scenes, and broadening the applicability of COD technologies.

Unsupervised camouflaged object segmentation (UCOS). Unsupervised learning is necessary due to the challenges in gaining extensive human labels for open-world applications, where supervised models often exhibit poor generalization. Domain adaptation is crucial because it allows models to effectively transfer knowledge from a source domain to a target domain. This adaptation helps bridge the gap between different data distributions, enhancing the model's performance in real-world scenarios where data characteristics may vary significantly. Therefore, a novel task, termed UCOS-DA[51], is formulated for unsupervised COD where both source and target labels are absent in the training phase. UCOS-DA leveraging foreground-background contrastive self-adversarial for pseudo-labels and domain-specific adaptation to address the UCOS task.

Collaborative camouflaged object detection (CoCOD). Zhang et al.[50] introduced this novel task to address the limitations of single-image analysis by jointly segmenting the same camouflaged object or objects belonging to the same class across multiple distinct images. This strategy leverages shared similarities and complementary cues inherent in the related images, thereby enhancing the accuracy of COD. To tackle CoCOD, BBNet[50] extracts and integrates camouflaged object features through inter-image collaborative feature exploration and intra-image object feature search. Additionally, BBNet enhances the representation of co-camouflaged features by employing the strategy of local-global feature refinement.

Referring camouflaged object detection (RefCOD). Standard COD aims to detect all camouflaged objects in a given scene. However, in certain real-world scenarios, such as ecological species protection and discovery, explorers may be interested only in locating specific camouflaged objects. To address this need, target references are introduced into COD, where referring text or images containing salient targets guide the detection of specified camouflaged objects. R2CNet[48] introduces reference and segmentation branches, combined with referring mask generation, to create pixel-level priors and enrich referring features, thereby enhancing detection capabilities. In contrast, MLKG[47] uses text as a reference in a multi-level knowledge-guided multimodal approach. By leveraging multimodal large language models (MLLMs), it achieves deep alignment, improves performance, and enables zero-shot generalization on unimodal datasets.

Open-vocabulary camouflaged object segmentation (OVCOS). Open-vocabulary refers to models that recognize and segment objects from novel classes not seen during training, leveraging vision-language models(VLM) like CLIP[168]. To fill in the gaps for COD in this field, Pang et al.[49] introduced the novel task OVCOS and built the corresponding baseline OVCoser, which integrates semantic guidance and visual structure cues in an iterative refinement manner via a transformer-based architecture attached to a frozen CLIP[168]. It also incorporates diverse sources of information, such as class semantic cues, spatial depth structures, object edge details, and iterative top-down guidance from the output space. To enhance task-relevant semantic context, OVCoser employs carefully crafted prompt templates, achieving robust and generalized performance.

Summary. The novel-task setting strategy aims to enhance generalization, handle complex scenes, and expand application scenarios. Techniques such as UCOS-DA, CoCOD, RefCOD, and OVCOS leverage unsupervised learning, collaborative segmentation, multimodal references, and vision-language

models. In the future, more advanced task settings could be introduced to address increasingly challenging problems in COD.

2 Video-Level COD Model

Unlike image-level COD techniques, which focus on single static images, video-level COD requires greater emphasis on motion cues to identify and localize camouflaged objects within continuous video frames. Video COD (VCOD) typically leverages temporal information, such as motion and changes across frames, to reveal objects that are difficult to detect in individual frames. However, this task presents significant challenges, including complex background noise, lighting variations, occlusions, and diverse camouflage strategies. Additionally, the high-dimensional nature of video data necessitates algorithms that are not only spatially accurate but also temporally consistent and stable. In this section, we categorize existing methods into two main types: traditional approaches and deep learning-based approaches.

2.1 Traditional VCOD methods

Compared to image-level methods, the video-level ones include more techniques, like optical flow analysis and motion detection, making them practical in video surveillance and security. As illustrated in Table 4, this section will delve into 12 representative traditional VCOD models, and we also categorize and introduce these methods based on the types of features they rely on.

Texture feature-based approaches. Malathi et al.[184] employed multi-camera codebooks for the detection of foreground objects. In their approach, texture pixels resembling the background are extracted and quantized into distinct codebooks. These codebooks are then integrated into a weighted framework to guide the abstraction of the foreground. To enhance the accuracy of predictions, disparity maps generated from the codebooks are used as supplementary information, particularly in cases where the color contrast between the foreground and background is weak. However, there remains the potential for shadows to be misclassified as targets.

Intensity feature-based approaches. When applied to video, these methods track dynamic changes between frames by leveraging temporal information, thereby improving the detection and localization of moving camouflaged objects. To tackle the challenge of foreground-background segmentation in video surveillance, particularly when complicated by camouflage, Guo et al.[181] combined Bayesian classification with Gaussian mixture models and apply temporal averaging across multiple frames in video sequences to reduce the bias of background models. However, detecting subtle differences remains difficult when the foreground and background are highly similar. To address this issue, Li et al.[20] extended COD to the wavelet domain, where subtle differences are emphasized in specific wavelet bands. They apply wavelet transforms to video sequences and estimate the probability of a wavelet coefficient belonging to the foreground by constructing foreground and background models within each individual wavelet band[187]. This approach detects camouflaged moving foregrounds through wavelet-domain multi-scale fusion.

Color feature-based approaches. Zhang et al.[188] employed computer-assisted detection of camouflaged targets, which first globally models the background of the input image, with the modeling of the foreground divided into global and local models. The global model captures the overall color and texture information of the foreground, while the local model focuses on details or changes that may exist in the foreground. Based on the models of the background and foreground, a factor measuring the degree of camouflage is introduced, determining whether a pixel is camouflaged by comparing the color differences between the background and foreground at that pixel. By comparing color contrast measurement results, true camouflaged regions are identified. Finally, the camouflage and identification models are fused under a Bayesian framework to perform complete target detection.

Motion feature-based approaches. These methods leverage the relative movement between the target object and the background between consecutive frames to identify camouflaged objects. Boult et al.[24] proposed a method for distinguishing background and foreground objects in visual surveillance systems using motion features. They introduced a conditional incremental model to update multiple background models in real-time and employed quasi-connected components to fill gaps caused by slight object movements, adapting to scene changes. However, background subtraction under camouflage conditions often results in fragmented, discontinuous object pieces, complicating subsequent classification or tracking processes. To address this, Conte et al.[182] developed an algorithm for detecting camouflaged personnel by

Table 4 12 representative traditional methods of VCOD. For more details, please refer to Section 2.1.

Index	Method	Year	Feature	Key technology
1	VSNCT[24]	2001	Motion	Dynamic & global thresholds
2	MBBS[26]	2004	Color & optical flow	The Normalized RGB color for estimating thedensity, the Optical flow feature for distinguishing motion exchange
3	CCMS[180]	2007	Color & motion	Background subtraction, integration of color and motion
4	TAMVF[181]	2008	Intensity	Bayesian classification, Gaussian mixture model
5	PCPD[182]	2009	Motion	Post-processing phase, integrating fragmented blocks to recover the original input
6	OTCID[25]	2010	Optical flow	Spatially coherent beam illuminates for identification
7	DBBOF[183]	2011	Optical flow	Clustering motion pattern, Kalman filter
8	FODCC[184]	2013	Texture	Multiple camera-based codebooks, disparity map
9	CDIM[185]	2013	Motion	Detection, identification & capture
10	USFS[186]	2015	Optical flow	Statistical distance, entropy-based spatial grouping, K-means clustering
11	BASCMOD[20]	2015	Color	Discriminative & camouflage modeling, Bayesian framework
12	FWFC[20]	2018	Intensity	Fusion in the wavelet domain, foreground & background models

aggregating fragmented detection blocks to reconstruct the complete shape of the target object during post-processing. This algorithm relies on the consistent segmentation of parts of the human body, such as the head, torso, and legs, across video sequences. By defining specific parameters to model the bounding boxes of human figures, the algorithm merges two or more boxes according to a set of rules. Considering perspective effects, a semi-automatic calibration phase dynamically adjusts parameters to ensure the bounding boxes accurately reflect the actual size of the individuals.

Optical flow feature-based approaches. Optical flow represents the motion of objects in a sequence as a vector field, providing crucial information on object speed and direction. To address challenges in classifying and identifying objects from low spatial resolution images—particularly in security-related applications—Beiderman et al.[25] introduced a technique that uses spatially coherent light beams to illuminate scenes and capture secondary speckle patterns from reflections. By tracking the temporal variations of these speckle patterns, the method extracts temporal feature signatures of objects. Comparing these signatures allows the algorithm to differentiate camouflaged objects from their surroundings, leveraging the unique physical properties of object surfaces for effective detection even under heavy camouflage. Building on this approach, Yin et al.[183] proposed a different method for detecting camouflaged objects in dynamic backgrounds using optical flow techniques. By simulating the motion patterns of objects and backgrounds and employing clustering analysis of optical flow features, this method accurately identifies objects. The algorithm further optimizes detection results through Kalman filtering, enhancing performance and accuracy while adapting to dynamic environments. Acknowledging the high data dependency of supervised methods and the difficulty of obtaining high-quality data in practical applications, Kim[186] proposed an unsupervised method using a visible-near-infrared hyperspectral camera. This method selects spectral and spatial features online through statistical distance to generate candidate feature bands, followed by entropy-based analysis to remove ineffective features. This approach achieves precise detection of camouflaged objects while reducing computational load.

Summary of the traditional VCOD methods. Traditional VCOD approaches rely on texture, intensity, color, motion, and optical flow features. While texture, intensity, and color-based methods enhance feature extraction and contrast analysis, motion techniques and optical flow leverage temporal information, dynamic changes, and motion tracking. However, compared to deep learning-based VCOD methods, these traditional approaches often lack adaptability, scalability, and accuracy, particularly in complex or high-dimensional environments and when handling diverse camouflage patterns or large-scale datasets.

2.2 Deep learning VCOD methods

As shown in Fig. 5, deep learning methods, compared to traditional VCOD techniques, verify a significant advantage by automatically learning complex feature representations from large datasets. These methods have a unique ability to capture intricate and subtle patterns, which in turn enhances the understanding of object dynamics within video sequences. However, video data presents greater complexity compared to image-based approaches. This added complexity stems from factors such as higher data dimensions, temporal continuity across frames, and the constantly evolving nature of dynamic changes in video content. These factors require models to possess not only strong spatial feature extraction capabilities but also a robust mechanism for accurately capturing subtle temporal variations. Despite promising advancements, existing work has largely focused on leveraging motion cues between different frames, and these approaches remain in their early stages of development, with considerable room for growth before they reach maturity.

In recent years, researchers have increasingly adopted a two-step framework, where optical flow maps or pseudo masks are pre-generated to serve as motion cues for video object detection. However, due to the challenges associated with cumulative errors and weak generalization[189], there is a growing trend towards employing end-to-end universal models to enhance reliability. The key characteristics of a total of 16 representative methods for VCOD are detailed in Table 5. We categorize these methods based on various criteria, e.g., network architecture, the use of optical flow maps, supervision level, and synthetic dataset generation. Besides, some methods provide links to their open-source projects.

2.2.1 Two-step VCOD framework

Since motion cues are crucial for distinguishing moving camouflaged objects from their backgrounds, researchers often incorporate a stage to pre-generate optical flow maps or pseudo masks for extracting motion information. Optical flow maps directly capture motion fields, i.e., the optical flow, between consecutive frames, providing compensation or registration to mitigate the camouflage effect. In contrast, pseudo masks implicitly learn temporal correspondences and motion patterns within the network, reducing reliance on external optical flow estimation and enabling the network to capture motion and maintain temporal consistency.

Explicit motion-based methods. Lamdouar et al.[191] adopted optical flow and a different image as inputs, and propose a

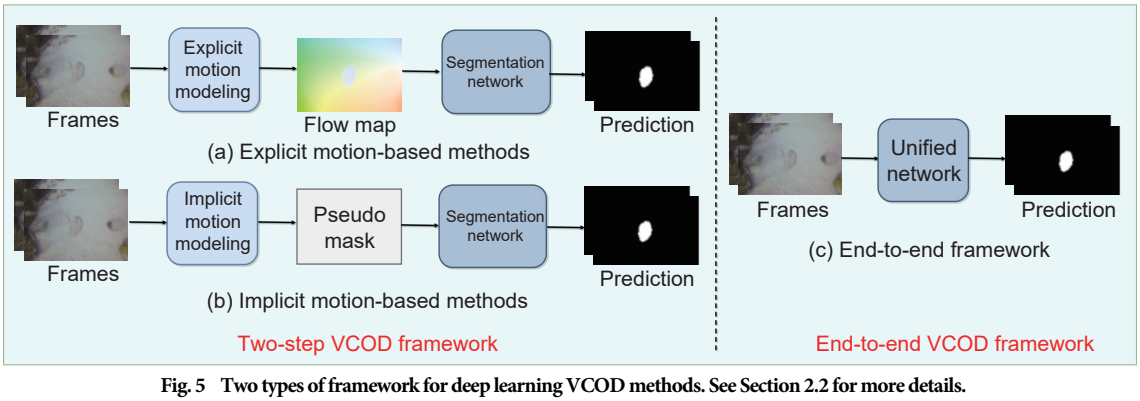

Fig. 5 Two types of framework for deep learning VCOD methods. See Section 2.2 for more details.

Table 5 16 representative deep learning methods of VCOD. N.A.: network architecture, including two-stage framework, specifically, EM (Explicit Motion-based methods) and IM (Implicit Motion-based methods), and End-to-End framework (EE). O.F.: whether pre-generating optical flow map. S.L.: supervision level, including F (Fully-supervision), S (Self-supervision) and U (Un-supervision). S.D.: whether generating synthetic dataset. Clicking on some method names may redirect you to their open-source codes or projects. "–" signifies no relevant information, e.g., backbone and S.D., of them. For more details, please refer to Section 2.2.

Index	Method	Year	Backbone	N.A.	O.F.	S.L.	S.D.	Core component
1	FMC[190]	2019	–	EM	√	F	–	Pixel-trajectory RNN, clustering
2	VRS[191]	2020	ResNet18	EM	√	F	–	Differentiable registration & motion segmentation
3	MG[192]	2021	–	EM	√	S	–	Motion grouping, self-supervised temporal consistency loss
4	SIMO[193]	2021	–	EM	√	F	√	Dual-head architecture
5	SLT-Net[194]	2022	PVT	IM	–	F	–	Short-term detection & long-term refinement
6	OFS[195]	2022	FakeModel	EM	√	U	–	Expectation-Maximization framework, motion augmentation
7	QSDI[196]	2022	ResNet, MobileNetV2	EM	√	F	–	Quantifying the static and dynamic biases
8	OCLR[197]	2022	ViT-S/8	EM	√	F	√	Object-centric layered representation
9	RCF[198]	2022	–	EM	√	F	–	Rotation-compensated flow, camera motion estimation
10	CAMEVAL[199]	2023	ResNet50	EM	√	F	√	Camouflage scores
11	ZoomNeXt[149]	2024	ResNet, PVTv2, EfficientNet	EE	–	F	–	Scale merging subnetwork, hierarchical difference propagation decoder
12	TMNet[200]	2024	SegFormer	EE	–	F	–	Learnable token selection
13	IMEX[201]	2024	ResNet50	EE	–	F	–	Implicit-explicit motion learning
14	TSP-SAM[189]	2024	SAM	EE	–	F	–	Temporal-spatial injection, prompt learning
15	SAM-PM[202]	2024	SAM-L	EE	–	F	–	Temporal fusion mask, memory prior affinity
16	EMIP[203]	2024	PVTv2	EE	–	F	–	Explicit motion modeling, interactive prompting

differentiable registration module for background alignment and a motion segmentation module with memory for moving object discovery. Facing the challenge of massive human annotation, Yang et al.[192] introduced a self-supervised method without any manual supervision, which groups motion with similar optical flow according to perceptual grouping principles. Meunier et al.[195] utilized unsupervised CNN-based motion segmentation from optical flow, leveraging the Expectation-Maximization framework for loss design and training, as well as designing data augmentation on the optical flow field, enabling real-time segmentation without annotations or iterative motion model estimation. Xie et al.[197] introduced object-centric layered representation and generate synthetic data for multi-object segmentation and tracking. However, reliance on external optical flow estimation can introduce errors that accumulate over time, potentially compromising the final mask prediction.

Implicit motion-based methods. SLT-Net[194] leverages short and long-term spatiotemporal relationships, specifically, utilizing the short-term motion capture between consecutive frames to produce pseudo masks and long-term temporal consistency to refine the former predictions, mitigating the flow estimation error. However, such implicit modeling may suffer from limited VCOD data.

2.2.2 End-to-end VCOD framework

This framework emerges as a solution to address the limitations of two-step approaches, particularly the cumulative errors stemming from intermediate stages and the weak generalization ability caused by limited training data. By integrating feature extraction, motion modeling, and segmentation into a unified network, this framework aims to minimize error propagation and maximize the utilization of scarce data, thereby enhancing robustness and performance. To extend the previous version[36] which zooms in and out on images with a shared triplet feature encoder, ZoomNeXt[149] implements image-video unified framework, and further integrates multi-head scale integration and rich granularity perception, enhancing the structural representation and discrimination. IMEX[201] unifies implicit and explicit motion learning in a cohesive framework, achieving inter-frame alignment and consistency preserving of camouflaged objects respectively. TSP-SAM[189] and EMIP[203] are both prompt learning-based methods for VCOD. The difference is that TSP-SAM utilizes frozen SAM embedded with temporal-spatial injection, motion-driven self-prompt learning and long-range consistency to learn reliable visual prompts, while EMIP adopts two-stream architecture, incorporating segmentation-to-motion and motion-to-segmentation prompts. Notably, although EMIP handles motion cues explicitly, it is not a two-step method, as it simultaneously conducts optical flow estimation and VCOD by interactive prompting. These methods demonstrate the evolving trend towards holistic, unified frameworks that not only address the shortcomings of traditional two-step approaches but also push the boundaries of performance and efficiency in VCOD.

Summary of the deep learning VCOD methods. The reviewed deep learning VCOD methods include both two-step and end-to-end frameworks. Two-step approaches often rely on optical flow or pseudo masks for motion cue extraction but are prone to cumulative errors and limited generalization. In contrast, end-to-end models integrate motion learning and segmentation into a unified framework, thereby reducing error propagation and enhancing robustness. Although these unified techniques are rapidly evolving, they remain relatively immature; nonetheless, their potential to overcome the limitations of traditional methods is promising.

3 Other Camouflaged Scenario Tasks

Beyond the fundamental tasks of COD and VCOD, the domain of

concealed scene understanding has evolved to encompass a wider array of high-level semantic tasks. These tasks aim to provide a deeper comprehension of camouflaged objects, extending from their classification to the generation of new camouflaged images. This section delves into the following advanced tasks, each addressing unique aspects of concealed scene understanding. The detail descriptions are illustrated in Fig. 6.

Camouflaged objects classification (COCls) focuses on distinguishing between different types of camouflaged objects. This task involves categorizing camouflaged objects into predefined or never seen classes, i.e., zero-shot[119] and open-vocabulary[49] learning, to handle the subtle differences and similarities among various camouflaged entities. Many COD methods integrate COCls[134, 130], training with datasets that are labeled with specific categories[12, 204], for better performance. Accurate classification of camouflaged objects can aid in better understanding and various ecological studies.

Camouflaged objects localization (COL) aims to identify the most detectable regions of camouflaged objects, and further localize the discriminative regions that make the camouflaged object stand out. Lv et al.[32, 117] is the first to propose this task and leverage an eye tracker to record human gaze patterns, pinpointing salient discriminative regions. Simultaneously, they relabel existing datasets with fixation annotation, providing a valuable resource for training and evaluating models. This task enhances the ability to pinpoint critical areas within a camouflaged scene, which is crucial for applications in wildlife monitoring, and search and rescue operations.

Camouflaged instance count (COCnt) is focused on quantifying the number of camouflaged objects within a given scene, even in complex environments where objects may overlap or partially obscure each other. Sun et al.[205] introduced a correlated task, indiscernible object counting (IOC), to count objects that blend seamlessly with their surroundings. To tackle this challenge, they propose a unified framework, IOCFormer, integrating density-based[206] and regression-based[207] counting methods. Due to the scarcity of suitable datasets, they also created IOCfish5K, full of high-resolution images for underwater IOC, with dense annotations. This emerging and promising task helps in assessing population densities and ensuring comprehensive area surveillance.

Camouflaged instance rank (CIR) addresses the challenge of ranking multiple camouflaged instances based on specific criteria such as visibility, detectability, or relevance. With the introduction of the COL task, Lv et al. proposed the CIR task, which is based on the difficulty level of camouflage. To this end, the proposed CAM-FR[32] and CAM-LDR[117] datasets also include ranking labels. Furthermore, they devised a triple-task learning framework that simultaneously localizes, segments and ranks camouflaged objects, efficiently utilizing the inner correlation among COD, COL, and CIR.

Camouflaged instance segmentation (CIS) is a more granular task that involves segmenting individual camouflaged objects, i.e., instances, from their background and from each other. CIS was proposed by Le et al.[208], who also introduce CIS dataset, i.e., CAMO++, by extending the previous CAMO dataset. They also conduct camouflage fusion learning, which fuses existing instance segmentation models, e.g., Cascade Mask RCNN[209], by learning to predict the best model per image. To better break the deceptive camouflage, Luo et al.[210] proposed a framework with a pixel-level camouflage decoupling module and an instance-level camouflage suppression module. Pei et al.[211] introduced the first one-stage transformer in CIS by fusing local features and long-range context dependencies. Recently, Vu et al.[212] leveraged text-to-image diffusion[213] and CLIP[168] for CIS, while utilizing open-vocabulary capabilities to learn multi-scale textual-visual features. This level of detail allows for more precise identification and interaction with each object, which is essential for applications that require accurate object differentiation, such as advanced robotic vision, detailed ecological studies, and targeted medical imaging.

4 Experiment

4.1 Datasets

Datasets play a pivotal role in the development and evaluation of COD algorithms. In this subsection, we provide an overview of prominent datasets relevant to COD tasks, categorized into image-level and video-level datasets. Table 6 summarizes essential information along with their respective links for access. Additionally, Figs. 7 and 8 showcases exemplar images from these datasets, offering a visual insight into the challenges posed by COD.

4.1.1 Characteristic COD datasets

CHAMELEON[215] is a small-scale, unpeer-reviewed dataset consisting of 76 camouflaged images collected from the internet using the keyword "camouflaged animals". Each image is manually annotated with labels focusing on animals camouflaged within complex ecological backgrounds. This dataset is typically used as a test dataset for model performance evaluation.

CAMO-COCO[40] consists of 2500 images across eight categories, created by merging the camouflaged dataset CAMO with the non-camouflaged dataset MS-COCO[10], each contributing 1250 images. In CAMO-COCO, 80% of the images are designated for training and the remaining for testing. CAMO includes both natural camouflaged objects, such as animals, and artificial camouflaged objects, such as human beings, featuring seven challenging attributes that complicate detection and segmentation.

NC4K[32], the largest image-level COD test dataset currently available, contains 4121 camouflaged scene images sourced from the internet, each annotated at both object and instance levels. The dataset covers predominantly natural scenes along with some

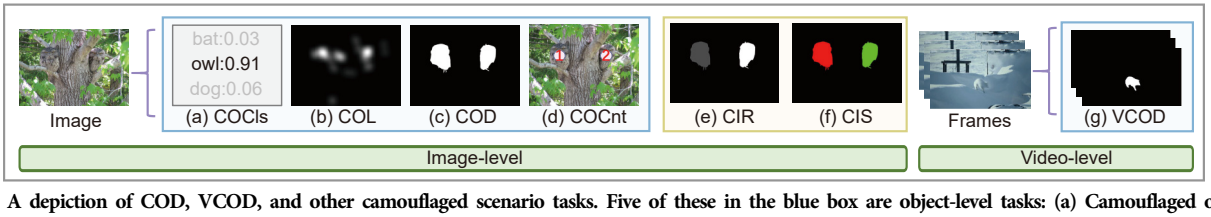

Fig. 6 A depiction of COD, VCOD, and other camouflaged scenario tasks. Five of these in the blue box are object-level tasks: (a) Camouflaged objects classification (COCLs), (b) Camouflaged objects localization (COL), (c) Camouflaged object detection (COD), (d) Camouflaged instance count (COCnt), and (g) Video-level camouflaged object detection (VCOD). The remaining two in the yellow box are instance-level tasks: (e) Camouflaged instance rank (CIR) and (f) Camouflaged instance segmentation (CIS). For further elaboration on these tasks, refer to Section 3.

Table 6 The basic information of both image-level and video-level COD datasets. Level: data type of dataset, i.e., image (I) and video (V). Train/Test: number of samples for training/testing, e.g., images for image dataset or frames for video dataset. N. Cam.: whether collecting non-camouflaged samples. Type: object categories of datasets. Cls.: whether providing classification labels for COCls. If so, the number of categories are provided. Fix.: whether providing fixation annotation for COL. B. Box: whether providing bounding box labels. Obj.: whether providing object-level segmentation masks. Ins.: whether providing instance-level segmentation masks for CIS. Ran.: whether providing ranking labels for CIR. Scr.: whether providing weakly-supervised labels in scribbled form. Gro.: whether providing corresponding category annotation within group images for CoCOD. Ref.: whether providing referring images for RefCOD. Uns.: whether providing unseen classes for OVCOS.

Index	Dataset	Year	Level	Statistics			Object		Label									
				Total	Train	Test	N.Cam.	Type	Cls.	Fix.	B.Box	Obj.	Ins.	Ran.	Scr.	Gro.	Ref.	Uns.
1	CAD2016[214]	2016	V	836	0	836	–	Animals	6	–	–	√	–	–	–	–	–	–
2	CHAMELEON[215]	2017	I	76	0	76	–	Animals	–	–	–	√	–	–	–	–	–	–
3	CPD1K[216]	2018	I	1000	0	1000	–	People	–	–	–	√	–	–	–	–	–	–
4	CAMO-COCO[40]	2019	I	2500	2000	500	√	Animals & humans	–	–	–	√	–	–	–	–	–	–
5	MoCA[191]	2020	V	37 250	0	37 250	–	Animals	67	–	√	–	–	–	–	–	–	–
6	NC4K[32]	2021	I	4121	0	4121	–	Animals	–	–	–	√	√	–	–	–	–	–
7	CAM-FR[32]	2021	I	2280	2000	280	–	Animals & humans	–	√	–	√	√	√	–	–	–	–
8	MoCA-Mask[194]	2022	V	22 939	19 313	3626	–	Animals	44	–	√	√	–	–	–	–	–	–
9	CAMO++[217]	2022	I	5500	3500	2000	√	Animals & humans	93	–	√	√	√	–	–	–	–	–
10	COD10K[12]	2022	I	10 000	6000	4000	√	Animals & humans	78	–	√	√	√	–	–	–	–	–
11	S-COD[92]	2023	I	4040	4040	0	–	Animals & humans	–	–	–	–	–	–	√	–	–	–
12	CAM-LDR[117]	2023	I	6066	4040	2026	–	Animals & humans	–	√	–	√	√	√	–	–	–	–
13	CDS2K[54]	2023	I	2492	0	2492	√	Industrial defect	18	–	√	√	–	–	–	–	–	–
14	CoCOD8K[50]	2023	I	8528	5933	2595	–	Animal & human	70	–	–	√	–	–	–	√	–	–
15	R2C7K[48]	2023	I	6615	–	–	√	Animals & humans	64	–	–	√	–	–	–	–	√	–
16	OVCamo[49]	2024	I	11 483	7713	3770	–	Animal & human	75	–	–	√	–	–	–	–	–	√

artificial camouflage, making it a preferred choice for evaluating the generalization capabilities of COD models.

COD10K[12] includes 10 superclasses and 78 subclasses, totaling 10 000 images, with 60% designated as training data. Within this largest image-level COD dataset to date, there are 5066 camouflaged images (3040 for training and 2026 for testing), 1934 non-camouflaged images, and 3000 background images. All camouflaged images are densely annotated with categories, bounding boxes, and object and instance level labels, supporting a wide range of research tasks, e.g., COD, COS, and CIS. The high-quality annotations and diversity of COD10K make it an essential dataset for COD research.

S-COD[92] is the first dataset created for weakly supervised learning based on scribble annotations. It uses a tagging process that relies on initial impressions to outline the rough structure of objects, including both foreground and background. The dataset comprises 3040 images from the COD10K[12] training set and 1000 images from the CAMO[40] training set, totaling 4040 samples. Compared to pixel-level annotations, the annotations in S-COD are simpler and more efficient.

CAM-LDR[117] facilitates research on COL and CIR by recording the time taken to detect camouflaged instances with an eye tracker, which correlates with the difficulty of detection. This dataset includes 4040 training images, derived from the CAMO[208] and COD10K[12] training sets, and 2026 testing images from the COD10K testing set. Detection times are used to rank camouflaged objects into six categories: background, easy, three medium levels, and hard. This approach offers a novel metric for understanding the detectability of camouflaged objects.

CoCOD8K[50] is the first dataset for CoCOD, which comprises 8 528 images reorganized from four COD datasets—CHAMELEON[215], CAMO[40], COD10K[12], and NC4K[32]. This dataset includes diverse natural and artificial scenes, classified into 5 superclasses and 70 subclasses, each annotated with object masks and category labels. Designed to foster research in detecting co-camouflaged objects across grouped images, CoCOD8K also filters images to fit specific criteria, supporting robust model training with 5933 training and 2595 test images.

R2C7K[48] encompasses 6615 images across 64 categories from real-world scenarios to facilitate RefCOD. It consists of a Camo-subset with 5015 camouflaged images, primarily derived from COD10K[12], and a Ref-subset with 1600 images of salient objects, uniformly sourced with 25 per category from Flickr and Unsplash webpages with no copyright disputes. For research, a referring split is provided where each category in the Ref-subset has 20 images for training and 5 for testing, while the distribution in the Camo-subset follows the original split from COD10K with additional samples from NC4K to ensure at least 6 samples in each category.

OVCamo[49] advances the task of OVCOS by offering a robust dataset featuring 11 483 images across 75 object classes, derived from merged public datasets such as Refs. [12, 194, 214, 216, 217]. This dataset uniquely tackles semantic ambiguities by redefining annotation standards, which ensures clear and distinct class definitions to improve segmentation accuracy. For realistic performance evaluation, OVCamo divides its dataset by allocating 14 seen classes to the training set, while the remaining 61 unseen classes are reserved for testing. This distribution maintains a training-to-testing sample ratio of 7∶3, simulating real-world application challenges and ensuring a rigorous assessment of model generalization.

In current research practices, the most representative datasets typically selected for experiments of image-level COD models are CHAMELEON[215], CAMO-COCO[40], COD10K[12], and NC4K[32].

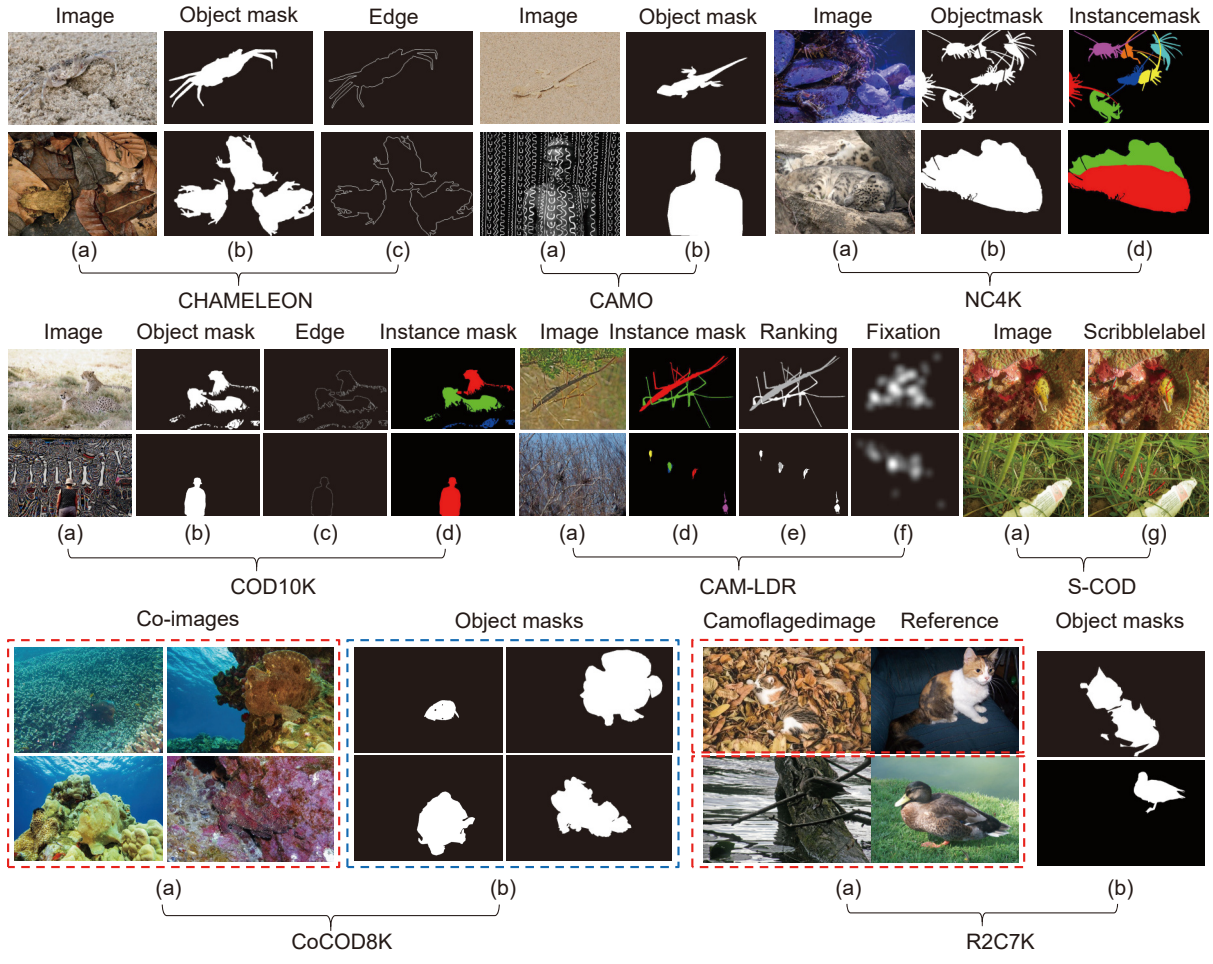

Fig. 7 Examples from eight characteristic COD datasets. Left to right: (a) RGB images, (b) Object-level ground-truths, (c) Edge maps, (d) Instance-level ground-truths, (e) Ranking maps, (f) fixation maps and (g) Scribble annotation. The red and blue boxes indicate a set of inputs and outputs, respectively. Refer to Section 4.1 for more details.

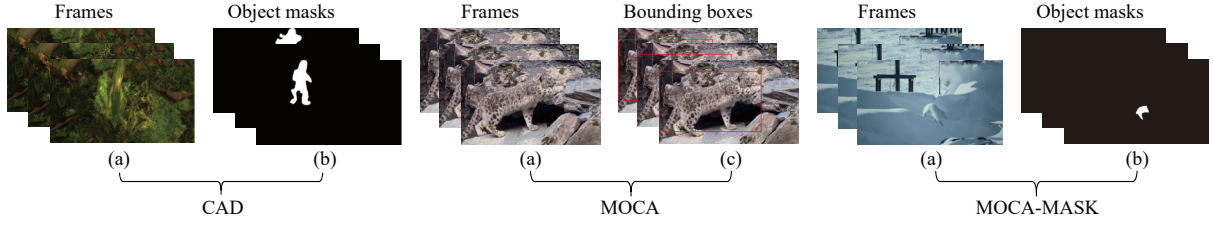

Fig. 8 Examples from three characteristic VCOD datasets. Left to right: (a) RGB Frames, (b) Object-level ground-truths, (c) Bounding boxes. Refer to Section 4.1 for more details.

The common setup[12] for training includes 1000 images from CAMO and 3040 images from COD10K. The remaining parts of these datasets are used to test the generalization ability and viability of the models. To maintain consistency in our analysis, we will adopt this setup in our subsequent performance comparison of image-level COD models.

4.1.2 Characteristic VCOD datasets

CAD2016[214] is composed of nine short video sequences sourced from YouTube, total having 836 frames. Each sequence is manually annotated every five frames. The camouflaged objects in these images are exclusively biological entities found in natural settings.

MoCA[191] is currently the largest dataset for camouflaged animal detection in video format. It consists of 141 video sequences, also sourced from YouTube, representing 67 different categories of animals found in natural scenarios. The dataset spans over 37 250 frames and 26 minutes of video content. Annotations include PWC-Net optical flow data for each frame and bounding boxes with motion labels provided every five frames, with linear interpolation used for the intervening frames.

MoCA-Mask[194] is an extension of MoCA[191] and includes 87 video sequences with a total of 22 939 frames, following the removal of irrelevant scenes. This dataset enhances MoCA by providing manually annotated masks every five frames, resulting in 4691 bounding boxes and pixel-level masks. The dataset is divided into a training set that comprises 71 videos (19 313 frames), along with a testing set consisting of 16 videos (3625 frames).

4.2 Evaluation metrics

We evaluate existing COD models using four common evaluation

metrics as recommended in Ref. [12]. These metrics, i.e., S-measure (S_α)[218], F-measure (F_β)[219], mean absolute error (MAE)[220], and E-measure (E_ϕ)[221], provide a comprehensive assessment of performance. Here, we detail these evaluation metrics:

Precision-recall (PR) curve is generated by transforming the input into a binary mask M, which is segmented across a range of thresholds from 0 to 255. Precision (P) and Recall (R) are calculated by comparing the binary mask M with the ground truth mask (G) at each threshold, producing the PR curve. P and R can be calculated as

$$P = \frac{|M(T) \cap G|}{|M(T)|}, \quad R = \frac{|M(T) \cup G|}{|G|} \tag{1}$$

where $M(T)$ is the mask obtained by thresholding the non-binary prediction map at threshold T.

S-measure (S_α) quantifies the spatial structural similarity between the predicted map (C) and the ground truth (G). It combines both object-aware (S_o) and region-aware (S_r) assessments with the following definition:

$$S_\alpha = \alpha S_o + (1-\alpha) S_r \tag{2}$$

where $\alpha \in [0,1]$ is a weighting factor, typically set at 0.5, that balances the contribution of S_o and S_r.

F-measure (F_β) is used to calculate the relationship between Precision (P) and Recall (R). Initially, the input is transformed into a binary mask, M, segmented over a range of thresholds from 0 to 255. P and R are calculated by comparing M with G across these thresholds. The formulas for P and R are defined as follows:

$$P = \frac{|M(T) \cap G|}{|M(T)|}, \quad R = \frac{|M(T) \cup G|}{|G|} \tag{3}$$

where $M(T)$ represents the binary mask obtained by thresholding the non-binary prediction map at threshold T, and $|\cdot|$ denotes the total area of the mask. But F_β further demonstrates the average harmonic mean value between them. The formula for F_β is defined as

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R} \tag{4}$$

with β^2 typically set to 0.3. From the range of thresholds, three variants of F_β are computed: the maximum ($F_\beta^{\max}$), the mean (F_β^{mean}) and the adaptive (F_β^{adj}). Besides, (F_β^{ω}) is also a widely used metric where the P and R are both weighted averages. F_β^{mean} is adopted in this paper.

E-measure (E_ϕ) evaluates both the local and global similarity between C and G. It is defined as follows:

$$E_\phi = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \phi[C(x,y), G(x,y)] \tag{5}$$

where ϕ represents an enhanced alignment matrix, and W and H are the width and height of the input image, respectively. E_ϕ also provides three indicative values: maximum ($E_\phi^{\max}$), mean (E_ϕ^{mean}), and adaptive (E_ϕ^{adj}). E_ϕ^{mean} is adopted for evaluation in this survey.

Mean absolute error (M) quantifies the average absolute difference per pixel between the normalized predicted map C and the ground truth map G, where $C, G \in [0,1]$. The mean absolute error M is formulated as follows:

$$M = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} |C(x,y) - G(x,y)| \tag{6}$$

where W and H denote the width and height of the input image, and (x,y) represents the pixel coordinates. Unlike F_β, E_ϕ, and S_α, a lower M suggests a more accurate model.

For VCOD, mean Dice (mDice)[19] for similarity evaluation and mean intersection over union (mIoU) for overlap measurement are also used for evaluation. Notice that larger mDice and mIoU scores indicate better performance.

4.3 Quantitative analysis

Results of deep COD models. In this section, we conduct experiments on 40 cutting-edge techniques, covering six strategies mentioned earlier, and categorize them based on their backbones into two types: convolution-based and transformer-based. As shown in Table 7.

PopNet[101] stands out as a top performer, achieving the highest S_α and F_β scores across multiple datasets. Its success is attributed to its innovative integration of source-free depth estimation and object-popping techniques, followed by the precise separation of objects from their contact surfaces. Overall, methods utilizing transformer-based backbones, such as FSNet[120] and HitNet[31], exhibit superior performance compared to those using convolution-based backbones. The self-attention mechanism inherent in transformers is highly effective in modeling long-range dependencies, which are crucial for accurate COD. Furthermore, the results reveal that certain strategies are particularly effective for specific datasets. For example, methods incorporating multi-scale context information, such as C^2F-Net-V2[86], tend to perform better on datasets with varying object sizes, such as COD10K-test. Similarly, mechanism simulation strategies, such as LSR-V2[117], are beneficial for datasets with complex backgrounds, like CAMO-test. Interestingly, GenSAM[68], despite incorporating advanced models like CLIP and BLIP2, does not consistently outperform other methods. This underscores that while powerful pre-trained models can be beneficial, their effectiveness also depends on their integration into the overall COD framework. Additionally, models designed for more challenging settings, such as UCOS-DA[51] for UCOS and MLKG[47] for RefCOD, do not always achieve outstanding performance. These models require more sophisticated approaches and larger datasets to generalize effectively. Consequently, some researchers are developing new datasets to train proposed models, such as R2CNet[48], OVCoser[49], and BBNet[50].

Results of deep VCOD models. We compare 17 cutting-edge methods across two camouflaged video datasets, as detailed in Table 8. This includes 7 image-based methods, 2 related video-oriented object segmentation methods, i.e., RCRNet[223] and PNS-Net[222], and 8 deep VCOD methods.

ZoomNeXt[149] stands out as the most effective VCOD approach, achieving state-of-the-art results in S_α, mDice and mIoU. This is mainly due to the zoom-in-and-out operation and scale integration for robust feature fusion and efficient temporal modeling in a unified framework. Video-based methods overall outperform their image-based counterparts across several key metrics, which is attributed to their ability to exploit temporal consistency across frames, which helps distinguish camouflaged objects from their backgrounds even when they are moving at a high speed. The end-to-end frameworks, e.g., TMNet[200], IMEX[201] and EMIP[203], demonstrate the efficacy of integrating temporal modeling directly into the network architecture. In contrast, two-stage frameworks, like SLT-Net[194] and MG[192], perform slightly worse, which follow an explicit or implicit motion-based approach, first estimating optical flow or pseudo masks before performing COD. While effective, this decoupled strategy may

Table 7 Quantitative comparison on the CHAMELEON, CAMO, COD10K and NC4K testing datasets, where the competing approaches are classified based on their use of convolution-based or transformer-based backbone. Red and blue indicate the best and the second best within each category, respectively.

	Model	Year	Backbone	CHAMELEON				CAMO-test				COD10K-test				NC4K			
				$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$
Convolution-based backbone	UR-SINet[70]	2021	ResNet50	0.876	0.824	0.942	0.031	0.741	0.649	0.804	0.091	0.775	0.643	0.869	0.041	0.806	0.731	0.873	0.057
	LSR[32]	2021	ResNet50	0.842	0.794	0.896	0.046	0.708	0.645	0.755	0.105	0.760	0.658	0.831	0.045	0.797	0.758	0.854	0.061
	PFNet[34]	2021	ResNet50	0.882	0.828	0.931	0.033	0.782	0.746	0.841	0.085	0.800	0.701	0.877	0.040	0.829	0.784	0.887	0.053
	MGL[71]	2021	ResNet50	0.893	0.833	0.917	0.031	0.775	0.726	0.812	0.088	0.814	0.711	0.852	0.035	0.833	0.782	0.867	0.052
	C^2F-Net[27]	2021	Res2Net50	0.888	0.828	0.935	0.032	0.796	0.762	0.854	0.080	0.813	0.723	0.890	0.036	0.838	0.795	0.897	0.049
	TINet[69]	2021	ResNet50	0.874	0.783	0.916	0.038	0.781	0.728	0.836	0.087	0.793	0.679	0.861	0.042	0.829	0.773	0.879	0.055
	BASNet[77]	2021	ResNet34	0.847	0.795	0.883	0.044	0.615	0.503	0.671	0.124	0.661	0.486	0.729	0.071	0.696	0.610	0.762	0.095
	UGTR[8]	2021	ResNet50	0.888	0.819	0.910	0.031	0.785	0.738	0.823	0.086	0.818	0.712	0.853	0.035	0.839	0.787	0.874	0.052
	FAP-Net[41]	2022	Res2Net50	0.893	0.842	0.940	0.028	0.815	0.776	0.865	0.076	0.822	0.731	0.888	0.036	0.851	0.810	0.899	0.047
	FindNet[87]	2022	Res2Net	0.895	0.845	0.944	0.027	0.803	0.763	0.862	0.077	0.811	0.706	0.883	0.036	0.841	0.802	0.895	0.048
	PreyNet[80]	2022	ResNet50	0.895	0.859	0.951	0.028	0.790	0.757	0.842	0.077	0.813	0.736	0.881	0.034	0.834	0.803	0.887	0.050
	BGNet[84]	2022	Res2Net50	0.901	0.860	0.943	0.027	0.812	0.789	0.870	0.073	0.831	0.753	0.901	0.033	0.851	0.820	0.907	0.044
	C^2F-Net-V2[86]	2022	Res2Net50	0.893	0.836	0.947	0.028	0.799	0.770	0.859	0.077	0.811	0.725	0.887	0.036	0.840	0.802	0.896	0.048
	SegMaR[35]	2022	ResNet50	0.906	0.872	0.951	0.025	0.815	0.795	0.874	0.071	0.833	0.757	0.899	0.034	0.841	0.821	0.896	0.046
	ZoomNet[36]	2022	ResNet50	0.902	0.864	0.943	0.023	0.820	0.794	0.877	0.066	0.838	0.766	0.888	0.029	0.853	0.818	0.896	0.043
	OCENet[90]	2022	ResNet50	0.901	0.843	0.940	0.028	0.802	0.766	0.852	0.080	0.827	0.741	0.894	0.033	0.853	0.818	0.902	0.045
	ERRNet[52]	2022	ResNet50	0.877	0.825	0.927	0.036	0.779	0.729	0.842	0.085	0.786	0.675	0.867	0.043	0.827	0.778	0.887	0.054
	SINet-V2[12]	2022	Res2Net50	0.888	0.835	0.942	0.030	0.820	0.782	0.882	0.070	0.815	0.718	0.887	0.037	0.847	0.805	0.903	0.048
	MRR-Net[122]	2023	ResNet50	0.882	0.821	0.931	0.033	0.811	0.772	0.869	0.076	0.822	0.730	0.889	0.036	0.848	0.801	0.898	0.049
	UJSCOD-V2[129]	2023	ResNet50	0.892	0.848	0.948	0.025	0.803	0.768	0.858	0.071	0.817	0.733	0.895	0.033	0.856	0.824	0.913	0.040
	PUENet[42]	2023	ResNet50	0.888	0.844	0.943	0.030	0.794	0.762	0.857	0.080	0.813	0.727	0.887	0.035	0.836	0.798	0.892	0.050
	WS-SAM[28]	2023	ResNet50	0.824	0.777	0.897	0.046	0.759	0.742	0.818	0.092	0.803	0.719	0.878	0.038	0.829	0.802	0.886	0.052
	FEDER[9]	2023	ResNet50	0.894	0.855	0.947	0.028	0.807	0.785	0.873	0.069	0.823	0.740	0.900	0.032	0.846	0.817	0.905	0.045
	PopNet[101]	2023	Res2net50	0.917	0.885	0.965	0.020	0.808	0.784	0.859	0.077	0.851	0.786	0.910	0.028	0.861	0.833	0.909	0.042
	DGNet[108]	2023	EfficientNet	0.890	0.834	0.938	0.029	0.839	0.806	0.901	0.057	0.822	0.728	0.896	0.033	0.857	0.814	0.911	0.042
	LSR-V2[117]	2023	ResNet50	0.878	0.828	0.929	0.034	0.789	0.751	0.840	0.079	0.805	0.711	0.880	0.037	0.840	0.801	0.896	0.048
	Camouflageator[19]	2024	ResNet50	0.903	0.863	0.952	0.026	0.829	0.805	0.891	0.066	0.843	0.763	0.920	0.028	0.869	0.835	0.922	0.041
Transformer-based backbone	CamoFormer[136]	2024	Swin-B	0.898	0.867	0.945	0.025	0.876	0.856	0.930	0.043	0.838	0.753	0.916	0.029	0.888	0.863	0.937	0.031
	FRINet[93]	2023	ViT	0.905	0.871	0.949	0.023	0.865	0.848	0.924	0.046	0.864	0.810	0.930	0.023	0.889	0.866	0.937	0.030
	UCOS-DA[51]	2023	DINO	0.715	0.629	0.802	0.095	0.701	0.646	0.784	0.127	0.689	0.546	0.740	0.086	0.755	0.689	0.819	0.085
	PUENet[42]	2023	ViT	0.910	0.869	0.957	0.022	0.878	0.860	0.930	0.045	0.873	0.812	0.938	0.022	0.898	0.874	0.945	0.028
	OAFormer[105]	2023	PVTv2	0.904	0.868	0.961	0.023	0.867	0.849	0.924	0.048	0.860	0.795	0.927	0.025	0.883	0.857	0.934	0.032
	XMSNet[96]	2023	PVT	0.904	0.895	0.950	0.025	0.864	0.871	0.923	0.048	0.861	0.828	0.927	0.024	0.879	0.877	0.933	0.034
	FSNet[120]	2023	SwinT	0.905	0.868	0.963	0.022	0.880	0.861	0.933	0.041	0.870	0.810	0.938	0.023	0.891	0.866	0.940	0.031
	FSPNet[30]	2023	ViT	0.908	0.867	0.943	0.023	0.857	0.830	0.899	0.050	0.851	0.769	0.895	0.026	0.879	0.843	0.915	0.035
	HitNet[31]	2023	PVTv2	0.921	0.900	0.967	0.019	0.849	0.831	0.906	0.055	0.871	0.823	0.935	0.023	0.875	0.853	0.926	0.037
	MLKG[47]	2023	SAM, CLIP	0.935	0.923	0.941	0.020	0.828	0.806	0.877	0.075	0.910	0.838	0.916	0.019	0.900	0.872	0.918	0.036
	VSCode[39]	2024	Swin-S	–	–	–	–	0.873	0.844	0.925	0.046	0.869	0.806	0.931	0.023	0.891	0.863	0.935	0.032
	CamoFocus[137]	2024	PVTv2	0.912	0.884	0.957	0.023	0.873	0.861	0.926	0.043	0.873	0.818	0.935	0.021	0.889	0.870	0.936	0.030
	GenSAM[68]	2024	CLIP, BLIP2	0.764	0.680	0.807	0.090	0.719	0.659	0.775	0.113	0.775	0.681	0.838	0.067	–	–	–	–

introduce errors that propagate through the pipeline. What's more, based on the powerful SAM, TSP-SAM[189] and SAM-PM[202] also perform well, which showcases the remarkable spatial segmentation versatility and potential of SAM when applied to the challenging task of VCOD.

4.4 Qualitative analysis

As depicted in Fig. 9, we present a comprehensive visual

Table 8 Quantitative comparison on CAD2016 and MoCA-Mask testing dataset. ∗ represents the related video-oriented object segmentation methods.

	Method	Year	Backbone	CAD2016						MoCA-Mask					
				$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$E_\phi \uparrow$	$M \downarrow$	mDice ↑	mIoU ↑	$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$E_\phi \uparrow$	$M \downarrow$	mDice ↑	mIoU ↑
Image-based	SINet[7]	2020	ResNet50	0.601	0.204	0.589	0.089	0.289	0.209	0.574	0.185	0.655	0.030	0.221	0.156
	SINet-V2[12]	2022	Res2Net50	0.544	0.181	0.546	0.049	0.170	0.110	0.571	0.175	0.608	0.035	0.211	0.153
	ZoomNet[36]	2022	ResNet50	0.587	0.225	0.594	0.063	0.246	0.166	0.582	0.211	0.536	0.033	0.224	0.167
	BGNet[84]	2022	Res2Net50	0.607	0.203	0.666	0.089	0.345	0.256	0.590	0.203	0.647	0.023	0.225	0.167
	FEDER[9]	2023	ResNet50	0.607	0.246	0.725	0.061	0.361	0.257	0.555	0.158	0.542	0.049	0.192	0.132
	FSPNet[30]	2023	ViT	0.539	0.220	0.553	0.145	0.309	0.212	0.594	0.182	0.608	0.044	0.238	0.167
	PUENet[42]	2023	ViT	0.673	0.427	0.803	0.034	0.499	0.389	0.594	0.204	0.619	0.037	0.302	0.212
Video-based	RCRNet*[222]	2019	ResNet50	0.627	0.287	0.666	0.048	0.309	0.229	0.597	0.174	0.583	0.025	0.194	0.137
	PNS-Net*[223]	2021	ResNet50	0.678	0.369	0.720	0.043	0.409	0.308	0.576	0.134	0.562	0.038	0.189	0.133
	MG[192]	2021	VGG-style	0.484	0.314	0.558	0.370	0.351	0.260	0.547	0.165	0.537	0.095	0.197	0.141
	SLT-Net[194]	2022	PVT	0.704	0.524	0.912	0.028	0.543	0.438	0.656	0.357	0.785	0.021	0.387	0.310
	ZoomNeXt[149]	2024	PVTv2	0.757	0.593	0.865	0.020	0.599	0.510	0.734	0.476	0.497	0.010	0.497	0.422
	TMNet[200]	2024	SegFormer	0.740	0.485	0.735	0.008	0.503	0.417	0.740	0.485	0.735	0.008	0.503	0.417
	IMEX[201]	2024	ResNet50	0.684	0.452	0.813	0.033	0.469	0.370	0.661	0.371	0.778	0.020	0.409	0.319
	TSP-SAM[189]	2024	SAM	0.704	0.524	0.912	0.028	0.543	0.438	0.689	0.444	0.808	0.008	0.458	0.388
	SAM-PM[202]	2024	SAM	0.729	0.602	0.746	0.018	0.594	0.493	0.695	0.464	0.732	0.011	0.497	0.416
	EMIP[203]	2024	PVTv2	0.719	0.514	–	0.028	0.536	0.425	0.675	0.381	–	0.015	0.426	0.333

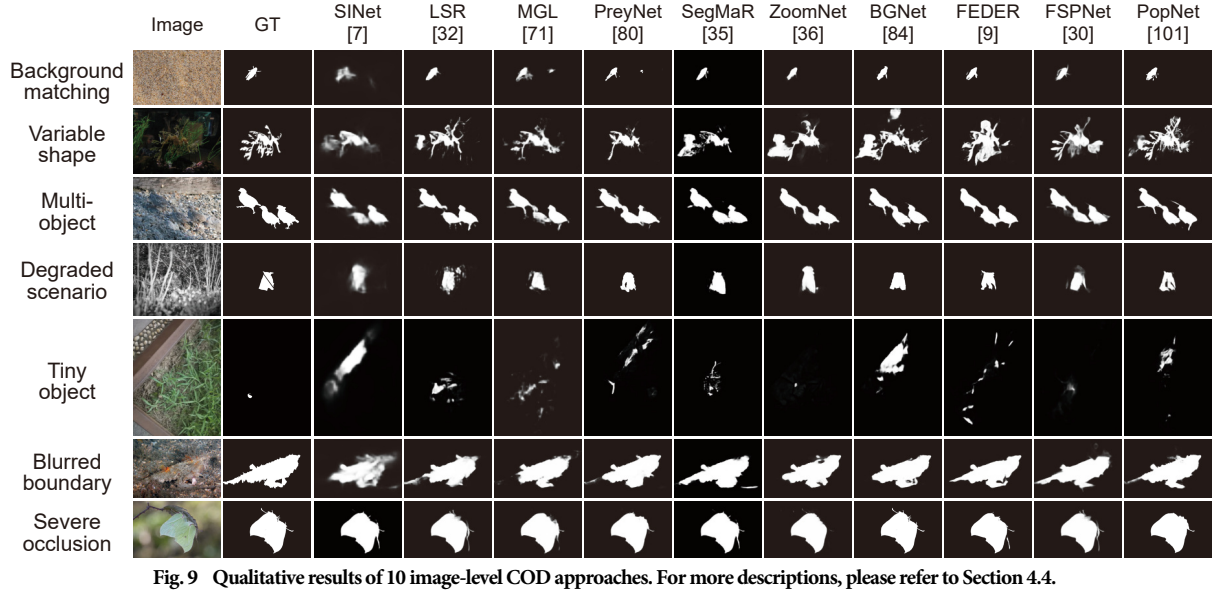

Fig. 9 Qualitative results of 10 image-level COD approaches. For more descriptions, please refer to Section 4.4.

comparison of 10 cutting-edge image-level COD methods. We selected various challenging camouflaged images across 7 typical complex scenarios, including background matching, variable shape, multi-object environments, degraded scenarios, tiny objects, blurred boundaries, and severe occlusion. Additional, more challenging scenarios for COD are discussed in Section 6, and typical examples can also be seen in Fig. 10.

In the background-matching scenario, where insects blend seamlessly with their surroundings, we observe that SINet[7] struggles to identify insects, often resulting in missed detections. Conversely, SegMaR[35] demonstrates robustness by effectively outlining insect contours, even under low-contrast conditions. In the variable shape scenario, such as with the leafy sea dragon, whose appendages resemble leaves, the adaptability of models is tested. While ZoomNet[36] misidentifies parts of these large appendages as separate objects, FPNet[95] exhibits a superior ability to segment more complete instances, illustrating its robustness to shape variations.

Under multi-object configurations, where the scene is crowded with numerous camouflaged entities, the models generally succeed in locating all targets but struggle with accurate segmentation. However, PopNet[101] performs commendably, achieving better coverage without excessive false positives. Degraded scenarios, including poor lighting or blurring, significantly impact the localization and identification processes of the models. The detection results across all models fall short of expectations, indicating considerable room for improvement in these challenging conditions.

For tiny objects, prediction precision degrades, with most methods failing to detect them accurately. However, ZoomNet[36],

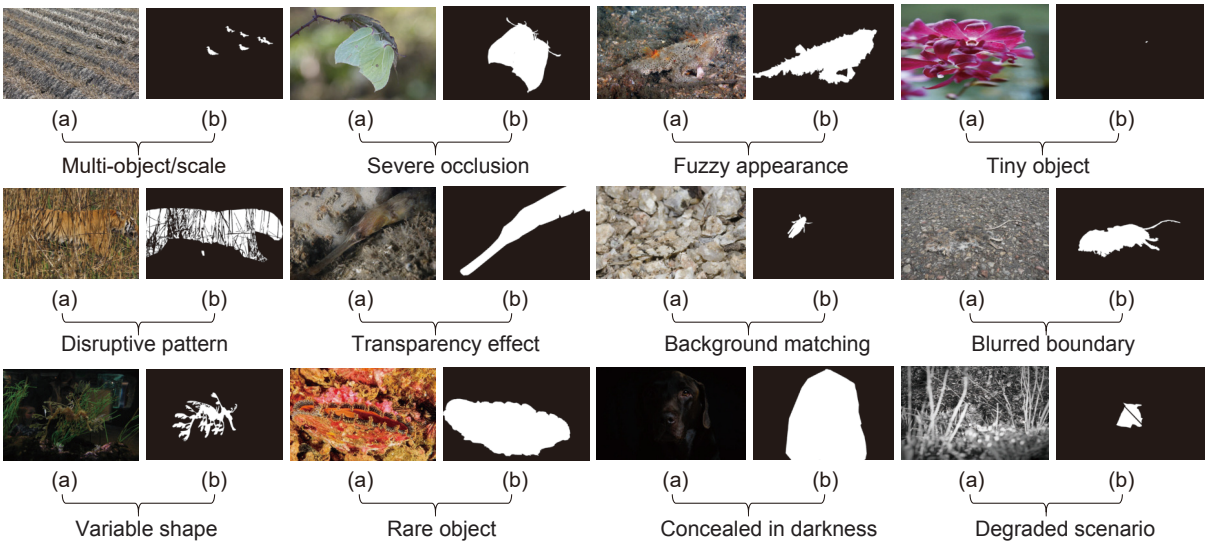

Fig. 10 Illustration of 12 exemplary challenging samples in extremely complex scenarios. Left to right: (a) RGB images, (b) Object-level ground-truths.

with its zoom-in-and-out operation, shows potential in highlighting even the smallest targets. In blurred boundary scenarios, defining precise contours becomes challenging. Here, FEDER[9] stands out with its ODE-inspired edge reconstruction for complete edge prediction, whereas other methods often produce incomplete boundary predictions. Lastly, in severe occlusion scenarios, where two insects partially obscure each other, the results from all methods are generally acceptable. Nonetheless, BGNet[84] displays an improved capacity through its edge-guidance feature module, though further enhancement in detail detection is needed.

In conclusion, while all cutting-edge methods demonstrate acceptable performance across various challenging COD scenarios, there are still notable limitations, particularly in degraded conditions, tiny objects, and blurred boundaries. Further research focusing on enhancing robustness and precision under these extreme conditions is crucial for advancing the field.

5 Future Direction

5.1 Mitigating current issues

Deep generative models for data scarcity. To mitigate the scarcity of data, leveraging deep generative models to synthesize diverse, realistic camouflaged images will bolster training effectiveness by dataset augmentation[224], enhancing model robustness in dealing with camouflaged scenarios. With the rise of image generation models represented by GANs[178] and diffusion models[213], diverse and high-quality images can now be controlled and generated using other multimodal inputs such as text. This advancement can effectively address the longstanding issue of dataset scarcity in this field. By training with both generated camouflaged object data and the original datasets, performance can be significantly improved. Adopting an adversarial manner, where the camouflaged sample generator and the segmentation network are pitted against each other, may help unlock further potential[19]. However, there are no quantitative metrics to evaluate the camouflage degree and sample quality in deep datasets, making the widespread use of generated samples for training a topic of debate. For VCOD datasets, the generation of such data still has a long way to go due to the current immaturity and lack of a general framework in video generation technology. Diffusion models have become a research hotspot in the field of computer vision due to their stability in algorithm training and high-quality sample generation. CamDiff[225] innovates by synthesizing salient objects within camouflage scenes, mitigating the scarcity of multi-pattern training data and enhancing robustness to salient misclassifications. The authors also use CamDiff to propose Diff-COD dataset from the original COD datasets[12, 32, 40, 215] to enhance the robustness to saliency. Meanwhile, LAKE-RED[226] tackles the limitations of dataset diversity and expensive data collection by automatically generating camouflage images without manual background specification, fostering scalability and interpretability. Both approaches emphasize the potential of diffusion models to address data-scarce vision tasks.

Tackling complex scenarios & challenging samples. Addressing the intricacies of COD requires grappling with various challenges posed by complex scenes and difficult samples. Key issues include extremely complex backgrounds and advanced camouflage techniques that render objects nearly invisible. Figure 10 depicts some extremely concealed scenarios with challenging samples. Multi-object, multi-scale scenarios present difficulties in capturing objects of varying sizes and scales, while severely occluded objects demand robust algorithms capable of disentangling overlapping structures. Objects with fuzzy appearances, particularly small ones, challenge the limits of detection frameworks due to their minimal visual cues. Disruptive patterns, transparency, and background matching further complicate detection by seamlessly blending objects into their surroundings. Blurred boundaries and structural ambiguity add to the difficulty of delineating object edges, while variable shapes require flexible recognition capabilities. Moreover, detecting extremely rare camouflaged objects necessitates specialized methods to distinguish them from the vast majority of non-camouflaged instances. Objects concealed in darkness, affected by drastic illumination variations, pose unique challenges that require innovative approaches to handle varying light conditions. Overcoming these obstacles demands the development of advanced detection models capable of adapting to diverse and extreme scenarios, ensuring continued progress and real-world applicability of COD research. Additionally, performance degradation in challenging environments, such as low-light conditions and foggy settings, exacerbates the difficulty of detecting camouflaged objects[154, 227, 228]. Low light intensifies contrast issues, making camouflaged objects nearly invisible against their surroundings, while fog introduces noise and

occlusion, further complicating feature extraction and localization. Enhancing robustness in such extreme conditions requires innovations in image enhancement techniques[159, 172, 229], coupled with adaptive feature learning strategies specifically tailored for degraded images. In actual application scenarios, the above problems are very common, for example, in the agricultural domain, detecting numerous small and concealed crops amidst severe occlusions highlights the need for precision and robustness. To address this, Wang et al.[14] proposed a depth-aware concealed crop detection method for dense agricultural scenes, along with the corresponding dataset ACOD-12K, specifically designed to tackle COD-related agricultural tasks.

Annotation-efficient learning in limited conditions. In the context of COD under constrained conditions, annotation-efficient learning emerges as a crucial approach to addressing the scarcity of labeled data. Strategies such as few/zero-shot learning, weakly supervised learning, and open-world learning aim to alleviate the challenge of exhaustive annotation requirements. Specifically, few/zero-shot learning[119, 134] addresses the detection of unseen object classes by leveraging knowledge transfer from related categories. Weakly supervised learning[28, 92], which relies on image-level labels to identify object locations, provides a less precise but still valuable alternative. Unsupervised learning[51], through techniques like clustering and representation learning, uncovers patterns in unlabeled data, thereby enhancing model generalization. Self-supervised learning further contributes by exploiting inherent data properties to generate pseudo-labels, fostering the development of robust representations. The limitation of training data also necessitates innovative data augmentation and domain adaptation techniques to combat overfitting. Open-world applications[49] introduce unique challenges, requiring models to detect novel objects while maintaining performance on known classes. Collectively, these approaches strive to overcome the difficulties posed by scarce annotations and unseen object classes, thereby pushing the boundaries of concealed object detection in practical, annotation-constrained scenarios.

Camouflage loss functions. The introduction of camouflage-specific loss functions enhances the training process for COD by directing model optimization toward accurately distinguishing camouflaged regions from salient ones. For instance, SCLoss[141] shows promise by incorporating spatial coherence into the learning process for ambiguous regions. This approach significantly improves the network's ability to discern boundaries and subtle nuances in camouflaged objects. By focusing on both individual pixel responses and their mutual interactions, SCLoss addresses the limitations of single-response loss functions, providing a more comprehensive solution to the challenges posed by camouflaged objects.

Real-time performance constraints. A significant challenge in COD is the lack of real-time performance, primarily due to the high computational and memory requirements of current models. This limitation impedes deployment on resource-constrained devices, restricting real-world applications such as surveillance and autonomous navigation. Train-free learning[68, 132] offers a solution by enabling rapid adaptation without retraining, thereby facilitating quicker deployment. Additionally, green learning-based gradient boosting[138] emphasizes the development of energy-efficient models that reduce computational costs by selectively focusing on challenging samples, thereby optimizing real-time performance. The development of efficient networks[100, 108] with lightweight architectures aims to minimize latency and memory usage, making real-time COD feasible in practical scenarios. Overcoming these constraints is essential for the broader adoption and practical impact of COD technologies.

5.2 Exploring expansive potentials

Embracing novel tasks. CoCOD[50], a collaborative approach to COD, holds the potential to significantly enhance performance by leveraging multi-source data and cross-modal interactions. However, challenges persist in efficiently fusing heterogeneous information and ensuring robust performance across diverse scenarios. Another promising avenue, RefCOD[47, 48], demands advancements in techniques for multi-modal alignment and the comprehension of complex textual or visual references, particularly for rare or ambiguous species. The primary difficulties include bridging cross-modal gaps and extracting fine-grained visual cues that are relevant to the provided textual or visual descriptions. Beyond CoCOD and RefCOD, there are numerous unexplored opportunities for novel tasks that could further empower COD. For instance, interactive COD (I-COD) could incorporate user feedback during detection, enabling iterative refinement and personalization. This would require the development of robust interaction mechanisms and ensuring the system's responsiveness to user inputs. By continually exploring and innovating in these novel tasks, we can substantially enhance the capabilities and applicability of COD, thereby pushing the boundaries of what is possible in this exciting field.

Differentialting SOD and COD. At the feature level, delving into the nuances between SOD and COD is crucial, as both, though under the umbrella of abnormal segmentation, target distinct abnormalities[230, 231]. Current COD techniques struggle with salient objects[9, 19, 46, 232, 233], as they misinterpret prominence as camouflage, thereby necessitating robustness enhancement. Transferring salient objects to concealed scenarios, or vice versa, could alleviate data scarcity and boost model generalization. The partial positive correlation between SOD and COD highlights opportunities for conversion strategies, increasing sample diversity. Integrating generative adversarial mechanisms between the two tasks fosters innovation. Ultimately, a multi-task unified framework that encapsulates domain self-generalization for saliency and camouflage detection within the broader AI landscape represents a compelling frontier.

Multimodal information fusion. Integrating multiple modalities, such as text[39], audio, video[149], optical[192], depth[14], infrared, and 3D, presents a promising yet challenging direction in COD. Each modality offers unique insights but also introduces additional complexities. For instance, RGB-T and thermal infrared modalities can enhance detection in low-light or obscured environments, yet they require robust fusion strategies to effectively combine visual cues with temperature variations. Audio cues, while informative for certain concealed objects, present challenges in accurately localizing and interpreting sounds amidst background noise. Video data[189], though rich in temporal information, necessitates efficient processing techniques to handle high dimensionality and real-time requirements. The key challenges lie in designing robust fusion mechanisms that can leverage the complementary strengths of these modalities while mitigating their individual limitations. Furthermore, annotating multi-modal datasets for concealed objects is both time-consuming and expensive, exacerbating the scarcity of labeled data. Addressing these challenges will require innovative approaches in multi-modal representation learning, fusion algorithms, and data augmentation techniques that are specifically tailored for COD.

Efficiency-oriented large-scale vision-language models. Large

vision-language models (VLMs) have demonstrated superior performance across various tasks, attracting significant attention from researchers. However, training VLMs specifically tailored for COD is not ideal due to limited data availability and the heavy computational demands involved. In this context, training-free prompt engineering has emerged as a promising direction to address the resource-intensive challenges associated with training large models. The primary challenge lies in designing effective prompts that can elicit nuanced understanding from these models without requiring extensive training. Recent advancements, such as SAM[174, 234–236] and LVLMs[68], offer compelling avenues for integrating visual and linguistic modalities, yet efficiency remains a major concern. Developing relatively lightweight models, combined with advanced prompt engineering techniques, could potentially mitigate efficiency issues while still handling diverse and challenging samples. SAM 2[237], as a pioneering effort, significantly advances the frontier of real-time video processing by leveraging a streaming memory transformer architecture. Its success in reducing interactions by 3x in video segmentation underscores the potential for enhancing efficiency in COD tasks, particularly within the video domain. To balance performance with efficiency, we could leverage the strengths of both large and lightweight models, potentially through techniques such as knowledge distillation[238], transfer learning[239], model compression[240], domain adaptation[51], meta-learning[241], or modularized architectures tailored for specific tasks. Additionally, when there is a significant domain gap between training data and test data, designing a plug-and-play component is a promising approach. For instance, adapters[130] enable efficient fine-tuning of large pre-trained models, allowing them to adapt to the nuances of camouflaged objects without the full cost of retraining and thereby significantly reducing computational overheads.

6 Practical Application

COD has found wide application across various fields due to its capability to identify objects that are otherwise difficult to discern. It is particularly valuable in medicine, industry, agriculture, environmental monitoring, and the arts. The significance of COD lies in its ability to enhance precision when detecting subtle or concealed patterns, making it indispensable for tasks that require high accuracy and reliability. As technology advances, the applications of COD continue to expand, emphasizing its critical role in both practical problem-solving and creative industries. Figure 11 provides an illustrative overview of COD's diverse applications[14, 226, 242–246], highlighting key areas where this technology is actively employed.

- **Medical diagnosis.** Polyp image segmentation[247], essential for early cancer detection, benefits significantly from COD's ability to differentiate polyps that are often camouflaged against surrounding tissue. Similarly, COD enhances lung infection segmentation[248], particularly for diseases like COVID-19, where infected areas can be challenging to isolate in medical imaging. By improving diagnostic precision, COD plays a critical role in facilitating more effective treatments and earlier interventions. The future of COD in medical diagnosis holds great promise, with the potential to revolutionize diagnostic accuracy and improve patient outcomes, facilitating disease diagnosis.
- **Industrial inspection.** COD plays a pivotal role in quality control by detecting subtle surface defects[249, 250] in materials such as wood, metal, and fabrics—imperfections that are often overlooked by traditional methods. It is also highly effective in detecting transparent objects[251, 252], such as glass and mirrors, which are challenging due to their low visibility. By accurately identifying these defects, COD minimizes production errors and improves

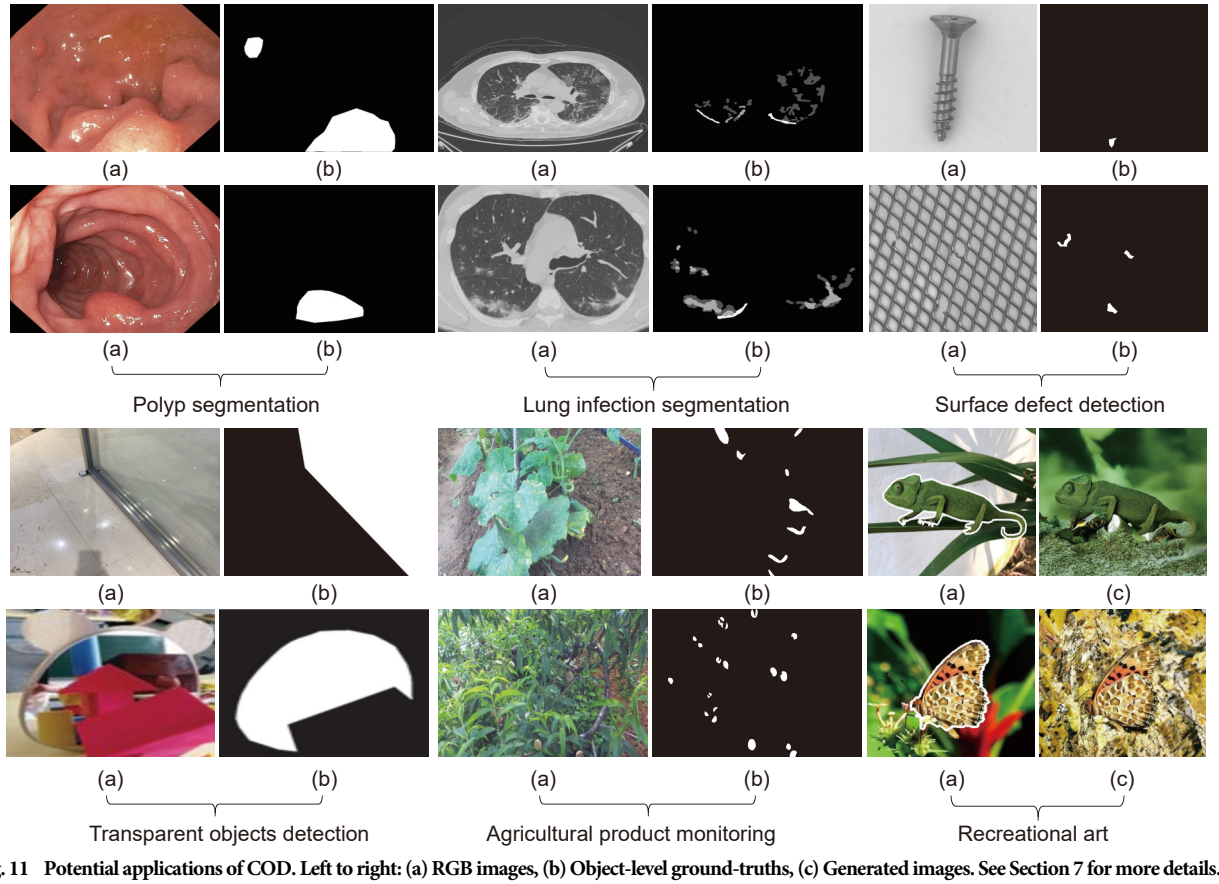

Fig. 11 Potential applications of COD. Left to right: (a) RGB images, (b) Object-level ground-truths, (c) Generated images. See Section 7 for more details.

safety standards, making it indispensable across various industries. Looking ahead, its potential lies in automating inspection processes[253], enhancing efficiency, and addressing more complex inspection tasks across diverse sectors.

- **Agricultural detection.** Monitoring agricultural products[14, 254] and detecting pests[255] present significant challenges due to the subtle visual differences in crops and the ability of pests to blend into their natural surroundings. Traditional detection methods are often slow and imprecise, hindering early intervention. COD addresses these issues by accurately identifying concealed or small-scale changes, thereby facilitating timely action. This early detection improves crop management and pest control, reducing losses and enhancing yields. The applications of COD hold great promise for increasing efficiency and sustainability in agricultural practices.
- **Environmental surveillance.** Detecting changes in biodiversity and responding to natural disasters present significant challenges due to the complexity of environments and the difficulty of identifying camouflaged or hidden objects. In tasks such as species monitoring[256], discovering new species[257], and locating individuals[216, 258] during rescue missions or autonomous driving scenarios, accurate and efficient detection is crucial. COD enhances the ability to discern subtle visual patterns in these contexts, improving the precision of conservation efforts and disaster response. The future of COD in this domain lies in its capacity to automate complex detection, making environmental monitoring faster and more reliable.
- **Artistic creation.** Artistic creation represents a fascinating and increasingly important domain for COD applications, particularly in areas such as style transfer[17, 259], AI-generated artwork[225, 226, 260], and authenticity verification. As digital art and AI-generated content gain traction, COD facilitates the creation of more sophisticated and lifelike blends of styles in recreational art[261]. More critically, it plays a crucial role in detecting forgeries and ensuring the originality of artwork by identifying subtle modifications or tampering in art pieces. In an era where digital media is proliferating, COD's role in safeguarding artistic authenticity holds significant future potential.

7 Conclusion

We present a comprehensive and exhaustive overview of the rapidly evolving field of COD, encompassing both traditional and deep learning methods in image and video domains. By reviewing approximately 150 relevant COD studies, we offer the most extensive and detailed survey to date, providing a concise yet thorough perspective for both newcomers and established scholars. Additionally, we provide both quantitative and qualitative benchmarks for representative image and video models across 6 characteristic datasets and 6 evaluation metrics. Through rigorous benchmarking, we have identified key limitations and challenges in current COD methods, thereby paving the way for future research directions. We hope this survey will inspire innovative solutions that push the boundaries of COD technology. Additionally, we have established a dedicated GitHub repository to house COD techniques, datasets, and resources, ensuring that the latest developments and insights are readily accessible to the research community for further exploration.

Acknowledgment

This work was supported by the STI 2030-Major Projects (No. 2021ZD0201404). The authors express their sincere appreciation to Dr. Deng-Ping Fan for his insightful comments, which greatly improved the quality of this paper.

References

[1] Y. Pu, Y. Han, Y. Wang, J. Feng, C. Deng, and G. Huang, Fine-grained recognition with learnable semantic data augmentation, *IEEE Trans. Image Process.*, vol. 33, pp. 3130–3144, 2024.

[2] Y. Pu, Y. Wang, Z. Xia, Y. Han, Y. Wang, W. Gan, Z. Wang, S. Song, and G. Huang, Adaptive rotated convolution for rotated object detection, in *Proc. 2023 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Paris, France, 2023, pp. 6566–6577.

[3] Y. Pu, W. Liang, Y. Hao, Y. Yuan, Y. Yang, C. Zhang, H. Hu, and G. Huang, Rank-DETR for high quality object detection, in *Proc. 37th Conf. Neural Information Processing Systems* (*NeurIPS 2023*), New Orleans, LA, USA, 2023, pp. 16100–16113.

[4] Y. Wang, Z. Li, J. Mei, Z. Wei, L. Liu, C. Wang, S. Sang, A. L. Yuille, C. Xie, and Y. Zhou, SwinMM: masked multi-view with swin transformers for 3D medical image segmentation, in *Proc. 26th Int. Conf. Medical Image Computing and Computer Assisted Intervention* (*MICCAI 2023*), Vancouver, Canada, 2023, pp. 486–496.

[5] J. Zhao, J. J. Liu, D. P. Fan, Y. Cao, J. Yang, and M. M. Cheng, EGNet: Edge guidance network for salient object detection, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Seoul, Republic of Korea, 2019, pp. 8779–8788.

[6] Z. Wu, L. Su, and Q. Huang, Cascaded partial decoder for fast and accurate salient object detection, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 3907–3916.

[7] D. P. Fan, G. P. Ji, G. Sun, M. M. Cheng, J. Shen, and L. Shao, Camouflaged object detection, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 2777–2787.

[8] 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 *Proc. 2021 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Montreal, Canada, 2021, pp. 4146–4155.

[9] C. He, K. Li, Y. Zhang, L. Tang, Y. Zhang, Z. Guo, and X. Li, Camouflaged object detection with feature decomposition and edge reconstruction, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 22046–22055.

[10] 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 *Proc. 13th European Conf. Computer Vision* (*ECCV*), Zurich, Switzerland, 2014, pp. 740–755.

[11] 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 *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 136–145.

[12] D. P. Fan, G. P. Ji, M. M. Cheng, and L. Shao, Concealed object detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6024–6042, 2022.

[13] A. Kumar, Computer-vision-based fabric defect detection: A survey, *IEEE Trans. Ind. Electron.*, vol. 55, no. 1, pp. 348–363, 2008.

[14] L. Wang, J. Yang, Y. Zhang, F. Wang, and F. Zheng, Depth-aware concealed crop detection in dense agricultural scenes, in *Proc. 2024 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2024, pp. 17201–17211.

[15] D. J. A. Rustia, C. E. Lin, J. Y. Chung, Y. J. Zhuang, J. C. Hsu, and T. T. Lin, Application of an image and environmental sensor network for automated greenhouse insect pest monitoring, *J. Asia Pac. Entomol.*, vol. 23, no. 1, pp. 17–28, 2020.

[16] D. P. Fan, G. P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, PraNet: Parallel reverse attention network for polyp segmentation,

in *Proc. 23rd Int. Conf. Medical Image Computing and Computer Assisted Intervention* (*MICCAI 2020*), Lima, Peru, 2020, pp. 263–273.

[17] H. K. Chu, W. H. Hsu, N. J. Mitra, D. Cohen-Or, T. T. Wong, and T. Y. Lee, Camouflage images, *ACM Trans. Graph.*, vol. 29, no. 4, pp. 1–8, 2010.

[18] X. Suo, Y. Jiang, P. Lin, Y. Zhang, M. Wu, K. Guo, and L. Xu, NeuralHumanFVV: Real-time neural volumetric human performance rendering using RGB cameras, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 6226–6237.

[19] C. He, K. Li, Y. Zhang, Y. Zhang, Z. Guo, X. Li, M. Danelljan, and F. Yu, Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects, in *Proc. 12h Int. Conf. Learning Representations* (*ICLR*), Vienna, Austria, 2024.

[20] S. Li, D. Florencio, W. Li, Y. Zhao, and C. Cook, A fusion framework for camouflaged moving foreground detection in the wavelet domain, *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3918–3930, 2018.

[21] Galun, Sharon, Basri, and Brandt, Texture segmentation by multiscale aggregation of filter responses and shape elements, in *Proc. 9th IEEE Int. Conf. Computer Vision*, Nice, France, 2003, pp. 716–723.

[22] A. Tankus and Y. Yeshurun, Detection of regions of interest and camouflage breaking by direct convexity estimation, in *Proc. 1998 IEEE Workshop on Visual Surveillance*, Bombay, India, 1998, pp. 42–48.

[23] C. Pulla Rao, A. Guruva Reddy, and C. B. Rama Rao, Camouflaged object detection for machine vision applications, *Int. J. Speech Technol.*, vol. 23, no. 2, pp. 327–335, 2020.

[24] T. E. Boult, R. J. Micheals, X. Gao, and M. Eckmann, Into the woods: Visual surveillance of noncooperative and camouflaged targets in complex outdoor settings, *Proc. IEEE*, vol. 89, no. 10, pp. 1382–1402, 2001.

[25] Y. Beiderman, M. Teicher, J. Garcia, V. Mico, and Z. Zalevsky, Optical technique for classification, recognition and identification of obscured objects, *Opt. Commun.*, vol. 283, no. 21, pp. 4274–4282, 2010.

[26] A. Mittal and N. Paragios, Motion-based background subtraction using adaptive kernel density estimation, in *Proc. 2004 IEEE Computer Society Conf. Computer Vision and Pattern Recognition* (*CVPR 2004*), Washington, DC, USA, 2004.

[27] Y. Sun, G. Chen, T. Zhou, Y. Zhang, and N. Liu, Context-aware cross-level fusion network for camouflaged object detection, in *Proc. 30th Int. Joint Conf. Artificial Intelligence* (*IJCAI-21*), virtual, pp. 1025–1031.

[28] C. He, K. Li, Y. Zhang, G. Xu, L. Tang, Y. Zhang, Z. Guo, and X. Li, Weakly-supervised concealed object segmentation with SAM-based pseudo labeling and multi-scale feature grouping, in *Proc. 37th Conf. Neural Information Processing Systems* (*NeurIPS 2023*), New Orleans, LA, USA, 2023, pp. 30726–30737.

[29] Z. Chen, Y. Zhang, J. Gu, Y. Zhang, L. Kong, and X. Yuan, Cross aggregation transformer for image restoration, in *Proc. 36th Conf. Neural Information Processing Systems* (*NeurIPS 2022*), New Orleans, LA, USA, 2022, pp. 25478–25490.

[30] Z. Huang, H. Dai, T. Z. Xiang, S. Wang, H. X. Chen, J. Qin, and H. Xiong, Feature shrinkage pyramid for camouflaged object detection with transformers, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 5557–5566.

[31] X. Hu, S. Wang, X. Qin, H. Dai, W. Ren, D. Luo, Y. Tai, and L. Shao, High-resolution iterative feedback network for camouflaged object detection, *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 1, pp. 881–889, 2023.

[32] 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 *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 11586–11596.

[33] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yang, and F. Yu, Dual aggregation transformer for image super-resolution, in *Proc. 2023 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Paris, France, 2023, 12312–12321.

[34] H. Mei, G. P. Ji, Z. Wei, X. Yang, X. Wei, and D. P. Fan, Camouflaged object segmentation with distraction mining, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 8768–8777.

[35] Q. Jia, S. Yao, Y. Liu, X. Fan, R. Liu, and Z. Luo, Segment, magnify and reiterate: Detecting camouflaged objects the hard way, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, LA, USA, 2022, pp. 4713–4722.

[36] Y. Pang, X. Zhao, T. Z. Xiang, L. Zhang, and H. Lu, Zoom in and out: A mixed-scale triplet network for camouflaged object detection, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, LA, USA, 2022, pp. 2160–2170.

[37] Y. Zhong, B. Li, L. Tang, S. Kuang, S. Wu, and S. Ding, Detecting camouflaged object in frequency domain, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, LA, USA, 2022, pp. 4494–4503.

[38] Z. Chen, Y. Zhang, D. Liu, B. Xia, J. Gu, L. Kong, and X. Yuan, Hierarchical integration diffusion model for realistic image deblurring, in *Proc. 37th Conf. Neural Information Processing Systems* (*NeurIPS 2023*), New Orleans, LA, USA, 2023, pp. 29114–29125.

[39] Z. Luo, N. Liu, W. Zhao, X. Yang, D. Zhang, D. P. Fan, F. Khan, and J. Han, VSCode: General visual salient and camouflaged object detection with 2D prompt learning, in *Proc. 2024 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2024, 17169–17180.

[40] T. N. Le, T. V. Nguyen, Z. Nie, M. T. Tran, and A. Sugimoto, Anabranch network for camouflaged object segmentation, *Comput. Vis. Image Underst.*, vol. 184, pp. 45–56, 2019.

[41] T. Zhou, Y. Zhou, C. Gong, J. Yang, and Y. Zhang, Feature aggregation and propagation network for camouflaged object detection, *IEEE Trans. Image Process.*, vol. 31, pp. 7036–7047, 2022.

[42] Y. Zhang, J. Zhang, W. Hamidouche, and O. Déforges, Predictive uncertainty estimation for camouflaged object detection, *IEEE Trans. Image Process.*, vol. 32, pp. 3580–3591, 2023.

[43] Q. Zhai, X. Li, F. Yang, Z. Jiao, P. Luo, H. Cheng, and Z. Liu, MGL: mutual graph learning for camouflaged object detection, *IEEE Trans. Image Process.*, vol. 32, pp. 1897–1910, 2023.

[44] A. Li, J. Zhang, Y. Lv, B. Liu, T. Zhang, and Y. Dai, Uncertainty-aware joint salient object and camouflaged object detection, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 10071–10081.

[45] Y. Yang and Q. Zhang, Finding camouflaged objects along the camouflage mechanisms, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2346–2360, 2024.

[46] C. Hao, Z. Yu, X. Liu, J. Xu, H. Yue, and J. Yang, A simple yet effective network based on vision transformer for camouflaged object and salient object detection, arXiv preprint arXiv: 2402.18922, 2024.

[47] S. Cheng, G. P. Ji, P. Qin, D. P. Fan, B. Zhou, and P. Xu, Large model based referring camouflaged object detection, arXiv preprint arXiv: 2311.17122, 2023.

[48] X. Zhang, B. Yin, Z. Lin, Q. Hou, D. P. Fan, and M. M. Cheng, Referring camouflaged object detection, arXiv preprint arXiv: 2306.07532, 2023.

[49] Y. Pang, X. Zhao, J. Zuo, L. Zhang, and H. Lu, Open-vocabulary camouflaged object segmentation, presented at 18th European Conf. Computer Vision (ECCV), Milan, Italy, 2024.

[50] C. Zhang, H. Bi, T. Z. Xiang, R. Wu, J. Tong, and X. Wang,

Collaborative camouflaged object detection: A large-scale dataset and benchmark, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 12, pp. 18470–18484, 2024.

[51] Y. Zhang and C. Wu, Unsupervised camouflaged object segmentation as domain adaptation, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 4334–4344.

[52] G. P. Ji, L. Zhu, M. Zhuge, and K. Fu, Fast camouflaged object detection *via* edge-based reversible re-calibration network, *Pattern Recognit.*, vol. 123, pp. 108414, 2022.

[53] H. Bi, C. Zhang, K. Wang, J. Tong, and F. Zheng, Rethinking camouflaged object detection: Models and datasets, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 5708–5724, 2022.

[54] D. P. Fan, G. P. Ji, P. Xu, M. M. Cheng, C. Sakaridis, and L. Van Gool, Advances in deep concealed scene understanding, *Vis. Intell.*, vol. 1, no. 1, pp. 16, 2023.

[55] Y. Liang, G. Qin, M. Sun, X. Wang, J. Yan, and Z. Zhang, A systematic review of image-level camouflaged object detection with deep learning, *Neurocomputing*, vol. 566, pp. 127050, 2024.

[56] A. Tankus and Y. Yeshurun, Convexity-based visual camouflage breaking, *Comput. Vis. Image Underst.*, vol. 82, no. 3, pp. 208–237, 2001.

[57] M. B. Neider and G. J. Zelinsky, Searching for camouflaged targets: Effects of target-background similarity on visual search, *Vis. Res.*, vol. 46, no. 14, pp. 2217–2235, 2006.

[58] N. U. Bhajantri and P. Nagabhushan, Camouflage defect identification: A novel approach, in *Proc. 9th Int. Conf. Information Technology* (*ICIT'06*), Bhubaneswar, India, 2006, pp. 145–148.

[59] P. Sengottuvelan, A. Wahi, and A. Shanmugam, Performance of decamouflaging through exploratory image analysis, in *Proc. 1st Int. Conf. Emerging Trends in Engineering and Technology*, Nagpur, India, 2008, pp. 6–10.

[60] W. R. Boot, M. B. Neider, and A. F. Kramer, Training and transfer of training in the search for camouflaged targets, *Atten. Percept. Psychophys.*, vol. 71, no. 4, pp. 950–963, 2009.

[61] L. Song and W. Geng, A new camouflage texture evaluation method based on WSSIM and nature image features, in *Proc. Int. Conf. Multimedia Technology*, Ningbo, China, 2010, pp. 1–4.

[62] P. Siricharoen, S. Aramvith, T. H. Chalidabhongse, and S. Siddhichai, Robust outdoor human segmentation based on color-based statistical approach and edge combination, in *Proc. 2010 Int. Conf. Green Circuits and Systems*, Shanghai, China, 2010, pp. 463–468.

[63] K. Chaduvula, B. P. Rao, and A. Govardhan, An efficient content based image retrieval using color and texture of image sub- blocks, https://www.researchgate.net/publication/50392348_An_efficient_content_based_image_retrieval_using_color_and_texture_of_image_sub-blocks, 2011.

[64] Y. Pan, Y. Chen, Q. Fu, P. Zhang, and X. Xu, Study on the camouflaged target detection method based on 3D convexity, *Mod. Appl. Sci.*, vol. 5, no. 4, pp. 152–157, 2011.

[65] F. Xue, G. Cui, R. Hong, and J. Gu, Camouflage texture evaluation using a saliency map, *Multimed. Syst.*, vol. 21, no. 2, pp. 169–175, 2015.

[66] S. Li, D. Florencio, Y. Zhao, C. Cook, and W. Li, Foreground detection in camouflaged scenes, in *Proc. IEEE Int. Conf. Image Processing* (*ICIP*), Beijing, China, 2017, pp. 4247–4251.

[67] Y. Liu, C. Q. Wang, and Y. J. Zhou, Camouflaged people detection based on a semi-supervised search identification network, *Def. Technol.*, vol. 21, pp. 176–183, 2023.

[68] J. Hu, J. Lin, S. Gong, and W. Cai, Relax image-specific prompt requirement in SAM: A single generic prompt for segmenting camouflaged objects, *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 11, pp. 12511–12518, 2024.

[69] J. Zhu, X. Zhang, S. Zhang, and J. Liu, Inferring camouflaged objects by texture-aware interactive guidance network, *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, pp. 3599–3607, 2021.

[70] N. Kajiura, H. Liu, and S. Satoh, Improving camouflaged object detection with the uncertainty of pseudo-edge labels, in *Proc. ACM Multimedia Asia*, Gold Coast, Australia, 2021, pp. 1–7.

[71] Q. Zhai, X. Li, F. Yang, C. Chen, H. Cheng, and D. P. Fan, Mutual graph learning for camouflaged object detection, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, 12997– 13007.

[72] J. Yan, T. N. Le, K. D. Nguyen, M. T. Tran, T. T. Do, and T. V. Nguyen, MirrorNet: bio-inspired camouflaged object segmentation, *IEEE Access*, vol. 9, pp. 43290–43300, 2021.

[73] J. Ren, X. Hu, L. Zhu, X. Xu, Y. Xu, W. Wang, Z. Deng, and P. A. Heng, Deep texture-aware features for camouflaged object detection, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1157–1167, 2023.

[74] K. Wang, H. Bi, Y. Zhang, C. Zhang, Z. Liu, and S. Zheng, D^{2}C-net: A dual-branch, dual-guidance and cross-refine network for camouflaged object detection, *IEEE Trans. Ind. Electron.*, vol. 69, no. 5, pp. 5364–5374, 2022.

[75] Y. Liu, D. Zhang, Q. Zhang, and J. Han, Integrating part-object relationship and contrast for camouflaged object detection, *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 5154–5166, 2021.

[76] M. Xiang, J. Zhang, Y. Lv, A. Li, Y. Zhong, and Y. Dai, Exploring depth contribution for camouflaged object detection, arXiv preprint arXiv: 2106.13217, 2021.

[77] X. Qin, D. P. Fan, C. Huang, C. Diagne, Z. Zhang, A. C. Sant' Anna, A. Suàrez, M. Jagersand, and L. Shao, Boundary-aware segmentation network for mobile and web applications, arXiv preprint arXiv: 2101.04704, 2021.

[78] G. Chen, X. Chen, B. Dong, M. Zhuge, Y. Wang, H. Bi, J. Chen, P. Wang, and Y. Zhang, Towards accurate camouflaged object detection with mixture convolution and interactive fusion, arXiv preprint arXiv: 2101.05687, 2021.

[79] H. Zhu, P. Li, H. Xie, X. Yan, D. Liang, D. Chen, M. Wei, and J. Qin, I can find you! boundary-guided separated attention network for camouflaged object detection, *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, pp. 3608–3616, 2022.

[80] M. Zhang, S. Xu, Y. Piao, D. Shi, S. Lin, and H. Lu, PreyNet: Preying on camouflaged objects, in *Proc. 30th ACM Int. Conf. Multimedia*, Lisboa, Portugal, 2022, pp. 5323–5332.

[81] C. Zhang, K. Wang, H. Bi, Z. Liu, and L. Yang, Camouflaged object detection *via* neighbor connection and hierarchical information transfer, *Comput. Vis. Image Underst.*, vol. 221, pp. 103450, 2022.

[82] M. C. Chou, H. J. Chen, and H. H. Shuai, Finding the Achilles heel: Progressive identification network for camouflaged object detection, in *Proc. IEEE Int. Conf. Multimedia and Expo* (*ICME*), Taipei, China, 2022, pp. 1–6.

[83] K. B. Park and J. Y. Lee, TCU-Net: Transformer and convolutional neural network-based advanced U-net for concealed object detection, *IEEE Access*, vol. 10, pp. 122347–122360, 2022.

[84] Y. Sun, S. Wang, C. Chen, and T. Z. Xiang, Boundary-guided camouflaged object detection, in *Proc. 31st Int. Joint Conf. Artificial Intelligence* (*IJCAI-22*), Vienna, Austria, pp. 1335–1341.

[85] M. Zhuge, X. Lu, Y. Guo, Z. Cai, and S. Chen, CubeNet: X-shape connection for camouflaged object detection, *Pattern Recognit.*, vol. 127, pp. 108644, 2022.

[86] G. Chen, S. J. Liu, Y. J. Sun, G. P. Ji, Y. F. Wu, and T. Zhou, Camouflaged object detection *via* context-aware cross-level fusion, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6981–6993, 2022.

[87] P. Li, X. Yan, H. Zhu, M. Wei, X. P. Zhang, and J. Qin, FindNet: can you find me? boundary-and-texture enhancement network for camouflaged object detection, *IEEE Trans. Image Process.*, vol. 31, pp. 6396–6411, 2022.

[88] W. Zhai, Y. Cao, H. Xie, and Z. J. Zha, Deep texton-coherence network for camouflaged object detection, *IEEE Trans. Multimedia*, vol. 25, pp. 5155–5165, 2023.

[89] J. Lin, X. Tan, K. Xu, L. Ma, and R. W. H. Lau, Frequency-aware

camouflaged object detection, *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 2, pp. 1–16, 2023.

[90] J. Liu, J. Zhang, and N. Barnes, Modeling aleatoric uncertainty for camouflaged object detection, in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision* (*WACV*), Waikoloa, HI, USA, 2022, pp. 1445–1454.

[91] W. Sun, C. Liu, L. Zhang, Y. Li, P. Wei, C. Liu, J. Zou, J. Jiao, and Q. Ye, DQnet: cross-model detail querying for camouflaged object detection, arXiv preprint arXiv: 2212.08296, 2022.

[92] R. He, Q. Dong, J. Lin, and R. W. H. Lau, Weakly-supervised camouflaged object detection with scribble annotations, *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 1, pp. 781–789, 2023.

[93] C. Xie, C. Xia, T. Yu, and J. Li, Frequency representation integration for camouflaged object detection, in *Proc. 31st ACM Int. Conf. Multimedia*, Ottawa, Canada, 2023, pp. 1789–1797.

[94] Q. Wang, J. Yang, X. Yu, F. Wang, P. Chen, and F. Zheng, Depth-aided camouflaged object detection, in *Proc. 31st ACM Int. Conf. Multimedia*, Ottawa, Canada, 2023, pp. 3297–3306.

[95] R. Cong, M. Sun, S. Zhang, X. Zhou, W. Zhang, and Y. Zhao, Frequency perception network for camouflaged object detection, in *Proc. 31st ACM Int. Conf. Multimedia*, Ottawa, Canada, 2023, pp. 1179–1189.

[96] Z. Wu, J. Wang, Z. Zhou, Z. An, Q. Jiang, C. Demonceaux, G. Sun, and R. Timofte, Object segmentation by mining cross-modal semantics, in *Proc. 31st ACM Int. Conf. Multimedia*, Ottawa, Canada, 2023, pp. 3455–3464.

[97] F. Xiao, P. Zhang, C. He, R. Hu, and Y. Liu, Concealed object segmentation with Hierarchical coherence modeling, in *Proc. 3rd CAAI Int. Conf. Artificial Intelligence*, Fuzhou, China, 2024, pp. 16–27.

[98] Q. Zhang, X. Sun, Y. Chen, Y. Ge, and H. Bi, Attention-induced semantic and boundary interaction network for camouflaged object detection, *Comput. Vis. Image Underst.*, vol. 233, pp. 103719, 2023.

[99] Z. Chen, R. Gao, T. Z. Xiang, and F. Lin, Diffusion model for camouflaged object detection, in *Proc. 26th European Conf. Artificial Intelligence*, Kraków, Poland, 2023, pp. 445–452.

[100] H. Xing, S. Gao, H. Tang, T. Q. Mok, Y. Kang, and W. Zhang, TINYCOD: Tiny and effective model for camouflaged object detection, in *Proc. ICASSP 2023 - 2023 IEEE Int. Conf. Acoustics, Speech and Signal Processing* (*ICASSP*), Rhodes Island, Greece, 2023, pp. 1–5.

[101] Z. Wu, D. P. Paudel, D. P. Fan, J. Wang, S. Wang, C. Demonceaux, R. Timofte, and L. Van Gool, Source-free depth for object pop-out, in *Proc. 2023 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Paris, France, 2023, pp. 1032–1042.

[102] D. Sun, S. Jiang, and L. Qi, Edge-aware mirror network for camouflaged object detection, in *Proc. IEEE Int. Conf. Multimedia and Expo* (*ICME*), Brisbane, Australia, 2023, pp. 2465–2470.

[103] Y. Song, X. Li, and L. Qi, Camouflaged object detection with feature grafting and distractor aware, in *Proc. IEEE Int. Conf. Multimedia and Expo* (*ICME*), Brisbane, Australia, 2023, pp. 2459–2464.

[104] Q. Zhang and W. Yan, CFANet: A cross-layer feature aggregation network for camouflaged object detection, in *Proc. IEEE Int. Conf. Multimedia and Expo* (ICME), Brisbane, Australia, 2023, pp. 2441–2446.

[105] X. Yang, H. Zhu, G. Mao, and S. Xing, OAFormer: Occlusion aware transformer for camouflaged object detection, in *Proc. IEEE Int. Conf. Multimedia and Expo* (*ICME*), Brisbane, Australia, 2023, pp. 1421–1426.

[106] X. Li, J. Yang, S. Li, J. Lei, J. Zhang, and D. Chen, Locate, refine and restore: A progressive enhancement network for camouflaged object detection, in *Proc. 32nd Int. Joint Conf. Artificial Intelligence*, Macau, China, 2023, pp. 1116–1124.

[107] H. Mei, K. Xu, Y. Zhou, Y. Wang, H. Piao, X. Wei, and X. Yang, Camouflaged object segmentation with omni perception, *Int. J. Comput. Vis.*, vol. 131, no. 11, pp. 3019–3034, 2023.

[108] G. P. Ji, D. P. Fan, Y. C. Chou, D. Dai, A. Liniger, and L. Van Gool, Deep gradient learning for efficient camouflaged object detection, *Mach. Intell. Res.*, vol. 20, no. 1, pp. 92–108, 2023.

[109] V. Asnani, A. Kumar, S. You, and X. Liu, PrObeD: Proactive object detection wrapper, in *Proc. 37th Conf. Neural Information Processing Systems* (*NeurIPS 2023*), New Orleans, LA, USA, 2023, pp. 77993–78005.

[110] Y. Liu, K. Zhang, Y. Zhao, H. Chen, and Q. Liu, Bi-RRNet: bi-level recurrent refinement network for camouflaged object detection, *Pattern Recognit.*, vol. 139, pp. 109514, 2023.

[111] X. Jiang, W. Cai, Y. Ding, X. Wang, Z. Yang, X. Di, and W. Gao, Camouflaged object detection based on ternary cascade perception, *Remote. Sens.*, vol. 15, no. 5, pp. 1188, 2023.

[112] J. Xiang, Q. Pan, Z. Zhang, S. Fu, and Y. Qin, Double-branch fusion network with a parallel attention selection mechanism for camouflaged object detection, *Sci. China Inf. Sci.*, vol. 66, no. 6, pp. 162403, 2023.

[113] X. Hu, X. Zhang, F. Wang, J. Sun, and F. Sun, Efficient camouflaged object detection network based on global localization perception and local guidance refinement, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 5452–5465, 2024.

[114] G. Yue, H. Xiao, H. Xie, T. Zhou, W. Zhou, W. Yan, B. Zhao, T. Wang, and Q. Jiang, Dual-constraint coarse-to-fine network for camouflaged object detection, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3286–3298, 2024.

[115] H. Xing, S. Gao, Y. Wang, X. Wei, H. Tang, and W. Zhang, Go closer to see better: Camouflaged object detection *via* object area amplification and figure-ground conversion, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 10, pp. 5444–5457, 2023.

[116] Y. Liu, H. Li, J. Cheng, and X. Chen, MSCAF-net: A general framework for camouflaged object detection *via* learning multi-scale context-aware features, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4934–4947, 2023.

[117] Y. Lv, J. Zhang, Y. Dai, A. Li, N. Barnes, and D. P. Fan, Toward deeper understanding of camouflaged object detection, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 7, pp. 3462–3476, 2023.

[118] X. Jiang, W. Cai, Y. Ding, X. Wang, D. Hong, Z. Yang, and W. Gao, Camouflaged object segmentation based on joint salient object for contrastive learning, *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–16, 2023.

[119] H. Li, C. M. Feng, Y. Xu, T. Zhou, L. Yao, and X. Chang, Zero-shot camouflaged object detection, *IEEE Trans. Image Process.*, vol. 32, pp. 5126–5137, 2023.

[120] Z. Song, X. Kang, X. Wei, H. Liu, R. Dian, and S. Li, FSNet: focus scanning network for camouflaged object detection, *IEEE Trans. Image Process.*, vol. 32, pp. 2267–2278, 2023.

[121] Y. Lyu, H. Zhang, Y. Li, H. Liu, Y. Yang, and D. Yuan, UEDG: uncertainty-edge dual guided camouflage object detection, *IEEE Trans. Multimed.*, vol. 26, pp. 4050–4060, 2024.

[122] X. Yan, M. Sun, Y. Han, and Z. Wang, Camouflaged object segmentation based on matching–recognition–refinement network, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 15993–16007, 2024.

[123] D. Zheng, X. Zheng, L. T. Yang, Y. Gao, C. Zhu, and Y. Ruan, MFFN: Multi-view feature fusion network for camouflaged object detection, in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision* (*WACV*), Waikoloa, HI, USA, 2023, pp. 6232–6242.

[124] J. Li, F. Lu, N. Xue, Z. Li, H. Zhang, and W. He, Cross-level attention with overlapped windows for camouflaged object detection, arXiv preprint arXiv: 2311.16618, 2023.

[125] T. Chen, J. Xiao, X. Hu, G. Zhang, and S. Wang, A bioinspired three-stage model for camouflaged object detection, arXiv preprint arXiv: 2305.12635, 2023.

[126] T. Chen, L. Zhu, C. Ding, R. Cao, Y. Wang, Z. Li, L. Sun, P. Mao, Y. Zang, T. Chen, et al., SAM fails to segment anything? -- SAM-Adapter: Adapting SAM in underperformed scenes: Camouflage,

shadow, medical image segmentation, and more, arXiv preprint arXiv: 2304.09148, 2023.
[127] Y. Dong, H. Zhou, C. Li, J. Xie, Y. Xie, and Z. Li, You do not need additional priors in camouflage object detection, arXiv preprint arXiv: 2310.00702, 2023.
[128] M. Q. Le, M. T. Tran, T. N. Le, T. V. Nguyen, and T. T. Do, CamoFA: A learnable Fourier-based augmentation for camouflage segmentation, arXiv preprint arXiv: 2308.15660, 2023.
[129] A. Li, J. Zhang, Y. Lv, T. Zhang, Y. Zhong, M. He, and Y. Dai, Joint salient object detection and camouflaged object detection *via* uncertainty-aware learning, arXiv preprint arXiv: 2307.04651, 2023.
[130] Y. Xing, D. Kong, S. Zhang, G. Chen, L. Ran, P. Wang, and Y. Zhang, Pre-train, adapt and detect: Multi-task adapter tuning for camouflaged object detection, arXiv preprint arXiv: 2307.10685, 2023.
[131] Z. Chen, K. Sun, and X. Lin, CamoDiffusion: camouflaged object detection *via* conditional diffusion models, *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 2, pp. 1272–1280, 2024.
[132] L. Tang, P. T. Jiang, Z. H. Shen, H. Zhang, J. W. Chen, and B. Li, Chain of visual perception: Harnessing multimodal large language models for zero-shot camouflaged object detection, in *Proc. 32nd ACM Int. Conf. Multimedia*, Melbourne, Australia, 2024, pp. 8805–8814.
[133] Z. Liu, P. Jiang, L. Lin, and X. Deng, Edge attention learning for efficient camouflaged object detection, in *Proc. 2024 IEEE Int. Conf. Acoustics, Speech and Signal Processing* (*ICASSP 2024*). Seoul, Republic of Korea, 2024, pp. 5230–5234.
[134] T. D. Nguyen, A. K N. Vu, N. D. Nguyen, V. T. Nguyen, T. D. Ngo, T. T. Do, M. T. Tran, and T. V. Nguyen, The art of camouflage: Few-shot learning for animal detection and segmentation, *IEEE Access*, vol. 12, pp. 103488–103503, 2024.
[135] X. Zhou, Z. Wu, and R. Cong, Decoupling and integration network for camouflaged object detection, *IEEE Trans. Multimedia*, vol. 26, pp. 7114–7129, 2024.
[136] B. Yin, X. Zhang, D. P. Fan, S. Jiao, M. M. Cheng, L. Van Gool, and Q. Hou, CamoFormer: masked separable attention for camouflaged object detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10362–10374, 2024.
[137] A. Khan, M. Khan, W. Gueaieb, A. El Saddik, G. De Masi, and F. Karray, CamoFocus: Enhancing camouflage object detection with split-feature focal modulation and context refinement, *in Proc. IEEE/CVF Winter Conf. Applications of Computer Vision*, no. WACV, pp. 1423, 1432.
[138] H. S. Chen, Y. Zhu, S. You, A. M. Madni, and C. C. Jay Kuo, GreenCOD: A green camouflaged object detection method, arXiv preprint arXiv: 2405.16144, 2024.
[139] Z. Chen, X. Zhang, T. Z. Xiang, and Y. Tai, Adaptive guidance learning for camouflaged object detection, arXiv preprint arXiv: 2405.02824, 2024.
[140] C. Guo and H. Huang, CoFiNet: unveiling camouflaged objects with multi-scale finesse, arXiv preprint arXiv: 2402.02217, 2024.
[141] Z. Yang, K. Choy, and S. Farsiu, Spatial coherence loss: All objects matter in salient and camouflaged object detection, arXiv preprint arXiv: 402.18698, 2024.
[142] C. He, X. Wang, L. Deng, and G. Xu, Image threshold segmentation based on GLLE histogram. In *CPSCom*, in *Proc. 2019 Int. Conf. Internet of Things* (*iThings*) *and IEEE Green Computing and Communications* (*GreenCom*) *and IEEE Cyber, Physical and Social Computing* (*CPSCom*) *and IEEE Smart Data* (*SmartData*), Atlanta, GA, USA, 2019, pp. 410–415.
[143] Y. Zhang, Y. Xie, C. Li, Z. Wu, and Y. Qu, Learn. all-in collaborative multiview binary representation for clustering, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 3, pp. 4260–4273, 2024.
[144] L. Xu, H. Wu, C. He, J. Wang, C. Zhang, F. Nie, and L. Chen, Multi-modal sequence learning for Alzheimer's disease progression prediction with incomplete variable-length longitudinal data, *Med. Image Anal.*, vol. 82, pp. 102643, 2022.
[145] Z. Chen, Y. Zhang, J. Gu, L. Kong, and X. Yang, Recursive generalization transformer for image super-resolution, arXiv preprint arXiv: 2303.06373, 2023.
[146] G. Xu, C. He, H. Wang, H. Zhu, and W. Ding, DM-fusion: Deep model-driven network for heterogeneous image fusion, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 10071–10085, 2024.
[147] C. He, C. Fang, Y. Zhang, T. Ye, K. Li, L. Tang, Z. Guo, X. Li, and S. Farsiu, Reti-diff: Illumination degradation image restoration with retinex-based latent diffusion model, arXiv preprint arXiv: 2311.11638, 2023.
[148] Y. Zhang, R. Hu, R. Li, Y. Qu, Y. Xie, and X. Li, Cross-modal match for language conditioned 3D object grounding, *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 7, pp. 7359–7367, 2024.
[149] Y. Pang, X. Zhao, T. Z. Xiang, L. Zhang, and H. Lu, ZoomNeXt: A unified collaborative pyramid network for camouflaged object detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 9205–9220, 2024.
[150] C. He, K. Li, G. Xu, Y. Zhang, R. Hu, Z. Guo, and X. Li, Degradation-resistant unfolding network for heterogeneous image fusion, in *Proc. 2023 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Paris, France, 2023, 12611–12621.
[151] Y. Zhang, Y. Qu, Y. Xie, Z. Li, S. Zheng, and C. Li, Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation, in *Proc. 2021 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Montreal, Canada, 2021, pp. 15520–15528.
[152] M. Ju, C. He, J. Liu, B. Kang, J. Su, and D. Zhang, IVF-net: An infrared and visible data fusion deep network for traffic object enhancement in intelligent transportation systems, *IEEE Trans. Intell. Transport. Syst.*, vol. 24, no. 1, pp. 1220–1234, 2023.
[153] Y. Lu, C. He, Y. F. Yu, G. Xu, H. Zhu, and L. Deng, Vector co-occurrence morphological edge detection for colour image, *IET Image Process.*, vol. 15, no. 13, pp. 3063–3070, 2021.
[154] C. He, K. Li, G. Xu, J. Yan, L. Tang, Y. Zhang, Y. Wang, and X. Li, HQG-net: Unpaired medical image enhancement with high-quality guidance, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 12, pp. 18404–18418, 2024.
[155] L. Deng, C. He, G. Xu, H. Zhu, and H. Wang, PcGAN: A noise robust conditional generative adversarial network for one shot learning, *IEEE Trans. Intell. Transport. Syst.*, vol. 23, no. 12, pp. 25249–25258, 2022.
[156] L. Tang, K. Li, C. He, Y. Zhang, and X. Li, Source-free domain adaptive fundus image segmentation with Class-balanced mean teacher, in *Proc. 26th Int. Conf. Medical Image Computing and Computer Assisted Intervention* (*MICCAI 2023*), Vancouver, Canada, 2023, pp. 684–694.
[157] L. Tang, K. Li, C. He, Y. Zhang, and X. Li, Consistency regularization for generalizable source-free domain adaptation, in *Proc. 2023 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Paris, France, 2023, pp. 23–33.
[158] L. Tang, Z. Tian, K. Li, C. He, H. Zhou, H. Zhao, X. Li, and J. Jia, Mind the interference: Retaining pre-trained knowledge in parameter efficient continual learning of vision-language models, arXiv preprint arXiv: 2407.05342, 2024.
[159] J. Huang, J. Pan, Z. Wan, H. Lyu, and J. Luo, Evolver: chain-of-evolution prompting to boost large multimodal models for hateful meme detection, arXiv preprint arXiv: 2407.21004, 2024.
[160] C. He, L. Xu, G. Lu, and L. Deng, GLLE entropic threshold segmentation based on fuzzy entropy, *J. Nanjing Univ. Sci. Technol.*, vol. 11, no. 6, pp. 757–763, 2019.
[161] M. Ju, C. He, C. Ding, W. Ren, L. Zhang, and K. K. Ma, All-inclusive image enhancement for degraded images exhibiting low-frequency corruption, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 45, no. 4, pp. 4462–4473, 2023.
[162] C. He, Y. Shen, C. Fang, F. Xiao, L. Tang, Y. Zhang, W. Zuo, Z. Guo, and X. Li, Diffusion models in low-level vision: A survey, arXiv preprint arXiv: 2406.11138, 2024.

[163] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv: 1409.1556, 2014.

[164] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, USA, 2016, pp. 770–778.

[165] S. H. Gao, M. M. Cheng, K. Zhao, X. Y. Zhang, M. H. Yang, and P. Torr, Res2Net: A new multi-scale backbone architecture, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, 2021.

[166] M. Tan and Q. V. Le, EfficientNet: rethinking model scaling for convolutional neural networks, in *Proc. 36th Int. Conf. Machine Learning*, Long Beach, CA, USA, 2019, pp. 6105–6114.

[167] 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, in *Proc. 9th Int. Conf. Learning Representations* (*ICLR*), virtual, 2021.

[168] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., Learning transferable visual models from natural language supervision, in *Proc. 38th Int. Conf. Machine Learning*, virtual, 2021, pp. 8748–8763.

[169] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in *Proc. 40th Int. Conf. Machine Learning*, Honolulu, HI, USA, 2023, pp. 19730–19742.

[170] W. Wang, E. Xie, X. Li, D. P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in *Proc. 2021 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Montreal, Canada, 2021, pp. 568–578.

[171] 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 *Proc. 2021 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Montreal, Canada, 2021, pp. 10012–10022.

[172] C. Fang, C. He, F. Xiao, Y. Zhang, L. Tang, Y. Zhang, K. Li, and X. Li, Real-world image dehazing with coherence-based label generator and cooperative unfolding network, arXiv preprint arXiv: 2406.07966, 2024.

[173] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, in *Proc. 31st Int. Conf. Neural Information Processing Systems* (*NIPS 2017*), Long Beach, CA, USA, 2017, pp. 6000–6010.

[174] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Y. Lo, et al., Segment anything, in *Proc. 2023 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Paris, France, 2023, pp. 4015–4026.

[175] J. Wang, Y. Pu, Y. Han, J. Guo, Y. Wang, X. Li, and G. Huang, GRA: detecting oriented objects through group-wise rotating and attention, arXiv preprint arXiv: 2403.11127, 2024.

[176] J. Wang, J. Pu, Z. Qi, J. Guo, Y. Ma, N. Huang, Y. Chen, X. Li, and Y. Shan, Taming rectified flow for inversion and editing, arXiv preprint arXiv: 2411.04746, 2024.

[177] J. Wang, Y. Ma, J. Guo, Y. Xiao, G. Huang, and X. Li, COVE: unleashing the diffusion feature correspondence for consistent video editing, arXiv preprint arXiv: 2406.08850, 2024.

[178] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial networks, in *Proc. 28th Int. Conf. Neural Information Processing Systems* (*NIPS 2014*), Montréal, Canada, 2014, pp. 2672–2680.

[179] Y. Xiao, Z. Luo, Y. Liu, Y. Ma, H. Bian, Y. Ji, Y. Yang, and X. Li, Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection, in *Proc. 2024 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2024, pp. 18709–18719.

[180] I. Huerta, D. Rowe, M. Mozerov, and J. Gonzàlez, Improving background subtraction based on a casuistry of colour-motion segmentation problems, in *Proc. 3rd Iberian Conf. Pattern Recognition and Image Analysis* (*IbPRIA 2007*), Girona, Spain, 2007, pp. 475–482.

[181] H. Guo, Y. Dou, T. Tian, J. Zhou, and S. Yu, A robust foreground segmentation method by temporal averaging multiple video frames, in *Proc. Int. Conf. Audio, Language and Image Processing*, Shanghai, China, 2008, pp. 878–882.

[182] D. Conte, P. Foggia, G. Percannella, F. Tufano, and M. Vento, An algorithm for detection of partially camouflaged people, in *Proc. 6th IEEE Int. Conf. Advanced Video and Signal Based Surveillance*, Genova, Italy, 2009, pp. 340–345.

[183] J. Yin, Y. Han, W. Hou, and J. Li, Detection of the mobile object with camouflage color under dynamic background based on optical flow, *Procedia Eng.*, vol. 15, pp. 2201–2205, 2011.

[184] T. Malathi and M. K. Bhuyan, Foreground object detection under camouflage using multiple camera-based codebooks, in *Proc. Annual IEEE India Conf.* (*INDICON*), Mumbai, India, 2013, pp. 1–6.

[185] J. R. Hall, I. C. Cuthill, R. Baddeley, A. J. Shohet, and N. E. Scott-Samuel, Camouflage, detection and identification of moving targets, *Proc. R. Soc. B.*, vol. 280, no. 1758, pp. 20130064, 2013.

[186] S. Kim, Unsupervised spectral-spatial feature selection-based camouflaged object detection using VNIR hyperspectral camera, *Sci. World J.*, vol. 2015, no. 1, pp. 834635, 2015.

[187] Y. Xiao, L. Song, S. Huang, J. Wang, S. Song, Y. Ge, X. Li, and Y. Shan, GrootVL: tree topology is all you need in state space model, arXiv preprint arXiv: 2406.02395, 2024.

[188] X. Zhang, C. Zhu, S. Wang, Y. Liu, and M. Ye, A Bayesian approach to camouflaged moving object detection, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 9, pp. 2001–2013, 2017.

[189] W. Hui, Z. Zhu, S. Zheng, and Y. Zhao, Endow SAM with keen eyes: Temporal-spatial prompt learning for video camouflaged object detection, in *Proc. 2024 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2024, pp. 19058–19067.

[190] C. Xie, Y. Xiang, Z. Harchaoui, and D. Fox, Object discovery in videos as foreground motion clustering, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 9994–10003.

[191] H. Lamdouar, C. Yang, W. Xie, and A. Zisserman, Betrayed by motion: Camouflaged object discovery *via* motion segmentation, in *Proc. 15th Asian Conference on Computer Vision* (*ACCV 2020*), Kyoto, Japan, 2020, pp. 488–503.

[192] C. Yang, H. Lamdouar, E. Lu, A. Zisserman, and W. Xie, Self-supervised video object segmentation by motion grouping, in *Proc. 2021 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Montreal, Canada, 2021, pp. 7177– 7188.

[193] H. Lamdouar, W. Xie, and A. Zisserman, Segmenting invisible moving objects, in *Proc. 32nd British Machine Vision Conf.* (*BMVC 2021*), virtual, 2021, pp. 1–14.

[194] X. Cheng, H. Xiong, D. P. Fan, Y. Zhong, M. Harandi, T. Drummond, and Z. Ge, Implicit motion handling for video camouflaged object detection, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, LA, USA, 2022, pp. 13854–13863.

[195] E. Meunier, A. Badoual, and P. Bouthemy, EM-driven unsupervised learning for efficient motion segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4462–4473, 2023.

[196] M. Kowal, M. Siam, M. A. Islam, N. D. B. Bruce, R. P. Wildes, and K. G. Derpanis, A deeper dive into what deep spatiotemporal networks encode: Quantifying static *vs*. dynamic information, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, LA, USA, 2022, pp. 3999–14009.

[197] J. Xie, W. Xie, and A. Zisserman, Segmenting moving objects *via*

an object-centric layered representation, in *Proc. 36th Conf. Neural Information Processing Systems* (*NeurIPS 2022*), New Orleans, LA, USA, 2022, pp. 28023–28036.

[198] P. Bideau, E. Learned-Miller, C. Schmid, and K. Alahari, The right spin: Learning object motion from rotation-compensated flow fields, *Int. J. Comput. Vis.*, vol. 132, no. 1, pp. 40–55, 2024.

[199] H. Lamdouar, W. Xie, and A. Zisserman, The making and breaking of camouflage, in *Proc. 2023 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Paris, France, 2023, pp. 832–842.

[200] Z. Yu, E. B. Tavakoli, M. Chen, S. You, R. Rao, S. Agarwal, and F. Ren, Tokenmotion: Motion-guided vision transformer for video camouflaged object detection VIA learnable token selection, in *Proc. 2024 IEEE Int. Conf. Acoustics, Speech and Signal Processing* (*ICASSP 2024)*, Seoul, Republic of Korea, 2024, pp. 2875–2879.

[201] W. Hui, Z. Zhu, G. Gu, M. Liu, and Y. Zhao, Implicit-explicit motion learning for video camouflaged object detection, *IEEE Trans. Multimedia*, vol. 26, pp. 7188–7196, 2024.

[202] M. N. Meeran, G. Adethya T, and B. P. Mantha, SAM-PM: Enhancing Video Camouflaged Object Detection using Spatio-Temporal Attention, in *Proc. 2024 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2024, pp. 1857–1866.

[203] X. Zhang, T. Xiao, G. Ji, X. Wu, K. Fu, and Q. Zhao, Explicit motion handling and interactive prompting for video camouflaged object detection, arXiv preprint arXiv: 2403.01968, 2024.

[204] H. Guo, Y. Guo, Y. Zha, Y. Zhang, W. Li, T. Dai, S. T. Xia, and Y. Li, MambaIRv2: attentive state space restoration, arXiv preprint arXiv: 2411.15269, 2024.

[205] G. Sun, Z. An, Y. Liu, C. Liu, C. Sakaridis, D. P. Fan, and L. Gool, Indiscernible object counting in underwater scenes, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 13791–13801.

[206] B. Wang, H. Liu, D. Samaras, and M. H. Nguyen, Distribution matching for crowd counting, in *Proc. 34th Conf. Neural Information Processing Systems* (*NeurIPS 2020*), Vancouver, Canada, 2020, pp. 1595–1607.

[207] D. Liang, W. Xu, and X. Bai, An end-to-end transformer model for Crowd localization, in *Proc. 17th European Conf. Computer Vision* (*ECCV*), Tel Aviv, Israel, 2022, pp. 38–54.

[208] T. N. Le, Y. Cao, T. C. Nguyen, M. Q. Le, K. D. Nguyen, T. T. Do, M. T. Tran, and T. V. Nguyen, Camouflaged instance segmentation in-the-wild: Dataset, method, and benchmark suite, *IEEE Trans. Image Process.*, vol. 31, pp. 287–300, 2022.

[209] Z. Cai and N. Vasconcelos, Cascade R-CNN: Delving into high quality object detection, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 6154–6162.

[210] N. Luo, Y. Pan, R. Sun, T. Zhang, Z. Xiong, and F. Wu, Camouflaged instance segmentation *via* explicit de-camouflaging, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 17918–17927.

[211] J. Pei, T. Cheng, D. P. Fan, H. Tang, C. Chen, and L. Van Gool, OSFormer: one-stage camouflaged instance segmentation with Transformers, in *Proc. 17th European Conf. Computer Vision* (*ECCV*), Tel Aviv, Israel, 2022, pp. 19–37.

[212] T. A. Vu, D. T. Nguyen, Q. Guo, B. S. Hua, N. M. Chung, I. W. Tsang, and S. K. Yeung, Leveraging open-vocabulary diffusion to camouflaged instance segmentation, arXiv preprint arXiv: 2312.17505, 2023.

[213] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, High-resolution image synthesis with latent diffusion models, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, LA, USA, 2022, pp. 10684–10695.

[214] P. Bideau and E. Learned-Miller, It's moving! A probabilistic model for causal motion segmentation in moving camera videos, in *Proc. 14th European Conf. Computer Vision* (*ECCV*), Amsterdam, The Netherlands, 2016, pp. 433–449.

[215] P. Skurowski, H. Abdulameer, J. Błaszczyk, T. Depta, A. Kornacki, and P. Kozieł, Animal camouflage analysis: Chameleon database https://www.polsl.pl/rau6/chameleon-database-animal-camouflage-analysis/, 2018.

[216] Y. Zheng, X. Zhang, F. Wang, T. Cao, M. Sun, and X. Wang, Detection of people with camouflage pattern *via* dense deconvolution network, *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 29–33, 2019.

[217] J. Yang, Q. Wang, F. Zheng, P. Chen, A. Leonardis, and D. P. Fan, PlantCamo dataset, https://github.com/yjybuaa/PlantCamo, 2023.

[218] M. M. Cheng and D. P. Fan, Structure-measure: A new way to evaluate foreground maps, *Int. J. Comput. Vis.*, vol. 129, no. 9, pp. 2622–2638, 2021.

[219] R. Margolin, L. Zelnik-Manor, and A. Tal, How to evaluate foreground maps, in *Proc. 2014 IEEE Conf. Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 248–255.

[220] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, Saliency filters: Contrast based filtering for salient region detection, in *Proc. 2012 IEEE Conf. Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 733–740.

[221] D. P. Fan, C. Gong, Y. Cao, B. Ren, and A. Borji, Enhanced-alignment measure for binary foreground map evaluation, in *Proc. 27th Int. Joint Conf. Artificial Intelligence* (*IJCAI'18*), Stockholm, Sweden, 2018, pp. 698–704.

[222] 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 *Proc. 7th Int. Conf. Learning Representations* (*ICLR*), New Orleans, LA, USA, 2019, pp. 7284–7293.

[223] G. P. Ji, Y. C. Chou, D. P. Fan, G. Chen, H. Fu, D. Jha, and L. Shao, Progressively normalized self-attention network for video polyp segmentation, in Proc. *24th Int. Conf. Medical Image Computing and Computer Assisted Intervention* (*MICCAI 2021*), Strasbourg, France, 2021, pp. 142–152.

[224] Y. Ma, Y. He, X. Cun, X. Wang, S. Chen, X. Li, and Q. Chen, Follow your pose: Pose-guided text-to-video generation using pose-free videos, *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 5, pp. 4117–4125, 2024.

[225] X. J. Luo, S. Wang, Z. Wu, C. Sakaridis, Y. Cheng, D. P. Fan, and L. Van Gool, CamDiff: camouflage image augmentation via diffusion, *CAAI Artif. Intell. Res.*, vol. 2, p. 9150021, 2023.

[226] P. Zhao, P. Xu, P. Qin, D. P. Fan, Z. Zhang, G. Jia, B. Zhou, and J. Yang, LAKE-RED: Camouflaged images generation by latent background knowledge retrieval-augmented diffusion, in *Proc. 2024 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2024, pp. 4092–4101.

[227] R. Hu, Y. Liu, K. Gu, X. Min, and G. Zhai, Toward a No-reference quality metric for camera-captured images, *IEEE Trans. Cybern.*, vol. 53, no. 6, pp. 3651–3664, 2023.

[228] R. Hu, Y. Liu, Z. Wang, and X. Li, Blind quality assessment of night-time image, *Displays*, vol. 69, pp. 102045, 2021.

[229] H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, and S. T. Xia, MambaIR: A Simple Baseline for Image Restoration with State-Space Model, in *Proc. 18th European Conf. Computer Vision* (*ECCV*), Milan, Italy, 2024, pp. 222–241.

[230] Z. Yang and S. Farsiu, Directional connectivity-based segmentation of medical images, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 11525–11535.

[231] Z. Yang, S. Soltanian-Zadeh, and S. Farsiu, BiconNet: An edge-preserved connectivity-based approach for salient object detection, *Pattern Recognit.*, vol. 121, pp. 108231, 2022.

[232] X. Zhao, Y. Pang, W. Ji, B. Sheng, J. Zuo, L. Zhang, and H. Lu, Spider: A unified framework for context-dependent concept segmentation, in *Proc. 41st Int. Conf. Machine Learning*, Vienna, Austria, 2024.

[233] G. Lu, C. He, L. Xu, J. Ren, G. Xu, and H. Zhao, Infrared and visible image fusion based on local gradient constraints, in *Proc.*

IEEE Cyber, Physical and Social Computing (*CPSCom*), Rhodes, Greece, 2020, pp. 571–575.

[234] L. Tang and B. Li, Evaluating SAM2's role in camouflaged object detection: From SAM to SAM2, arXiv preprint arXiv: 2407.21596, 2024.

[235] G. P. Ji, D. P. Fan, P. Xu, B. Zhou, M. M. Cheng, and L. Van Gool, SAM struggles in concealed scenes—Empirical study on "segment anything", *Sci. China Inf. Sci.*, vol. 66, no. 12, p. 226101, 2023.

[236] W. Ji, J. Li, Q. Bi, T. Liu, W. Li, and L. Cheng, Segment anything is not always perfect: An investigation of SAM on different real-world applications, *Mach. Intell. Res.*, vol. 21, no. 4, pp. 617–630, 2024.

[237] N. Ravi, V. Gabeur, Y. T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al., SAM 2: Segment anything in images and videos, arXiv preprint arXiv: 2408.00714, 2024.

[238] G. Hinton, O. Vinyals, and J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv: 1503.02531, 2015.

[239] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, A survey on deep transfer learning, in *Proc. 27th Int. Conf. Artificial Neural Networks* (*ICANN*), Rhodes, Greece, 2018, pp. 270–279.

[240] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, Model compression, in *Proc. 12th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Philadelphia, PA, USA, 2006, pp. 535–541.

[241] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, Meta-learning in neural networks: A survey, arXiv preprint 2004.05439, 2020.

[242] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, Kvasir-SEG: A segmented polyp dataset, in *Proc. 26th Int. Conf. Multimedia Modeling*, Daejeon, Republic of Korea, 2020, pp. 451–462.

[243] H. B. Jenssen and T. Sakinis, MedSeg Covid Dataset 1, https://doi.org/10.6084/m9.figshare.13521488.v2, 2021.

[244] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection, *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1038–1059, 2021.

[245] H. Mei, X. Yang, Y. Wang, Y. A. Liu, S. He, Q. Zhang, X. Wei, and R. W. H. Lau, Don't hit me! glass detection in real-world scenes, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 3684–3693.

[246] J. Lin, G. Wang, and R. W. H. Lau, Progressive mirror detection, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 3694–3702.

[247] B. Dong, W. Wang, D. P. Fan, J. Li, H. Fu, and L. Shao, Polyp-PVT: Polyp segmentation with pyramid vision transformers, *CAAI Artif. Intell. Res.*, vol. 2, p. 9150015, 2023.

[248] D. P. Fan, T. Zhou, G. P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, Inf-net: Automatic COVID-19 lung infection segmentation from CT images, *IEEE Trans. Med. Imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.

[249] H. Dong, K. Song, Y. He, J. Xu, Y. Yan, and Q. Meng, PGA-net: Pyramid feature fusion and global context attention network for automated surface defect detection, *IEEE Trans. Ind. Inf.*, vol. 16, no. 12, pp. 7448–7458, 2020.

[250] H. Zhou, R. Yang, R. Hu, C. Shu, X. Tang, and X. Li, ETDNet: Efficient transformer-based detection network for surface defect detection, *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2023.

[251] D. Han, C. Zhang, Y. Qiao, M. Qamar, Y. Jung, S. Lee, S. H. Bae, and C. S. Hong, Segment anything model (SAM) meets glass: Mirror and transparent objects cannot be easily detected, arXiv preprint arXiv: 2305.00278, 2023.

[252] H. Zhou, R. Yang, Y. Zhang, H. Duan, Y. Huang, R. Hu, X. Li, and Y. Zheng, UniHead: unifying multi-perception for detection heads, *IEEE Trans. Neural Netw. Learn. Syst.*, DOI: https://doi.org/10.1109/TNNLS.2024.3412947.

[253] J. Shi, A. Yong, Y. Jin, D. Li, H. Niu, Z. Jin, and H. Wang, ASGrasp: Generalizable transparent object reconstruction and grasping from RGB-D active stereo camera, in *Proc. 2024 IEEE Int. Conf. Robotics and Automation* (*ICRA*), Yokohama, Japan, 2024.

[254] H. Zhou, L. Tang, R. Yang, G. Qin, Y. Zhang, R. Hu, and X. Li, UniQA: unified vision-language pre-training for image quality and aesthetic assessment, arXiv preprint arXiv: 2406.01069, 2024.

[255] X. Wu, C. Zhan, Y. K. Lai, M. M. Cheng, and J. Yang, IP102: A large-scale benchmark dataset for insect pest recognition, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019, pp. 8787–8796.

[256] A. C. C. Viodor, MangroveLens: A smart solution for mangrove species identification through MobileNetV3 network architecture and biodiversity monitoring, in *Proc. IEEE Open Conf. Electrical, Electronic and Information Sciences* (*eStream*), Vilnius, Lithuania, 2024, pp. 1–6.

[257] J. Wäldchen and P. Mäder, Plant species identification using computer vision techniques: A systematic literature review, *Arch. Comput. Meth. Eng.*, vol. 25, no. 2, pp. 507–543, 2018.

[258] B. Jin, Y. Zheng, P. Li, W. Li, Y. Zheng, S. Hu, X. Liu, J. Zhu, Z. Yan, H. Sun et al., TOD3Cap: Towards 3D dense captioning in outdoor scenes, in *Proc. 18th European Conf. Computer Vision* (*ECCV*), Milan, Italy, 2024, pp. 367–384.

[259] C. Zhu, K. Li, Y. Ma, L. Tang, C. Fang, C. Chen, Q. Chen, and X. Li, InstantSwap: fast customized concept swapping across sharp shape differences, arXiv preprint arXiv: 2412.01197, 2024.

[260] C. Zhu, K. Li, Y. Ma, C. He, and L. Xiu, MultiBooth: towards generating all your concepts in an image from text, arXiv preprint arXiv: 2404.14239, 2024.

[261] Q. Zhang, G. Yin, Y. Nie, and W. S. Zheng, Deep camouflage images, *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 12845–12852, 2020.