Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.055327

REVIEW

Confusing Object Detection: A Survey

Kunkun Tong[1,#], Guchu Zou[2,#], Xin Tan[1,*], Jingyu Gong[1], Zhenyi Qi[2], Zhizhong Zhang[1], Yuan Xie[1] and Lizhuang Ma[1]

[1]School of Computer Science and Technology, East China Normal University, Shanghai, 200062, China

[2]Shanghai Institute of Ceramics, Chinese Academy of Sciences, Shanghai, 200050, China

*Corresponding Author: Xin Tan. Email: xtan@cs.ecnu.edu.cn

[#]These authors contributed equally to this work

Received: 23 June 2024 Accepted: 07 August 2024 Published: 12 September 2024

ABSTRACT

Confusing object detection (COD), such as glass, mirrors, and camouflaged objects, represents a burgeoning visual detection task centered on pinpointing and distinguishing concealed targets within intricate backgrounds, leveraging deep learning methodologies. Despite garnering increasing attention in computer vision, the focus of most existing works leans toward formulating task-specific solutions rather than delving into in-depth analyses of methodological structures. As of now, there is a notable absence of a comprehensive systematic review that focuses on recently proposed deep learning-based models for these specific tasks. To fill this gap, our study presents a pioneering review that covers both the models and the publicly available benchmark datasets, while also identifying potential directions for future research in this field. The current dataset primarily focuses on single confusing object detection at the image level, with some studies extending to video-level data. We conduct an in-depth analysis of deep learning architectures, revealing that the current state-of-the-art (SOTA) COD methods demonstrate promising performance in single object detection. We also compile and provide detailed descriptions of widely used datasets relevant to these detection tasks. Our endeavor extends to discussing the limitations observed in current methodologies, alongside proposed solutions aimed at enhancing detection accuracy. Additionally, we deliberate on relevant applications and outline future research trajectories, aiming to catalyze advancements in the field of glass, mirror, and camouflaged object detection.

KEYWORDS

Confusing object detection; mirror detection; glass detection; camouflaged object detection; deep learning

1 Introduction

Object detection is the task that aims to detect and locate the object in the images or videos, which has attracted considerable research attention in recent years. The methods for object detection have achieved significant advancements since the introduction of deep convolutional neural networks, e.g., AlexNet [1]. While general object detection methods [2,3] perform well on most regular objects, there exist many tricky objects that it cannot detect reliably.

Easily confused objects such as glass, mirrors, and camouflaged objects pose a significant challenge to commonly used general object detection methods. The left three columns of Fig. 1 show some examples of confusing objects. For instance, the inherent optical properties of mirrors pose considerable difficulties for reliable object detection. This challenge arises from the striking resemblance between the content within the mirror and the surrounding environment. Moreover, the variability in both the size and shape of mirrors, coupled with the potential presence of any object within their reflections, poses a substantial hurdle for conventional object detection systems. Similar complexities are observed in the task of detecting glass. Unlike mirrors that reflect real-world objects, glass transmits light, allowing for the visualization of objects positioned behind it. This unique characteristic of glass introduces inherent complexities distinct from those associated with mirror detection. Compared to glass and mirrors, which we humans can easily recognize, we cannot realize the existence of camouflaged objects without paying enough attention to the objects and their surroundings. Therefore, it is difficult to detect confusing objects using general or salient object detection methods.

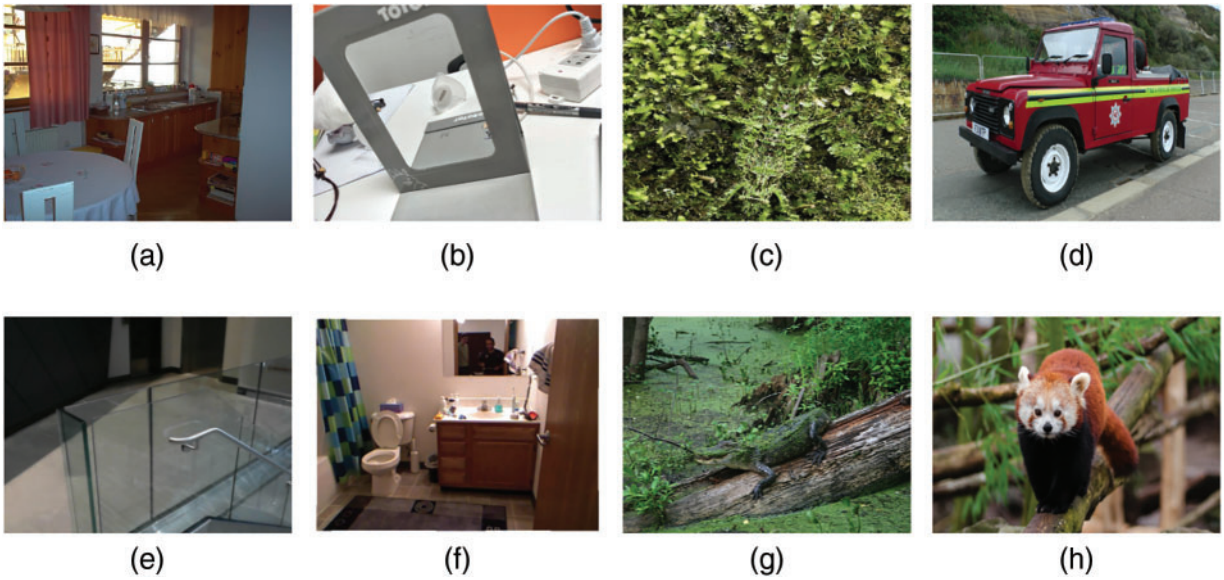

Figure 1: Representative samples of confusing objects sourced from popular datasets are as follows: Image (a) depicts a glass surface sourced from the glass detection dataset (GDD). Reprinted with permission from Reference [4]. Copyright 2020, Copyright Haiyang Mei. Image (e) depicts a glass surface sourced from the glass surface dataset (GSD). Reprinted with permission from Reference [5]. Copyright 2021, Copyright Jiaying Lin. Image (b) represents a mirror sourced from the Mirror Segmentation Dataset (MSD). Reprinted with permission from Reference [6]. Copyright 2019, Copyright Xin Yang. Image (f) represents a mirror sourced from the Progressive Mirror Detection (PMD) dataset. Reprinted with permission from Reference [7]. Copyright 2020, Copyright Jiaying Lin. Image (c) illustrates a camouflaged object sourced from the camouflaged object images (CAMO) dataset. Reprinted with permission from Reference [8]. Copyright 2020, Copyright Elsevier. Image (g) illustrates a camouflaged object sourced from the COD10K dataset. Reprinted with permission from Reference [9]. Copyright 2020, Copyright Dengping Fan. Image (d, h) is non-confusing object images in the public domain that were downloaded from the Internet

In various applications integrating object detection technology, the inability to detect confusing objects such as glasses, mirrors and camouflaged objects presents a significant concern. Given the widespread prevalence of these objects in both indoor settings, like vanity mirrors, and outdoor environments, including glass doors in public spaces, accurate detection holds paramount importance in averting undesirable incidents. The ability to detect these objects correctly is crucial in preventing potential mishaps. For instance, a failure to recognize the presence of glass doors by a robot may result in collisions, leading to property damage and posing safety hazards to pedestrians. While camouflaged object detection, a task aimed at accurately detecting a target object from an environment that blends perfectly with the target, has long been a research hotspot [10–14] in biology and the medical aspect. Mainstream protruding object detection methods make extensive use of discriminative features, while confusing targets have relatively few discriminative features. Accordingly, due to the opposed properties of protruding and camouflaged objects, it is not a good idea to impose related methods on protruding object detection tasks. Hence, establishing an object detection system resilient enough to reliably detect the existence of confusing objects stands as a matter of substantial significance.

Historically, research in object detection predominantly emphasized discerning prominent entities or salient objects like humans [15], animals [16], and vehicles [17]. These studies showcased cutting-edge performance by harnessing extensive datasets and leveraging deep neural networks. Instead, for confusing objects like glasses, mirrors, and camouflaged objects, it lacks both datasets and efficient methods to work on them. In recent years, this situation has begun to change with the prevalence of deep learning. Numerous deep learning-based models have been proposed for the challenging task of COD. This progress is accompanied by the availability of large-scale datasets for COD with professionally annotated tags. The research on confusing object has a long and rich history since the early days [18]. However, early approaches to confusing object segmentation predominantly centered on low-level features encompassing color [12], texture [19], shape [11], and edge [18]. However, these methodologies exhibited limitations, being primarily suited for simple scenes and proving inadequate when confronted with intricate environmental contexts.

Indeed, detecting confusing objects like mirrors, glass, and camouflaged targets is a challenging task. Nonetheless, these areas not only have research value but also have broad application prospects. For example, these specialized object detection accuracy improvements can address edge cases in general object detection methods, thereby avoiding unnecessary failure conditions. The following section covers the literature on deep learning-based glass, mirror, and camouflaged object detection. Since object detection for these specialized subjects is relatively new, the number of papers is smaller than in the general area of object detection and semantic segmentation. We organize the subsequent section as follows: Section 2 provides a brief review comparing our work with previous surveys in the related field. Section 3 briefly introduces the general object detection method. Section 4 dives into the popular methods for confusing object detection, focusing on mirror, glass, and camouflaged objects. Section 5 provides an overview of the mainstream datasets used for our task, with a primary focus on open-source ones. Section 6 overviews the performance of various models under the same settings. Section 7 summarizes the recent work and the future direction of COD.

2 Comparison with Previous Review

To the best of our knowledge, this study is the first to systematically address the detection and segmentation of confusing objects, encompassing glass/transparent objects, mirrors, and camouflaged objects at the same time. There is no prior survey literature on glass and mirror detection. Therefore, our work represents a novel and comprehensive review of this subject area. While camouflaged object

has a longer history and is supported by a few existing studies, the research on glass and mirror detection is relatively limited.

Previous studies, such as [20], have primarily focused on non-deep learning approaches to camouflage detection. This is a very limiting scope, considering the pervasive influence of deep learning in recent years. Another study [21,22] summarizes image-level models for camouflaged object detection, although it includes only a small amount of literature. In addition, literature [23] provides a comprehensive review on model structure and paradigm classification, public benchmark datasets, evaluation metrics, model performance benchmarks, and potential future development directions. Specifically, a number of existing deep learning algorithms are reviewed, offering researchers an extensive overview of the latest methods in this field.

Meanwhile, the published literature [24] reviews relevant work in the broader field of concealed scene understanding. This comprehensive review summarizes a total of 48 existing models for the image-level task. Unlike previous surveys, this study systematically and comprehensively examines the integration of deep learning into the image-level camouflaged scene understanding. It provides an in-depth analysis and discussion on various aspects, including model structure, learning paradigms, datasets, evaluation metrics, and performance comparisons.

In particular, we have summarized and cataloged a substantial number of existing deep learning-based methods for COD. We compare the performance of relevant mainstream models using core metrics to enhance understanding of these approaches. Furthermore, we provide insights into the challenges, key open issues, and future directions of image-level COD.

3 Overview of General Object Detection

Object detection or segmentation has a long history since the early days of computer vision, and image segmentation plays a vital role in different real-world applications. Before the era of deep learning, image segmentation methods used techniques such as k-mean clustering, normalized cuts, region growing, and threshold, which usually yield bad performance. As deep learning comes onto the scene, models based on convolutional neural networks (CNNs) [25] have achieved outstanding performance never achieved by earlier methods. Previous work [26] already has a comprehensive depiction of this topic. The most common deep learning-based method in object detection is the CNN-based model, CNN is one of the most successful and widely used architectures in the field of computer vision. As the transformer model prevails in natural language processing, researchers in computer vision find it also achieves promising results when it is incorporated into computer vision tasks. Different from CNN-based approaches, transformers rely on the attention mechanism at its heart, which enhances their performance in various vision tasks. Simply put it, this mechanism allows transformers to capture long-range dependencies and contextual information more effectively. Besides object detection, transformers have demonstrated their effectiveness in a wide range of vision applications, including image classification, segmentation, and even generative tasks. Generative Adversarial Networks (GANs) [27] is a newly proposed model, using GANs to solve the segmentation task has been a research interest in recent times. The dilated convolution is a slightly modified version of convolution with an additional parameter called dilation rate. The working mechanism of dilation convolution expands the receptive field, therefore capturing more information with less computation cost. Dilation convolution is a widespread technique in the segmentation model. Besides, the probabilistic graphic model is helpful to exploit scene-level semantic information. While challenges exist, probabilistic graphical model Conditional Random fields and Markov Random Fields still achieve promising outcomes in recent works. Encoder-decoder architecture is applied in most object detection models

explicitly or implicitly nowadays. It has gained popularity since Badrinarrayann et al. [3] proposed SegNet. These general object techniques are also common components of the model for COD.

However, given the special nature of confusing objects, a different approach from general object detection is required. Fig. 2 provides an overview summary of the overall architecture present in the literature. The following section aims to give a comprehensive review of popular methods in recent years.

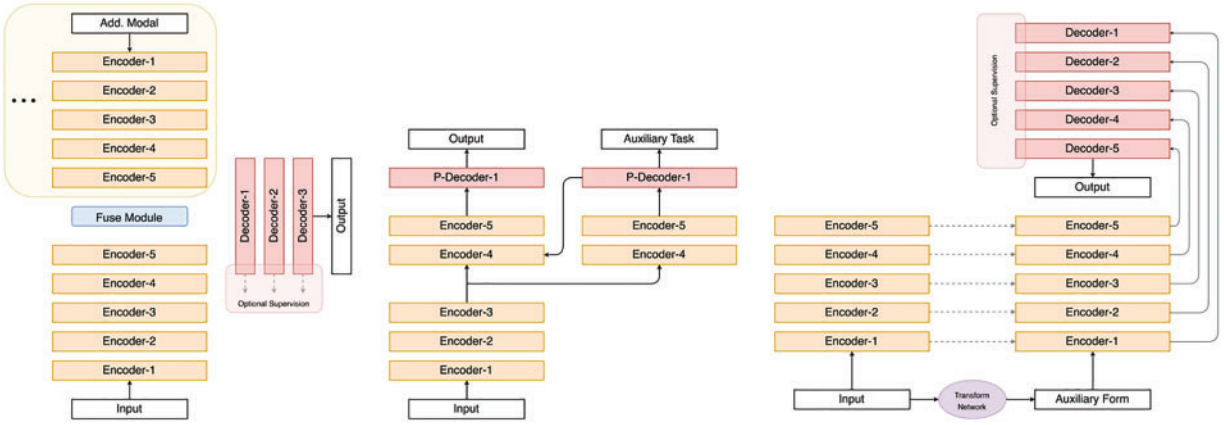

Figure 2: Overview of network architectures for confusing object detection. Three common frameworks are presented in a sequential arrangement from left to right

4 Confusing Object Detection

Confusing objects are intrinsically different from general or salient object detection. Confusing object reference to objects like shadow, water, glass, mirrors, and camouflaged objects. Confusing object has internal and optical characteristics that are completely different from ordinary or prominent objects. Using generic object parts directly will certainly not yield good performance, so it is necessary to design a model specifically for obfuscated objects. In recent years, many researchers have published a large amount of work using deep learning-based models on this topic, and it is time to summarize their outstanding work. Fig. 3 demonstrates the amount of work published within each category. With the missing of some topics, here we focus on glass, mirrors, and camouflaged objects.

4.1 CNN-Based Models

For an overall understanding of CNN-based architecture, refer to subfigure (a) in Fig. 4. For the differences between Transformer-based models and multi-task learning frameworks, see subfigure (b,c). Yang et al. [6] propose the first deep learning-based model MirrorNet to apply to the mirror detection task. It is inspired by human biology and detects entire mirrored areas by identifying content discontinuities. Notably, MirrorNet employs ResNeXt101 as the backbone feature extraction network. Lin et al. [7] propose the PMDNet, which uses not only discontinuities but also the relational content inside and outside the mirror, and then leverages the extracted features to construct mirror edges. The methods that exploit discontinuities and the correspondence between mirror contents will fail in certain scenarios. Tan et al. [28] realize the mirror image exhibits visual chirality property and propose the model called VCNet makes use of visual chirality cues to detect mirror region. Guan et al. observe that there is a semantic connection in the position mirror (i.e.,

people usually place mirrors in several fixed positions), and propose the SANet to leverage the semantic association for mirror detection. Aiming at the shortcomings of previous methods that consume a lot of computing resources, He et al. [29] introduce the HetNet which applies heterogeneous modules in different levels of features, which outperforms the previous work in accuracy and efficiency. The key difference with the previous method lies in that it treats the backbone feature differently with different specially design modules to utilize the characteristics of varying level features fully. The model proposed by Gonzales et al. [30] incorporates parallel convolutional layers alongside a lightweight convolutional block attention module to capture both low-level and high-level features for edge extraction. Mei et al. [31] propose a novel network MirrorNet+ which models both contextual contrasts and semantic associations. While Mei et al. [4] propose the glass detection net (GDNet) for glass object detection. By employing ResNeXt101 as the feature extractor, GDNet utilizes both low-level cues and high-level semantic information for high-accuracy glass detection. The GDNet employs a cascade strategy by embedding multiple modules in the last four layers of the backbone network and dealing with low-level and high-level features separately. The GDNet outperforms existing semantic segmentation and object detection methods. GDNet-B [32] is the successor to GDNet, with the additional boundary feature enhancement module incorporated into the original design to boost performance.

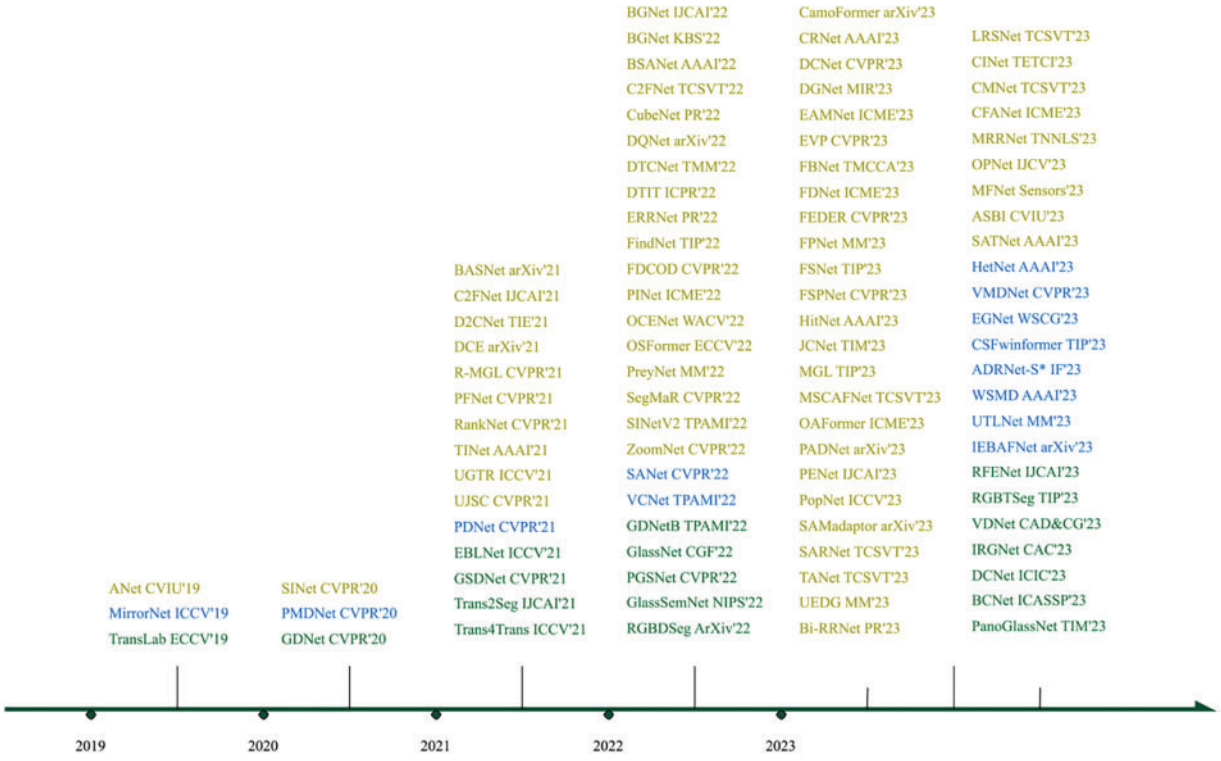

Figure 3: Timeline of the deep learning-based COD methods from 2019 to 2023

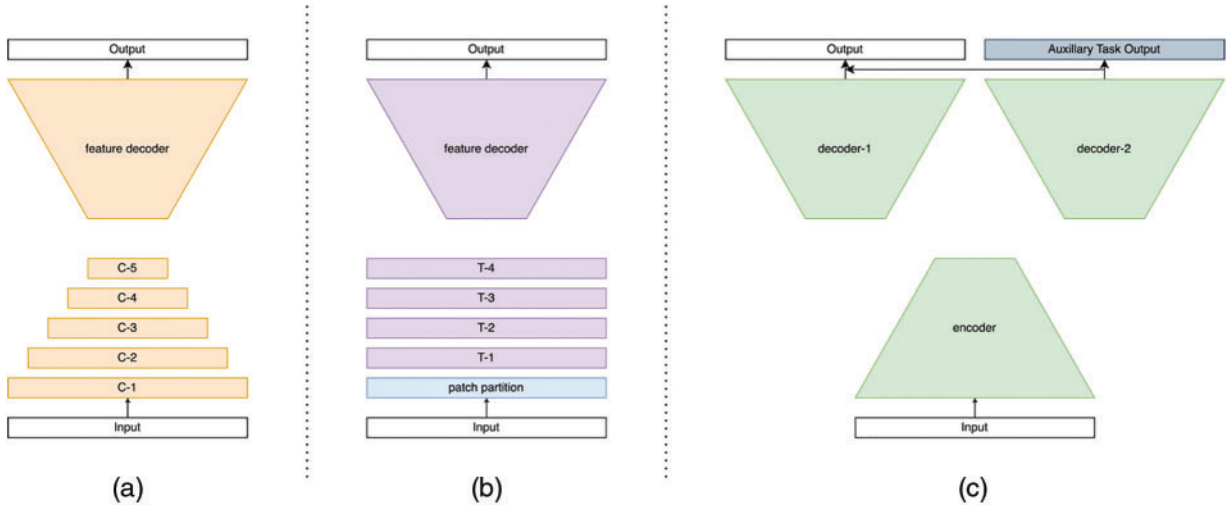

Figure 4: Simplified architectures of CNN-based models, Transformer-based models, and multi-task learning frameworks, where (a) refers to the CNN-based architecture, the lower part refers to the pre-trained CNN backbone, and the upper part refers to the specialized decoder architecture. (b) has a similar architecture, but the backbone is replaced by the Transformer backbone. (c) uses two decoder architectures, and the final result is a fusion of its own decoder and the auxiliary task

Xie et al. [33] propose TransLab, named after "Looking at the Boundary", focusing on transparent object detection. As the name suggests, TransLab takes the inspiration that transparent object often has a clear boundary, it adopts a dual path scheme with backbone features sending into two different modules for calculating boundary loss and segment loss separately. Besides, Xie et al. [34] present Trans2Seg as a reformed version of TransLab. Although Trans2Seg adopts a Transformer-based encoder-decoder architecture, it uses a CNN-based network as the feature extractor. Due to the hybrid CNN-Transformer architecture, Trans2Seg obtains a wider receptive field, thereby showing more advantages than previous CNN-based models. Like TransLab, EBLNet introduced by He et al. [35] also extensively uses boundary information. EBLNet adopts a powerful module called the fine differential module, which works in a coarse-to-fine manner to reduce the impact of complex internal components and obtain accurate boundary predictions. Besides, they use a point-based graph convolution network module to leverage accurate edge prediction to enhance global feature learning around edges, thereby improving the final prediction. Another model based on CNN is GSDNet [5], GSDNet uses glass reflections and boundaries as the two main cues for locating and segmenting glass objects. GSDNet reliably extracts boundary features and detects glass reflections from inputs to segment mirrors in images. Han et al. [36] propose a structure similar to EBLNet, but in their work, they differentiated the boundary regions into internal and external boundaries, based on which they optimized the internal and external features of transparent surfaces.

Yu et al. [37] introduce IRGNet, a lightweight RGB-Infrared fusion glass detection network specifically designed to satisfy low power consumption requirements and ensure high real-time performance for mobile robots. This network incorporates an information fusion module that amalgamates complementary feature information from RGB and infrared images at multiple scales. Zhang et al. [38] propose a novel detail-guided and cross-level fusion network, termed DCNet, to utilize label decoupling to obtain detail labels explore finer detail cues. This approach leverages discontinuities and correlations to refine the glass boundary, thereby effectively extracting local pixel and global semantic

cues from glass-like object regions and fusing features from all stages. Xiao et al. [39] propose BCNet, a network for efficient and accurate glass segmentation. It features a multi-branch boundary extraction module for precise boundary cues and a boundary cue guidance module that integrates these cues into representation learning. This approach captures contextual information across different receptive fields to detect glass objects of varying sizes and shapes. Zheng et al. [40] propose a novel glass segmentation network, termed GlassSegNet, for detecting transparent glass. This two-stage network comprises an identification stage and a correction stage. The former stage simulates human recognition by utilizing global context and edge information to identify transparent glass. While the correction stage refines the coarse predictions by correcting erroneous regions based on the information gathered in the identification stage. The transparent object segmentation network, ShuffleTrans [41], is designed with a Patch-wise Weight Shuffle operation combined with dynamic convolution to incorporate global context cues.

Le et al. [8] present ANet, the first deep learning-based network for detecting camouflaged objects. The idea of ANet is straightforward, using a salient object segmentation method to accomplish the task and an additional classification stream to determine if the image contained camouflaged objects. As a result, it does not design a deep learning-based method to specifically address the problem of camouflaged object segmentation, which is a clear departure from later work. The SINet [9] propose by Fan et al. uses a partial decoder structure that is roughly divided into two sub-modules (i.e., search module and recognition module). SINet also is the most common baseline for the camouflaged object detection task. The subsequent iteration of SINet, known as SINet v2 [42] in journal publications, advances visual outcomes by enhancing adaptability to various lighting conditions, alterations in appearance, and addressing ambiguous or undefined boundaries more effectively. The PFNet [43] employs a positioning and focusing strategy, using a positioning module to locate the position and the cascading focusing modules to refine the segmentation map using features obtained at different stages of ResNet-50. Zhang et al. [44] endeavor to unravel the intricacies of accurate detection and propose PreyNet, a model that emulates two fundamental facets of predation: initial detection and predator learning, akin to cognitive mechanisms. To harness the sensory process effectively, PreyNet integrates a bidirectional bridging interaction module, specifically crafted to discern and consolidate initial features through attentive selection and aggregation. The process of predator learning is delineated through a policy and calibration paradigm, aimed at identifying uncertain regions and fostering targeted feature refinement. Taking inspiration from biological search and recognition mechanisms, Yue et al. [45] introduce a novel framework named DCNet, which leverages two specific constraints—object area and boundary—to explore candidate objects and additional object-related edges. Employing a coarse-to-fine approach, it detects camouflaged objects by progressively refining the identification process. Utilizing a deep supervision strategy, DCNet achieves precise localization of camouflaged objects, thereby enhancing its accuracy in COD tasks.

Sun et al. [46] propose the C^2FNet represents a novel approach that capitalizes on contextual information leveraging an attention mechanism. Central to its design is an attention-induced cross-level fusion module, strategically engineered to amalgamate multiscale features effectively. C^2FNet cascades these two modules into the network to get the final prediction map. D^2CNet [47] bears a striking resemblance to that of SINet, albeit with notable distinctions. D^2CNet delineates itself through the adoption of a U-shaped network structure, where the first stage of the network operation does not incorporate the underlying information. Moreover, D^2CNet introduces supplementary modules, notably the self-refining attention unit and the cross-refining unit, augmenting its functionality and enabling enhanced information refinement and integration throughout the network's processing stages. Zhai et al. [48] devise a novel model of Mutual Graph Learning (MGL) that extends the

idea of mutual learning from the regular grid to the graph domain. MGL is equipped with typed functions to handle different complementary relationships, thus maximizing information interactions and obtaining superior performance gains. The updated iteration of MGL, named R-MGLv2 [49], brings forth multiple enhancements. These improvements encompass the integration of a new multi-source attention block explicitly engineered for conducting attention within the realm of COD. Moreover, Zhai et al. incorporate side-output features intended to collaboratively guide the learning process. This innovation effectively alleviates the burdens associated with recurrent learning overhead and mitigates the accuracy reduction during inference at lower resolutions.

POCINet [50] propose by Liu et al. has a similar procedure to SINet, which can also be divided into search and recognition phases. The difference is that POCINet adopts a novel scheme to decode camouflaged objects using contrast information and part-object relationship knowledge. TANet [51] exploits the subtleties inherent in the texture distinction between camouflaged objects and their backgrounds. This strategic approach enables the model to cultivate texture-aware features, delve into intricate object structural details, and amplify texture differences. As a result, TANet enhances the ability to recognize texture differences, thereby improving the overall efficacy and performance. Inspired by the complementary relationship between texture labels and camouflaged object labels, Zhu et al. [52] design TINet as an interactive guidance framework that focuses on finding uncertain boundaries and texture differences through progressive interactive guidance. It maximizes the usefulness of fine-grained multilevel texture cues for guiding segmentation. The work of Chen et al. [53] is an improved version of the work of Sun et al. [46] with similar design ideas and follows the name C^2FNet. The advancements made in this work significantly enhance the previous model by implementing a multi-step refinement process that involves the iterative refining of low-level features through the utilization of preliminary maps. This iterative refinement strategy culminates in the prediction of the final outcome. The improvements introduced in this updated version showcase notable enhancements in performance compared to the preceding model. Li et al. [54] introduce the Progressive Enhancement Network (PENet), a novel system that mirrors the human visual detection system. PENet adopts a three-stage detection process, comprising object localization, texture refinement, and boundary restoration.

Ji et al. [55] introduce a novel network architecture named ERRNet, which stands out for its innovative edge-based reversible re-calibration mechanism. Compared to prevailing methods, ERRNet demonstrates substantial performance enhancements coupled with notably higher processing speeds. Its ability to achieve superior performance while maintaining efficiency positions ERRNet as a promising solution with broad applicability. Chen et al. [56] propose a novel boundary-guided network (BgNet), to tackle the challenging task in a systematic coarse-to-fine fashion. The architecture of BgNet is characterized by the localization module and the boundary-guided fusion module. The dual-module strategy empowers BgNet to achieve accurate and expeditious segmentation of camouflaged regions, thereby establishing its efficacy in addressing this intricate problem. The BGNet [57] shares a parallel motivation akin to BgNet, aiming to strategically incorporates essential object-related edge semantics into the representation learning process, compelling the model to prioritize the generation of features that accentuate the object's structural elements. By explicitly integrating edge semantics and extensively leveraging boundary features, BGNet enables the model to discern and emphasize critical structural components of camouflaged objects, thereby elevating its capability for accurately localizing boundaries, a crucial aspect within this domain.

Obviously, COD methods heavily rely on boundary information, and BSA-Net [58] introduces a novel boundary-guided separation of attention mechanism. The network design is based on procedural steps observed in the human way to detect camouflaged objects, where object boundaries are

delineated by recognizing subtle differences between foreground and background. Distinguishing itself from existing networks, BSA-Net adopts a dual-stream separated attention module specifically crafted to emphasize the separation between the image background and foreground. This unique design incorporates a reverse attention stream, facilitating the exclusion of the interior of camouflaged objects to prioritize the background. Conversely, the normal attention stream works to restore the interior details, emphasizing foreground elements. Both attention streams are steered by a boundary-guiding module, collaboratively combining their outputs to augment the model's understanding and refinement of object boundaries. FAPNet [59] adopts a multifaceted approach. FAPNet capitalizes on cross-level correlations, enhancing the overall contextual understanding. One notable advantage of FAPNet lies in its adaptability and scalability to the polyp segmentation task, demonstrating its versatility and robustness beyond the realm of COD. Li et al. [60] propose another boundary-guided network (FindNet) with the utilization of texture cues from a single image. The capability of FindNet to perform accurate detection across diverse visual conditions characterized by varying textures and boundaries underscores its versatility and robustness. Sun et al. [61] introduce the Edge-aware Mirror Network (EAMNet), which implements a two-branch architecture facilitating mutual guidance between the segmentation and edge detection branches, establishing a cross-guidance mechanism to amplify the extraction of structural details from low-level features.

In the realm of COD, accurate annotations prove challenging due to the resemblance between camouflaged foreground and background elements, particularly around object boundaries. Liu et al. [62] highlight concerns regarding direct training with noisy camouflage maps, positing that this approach may result in models lacking robust generalization capabilities. To address this, they introduce an explicitly designed aleatoric uncertainty estimation technique to account for predictive uncertainty arising from imperfect labeling. Their proposed framework, OCENet, positions itself as a confidence-aware solution. This framework leverages dynamic supervision to generate both precise camouflage maps and dependable aleatoric uncertainty estimations. Once OCENet is trained, the embedded confidence estimation network is capable of assessing pixel-wise prediction accuracy autonomously, reducing reliance on ground truth camouflage maps for evaluation purposes.

HeelNet [63] employing cascading decamouflage modules to iteratively refine the prediction graph. These modules are composed of distinct components: a region enhancement module and a reverse attention mining module, strategically designed for precise detection and the thorough extraction of target objects. Additionally, HeelNet introduces a novel technique called classification-based label reweighting. This method generates a gated label graph, serving as supervisory guidance for the network. Its purpose is to aid in identifying and capturing the most salient regions of camouflaged objects, thereby facilitating the complete acquisition of the target object. The objective of CubeNet [64] revolves around harnessing hierarchical features extracted from various layers and diverse input supervisions. Specifically, CubeNet employs X-connections to facilitate multi-level feature fusion, utilizes different supervised learning methods, and employs a refinement strategy to elaborate on the complex details of camouflaged objects. To this end, it makes full use of edge information and considers it as an important cue for capturing object boundaries. By incorporating these methodologies, CubeNet endeavors to enhance the detection and delineation of camouflaged objects, effectively leveraging hierarchical features, diverse supervisory signals, and edge information for comprehensive object understanding and boundary delineation. Zhai et al. [65] present a novel approach in their paper, introducing the Deep Texton-Coherence Network (DTC-Net). The primary focus of DTC-Net revolves around extracting discriminative features through the comprehensive understanding of spatial coherence within local textures. This understanding facilitates the effective detection of camouflaged objects within scenes. DTC-Net implements a deep supervision mechanism across multiple layers.

This mechanism involves applying supervision at various stages of the network, facilitating the iterative refinement of network parameters through continuous feedback from these supervisory signals. This approach serves to promote the effective updating and optimization of the network's parameters, enhancing the model's performance in camouflage object detection. Zhai et al. [66] advocate for leveraging the intricate process of figure-ground assignment to enhance the capabilities of CNNs in achieving robust perceptual organization, even in the presence of visual ambiguity. By integrating the figure-ground assignment mechanism into CNN architectures, the model's success in various challenging applications, notably in COD tasks, underscores the potential efficacy of integrating cognitive-inspired mechanisms into CNN architectures. DGNet [67], a newly introduced deep framework, revolutionizes COD by leveraging object gradient supervision. The linchpin of this architecture lies in the gradient-induced transition, symbolizing a fluid connection between context and texture features. This mechanism essentially establishes a soft grouping between the two feature sets. The application of DGNet in various scenarios, including polyp segmentation, defect detection, and transparent object segmentation, has demonstrated remarkable efficacy and robustness, further validating its superiority and versatility in diverse visual detection tasks. Hu et al. [68] aim to address the challenges by prioritizing the extraction of high-resolution texture details, thereby mitigating the problem of detail degradation that leads to blurred edges and boundaries. Their approach includes the introduction of HitNet, a novel framework designed to enhance low-resolution representations by incorporating high-resolution features into iterative feedback loops. In addition, the authors propose an iterative feedback loss that adds an extra layer of constraints to each feedback connection. This iterative feedback loss approach aims to further refine the iterative connections, thereby enhancing the model's ability to capture the fine-grained details that are critical. Wang et al. [69] propose a new framework named FLCNet, which includes an underlying feature mining module, a texture-enhanced module, and a neighborhood feature fusion module. Deng et al. [70] propose a new ternary symmetric fusion network for detecting camouflaged objects by fully fusing features from different levels and scales. To effectively enhance detection performance, Shi et al. [71] propose a novel model featuring context-aware detection and boundary refinement.

Motivated by the complementary relationship between boundaries and camouflaged object regions, Yu et al. [72] propose an alternate guidance network named AGNet for enhanced interaction. They introduce a feature selective module to choose highly discriminative features while filtering out noisy background features. Liu et al. [73] propose MFNet, a novel network for multi-level feature integration. Xiang et al. [74] propose a double-branch fusion network with a parallel attention selection mechanism. Yan et al. [75] propose a matching-recognition-refinement network (MRR-Net) to break visual wholeness and see through camouflage by matching the appropriate field of view. Zhang et al. [76] design a novel cross-layer feature aggregation network (CFANet). CFANet effectively aggregates multi-level and multi-scale features from the backbone network by exploring the similarities and differences of features at various levels.

To sum up, CNN-based methods can offer high accuracy and robust performance due to their ability to automatically learn and extract complex features from images, making them effective in diverse and challenging COD settings. Fig. 5 indicates the number of publications on CNN-based literature from 2019 to 2023. On the other hand, CNN-based architecture facilitates end-to-end learning and adaptability, particularly through transfer learning, which allows the use of pre-trained models to enhance efficiency and accuracy. Yet, these methods are computational resources-consuming, which could be an obstacle for real-time applications. Besides, CNNs are less effective on unseen data and vulnerable to adversarial attacks. Despite these challenges, CNN-based methods remain a reliable tool for COD in complex scenarios with sufficient data.

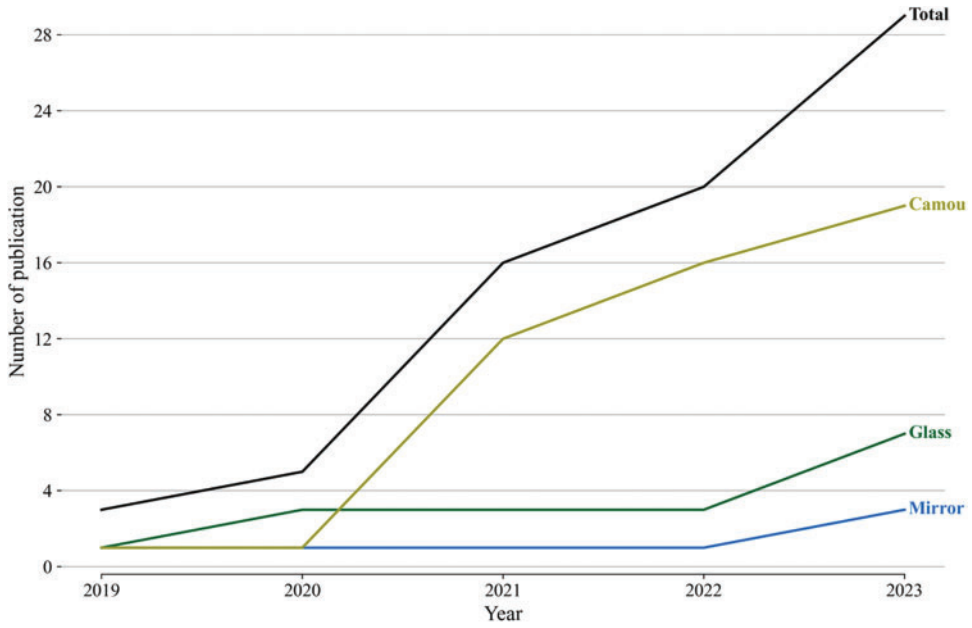

Figure 5: Number of CNN-based image-level COD methods published annually from 2019 to 2023

4.2 Transformer-Based Models

Huang et al. [77] present the SATNet which applies a transformer as the backbone network. Observing the content in the mirror is not strictly symmetry with the real-world object, i.e., loose symmetry, they proposed a network that takes as input an input image and its flipped image to provide data augmentation techniques and further fully exploits symmetry features and proposed a network with takes input images and its flipped images as input to fully exploit symmetry features as well as serves a data augmentation technique. Liu et al. [78] focus on the phenomenon of reflections on mirror surfaces, introducing a frequency domain feature extraction module. This module maps multi-scale features of the mirror to the frequency domain, extracts mirror-specific features, and suppresses interference caused by reflections of external objects. Additionally, they propose a cross-level fusion module based on reverse attention, which integrates features from different levels to enhance overall performance. To identify mirror features in more diverse scenes, Xie et al. [79] introduce a cross-space-frequency window transformer, which is designed to extract both spatial and frequency characteristics for comprehensive texture analysis.

Furthermore, some transparent object detection methods employ transformer architecture to detect glass-like objects. Zhang et al. [80] contribute Trans4Trans, which means Transformer for Transparent, an encoder-decoder model based entirely on the Transformer architecture. With a carefully designed Transformer pairing module and dual-path structure, Trans4Trans outperforms the then-SOTA transparent object segmentation method, namely Trans2Seg. Xin et al. [81] observe glass-induced image distortion and introduced a visual distortion-aware module to mitigate this problem. This module captures multi-scale visual distortion information and integrates it effectively, allowing the Swin-B backbone network to focus on regions affected by glass-induced distortion, thereby accurately identifying these surfaces. In addition, the inclusion of glass surface centroid information through a classification sub-task improves the accuracy of glass mask predictions. To differentiate between glass and non-glass regions, the transformer architecture leverages two crucial visual cues:

boundary and reflection feature learning. Vu et al. [82] propose TransCues, a pyramidal transformer encoder-decoder architecture designed for the segmentation of transparent objects from color images.

The study [83] constructs a contextual attention module to extract backbone features using a self-attention approach and proposes a new enhanced feature fusion algorithm for detecting glass regions in a single RGB image. Additionally, it introduces a VIT-based deep semantic segmentation architecture that associates multilevel receptive field features and retains the feature information captured at each level. Hu et al. [84] introduce a novel convolutional attention glass segmentation network, designed to minimize the number of training cycles and iterations, thereby enhancing performance and efficiency. The network employs a custom edge-weighting scheme to optimize glass detection within images, further improving segmentation precision. Inspired by the scale integration strategy and refinement method, Xu et al. [85] propose MGNet, featuring a fine-rescaling and merging module to enhance spatial relationship extraction and a primary prediction guiding module to mine residual semantics from fused features. An uncertainty-aware loss supervises the model to produce high-confidence segmentation maps. Observing glass naturally results in blurs. Building on this intrinsic visual blurriness cue, Qi et al. [86] propose a novel visual blurriness aggregation module that models blurriness as a learnable residual. This approach extracts and aggregates valuable multiscale blurriness features, which guide the backbone features to detect glass with high precision. The Progressive Glass Segmentation Network (PGSNet) [87] is constructed using multiple discriminability enhancement modules and a focus-and-exploration-based fusion strategy. This design progressively aggregates features from high-level to low-level, enabling a coarse-to-fine glass segmentation approach.

Yang et al. [88] introduce an innovative method named Uncertainty-Guided Transformer Reasoning (UGTR), utilizing probabilistic representational models in conjunction with a transformer architecture to facilitate explicit reasoning under uncertainty. The foundational concept involves the initial acquisition of an estimate and its associated uncertainty by learning the conditional distribution of the backbone output. Subsequently, the attention mechanism is employed to deliberate over these uncertain regions, culminating in a refined and definitive prediction. This method strategically amalgamates the strengths of Bayesian learning and transformer-based inference, leveraging both deterministic and probabilistic information. The synergistic integration of these elements enhances the model's capability to discern camouflaged objects effectively. Inherent uncertainty poses a significant challenge which is compounded by two primary biases observed in the training data. The "center bias" prevalent in the dataset causes models to exhibit poor generalization, as they tend to focus on detecting camouflaged objects primarily around the image center, termed as "model bias." Additionally, accurately labeling the boundaries of camouflaged objects is challenging due to the resemblance between the object and its surroundings, leading to inaccuracies in defining the object's scope, known as "data bias." To effectively address these biases, Zhang et al. [89] propose leveraging uncertainty estimation techniques. They introduce a predictive uncertainty estimation approach, combining model uncertainty and data uncertainty. Their proposed solution, PUENet, comprises a Bayesian conditional VAE to achieve predictive uncertainty estimation. PUENet aims to mitigate the impact of model and data biases by estimating predictive uncertainty. Liu et al. [90] challenge the conventional bio-inspired framework used in object detection methodologies, highlighting the inherent limitations in the recurrent search for objects and boundaries, which can be taxing and limiting for human perception. Their proposed solution involves a transformer-based model that enables the simultaneous detection of the object's accurate position and its intricate boundary by extracting features related to the foreground object and its surrounding background, enabling the acquisition of initial object and boundary features. Pei et al. [91] introduce OSFormer, a one-stage transformer framework which is founded on two fundamental architectural components. OSFormer adeptly combines local feature

extraction with the assimilation of extensive contextual dependencies. This integration enhances its capability to accurately predict camouflaged instances by efficiently leveraging both local and long-range contextual information. Yin et al. [92] decompose the multi-head self-attention mechanism into three distinct segments. Each segment is tasked with discerning camouflaged objects from the background by employing diverse mask strategies. Additionally, they introduce CamoFormer, a novel framework designed to progressively capture high-resolution semantic representations. This is achieved through a straightforward top-down decoder augmented by the proposed masked separable attention mechanism, enabling the attainment of precise segmentation outcomes.

Existing methods often replicate the predator's sequential approach of positioning before focusing, but they struggle to locate camouflaged objects within cluttered scenes or accurately outline their boundaries. This limitation arises from their lack of holistic scene comprehension while concentrating on these objects. Mei et al. [93] contend that an ideal model for COD should concurrently process both local and global information, achieving a comprehensive perception of the scene throughout the segmentation process. Their proposed Omni-Perception Network (OPNet) aims to amalgamate local features and global representations, facilitating accurate positioning of camouflaged objects and precise focus on their boundaries, respectively. Huang et al. [94] introduce the FSPNet, a transformer-based framework which employs a hierarchical decoding strategy aimed at enhancing the local characteristics of neighboring transformer features by progressively reducing their scale. The objective is to accumulate subtle yet significant cues progressively, aiming to decode critical object-related information that might otherwise remain imperceptible. Drawing inspiration from query-based transformers, Dong et al. [95] present a unified query-based multi-task learning framework, UQFormer, specifically designed for camouflaged instance segmentation. UQFormer approaches instance segmentation as a query-based direct set prediction task, eliminating the need for additional post-processing techniques like non-maximal suppression. Jiang et al. [96] introduce a novel joint comparative network (JCNet), leveraging joint salient objects for contrastive learning. The key innovation within JCNet lies in the design of the contrastive network, which generates a distinct feature representation specifically for the camouflaged object, setting it apart from others. The methodology involves the establishment of positive and negative samples, alongside the integration of loss functions tailored to different sample types, contributing to the overall efficacy of the approach.

Liu et al. [97] present the MSCAF-Net, a comprehensive COD framework. This framework concentrates on acquiring multi-scale context-aware characteristics by employing the enhanced Pyramid Vision Transformer (PVTv2) model as the primary extractor for global contextual data across various scales. An improved module for expanding the receptive field is developed to fine-tune characteristics at each scale. Additionally, they introduce a cross-scale feature fusion module to effectively blend multi-scale information, enhancing the variety of features extracted. Moreover, a dense interactive decoder module is formulated to generate a preliminary localization map, utilized for fine-tuning the fused features, leading to more precise detection outcomes. Drawing inspiration from human behavior, which involves approaching and magnifying ambiguous objects for clearer recognition, Xing et al. [98] introduce a novel three-stage architecture termed the SARNet which alternate their focus between foreground and background, employing attention mechanisms to effectively differentiate highly similar foreground and background elements, thereby achieving precise separation. Song et al. [99] present an innovative approach centered on focus areas, which signify regions within an image that exhibit distinct colors or textures. They introduce a two-stage focus scanning network named FSNet. This process aims to enhance the model's ability to discern camouflaged objects by capturing detailed information within these distinctive regions. Yang et al. [100] introduce an innovative occlusion-aware transformer network (OAFormer) aimed at precise identification of occluded camouflaged

objects. Within OAFormer, a hierarchical location guidance module is developed to pinpoint potential positions of camouflaged objects. This approach enables OAFormer to cleverly capture the full picture of camouflaged objects and integrates an assisted supervision strategy to enhance the learning capability of the model. Bi et al. [101] propose a novel architecture which comprising an in-layer information enhancement module and a cross-layer information aggregation module. By combining shallow texture information with deep semantic information, this architecture accurately locates target objects while minimizing noise and interference. Liu et al. [102] present Bi-level Recurrent Refinement Network (Bi-RRNet. It includes a Lower-level RRNet (L-RRN) that refines high-level features with low-level features in a top-down manner, and an Up-level RRNet (U-RRN) that polishes these features recurrently, producing high-resolution semantic features for accurate detection.

The Transformer architecture excels in handling long-range dependencies and can easily extract contextual information, which makes it an ideal choice for our COD task. Transformer-based models have considerable ability to capture global dependencies in images, so they can help reliably detect confusing objects without being affected by complex conditions and environments. The complexity of such models may add obstacles to the tuning part. Despite these shortcomings, the Transformer-based models in the environment can surpass CNN-based models in terms of accuracy in detecting confusing objects. Fig. 6 presents the number of publications on Transformer-based methods from 2021 to 2023.

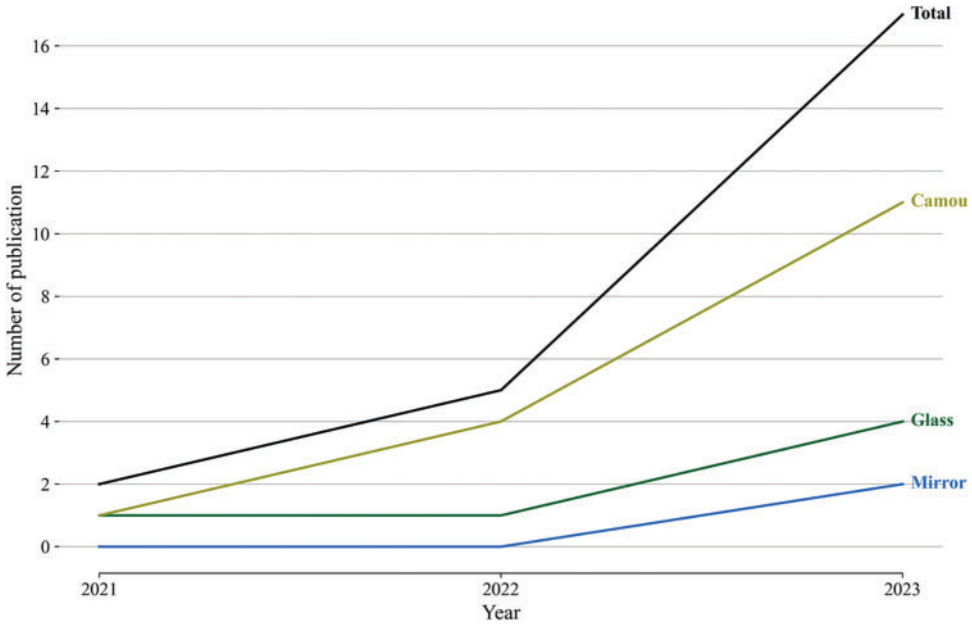

Figure 6: Number of Transformer-based image-level COD methods published annually from 2021 to 2023

4.3 Multi-Task Learning Framework

GlassNet [103] utilizes the label decoupling framework in the glass detection task, which has the advantage of better predicting the labels of pixels near the actual boundaries. Specifically, they use the label decoupling procedure on the Ground Truth (GT) map to obtain internal diffusion maps and boundary diffusion maps, which in turn serve as the GT maps of additional supervision streams in a multi-task learning manner. The RFENet designed by Fan et al. [104] also utilizes the

boundary as an additional supervisory stream to work in an independent but consistent manner. It introduces semantic and boundary-supervised losses at different feature stages in a cascading manner to achieve synergistic feature enhancement. Zhang et al. [105] propose an algorithm for staged feature extraction that employs a multitype backbone network, integrating features from both CNNs and transformers. Additionally, a multiview collector is used to extract cross-modal fusion features from diverse perspectives. Wan et al. [106] propose a novel bidirectional cross-modal fusion framework incorporating shift-window cross-attention for glass segmentation which includes a feature exchange module and a shifted-window cross-attention feature fusion module within each transformer block stage to calibrate, exchange, and fuse cross-modal features. Chang et al. [107] propose PanoGlassNet for panoramic images which uses a novel module with four branches of varying kernel sizes and deformable convolutions to capture the wide field of view and irregular boundaries. Given that most existing structures are complex and heavy, while lightweight structures often lack accuracy, Zhou et al. [108] propose a novel Asymmetric Depth Registration Network student model. This model, trained with distilled knowledge, is designed to address these limitations effectively. Zhou et al. [109] propose a novel uncertainty-aware transformer localization network for RGB-D mirror segmentation. This approach is inspired by biomimicry, particularly the observational behavior patterns of humans. It aims to explore features from various perspectives and concentrate on complex features that are challenging to discern during the coding stage.

Lv et al. [110] devise a multi-task learning framework termed RankNet, aimed at concurrent localization, segmentation, and ranking of camouflaged objects. The motivation is that explicitly capturing the unique properties of camouflaged objects in their surroundings not only enriches the understanding of camouflage and animal evolutionary strategies but also provides valuable insights for the development of more sophisticated camouflage techniques. Notably, specific components or features of camouflaged objects play a key role in predator discrimination of camouflaged objects in their surroundings. The proposed network consists of three interrelated models, a localization model designed to identify discriminatory regions that make camouflaged objects conspicuous, a segmentation model responsible for delineating the complete range of camouflaged objects, and a novel ranking model designed to assess the ability to detect differences between different camouflaged entities. The collaborative operation of these three modules within RankNet yields promising results when tested on widely used datasets.

He et al. [111] introduce the FEDER model, specifically addressing the inherent similarity between foreground and background elements. FEDER employs learnable wavelets to decompose features into distinct frequency bands, mitigating this similarity. To tackle the issue of ambiguous boundaries, FEDER adopts an auxiliary edge reconstruction task concurrent with the primary objective. The FEDER model achieves enhanced precision in generating prediction maps with precise object boundaries through simultaneous learning of both tasks. Lv et al. [112] introduce a triple-task learning framework capable of concurrently localizing, segmenting, and ranking camouflaged objects, thus quantifying the level of conspicuousness in camouflage. Due to the absence of datasets for the localization and ranking models, the authors employ an eye tracker to generate localization maps. These maps are subsequently aligned with instance-level labels to create their ranking-based training and testing dataset, providing a pioneering approach to comprehensively evaluating and ranking.

Yang et al. [113] propose a novel perspective, emphasizing its potential to deepen the understanding of camouflage and revolutionize the approach to detecting camouflaged objects. Subsequently, acknowledging the intrinsic connections between salient object detection (SOD) and COD, introducing a multi-task learning framework. This framework captures the inherent relationships between the two tasks from diverse angles. The task-consistent attribute, established through an adversarial

learning scheme, seeks to accentuate the boundary disparities between camouflaged objects and backgrounds, thereby achieving comprehensive segmentation of the camouflaged objects. Xing et al. [114] introduce a groundbreaking paradigm termed the "pre-train, adapt, and detect" approach. This method leverages a significantly pre-trained model, allowing the direct transfer of extensive knowledge obtained from vast multi-modal datasets. To tailor the features for the downstream task, a lightweight parallel adapter is integrated, facilitating necessary adjustments. Furthermore, a multi-task learning scheme is implemented to fine-tune the adapter, enabling the utilization of shared knowledge across various semantic classes for enhanced performance. Lyu et al. [115] introduce the UEDG architecture, adept at amalgamating probabilistic-derived uncertainty and deterministic-derived edge information to achieve precise detection of concealed objects. UEDG harnesses the advantages of both Bayesian learning and convolution-based learning, culminating in a robust multitask-guided approach.

The multi-task learning framework allows models to be trained on related tasks, which can bring multiple benefits. By training models on related tasks simultaneously, such as depth estimation and edge detection, our model can exploit complementary information in shared features, thereby improving overall performance. This framework can be easily extended from existing methods, achieving stronger generalization capabilities across different scenarios and reducing the risk of overfitting. However, it also presents challenges, such as the need for carefully balanced loss functions to ensure that all tasks are learned effectively without one dominating the others. They require more complex architectures and hyperparameter optimization, and often require the help of additional data or pseudo-labels.

4.4 Models Using Multimodal Inputs

Mei et al. [116] present the first mirror segmentation model called PDNet that leverages the information from the depth map. PDNet subdivides the mirror segmentation task into two separate stages (i.e., positioning and delineating). The previous stage is performed by exploiting semantic and depth discontinuities from RGB and depth maps respectively. The delineating stage utilizes lower-level features to refine the mirrored area progressively to obtain the final map. Kalra et al. [117] introduce polarization cues to the task of transparent object segmentation, naming the model Polarized Mask R-CNN. Specifically, polarized CNNs use three separated CNN backbones (RGB images) and two polarization cues to extract features separately and fuse the features with a self-designed attention module to better utilize features from different sources. GlassSemNet [118] introduces an additional semantic ground truth mask that serves as additional input. GlassSemNet uses two independent backbone networks (i.e., SegFormer and ResNet50) on different inputs to extract spatial and semantic features, and then uses an attention mechanism to segment glass objects from the enhanced features. PGSNet [119] is another network that takes Angle of Linear Polarization (AoLP) and Degree of Linear Polarization (DoLP) information as additional input. Unlike the Polarized Mask R-CNN, PGSNet is a more complex structure that makes full use of multimodal cues with the Conformer as its backbone. Noting the subtle physical differences in the response of glass-like objects to thermal radiation and visible light, Huo et al. [120] introduce another modal input (i.e., thermal images.) RGBTSeg builds the network in an encoder-decoder fashion and implements an efficient fusion strategy with a novel multimodal fusion module.

Yan et al. [121] present MirrorNet, a bio-inspired network tailored for camouflage body measurements, in which the bio-inspired representation uses flipped images to reveal more information, uniquely exploiting synergies between instance segmentation and bio-inspired attack streams. Unlike traditional models that rely on a single input stream, this innovative model integrates two segmentation streams. Pang et al. [122] introduce ZoomNet, a mixed-scale triplet network designed to emulate

the human behavior of zooming in and out when observing ambiguous images. The central strategy employed by ZoomNet involves utilizing zooming techniques to acquire discriminative mixed-scale semantics. ZoomNet aims to capture nuanced visual details across multiple scales, thereby enhancing its ability to discern and accurately predict objects in the presence of ambiguous or vague visual information. Jia et al. [123] align with ZoomNet's approach by employing human attention principles alongside a coarse-to-fine detection strategy. Their proposed framework, named SegMaR, operates through an iterative refinement process involving Segmentation, Magnification, and Reiteration across multiple stages for detection. Specifically, SegMaR introduces a novel discriminative mask that directs the model's focus toward fixation and edge regions. Notably, the model utilizes an attention-based sampler to progressively amplify object regions without requiring image size enlargement. Comprehensive experimentation demonstrates SegMaR's remarkable and consistent enhancements in performance.

Unlike existing approaches utilizing contextual aggregation techniques developed primarily for SOD, Lin et al. [124] approach introduces a new method, FBNet, which aims to address a key challenge, i.e., the prevalent contextual aggregation strategy tends to prioritize distinctive objects while potentially attenuating the features of less discriminative objects. The FBNet approach incorporates frequency learning to effectively suppress high-frequency texture information. In addition, the proposed FBNet integrates a gradient-weighted loss function that strategically directs the method to emphasize the contour details, thus refining the learning process. Luo et al. [125] introduce a De-camouflaging Network (DCNet) comprising a pixel-level camouflage decoupling module and an instance-level camouflage suppression module, marking a novel approach in the field. The authors introduce reliable reference points to establish a more robust similarity measurement, aiming to diminish the impact of background noise during segmentation. By integrating these two modules, the DCNet models de-camouflaging, enabling precise segmentation of camouflaged instances.

Zhong et al. [126] assert that the objective of the task surpasses replicating human visual perception within a singular RGB domain; rather, it aims to transcend human biological vision. They present FDCOD, a robust network integrating two specialized components to effectively incorporate frequency clues into CNN models. The frequency enhancement module encompasses an offline discrete cosine transform enabling the extraction and refinement of significant information embedded within the frequency domain. Subsequently, a feature alignment step is employed to fuse the features derived from both the RGB and frequency domains. Furthermore, to maximize the utilization of frequency information, Zhong et al. propose the high-order relation module which is designed to handle the intricate fusion of features, leveraging the rich information obtained from the fused RGB and frequency domains. This module thereby facilitates the comprehensive integration and exploitation of frequency clues, augmenting the network's capacity for accurate and nuanced detection. Cong et al. [127] introduce FPNet, integrating a learnable and separable frequency perception mechanism driven by semantic hierarchy within the frequency domain.

PopNet [128] integrates depth cues into the task. Rather than directly deriving depth maps from RGB images, Wu et al. employ modern learning-based techniques to infer reliable depth maps in real-world scenarios. This approach utilizes pre-trained depth inference models to establish the "pop-out" prior for objects in a 3D context. The "pop-out" prior assumes object placement on the background surface, enabling reasoning about objects in 3D space. Xiang et al. [129] investigate the role of depth information, utilizing depth maps generated through established monocular depth estimation methodologies. However, due to inherent discrepancies between the MDE dataset and the camouflaged obejct dataset, the resulting depth maps lack the necessary accuracy for direct utilization. Zheng et al. [130] introduce a behavior-inspired framework termed the MFFN, drawing inspiration

from human approaches to identifying ambiguous objects in images. This framework mirrors human behavior by employing multiple perspectives, angles, and distances to observe such objects. MFFN leverages the interplay between views and channels to explore channel-specific contextual information across diverse feature maps through iterative processes.

4.5 Detection in Videos

Lin et al. [131] propose the first video mirror detection model VCNet which takes video as input. To make use of the properties of video data, VCNet applies a novel Dual Correspondence (DC) module to leverage both spatial and temporal correspondence inside videos. Qiao et al. [132] introduce the first polarization-guided video glass segmentation propagation solution, capable of robustly propagating glass segmentation in RGB-P video sequences. This method leverages spatiotemporal polarization and color information, combining multi-view polarization cues to reduce the view dependence of single-input intensity variations on glass objects.

Lamdouar et al. [133] introduce a novel approach to camouflaged animal detection in video sequences leveraging optical flow between consecutive frames. The model is divided into two key modules, the first of which consists of a differentiable registration module responsible for the alignment of consecutive frames, and the second of which consists of a motion segmentation module characterized by a modified U-shaped network structure with an additional memory component aimed at segmenting camouflaged animals. Yang et al. [134] introduce a self-supervised model specifically designed to address the intricate task of segmenting camouflaged objects within video sequences. Their approach hinges upon the strategic exploitation of motion grouping mechanisms. The methodology commences by employing a modified Transformer framework, serving as the initial stage in segmenting optical flow frames into distinct primary objects and background components within the video context. SIMO [135] is another work on RGB camouflaged object segmentation in video sequences. In this work, Lamdouar et al. designed a two-path architecture consisting of ConvNets and Transformer that accepts optical flow input sequences and is designed to learn to segment moving objects in difficult scenarios such as partial occlusion or stationary states.

The research conducted by Meunier et al. [136] operates under the assumption that the input optical flow can be effectively represented as a set of parametric motion models, commonly characterized by affine or quadratic forms. Notably, this approach eliminates the need for ground truth or manual annotation during the training phase. The research introduces an efficient data augmentation technique tailored for optical flow fields which is applicable to any network utilizing optical flow as input and is inherently designed to segment multiple motions. The resulting Expectation-Maximization (EM)-driven motion segmentation network was evaluated on both camouflaged object and salient object video datasets, demonstrating high performance while maintaining efficiency during test time.

Since current video camouflaged object detection methods usually utilize isomorphic or optical flow to represent motion, the detection error may be accumulated by motion estimation error and segmentation error. Cheng et al. [137] propose a new video detection framework that can detect camouflaged objects from video frames using short-term dynamics and long-term temporal coherence. Lamdouar et al. [138] utilize a transformer-based architecture trained on synthetic datasets, showcasing its efficacy in identifying concealed objects within real video content, such as in the case of MoCA. Their proposed model amalgamates elements from two pre-existing architectures—motion segmentation and SINet. This combined framework facilitates the production of high-resolution

segmentation masks derived from the motion stream, enhancing the model's ability to reveal concealed objects within the visual content.

Video COD research is making significant progress, but challenges remain for a number of reasons. Contemporary approaches use deep learning-based models to combine object detection and optical flow analysis to exploit information between different frames. Such approaches typically do not handle challenges such as static objects, rapid scene changes, changing lighting conditions, and reflections well. In addition, real-time detection still requires a lot of computation, especially on edge devices, and the price of high-quality data annotation is quite expensive and difficult to deploy in real scenarios. To address the above issues, future research on video COD should develop more advanced and efficient architectures. Another promising area is to combine unsupervised and semi-supervised learning techniques to exploit the large amount of unlabeled video data and reduce the dependence on large annotated datasets.

4.6 Other Methods

Costanzino et al. [139] present a simple pipeline for neural networks to estimate depth accurately for reflective surfaces without ground-truth annotations. They generate reliable pseudo labels by in-painting mirror objects and using a monocular depth estimation model. Li et al. [140] develop UJSC, a combined SOD and COD segmentation model which leverages the connection between them. The authors constructed a similarity measurement module to refine the feature encoder since the saliency and camouflage streams' predictions should not intersect. In addition, they introduce adversarial learning to train the predictive encoder for generating the final result and the confidence estimation module for modeling uncertainty in certain regions of the image. The performance achieved a satisfactory result despite using only optical flow as input. To delve into cross-task relevance through a "contrast" approach, Li et al. [141] incorporate contrast learning into their dual-task learning framework. By injecting a structure based on adversarial training, they explored multiple training strategies specialized for discriminators, thus improving the stability of training. Zhao et al. [142] propose to utilize existing successful SOD models for camouflage object detection to reduce the development costs associated with COD models. Their central premise lies in the fact that SOD and COD share two aspects of information, namely the semantic representation of objects used to differentiate between objects and context, and the contextual attributes that play a key role in determining the class of an object.

Observing mirror reflections is crucial to how people perceive the presence of mirrors, and such mid-level features can be effectively transferred from self-supervised pre-trained models. Lin et al. [143] aim to enhance mirror detection methods by proposing a novel self-supervised learning pre-training framework. This framework progressively models the representation of mirror reflections during the pre-training process. Le et al. [144] propose a simple yet efficient CFL framework that undergoes a dual-stage training process to achieve its optimized performance. This scene-driven framework capitalized on the diverse advantages offered by different methods, adaptive selecting the most suitable models for each image. Song et al. [145] propose FDNet, strategically combines Convolutional Neural Networks and Transformer architectures to encode multi-scale images simultaneously. To synergistically exploit the advantages of both encoders, the authors designed a feature grafting module based on a cross-attention mechanism.

The primary aim of camouflaged object detection techniques is the identification of objects seamlessly blending into their environments visually. While prevailing COD methodologies concentrate solely on recognizing camouflaged objects within familiar categories in the training dataset,

they confront challenges in accurately identifying objects from unfamiliar categories, resulting in diminished performance. Real-world implementation proves arduous due to the complexities in amassing adequate data for recognized categories, compounded by the demanding expertise needed for accurate labeling, rendering these established approaches unfeasible. Li et al. [146] introduce a novel zero-shot framework tailored to proficiently detect previously unseen categories of camouflaged objects. The incorporation of Li's graph reasoning, underpinned by a dynamic searching strategy, prioritizes object boundaries, effectively mitigating the influence of background elements.

Chen et al. [147] observe that despite the remarkable success of the Segment Anything Model (SAM) large model in various image segmentation tasks, it encountered limitations in tasks like shadow detection and camouflaged object detection. Rather than opting for fine-tuning SAM directly, they introduced SAM-Adapter, a novel approach aiming to enhance SAM's performance in these challenging tasks. SAM-Adapter integrates domain-specific information into the segmentation network through simple yet effective adapters. He et al. [148] address the constraints of prevalent techniques relying on extensive datasets with pixel-wise annotations, a process inherently laborious due to the intricate and ambiguous object boundaries. Their introduced CRNet framework via scribble learning underscores the importance of structural insights and semantic relationships to augment the model's comprehension and detection capabilities in scenarios with limited annotation information. Primarily, they propose a new consistency loss function, leveraging scribble annotations outlining object structures without detailing pixel-level information. This function aims to guide the network in accurately localizing camouflaged object boundaries. Additionally, CRNet incorporates a feature-guided loss, using both directly extracted visual features from images and semantically significant model-captured features.

Zhang et al. [149] introduce the concept of referring camouflaged object detection, a novel task focused on segmenting specified camouflaged objects. This segmentation is achieved using a small set of referring images containing salient target objects for reference. Ma et al. [150] propose a cross-level interaction network with scale-aware augmentation. The scale-aware augmentation module calculates the optimal receptive field to perceive object scales, while the cross-level interaction module enhances feature map context by integrating scale information across levels.

Le et al. [151] aim to automatically learn transformations that reveal the underlying structure of camouflaged objects, enabling the model to better identify and segment them. They propose a learnable augmentation method in the frequency domain via a Fourier transform approach, dubbed CamoFourier. Chen et al. [152] propose a new paradigm that treats camouflaged object detection as a conditional mask-generation task by leveraging diffusion models. They employ a denoising process to progressively refine predictions while incorporating image conditions. The stochastic sampling process of diffusion allows the model to generate multiple possible predictions, thus avoiding the issue of overconfident point estimation. Chen et al. [153] propose a diffusion-based framework. This novel framework treats the camouflaged object segmentation task as a denoising diffusion process, transforming noisy masks into precise object masks. Zhang et al. [154] formulate unsupervised camouflaged object segmentation as a source-free unsupervised domain adaptation task, where both source and target labels are absent during the entire model training process. They define a source model comprising self-supervised vision transformers pre-trained on ImageNet. In contrast, the target domain consists of a simple linear layer and unlabeled camouflaged objects. Liu et al. [155] present the first systematic work on military high-level camouflage object detection, targeting objects embedded in chaotic backgrounds. Inspired by biological vision, which first perceives objects through global search and then strives to recover the complete object, they propose a novel detection network called MHNet. Li et al. [156] recognize the ambiguous semantic biases in camouflaged object datasets that

affect detection results. To address this challenge, they design a counterfactual intervention network (CINet) to mitigate these biases and achieve accurate results.

This section introduces some novel methods in COD, most of which cannot be directly benchmarked with previous methods. Yet it presents the probable direction for future research, such as the zero-shot framework for COD, the novel sam-adaptor to fine-tune the segment anything model, and more recently the diffusion model-based COD framework. We summarize notable characteristics of the reviewed models in Tables 1 and 2.

Table 1: Characteristics of reviewed image-based methods for glass-like and mirror object detection

Categories	Number	Model	Pub.	Code	Backbone
Glass	G1	EBLNet	ICCV'21	Yes	ResNeXt101
Glass	G2	GDNetB	TPAMI'22	No	ResNeXt101
Glass	G3	GDNet	CVPR'20'	–	ResNeXt101
Glass	G4	GlassNet	CGF'22	No	ResNet50
Glass	G5	GSDNet	CVPR'21	–	ResNeXt101
Glass	G6	RFENet	IJCAI'23	Yes	ResNet50
Glass	G7	Trans2Seg	IJCAI'21	Yes	ResNet50
Glass	G8	Trans4Trans	ICCV'21	Yes	PVT
Glass	G9	TransLab	ECCV'19	Yes	ResNeXt101
Glass	G10	RGBTSeg	TIP'23	Yes	ResNeXt101
Glass	G11	PGSNet	CVPR'22	–	Conformer
Glass	G12	GlassSemNet	NeurIPS'22	–	SegFormer & ResNet50
Glass	G13	RGBDSeg	ArXiv'22	No	ResNeXt101
Glass	G14	VDNet	CAD&CG'23	–	Swin Transformer
Glass	G15	IRGNet	CAC'23	–	ResNet50
Glass	G16	DCNet	ICIC'23	–	ResNeXt101
Glass	G17	BCNet	ICASSP'23	–	PVTv2
Glass	G18	PanoGlassNet	TIM'23	Yes	–
Mirror	M1	MirrorNet	ICCV'19	–	ResNeXt101
Mirror	M2	PMDNet	CVPR'20	–	ResNeXt101
Mirror	M3	SANet	CVPR'22	–	ResNeXt101
Mirror	M4	VCNet	TPAMI'22	Yes	ResNeXt101
Mirror	M5	SATNet	AAAI'23	Yes	Swin Transformer
Mirror	M6	HetNet	AAAI'23	Yes	ResNeXt101
Mirror	M7	VMDNet	CVPR'23	Yes	ResNeXt101
Mirror	M8	PDNet	CVPR'21	–	ResNet50
Mirror	M9	EGNet	WSCG'23	Yes	EfficientNetV2-Medium
Mirror	M10	SEMCNet	SPL'24	No	SegFormer & ResNet50
Mirror	M11	CSFwinformer	TIP'23	Yes	Swin Transformer
Mirror	M12	ADRNet-S_*	IF'23	–	SegFormer
Mirror	M13	WSMD	AAAI'23	Yes	PVT
Mirror	M14	UTLNet	MM'23	Yes	ConvNeXt
Mirror	M15	IEBAFNet	arXiv'23	No	DeeplabV3+

Table 2: Characteristics of reviewed image-based methods for camouflaged object detection

Number	Model	Pub.	Code	Backbone
CO1	BASNet	arXiv'21	No	–
CO2	BGNet	IJCAI'22	Yes	Res2Net-50
CO3	BGNet	KBS'22	Yes	ResNet-50
CO4	BSANet	AAAI'22	Yes	Res2Net-50
CO5	C^2FNet	IJCAI'21	Yes	Res2Net-50
CO6	C^2FNet	TCSVT'22	Yes	Res2Net-50
CO7	CamoFormer-S	arXiv'23	–	Swin-B
CO8	CRNet	AAAI'23	Yes	ResNet-50
CO9	CubeNet	PR'22	No	ResNet-50
CO10	D2CNet	TIE'21	No	Res2Net-50
CO11	DCNet	CVPR'23	Yes	ResNet-50
CO12	DCE	arXiv'21	No	ResNet-50
CO13	DGNet	MIR'23	Yes	EffNet-B4
CO14	DQNet	arXiv'22	Yes	ResNet-50 & ViT
CO15	DTCNet	TMM'22	No	ResNet-50
CO16	DTIT	ICPR'22	Yes	SegFormer
CO17	EAMNet	ICME'23	Yes	Res2Net-50
CO18	ERRNet	PR'22	–	ResNet-50
CO19	EVP	CVPR'23	Yes	SegFormer
CO20	FBNet	TMCCA'23	No	ResNet-50
CO21	FDNet	ICME'23	Yes	Res2Net-50 & PVT
CO22	FEDER	CVPR'23	Yes	ResNet-50
CO23	FindNet	TIP'22	No	Res2Net-50
CO24	FPNet	MM'23	Yes	Res2Net-50
CO25	FDCOD	CVPR'22	No	Res2Net-50
CO26	FSNet	TIP'23	Yes	Swin
CO27	FSPNet	CVPR'23	Yes	ViT
CO28	PINet	ICME'22	Yes	ResNet-50
CO29	HitNet	AAAI'23	Yes	PVT
CO30	JCNet	TIM'23	No	Swin-S
CO31	R-MGL	CVPR'21	Yes	ResNet-50
CO32	MGL	TIP'23	No	ResNet-50
CO33	MSCAFNet	TCSVT'23	Yes	PVTv2
CO34	OAFormer	ICME'23	No	PVTv2
CO35	OCENet	WACV'22	Yes	ResNet-50
CO36	OSFormer	ECCV'22	Yes	ResNet-50
CO37	PADNet	arXiv'23	No	ViT
CO38	PENet	IJCAI'23	No	Res2Net-50
CO39	PFNet	CVPR'21	Yes	Res2Net-50
CO40	PopNet	ICCV'23	Yes	–
CO41	PreyNet	MM'22	Yes	Res2Net-50
CO42	RankNet	CVPR'21	Yes	ResNet-50

(Continued)

Table 2 (continued)

Number	Model	Pub.	Code	Backbone
CO43	SAMadaptor	arXiv'23	Yes	SAM
CO44	SARNet	TCSVT'23	Yes	PVT
CO45	SegMaR	CVPR'22	Yes	Res2Net-50
CO46	SINetV2	TPAMI'22	Yes	Res2Net-50
CO47	SINet	CVPR'20	Yes	ResNet-50
CO48	TANet	TCSVT'23	No	Res2Net50
CO49	TINet	AAAI'21	No	ResNet-50
CO50	UEDG	MM'23	Yes	PVT
CO51	UGTR	ICCV'21	Yes	ResNet-50
CO52	UJSC	CVPR'21	Yes	ResNet-50
CO53	ZoomNet	CVPR'22	Yes	ResNet-50
CO54	Bi-RRNet	PR'23	Yes	VAN-small
CO55	LRSNet	TCSVT'23	Yes	ResNet-50
CO56	CINet	TETCI'23	Yes	EfficientNet
CO57	CamoDiff	AAAI'24	Yes	–
CO58	CMNet	TCSVT'23	Yes	Res2Net-50
CO59	CFANet	ICME'23	Yes	Res2Net-50
CO60	MRRNet	TNNLS'23	Yes	ResNet-50
CO61	OPNet	IJCV'23	Yes	Conformer-B
CO62	MFNet	Sensors'23	Yes	Res2Net-50
CO63	ASBI	CVIU'23	Yes	ResNet-50

5 Dataset

5.1 Glass Dataset

Table 3 lists the characteristics of the related COD dataset. Mei et al. [4] construct the first large benchmark dataset called the GDD for glass detection. GDD contains a total of 3916 images, of which 2980 are used for training and 936 for testing. To overcome the limitations of GDD, Lin et al. [5] contribute a more challenging dataset called GSD, composed of 4012 images collected from existing datasets and the Internet with manually annotated masks. Similarly, Xie et al. [33] contribute to the first dataset Trans10K specifically for transparent object segmentation in the same year. Trans10K is composed of 10,428 transparent object images collected from the real world, with different categories including glass objects and other transparent objects like plastic bottles. Furthermore, Xie et al. [34] extend the Trans10K dataset with a new dataset, Trans10K-v2, designed to overcome the previous version's lack of detailed transparent object categories.

Table 3: Characteristics of the related dataset

Obj. type	Dataset	Accessibility	Train	Test	Tasks
Glass	GDD	–	2980	936	Image
Glass	GSD	Yes	3202	810	Image

(Continued)

Table 3 (continued)

Obj. type	Dataset	Accessibility	Train	Test	Tasks
Glass	GSD-S	Yes	3911	608	Image
Glass	Trans10k	Yes	5000	4428	Image
Glass	RGBP-Glass	Yes	3207	1304	Image
Glass	RGBT-Glass	Yes	4427	1124	Image
Mirror	MSD	Yes	3063	955	Image
Mirror	PMD	Yes	5096	571	Image
Mirror	RGBD-Mirror	–	2000	1049	Image
Mirror	VMD-D	Yes	7835	7152	Video
CO	CHAMELEON	Yes	0	76	Image
CO	CAMO	Yes	2000	500	Image
CO	CAMO++	Yes	3500	2000	Image
CO	COD10K	Yes	6000	4000	Image
CO	NC4K	Yes	0	4121	Image
CO	ACOD2K	Yes	–	–	Image
CO	MoCA	Yes	6000	6000	Video
CO	MoCA-Mask	Yes	19,313	3626	Video

Lin et al. [118] present a dataset with semantic annotations that contains 4519 images directly collected from existing datasets that are larger than previously popular datasets such as GDD and GSD. Similarly, Mei et al. [119] propose a polarised glass dataset containing both AoLP and DoLP information. Their main observation is that glass reacts differently to light than normal objects, so AoLP and DoLP often provide useful cues about mirrors. It is verified that thermal images also reveal unique information concerned with the mirror, Huo et al. [120] present the RGB thermal image dataset called RGB-T, which is the first large-scale RGB thermal image dataset for transparent object detection. The RGB-T dataset consists of 5551 images acquired by RGB thermal cameras from a variety of scenes, manually labeled with truth maps.

5.2 Mirror Dataset

Yang et al. [6] construct the first large-scale mirror datasets named MSD with a total of 4018 manually annotated images. To overcome the shortcomings of MSD objects being too monotonous and containing only simple scenes, Lin et al. [7] propose a more challenging dataset PMD, which contains various images collected from existing datasets. As PDNet takes depth maps as additional inputs, the normal RGB dataset does not satisfy the requirement.

Mei et al. [116] construct the first RGBD dataset specifically for the mirror called RGBD-Mirror which contains 3049 RGB images with corresponding depth maps and annotated tags. As Lin et al. [157] propose the video mirror detection model, the first obstacle that needs to be overcome is the lack of the mirror dataset in video form. For this reason, they constructed the first large-scale video mirror dataset VMD-D.

5.3 Camouflaged Object Dataset

Le et al. [8] develop the CAMO dataset specifically tailored for the segmentation of camouflaged objects. Comprising 1250 images sourced from diverse online repositories, CAMO features meticulous manual annotations of semantic maps. The dataset spans a wide spectrum of categories, encompassing not only natural camouflage scenes but also diverse artificial contexts. The COD10K [9] presently stands as the most common benchmark dataset designed for camouflaged object detection. It comprises a total of 10,000 images, delineated into 5066 instances of camouflaged objects, 3000 background images, and 1934 occurrences of non-camouflaged objects. In practice, the COD10K dataset undergoes a random division where 6000 images are allocated for training purposes, while the remaining 4000 images are reserved for testing. Lv et al. [110] curate the NC4K dataset, presently recognized as the largest and most comprehensive test dataset tailored specifically for camouflaged object segmentation. This extensive dataset encompasses a total of 4121 meticulously collected images sourced from various online resources. Its vast and diverse collection serves as a popular benchmark of camouflaged object segmentation algorithms. The CAMO++ [144] represents a significant advancement in the realm of camouflaged instance segmentation datasets, surpassing its predecessor, the CAMO dataset, in both scale and scope. This new iteration encompasses 5500 images featuring individuals alongside over 90 distinct animal species, meticulously annotated at a pixel level hierarchically. CAMO++ serves as a versatile benchmark not only for camouflaged instance segmentation but also for conventional camouflaged object segmentation tasks. Its expansive coverage, balanced representation of camouflage scenarios, and meticulously detailed annotations establish it as a valuable resource for benchmarking and advancing research in the fields of camouflaged instance segmentation.

The inaugural video dataset for camouflaged objects is the Moving Camouflaged Animals (MoCA) [133] dataset, which comprises 37K image frames from 141 video sequences found on YouTube. The dataset comprises 67 animals that move in natural surroundings. The use of a bounding box in the original dataset instead of a dense segmentation mask poses a challenge when evaluating video camouflaged obejct segmentation performance. Cheng et al. [137] revamp the dataset into MoCA-Mask and develop a comprehensive benchmark with more inclusive evaluation criteria. MoCA-Mask annotates the dataset using fragment masks supported by humans and generates pseudo-ground truth masks via a bidirectional optical flow-based approach.

Artificial camouflage, a deliberate design approach employing methods like painting and camouflage uniforms, aims to exploit human visual perception traits for enhanced deception of the human visual system. Its practical utility extends notably to tasks such as aiding disaster-assisted search and rescue operations. In light of this advantage, Song et al. [145] curate ACOD2K, recognized as the most extensive artificial camouflage dataset available. Notably, camouflaged object detection methods are exclusively trained on natural camouflage images due to the predominant presence of natural camouflaged animals in existing datasets. This limitation hinders the training of models proficient in accurately detecting artificial camouflage. ACOD2K comprises 2000 images, including 1500 featuring camouflaged objects, 400 displaying non-camouflaged objects, and 100 background images. Each image underwent meticulous pixel-level matting annotations of high quality and precision. To ensure the annotation accuracy, an additional researcher performed thorough verification of all annotations.

6 Performance Measurement

This study establishes a performance investigation focusing on confusing object detection tasks, chosen due to their established nature and the availability of diverse competing methodologies. Subsequent sections will elaborate on the evaluation metrics employed in this analysis.

6.1 Metrics

To provide a comprehensive assessment of each model's accuracy and generalization capability, this survey employs a range of metrics. These metrics include widely recognized benchmarks in segmentation tasks like Intersection over Union (IoU) and pixel accuracy, complemented by less common metrics such as S measure and E measure. These metrics aim to offer a holistic evaluation of model performance across various facets.

a) IoU

The Intersection over Union (IoU) metric stands as a prevalent evaluation measure in object detection and image segmentation endeavors, serving to assess the precision of algorithmic predictions by analyzing the intersection between the segmentation mask and the ground truth. The provided IoU equation is shown in Eq. (1):

$$\text{IoU} = J(A,\ B) = \frac{P \cap T}{P \cup T}, \tag{1}$$

where P represents the prediction mask, while T denotes the ground truth. IoU assumes critical significance in appraising algorithmic accuracy in tasks necessitating spatial agreement between predicted and ground truth regions.

b) Pixel Accuracy

Pixel Accuracy (Acc) serves as a semantic segmentation metric, indicating the proportion of correctly classified pixels within an image. This metric quantifies the ratio between the number of accurately classified pixels and the total pixel count within the image.

$$\text{Acc} = \frac{\sum_{k=1}^{K} n_{kk}}{N}. \tag{2}$$

where the variable n_{kk} represents the total number of pixels classified and correctly labeled as class k. Put differently, it signifies the count of true positives for class k. On the other hand, N denotes the total number of pixels.

c) MAE

The Mean Absolute Error (MAE) stands as a metric employed to gauge the average magnitude of discrepancies between predicted and actual values. It offers a direct means to comprehend the typical deviation of predictions from true values.

MAE's interpretability arises from its representation of the average absolute variance between predicted and actual values. This attribute proves particularly valuable when evaluating models, especially in scenarios where outliers or substantial errors might exert a pronounced influence on the assessment of model performance.

$$\text{MAE} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \left| P\left(i,j\right) - T\left(i,j\right) \right|, \tag{3}$$

where W and H are the width and height of T, respectively, and i, j are pixel coordinates in T.

d) BER

The Balanced Error Rate (BER) serves as a widely adopted performance metric in binary classification tasks, particularly beneficial for evaluating models in the presence of imbalanced datasets.

BER considers both the false positive rate (FPR) and the false negative rate (FNR) to derive a more comprehensive assessment of a model's accuracy. This metric offers a balanced evaluation by simultaneously accounting for errors in positive and negative classifications, providing a nuanced understanding of the model's overall performance.

$$\mathrm{BER} = \frac{1}{2}\left(\mathrm{FPR} + \mathrm{FNR}\right) = \frac{1}{2}\left(\frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}} + \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}\right). \tag{4}$$

e) F-Measure

The F-measure, or F1 score offers an evaluation of a classification model's accuracy. It consolidates precision and recall metrics into a singular value, rendering a more equitable assessment of a model's performance, particularly in scenarios involving imbalanced datasets.

In elucidating the F-measure, it's pivotal to introduce the concepts of precision and recall. Precision quantifies the accuracy of positive predictions made by a model, while recall assesses the model's ability to correctly identify all positive instances.

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}; \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{5}$$

The F_β score represents a broader variant of the F1 score, offering control over the balance between precision and recall by incorporating a parameter, β. It is formulated as a weighted harmonic mean of precision and recall, with the beta parameter governing the emphasis on recall in comparison to precision. Adjusting the beta value enables a nuanced adjustment of the F_β score, allowing practitioners to tailor the evaluation based on the specific importance accorded to precision and recall in a given context.

$$\mathrm{F}_\beta = \frac{\left(1 + \beta^2\right) \mathrm{Precision} \times \mathrm{Recall}}{\beta^2\, \mathrm{Precision} + \mathrm{Recall}}. \tag{6}$$

f) S-Measure

The Structure measure S_α [158] is a metric employed to quantify the structural similarity existing between a non-binary prediction map and a corresponding ground-truth mask.

$$\mathrm{S}_\alpha = (1 - \alpha)\, S_0\left(P, T\right) + \alpha S_r\left(P, T\right). \tag{7}$$

where the parameter α serves as a weighting factor, governing the balance between object-aware similarity S_0 and region-aware similarity S_α within the Structure measure S_α calculation. The default setting for α, as per the original paper, is 0.5.

g) E-Measure

The Enhanced-alignment measure [159] stands as a novel evaluation metric specifically designed for assessing binary foregrounds. It incorporates considerations of both local and global similarity between two binary maps. The formulation of this metric is defined as follows:

$$\mathrm{E}_\phi = \frac{1}{W \times H} \sum\nolimits_{i=1}^{W} \sum\nolimits_{j=1}^{H} \phi\left(P\left(i, j\right) - T\left(i, j\right)\right), \tag{8}$$

where the symbol ϕ represents the enhanced alignment matrix defined in the literature.

6.2 Evaluation

We conducted distinct experiments for COD tasks. The outcomes of these experiments are elaborated upon in the subsequent subsections.

6.2.1 Glass

Table 4 showcases the assessment of several deep learning-based methodologies within the GDD [4] datasets, utilizing five widely adopted metrics elaborated upon in the previous section. To ensure fairness and comparability, all models underwent training on the respective training set and were subsequently evaluated on the testing set. The findings suggest that GlassNet [103] and EBLNet [35] demonstrate closely aligned performance concerning Acc metric. In contrast, for the IoU, F_β, MAE score, VDNet [91] emerges as a formidable competitor to EBLNet. Notably, these distinctions in performance are partially influenced by the constraints inherent in the GDD dataset. Table 5 illustrates the assessment of available model within the multi-modal datasets.

Table 4: Quantitative comparison on GDD [4] test set for glass detection task

Number	Methods	IoU	Acc	F_β	MAE	BER
G1	EBLNet	0.887	0.944	0.94	0.055	5.36
G2	GDNetB	0.878	0.941	0.939	0.061	5.52
G3	GDNet	0.876	0.939	0.937	0.063	5.62
G4	GlassNet	0.887	0.946	0.937	0.054	5.42
G8	Trans4Trans	0.844	0.922	0.905	0.078	7.36
G9	TransLab	0.816	0.903	0.892	0.097	9.7
G14	VDNet	0.918	–	0.951	0.039	3.9
G16	DCNet	0.884	–	0.905	0.058	5.81
G17	BCNet	0.892	–	0.947	0.055	5.00

Table 5: Quantitative comparison on RGBP-Glass [129] test set for glass detection task

Number	Model	mIOU	F_β	mAE	mBER
G3	GDNet	77.64	0.807	0.119	11.79
G5	GSDNet	78.11	0.806	0.122	12.61
G7	Trans2Seg	75.21	0.799	0.122	13.23
G9	TransLab	73.59	0.772	0.148	15.73
G11	PGSNet	81.08	0.842	0.091	9.63

6.2.2 Mirror

Table 6 presents the evaluation of several deep learning-based methodologies across three distinct testing datasets, utilizing three widely adopted metrics. The outcomes indicate that CSFwinformer [89] showcases superior performance compared to alternative methodologies. This advantage is attributed to the utilization of a transformer as the backbone network. Given that the content observed in the

mirror does not strictly align with the real-world object, displaying a loose symmetry, their approach involves a network that leverages both an input image and its mirrored counterpart. This methodology is proposed as a means of data augmentation, effectively harnessing symmetry features to their fullest extent. The network architecture is designed to accept input images and their corresponding flipped versions, enabling comprehensive exploration and utilization of symmetry features while serving as a data augmentation technique.

Table 6: Quantitative comparison on MSD [6], PMD [7], and RGBD-Mirror [126] test set for mirror detection

Number	Methods	MSD			PMD			RGBD-Mirror		
		MAE	IoU	F_β	MAE	IoU	F_β	MAE	IoU	F_β
M1	MirrorNet	0.065	0.79	0.857	0.043	0.585	0.741	0.062	0.684	0.723
M2	PMDNet	0.047	0.815	0.892	0.032	0.66	0.794	0.054	0.723	0.775
M3	SANet	0.054	0.798	0.877	0.032	0.668	0.795	0.048	0.75	0.873
M4	VCNet	0.044	0.801	0.898	0.028	0.64	0.815	0.052	0.73	0.849
M5	SATNet	0.033	0.854	0.922	0.025	0.694	0.847	0.031	0.784	0.906
M6	HetNet	0.043	0.828	0.906	0.029	0.69	0.814	0.048	0.739	0.853
M9	EGNet	0.081	–	0.850	0.036	–	0.790	–	–	–
M11	CSFwinformer	0.045	0.821	0.896	0.024	0.701	0.838	0.031	0.787	0.900
M14	UTLNet	0.040	0.830	0.892	–	–	–	0.032	0.805	0.858

6.2.3 Camouflaged Object

Camouflaged object stands out as a more pervasive topic compared to the previously discussed glass and mirror detection. The field witnesses a notably substantial influx of proposed work annually. Notably, among these methodologies as illustrated in Table 7, the CamoDiff [152] demonstrates a significant performance advantage across various datasets, outperforming other models across four key metrics.

Table 7: Quantitative comparison on CAMO, COD10K test set for camouflaged object detection

Number	Methods	CAMO-Test				COD10K-Test			
		S_α	E_ϕ	wF_β	M	S_α	E_ϕ	wF_β	M
CO1	BASNet	0.749	0.796	0.646	0.096	0.802	0.855	0.677	0.038
CO2	BGNet	0.812	0.87	0.749	0.073	0.831	0.901	0.722	0.033
CO3	BGNet	0.804	0.859	0.719	0.075	0.804	0.881	0.663	0.039
CO4	BSANet	0.796	0.851	0.717	0.079	0.818	0.891	0.699	0.034
CO5	C2FNet	0.796	0.854	0.719	0.08	0.813	0.89	0.686	0.036
CO6	C2FNet	0.8	0.869	0.73	0.077	0.811	0.891	0.691	0.036
CO7	CamoFormer-S	0.876	0.935	0.832	0.043	0.862	0.932	0.772	0.024
CO8	CRNet	0.818	0.897	0.744	0.046	0.733	0.832	0.576	0.049
CO9	CubeNet	0.788	0.838	0.682	0.085	0.795	0.864	0.644	0.041

(Continued)

Table 7 (continued)

Number	Methods	CAMO-Test				COD10K-Test			
		S_α	E_ϕ	wF_β	M	S_α	E_ϕ	wF_β	M
CO10	D2CNet	0.774	0.818	0.735	0.087	0.807	0.876	0.72	0.037
CO11	DCNet	0.87	0.922	0.831	0.05	0.873	0.934	0.81	0.022
CO12	DCE	0.819	0.881	0.798	0.069	0.829	0.903	0.751	0.032
CO13	DGNet	0.839	0.901	0.769	0.057	0.822	0.896	0.693	0.033
CO14	DQNet	0.898	0.944	0.898	0.034	0.882	0.93	0.801	0.021
CO15	DTCNet	0.778	0.804	0.667	0.084	0.79	0.821	0.616	0.041
CO16	DTIT	0.857	0.916	0.796	0.05	0.824	0.896	0.695	0.034
CO17	EAMNet	0.831	0.89	0.763	0.064	0.839	0.907	0.733	0.029
CO18	ERRNet	0.761	0.817	0.66	0.088	0.78	0.867	0.629	0.044
CO19	EVP	0.846	0.895	0.777	0.059	0.843	0.907	0.742	0.029
CO20	FBNet	0.783	0.839	0.702	0.081	0.809	0.889	0.684	0.035
CO21	FDNet	0.836	0.886	0.777	0.066	0.857	0.918	0.763	0.028
CO22	FEDER	0.807	0.873	0.785	0.069	0.823	0.9	0.74	0.032
CO23	FindNet	0.895	0.944	0.839	0.027	0.811	0.883	0.688	0.036
CO24	FPNet	0.852	0.905	0.806	0.056	0.85	0.913	0.748	0.029
CO25	FDCOD	0.844	0.898	0.778	0.062	0.837	0.918	0.731	0.03
CO26	FSNet	0.88	0.933	0.861	0.041	0.87	0.938	0.81	0.023
CO27	FSPNet	0.856	0.899	0.799	0.05	0.851	0.895	0.735	0.026
CO28	PINet	0.814	0.868	0.737	0.073	0.825	0.891	0.704	0.035
CO29	HitNet	0.844	0.904	0.806	0.056	0.869	0.936	0.804	0.023
CO30	JCNet	0.85	0.8	0.913	0.054	0.852	0.763	0.927	0.054
CO31	R-MGL	0.775	0.847	0.673	0.088	0.814	0.865	0.666	0.035
CO32	MGL	0.775	0.847	0.673	0.088	0.814	0.865	0.666	0.035
CO33	MSCAFNet	0.873	0.929	0.828	0.046	0.865	0.927	0.775	0.024
CO34	OAFormer	0.866	0.924	0.826	0.048	0.86	0.927	0.773	0.025
CO35	OCENet	0.807	0.866	0.767	0.075	0.832	0.89	0.745	0.033
CO36	OSFormer	0.799	0.858	0.767	0.073	0.811	0.881	0.701	0.034
CO37	PADNet	0.836	0.886	0.777	0.066	0.857	0.918	0.763	0.028
CO38	PENet	0.828	0.89	0.771	0.063	0.831	0.908	0.723	0.031
CO39	PFNet	0.782	0.852	0.695	0.085	0.8	0.868	0.66	0.04
CO40	PopNet	0.806	0.869	0.821	0.073	0.827	0.897	0.789	0.031
CO41	PreyNet	0.79	0.842	0.708	0.077	0.813	0.881	0.697	0.034
CO42	RankNet	0.787	0.854	0.696	0.08	0.804	0.892	0.673	0.037
CO43	SAMadaptor	0.847	0.873	0.765	0.07	0.883	0.918	0.801	0.025
CO44	SARNet	0.815	0.872	0.742	0.071	0.831	0.901	0.722	0.033
CO45	SegMaR	0.815	0.874	0.753	0.071	0.833	0.899	0.724	0.034
CO46	SINetV2	0.82	0.882	0.743	0.07	0.815	0.887	0.68	0.037
CO47	SINet	0.751	0.771	0.606	0.1	0.771	0.806	0.551	0.051
CO48	TANet	0.823	0.882	0.763	0.066	0.829	0.902	0.725	0.03
CO49	TINet	0.781	0.848	0.678	0.087	0.793	0.861	0.635	0.042

(Continued)

Table 7 (continued)

Number	Methods	CAMO-Test				COD10K-Test			
		S_α	E_ϕ	wF_β	M	S_α	E_ϕ	wF_β	M
CO50	UEDG	0.868	0.922	0.819	0.048	0.858	0.924	0.766	0.025
CO51	UGTR	0.784	0.851	0.684	0.086	0.817	0.89	0.666	0.036
CO52	UJSC	0.803	0.853	0.728	0.076	0.817	0.892	0.684	0.035
CO53	ZoomNet	0.82	0.892	0.752	0.066	0.838	0.911	0.729	0.029
CO54	Bi-RRNet	0.843	0.909	0.802	0.054	0.84	0.912	0.746	0.026
CO55	LSRNet	0.789	0.840	0.751	0.079	0.805	0.880	0.711	0.037
CO56	CINet	0.847	0.899	0.794	0.055	0.841	0.914	0.744	0.028
CO57	CamoDiff	0.879	0.940	0.854	0.042	0.880	0.943	0.815	0.020
CO58	CMNet	0.835	0.902	0.828	0.063	0.906	0.966	0.888	0.024
CO59	CFANet	0.815	0.876	0.761	0.073	0.898	0.944	0.846	0.025
CO60	MRRNet	0.826	0.880	0.797	0.070	0.835	0.901	0.753	0.032
CO61	OPNet	0.858	0.915	0.817	0.050	0.857	0.919	0.767	0.026
CO62	MFNet	0.824	0.883	0.763	0.067	0.834	0.901	0.726	0.032
CO63	ASBI	0.871	0.931	0.845	0.050	0.868	0.938	0.802	0.024

Our findings reveal that the testing set for COD10K exhibits the highest overall resolution compared to the CAMO dataset. This observation suggests that models incorporating higher resolutions or employing multi-scale modeling techniques would derive advantages from this characteristic. Numerous models demonstrate sensitivity to the resolution of input images, implying the potential utilization of specific models tailored for distinct resolutions in real-world applications.

7 Summary

This paper endeavors to offer an extensive overview of deep learning methodologies tailored for the detection of confusing objects. To comprehensively survey this field's global landscape, we contribute in four significant ways. Initially, we present an in-depth survey encompassing these specific tasks, delineating their background, taxonomy, task-specific challenges, and advancements achieved in the era of deep learning. Notably, this survey represents the most comprehensive compilation available to date. Additionally, we establish the most current benchmark for glass, mirror and camouflaged object detection, a pivotal and thriving area at deep learning community. We thoroughly evaluate the latest deep learning methods to facilitate future advancement. This benchmark facilitates quantitative comparisons among state-of-the-art techniques.

Finally, to deliberate on potential avenues for this research field and stimulate further research and development in this domain, we have noticed that diffusion models and large language models (LLMs) have gained significant research interest in the deep learning community. Not only in their specific domains, such as generation tasks and natural language processing, but the capabilities of these models have also drawn substantial attention to various tasks.

The integration of diffusion models and LLMs with previous COD methods presents a promising perspective for enhancing the detection of challenging and confusing objects. To sum up, diffusion models excel in generating diverse, high-quality images and identifying anomalies, while LLMs provide

contextual understanding and reasoning. Combining these capabilities can lead to robust multimodal systems that improve capability and adaptability in complex environments. Future advancements may focus on leveraging these models for enhanced dataset augmentation, contextual labeling, and adaptive learning, leading to more accurate and context-aware object detection solutions.

Acknowledgement: We are grateful to our families and friends for their unwavering understanding and encouragement.

Funding Statement: This work is supported by the National Natural Science Foundation of China Nos. 62302167, U23A20343, Shanghai Sailing Program (23YF1410500), Chenguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (23CGA34).

Author Contributions: Study conception and design: Kunkun Tong, Xin Tan, Jingyu Gong; analysis and interpretation of results: Guchu Zou, Zhenyi Qi, Zhizhong Zhang; draft manuscript preparation: Kunkun Tong, Guchu Zou, Yuan Xie, Lizhuang Ma. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, Jun. 2017. doi: 10.1145/3065386.

[2] L. -C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017. doi: 10.1109/TPAMI.2016.2644615.

[4] H. Me *et al.*, "Don't hit me! glass detection in real-world scenes," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Seattle, WA, USA, Jun. 13–19, 2020.

[5] J. Lin, Z. He, and R. W. H. Lau, "Rich context aggregation with reflection prior for glass surface detection," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Nashville, TN, USA, Jun. 20–25, 2021.

[6] X. Yang, H. Mei, K. Xu, X. Wei, B. Yin and R. W. Lau, "Where is my mirror?" presented at the Proc. IEEE Int. Conf. Comput. Vis., Seoul, Republic of Korea, Oct. 27–Nov. 2, 2019.

[7] J. Lin, G. Wang, and R. W. Lau, "Progressive mirror detection," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Seattle, WA, USA, Jun. 13–19, 2020.

[8] 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. doi: 10.1016/j.cviu.2019.04.006.

[9] D. -P. Fan, G. -P. Ji, G. Sun, M. -M. Cheng, J. Shen and L. Shao, "Camouflaged object detection," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Seattle, WA, USA, Jun. 13–19, 2020.

[10] 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. doi: 10.1109/5.959337.

[11] J. Yin, Y. Han, W. Hou, and J. Li, "Detection of the mobile object with camouflage color under dynamic background based on optical flow," *Proc. Eng.*, vol. 15, pp. 2201–2205, 2011. doi: 10.1016/j.proeng.2011.08.412.

[12] 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, Sep. 2017. doi: 10.1109/TCSVT.2016.2555719.

[13] Z. Xie, R. Qiu, S. Wang, X. Tan, Y. Xie and L. Ma, "PIG: Prompt images guidance for night-time scene parsing," *IEEE Trans. Image Process.*, vol. 33, pp. 3921–3934, 2024.

[14] Galun, Sharon, Basri, and Brandt, "Texture segmentation by multiscale aggregation of filter responses and shape elements," presented at the IEEE Int. Conf. Comput. Vis., Nice, France, Oct. 13–16, 2003.

[15] D. T. Nguyen, W. Li, and P. O. Ogunbona, "Human detection from images and videos: A survey," *Pattern Recognit.*, vol. 51, pp. 148–175, Mar. 2016. doi: 10.1016/j.patcog.2015.08.027.

[16] J. Parham, C. Stewart, J. Crall, D. Rubenstein, J. Holmberg and T. Berger-Wolf, "An animal detection pipeline for identification," presented at the 2018 IEEE Winter Conf. Appl. Comput. Vis., Lake Tahoe, NV, USA, Mar. 12–15, 2018.

[17] Y. Cai *et al.*, "YOLOv4-5D: An effective and efficient object detector for autonomous driving," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021. doi: 10.1109/TIM.2021.3065438.

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

[19] 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, Jul. 2006. doi: 10.1016/j.visres.2006.01.006.

[20] A. Mondal, "Camouflaged object detection and tracking: A survey," *Int. J. Image Graph.*, vol. 20, no. 4, 2020, Art. no. 2050028. doi: 10.1142/S021946782050028X.

[21] S. C. J. Shi, B. J. Ren, Z. W. Wang, J. W. Yan, and Z. Shi, "Survey of camouflaged object detection based on deep learning," , (in Chinese), *J. Front. Comput. Sci. Technol.*, vol. 16, no. 12, pp. 2734–2751, 2022. doi: 10.3778/j.issn.1673-9418.2206078.

[22] 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, 2021.

[23] 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. 556, 2023, Art. no. 127050.

[24] 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, 2023, Art. no. 16. doi: 10.1007/s44267-023-00019-6.

[25] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989. doi: 10.1162/neco.1989.1.4.541.

[26] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, 2021. doi: 10.1109/TPAMI.2021.3059968.

[27] I. Goodfellow *et al.*, "Generative adversarial nets," presented at the Adv. Neural Inf. Process. Syst., Montreal, QC, Canada, Dec. 8–13, 2014.

[28] X. Tan, J. Lin, K. Xu, P. Chen, L. Ma and R. W. Lau, "Mirror detection with the visual chirality cue," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3492–3504, 2022. doi: 10.1109/TPAMI.2022.3181030.

[29] R. He, J. Lin, and R. W. Lau, "Efficient mirror detection via multi-level heterogeneous learning," presented at the AAAI Conf. Artif. Intell., Washington, DC, USA, Feb. 7–14, 2023.

[30] M. E. M. Gonzales, L. C. Uy, and J. P. Ilao, "Designing a lightweight edge-guided convolutional neural network for segmenting mirrors and reflective surfaces," *Comput. Sci. Res. Notes*, vol. 3301, pp. 107–116, 2023. doi: 10.24132/CSRN.3301.14.

[31] H. Mei *et al.*, "Mirror segmentation via semantic-aware contextual contrasted feature learning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 2s, pp. 1–22, 2023. doi: 10.1145/3566127.

[32] H. Mei, X. Yang, L. Yu, Q. Zhang, X. Wei and R. W. Lau, "Large-field contextual feature learning for glass detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3329–3346, 2022. doi: 10.1109/TPAMI.2022.3181973.

[33] E. Xie, W. Wang, W. Wang, M. Ding, C. Shen and P. Luo, "Segmenting transparent objects in the wild," presented at the Eur. Conf. Comput. Vis., Glasgow, UK, Aug. 23–28, 2020.

[34] E. Xie *et al.*, "Segmenting transparent object in the wild with transformer," *arXiv preprint arXiv: 2101.08461*, 2021.

[35] H. He *et al.*, "Enhanced boundary learning for glass-like object segmentation," presented at the IEEE Int. Conf. Comput. Vis., Montreal, QC,Canada, Oct. 11–17, 2021.

[36] D. Han and S. Lee, "Internal-external boundary attention fusion for glass surface segmentation," *arXiv preprint arXiv:2307.00212*, 2023.

[37] N. Yu and Z. Tian, "Glass detection network based on RGB-infrared image fusion," presented at the 2023 China Automat. Congress, Chongqing, China, Nov. 17–19, 2023.

[38] J. Zhang, G. Yang, and C. Liu, "DCNet: Glass-like object detection via detail-guided and cross-level fusion," presented at the Int. Conf. Intell. Comput., Comput., Zhengzhou, China, Aug. 10–13, 2023.

[39] J. Xiao, T. Chen, X. Hu, G. Zhang, and S. Wang, "Boundary-guided context-aware network for camouflaged object detection," *Neural Comput. Appl.*, vol. 35, no. 20, pp. 15075–15093, 2023. doi: 10.1007/s00521-023-08502-3.

[40] C. Zheng, P. Li, X. -P. Zhang, X. Lu, and M. Wei, "Don't worry about mistakes! glass segmentation network via mistake correction," *arXiv preprint arXiv:2304.10825*, 2023.

[41] B. Zhang *et al.*, "ShuffleTrans: Patch-wise weight shuffle for transparent object segmentation," *Neural Netw.*, vol. 167, pp. 199–212, 2023. doi: 10.1016/j.neunet.2023.08.011.

[42] D. -P. Fan, G. -P. Ji, M. -M. Cheng, L. Shao, "Concealed object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6024–6042, 2021. doi: 10.1109/TPAMI.2021.3085766.

[43] H. Mei, G. -P. Ji, Z. Wei, X. Yang, X. Wei and D. -P. Fan, "Camouflaged object segmentation with distraction mining," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Nashville, TN, USA, Jun. 20–25, 2021.

[44] M. Zhang, S. Xu, Y. Piao, D. Shi, S. Lin and H. Lu, "PreyNet: Preying on camouflaged objects," presented at the ACM Int. Conf. Multimedia, Lisboa, Portugal, Oct. 10–14, 2022.

[45] G. Yue *et al.*, "Dual-constraint coarse-to-fine network for camouflaged object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3286–3298, 2024. doi: 10.1109/TCSVT.2023.3318672.

[46] Y. Sun, G. Chen, T. Zhou, Y. Zhang, and N. Liu, "Context-aware cross-level fusion network for camouflaged object detection," presented at the Int. Joint Conf. Artif. Intell., Montreal, QC, Canada, Aug. 19–26, 2021.

[47] K. Wang, H. Bi, Y. Zhang, C. Zhang, Z. Liu and S. Zheng, "D^2C-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, 2021.

[48] Q. Zhai, X. Li, F. Yang, C. Chen, and D. -P. Fan, "Mutual graph learning for camouflaged object detection," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Nashville, TN, USA, Jun. 20–25, 2021.

[49] Q. Zhai *et al.*, "MGL: Mutual graph learning for camouflaged object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1897–1910, 2022.

[50] 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.

[51] J. Ren *et al.*, "Deep texture-aware features for camouflaged object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, pp. 1157–1167, 2021.

[52] J. Zhu, X. Zhang, S. Zhang, and J. Liu, "Inferring camouflaged objects by texture-aware interactive guidance network," presented at the AAAI Conf. Artif. Intell., Feb. 2–9, 2021.

[53] 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, pp. 6981–6993, 2022. doi: 10.1109/TCSVT.2022.3178173.

[54] 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," presented at the Int. Joint Conf. Artif. Intell., Macao, China, Aug. 19–25, 2023.

[55] 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, 2021, Art. no. 108414. doi: 10.1016/j.patcog.2021.108414.

[56] T. Chen, J. Xiao, X. Hu, G. Zhang, and S. Wang, "Boundary-guided network for camouflaged object detection," *Knowl. Based Syst.*, vol. 248, 2022, Art. no. 108901. doi: 10.1016/j.knosys.2022.108901.

[57] Y. Sun, S. Wang, C. Chen, and T. -Z. Xiang, "Boundary-guided camouflaged object detection," *arXiv preprint arXiv:2207.00794*, 2022.

[58] H. Zhu *et al.*, "I can find you! boundary-guided separated attention network for camouflaged object detection," presented at the AAAI Conf. Artif. Intell., Feb. 22– Mar. 1, 2022, pp. 22.

[59] 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. doi: 10.1109/TIP.2022.3217695.

[60] P. Li, X. Yan, H. Zhu, M. Wei, X. -P. Zhang 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. doi: 10.1109/TIP.2022.3189828.

[61] D. Sun, S. Jiang, and L. Qi, "Edge-aware mirror network for camouflaged object detection," presented at the IEEE Int. Conf. Multimedia Expo, Brisbane, Australia, Jul. 10–14, 2023.

[62] J. Liu, J. Zhang, and N. Barnes, "Modeling aleatoric uncertainty for camouflaged object detection," presented at the IEEE Winter Conf. Appl. Comput. Vis., Waikoloa, HI, USA, Jan. 3–8, 2022.

[63] M. -C. Chou, H. -J. Chen, and H. -H. Shuai, "Finding the achilles heel: Progressive identification network for camouflaged object detection," presented at the IEEE Int. Conf. Multimedia Expo, Taipei, Taiwan, China, Jul. 18–22, 2022.

[64] M. Zhuge, X. Lu, Y. Guo, Z. Cai, and S. Chen, "CubeNet: X-shape connection for camouflaged object detection," *Pattern Recognit.*, vol. 127, 2022, Art. no. 108644. doi: 10.1016/j.patcog.2022.108644.

[65] W. Zhai, Y. Cao, H. Xie, Z. -J. Zha, "Deep texton-coherence network for camouflaged object detection," *IEEE Trans. Multimedia*, vol. 25, pp. 5155–5165, 2023. doi: 10.1109/TMM.2022.3188401.

[66] W. Zhai, Y. Cao, J. Zhang, and Z. -J. Zha, "Exploring figure-ground assignment mechanism in perceptual organization," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 17030–17042, 2022.

[67] 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. doi: 10.1007/s11633-022-1365-9.

[68] X. Hu *et al.*, "High-resolution iterative feedback network for camouflaged object detection," in presented at the AAAI Conf. Artif. Intell., Washington, DC, USA, Feb. 7–14, 2023.

[69] T. Wang, J. Wang, and R. Wang, "Camouflaged object detection with a feature lateral connection network," *Electronics*, vol. 12, no. 12, 2023, Art. no. 2570. doi: 10.3390/electronics12122570.

[70] Y. Deng, J. Ma, Y. Li, M. Zhang, and L. Wang, "Ternary symmetric fusion network for camouflaged object detection," *Appl. Intell.*, vol. 53, no. 21, pp. 25216–25231, 2023. doi: 10.1007/s10489-023-04898-6.

[71] C. Shi, B. Ren, H. Chen, L. Zhao, C. Lin and Y. Zhao, "Camouflaged object detection based on context-aware and boundary refinement," *Appl. Intell.*, vol. 53, no. 19, pp. 22429–45, 2023. doi: 10.1007/s10489-023-04645-x.

[72] J. Yu *et al.*, "Alternate guidance network for boundary-aware camouflaged object detection," *Mach. Vision Appl.*, vol. 34, no. 4, 2023, Art. no. 69. doi: 10.1007/s00138-023-01424-z.

[73] K. Liu, T. Qiu, Y. Yu, S. Li, and X. Li, "Edge-guided camouflaged object detection via multi-level feature integration," *Sensors*, vol. 23, no. 13, 2023, Art. no. 5789. doi: 10.3390/s23135789.

[74] 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, 2023, Art. no. 162403. doi: 10.1007/s11432-022-3592-8.

[75] X. Yan, M. Sun, Y. Han, and Z. Wang, "Camouflaged object segmentation based on matching-recognition–refinement network," *IEEE Trans. Neural Netw. Learn. Syst.*, 2023. doi: 10.1109/TNNLS.2023.3291595.

[76] Q. Zhang and W. Yan, "CFANet: A cross-layer feature aggregation network for camouflaged object detection," presented at the IEEE Int. Conf. Multimedia Expo, Brisbane, Australia, Jul. 10–14, 2023.

[77] T. Huang, B. Dong, J. Lin, X. Liu, R. W. Lau and W. Zuo, "Symmetry-aware transformer-based mirror detection," presented at the AAAI Conf. Artif. Intell., Washington, DC, USA, Feb. 7–14, 2023, pp. 7–14.

[78] T. Liu and Z. Liu, "Mirror detection in frequency domain," presented at the Int. Conf. Commun. Netw. Machine Learn., Zhengzhou, China, Oct. 27–28, 2023.

[79] Z. Xie, S. Wang, Q. Yu, X. Tan, and Y. Xie, "CSFwinformer: Crossspace-frequency window transformer for mirror detection," *IEEE Trans. Image Process.*, vol. 33, pp. 1853–1867, 2024. doi: 10.1109/TIP.2024.3372468.

[80] J. Zhang, K. Yang, A. Constantinescu, K. Peng, K. Muller and R. Stiefelhagen, "Trans4Trans: Efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world," presented at the IEEE Int. Conf. Comput. Vis., Montreal, QC, Canada, Oct. 11–17, 2021.

[81] T. Xin, F. Qi, N. Wang, Z. Zhang, Y. Xie and L. Ma, "Glass surface detection method based on visual distortion," *J. Comput. Aided Design Comput. Graph.*, 2023. doi: 10.3724/SP.J.1089.2023-00342.

[82] T. -A. Vu *et al.*, "TransCues: Boundary and reflection-empowered pyramid vision transformer for semantic transparent object segmentation," in *ICLR 2024*, 2023.

[83] X. Hu, R. Gao, S. Yang, and K. Cho, "TGSNet: Multi-field feature fusion for glass region segmentation using transformers," *Mathematics*, vol. 11, no. 4, 2023, Art. no. 843. doi: 10.3390/math11040843.

[84] X. Hu, R. Gao, S. Yang, and K. Cho, "CAGNet: A multi-scale convolutional attention method for glass detection based on transformer," *Mathematics*, vol. 11, no. 19, 2023, Art. no. 4084. doi: 10.3390/math11194084.

[85] Z. Xu and Q. Chen, "Glass segmentation with multi scales and primary prediction guiding," *arXiv preprint arXiv:2402.08571*, 2024.

[86] F. Qi, X. Tan, Z. Zhang, M. Chen, Y. Xie and L. Ma, "Glass makes blurs: Learning the visual blurriness for glass surface detection," *IEEE Trans. Ind. Inform.*, vol. 20, no. 4, pp. 6631–6641, Apr. 2024. doi: 10.1109/TII.2024.3352232.

[87] L. Yu *et al.*, "Progressive glass segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 2920–2933, 2022. doi: 10.1109/TIP.2022.3162709.

[88] F. Yang *et al.*, "Uncertainty-guided transformer reasoning for camouflaged object detection," presented at the IEEE Int. Conf. Comput. Vis., Montreal, QC, Canada, Oct. 11–17, 2021, pp. 11–17.

[89] Y. Zhang, J. Zhang, W. Hamidouche, and O. Deforges, "Predictive uncertainty estimation for camouflaged object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 3580–3591, 2023. doi: 10.1109/TIP.2023.3287137.

[90] Z. Liu, Z. Zhang, and W. Wu, "Boosting camouflaged object detection with dual-task interactive transformer," presented at the Int. Conf. Pattern Recognit., Montreal, QC, Canada, Aug. 21–25, 2022.

[91] J. Pei, T. Cheng, D. -P. Fan, H. Tang, C. Chen and L. Van Gool, "OSFormer: One-stage camouflaged instance segmentation with transformers," presented at the Eur. Conf. Comput. Vis., Tel-Aviv, Israel, Oct. 23–27, 2022.

[92] B. Yin, X. Zhang, Q. Hou, B. -Y. Sun, D. -P. Fan and L. V. Gool, "CamoFormer: Masked separable attention for camouflaged object detection," *arXiv preprint arXiv:2212.06570*, 2022.

[93] H. Mei *et al.*, "Camouflaged object segmentation with omni perception," *Int. J. Comput. Vis.*, vol. 131, no. 11, pp. 3019–3034, 2023. doi: 10.1007/s11263-023-01838-2.

[94] Z. Huang *et al.*, "Feature shrinkage pyramid for camouflaged object detection with transformers," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Vancouver, QC, Canada, Jun. 18–22, 2023.
[95] D. Dong, J. Pei, R. Gao, T. -Z. Xiang, S. Wang, and H. Xiong, "A unified query-based paradigm for camouflaged instance segmentation," *arXiv preprint arXiv:2308.07392*, 2023.
[96] X. Jiang *et al.*, "Camouflaged object segmentation based on joint salient object for contrastive learning," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–16, 2023. doi: 10.1109/TIM.2023.3306520.
[97] Y. Liu, H. Li, J. Cheng, and X. Chen, "MSCAF-Net: A general framework for camouflaged object detection via learning multiscale context-aware features," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4934–4947, Sep. 2023. doi: 10.1109/TCSVT.2023.3245883.
[98] H. Xing, Y. Wang, X. Wei, H. Tang, S. Gao 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. doi: 10.1109/TCSVT.2023.3255304.
[99] 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. doi: 10.1109/TIP.2023.3266659.
[100] X. Yang, H. Zhu, G. Mao, and S. Xing, "OAFormer: Occlusion aware transformer for camouflaged object detection," presented at the IEEE Int. Conf. Multimedia Expo, Brisbane, Australia, Jul. 10–14, 2023.
[101] H. Bi, C. Zhang, K. Wang, and R. Wu, "Towards accurate camouflaged object detection with in-layer information enhancement and cross-layer information aggregation," *IEEE Trans. Cogn. Dev. Syst.*, vol. 15, no. 2, pp. 615–624, Jun. 2023. doi: 10.1109/TCDS.2022.3172331.
[102] 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, 2023, Art. no. 109514. doi: 10.1016/j.patcog.2023.109514.
[103] C. Zheng *et al.*, "GlassNet: Label decoupling-based three-stream neural network for robust image glass detection," *Comput. Graph. Forum*, vol. 41, no. 1, pp. 377–388, 2022. doi: 10.1111/cgf.14441.
[104] K. Fan, C. Wang, Y. Wang, C. Wang, R. Yi and L. Ma, "RFENet: Towards reciprocal feature evolution for glass segmentation," *arXiv preprint arXiv:2307.06099*, 2023.
[105] H. Zhang, X. Ran, and W. Zhou, "Self-knowledge distillation-based staged extraction and multiview collection network for RGB-D mirror segmentation," *IEEE Signal Process. Lett.*, vol. 31, pp. 1029–1033, 2024. doi: 10.1109/LSP.2024.3386470.
[106] Y. Wan, Q. Zhao, J. Xu, H. Wang, and L. Fang, "DAGNet: Depthaware glass-like objects segmentation via cross-modal attention," *J. Vis. Commun. Image Represent.*, vol. 100, no. 9, 2024, Art. no. 104121. doi: 10.1016/j.jvcir.2024.104121.
[107] Q. Chang, H. Liao, X. Meng, S. Xu, and Y. Cui, "PanoGlassNet: Glass detection with panoramic RGB and intensity images," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–15, 2024. doi: 10.1109/TIM.2024.3390163.
[108] W. Zhou, Y. Cai, X. Dong, F. Qiang, and W. Qiu, "ADRNet-S$_*$: Asymmetric depth registration network via contrastive knowledge distillation for RGB-D mirror segmentation," *Inf. Fusion*, vol. 108, 2024, Art. no. 102392. doi: 10.1016/j.inffus.2024.102392.
[109] W. Zhou, Y. Cai, L. Zhang, W. Yan, and L. Yu, "UTLNet: Uncertainty-aware transformer localization network for RGB-depth mirror segmentation," *IEEE Trans. Multimedia*, vol. 26, pp. 4564–4574, 2024. doi: 10.1109/TMM.2023.3323890.
[110] Y. Lv *et al.*, "Simultaneously localize, segment and rank the camouflaged objects," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Nashville, TN, USA, Jun. 20–25, 2021.
[111] C. He *et al.*, "Camouflaged object detection with feature decomposition and edge reconstruction," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Vancouver, BC, Canada, Jun. 18–22, 2023.
[112] Y. Lv, J. Zhang, Y. Dai, A. Li, N. Barnes and D. -P. Fan, "Towards deeper understanding of camouflaged object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 7, pp. 3462–3476, Jul. 2023. doi: 10.1109/TCSVT.2023.3234578.

[113] 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, Apr. 2024. doi: 10.1109/TCSVT.2023.3308964.
[114] Y. Xing *et al.*, "Pre-train, adapt and detect: Multi-task adapter tuning for camouflaged object detection," *arXiv preprint arXiv:2307.10685*, 2023.
[115] Y. Lyu, H. Zhang, Y. Li, H. Liu, Y. Yang and D. Yuan, "UEDG: Uncertainty-edge dual guided camouflage object detection," *IEEE Trans. Multimedia*, vol. 26, pp. 4050–4060, 2024. doi: 10.1109/TMM.2023.3295095.
[116] H. Mei *et al.*, "Depth-aware mirror segmentation," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Nashville, TN, USA, Jun. 20–25, 2021.
[117] A. Kalra, V. Taamazyan, S. K. Rao, K. Venkataraman, R. Raskar and A. Kadambi, "Deep polarization cues for transparent object segmentation," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Seattle, WA, USA, Jun. 13–19, 2020.
[118] J. Lin, Y. -H. Yeung, and R. Lau, "Exploiting semantic relations for glass surface detection," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 22490–22504, 2022.
[119] H. Mei *et al.*, "Glass segmentation using intensity and spectral polarization cues," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., New Orleans, LA, USA, Jun. 19–24, 2022.
[120] D. Huo, J. Wang, Y. Qian, and Y. -H. Yang, "Glass segmentation with RGB-thermal image pairs," *IEEE Trans. Image Process.*, vol. 32, pp. 1911–1926, 2023. doi: 10.1109/TIP.2023.3256762.
[121] 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. doi: 10.1109/ACCESS.2021.3064443.
[122] 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," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., New Orleans, LA, USA, Jun. 19–24, 2022.
[123] Q. Jia, S. Yao, Y. Liu, X. Fan, R. Liu and Z. Luo, "Segment, magnify and reiterate: Detecting camouflaged objects the hard way," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., New Orleans, LA, USA, Jun. 19–24, 2022.
[124] J. Lin, X. Tan, K. Xu, L. Ma, and R. W. Lau, "Frequency-aware camouflaged object detection," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 2, pp. 1–16, 2023. doi: 10.1145/3545609.
[125] N. Luo, Y. Pan, R. Sun, T. Zhang, Z. Xiong and F. Wu, "Camouflaged instance segmentation via explicit de-camouflaging," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Vancouver, BC, Canada, Jun. 18–22, 2023.
[126] Y. Zhong, B. Li, L. Tang, S. Kuang, S. Wu and S. Ding, "Detecting camouflaged object in frequency domain," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., New Orleans, LA, USA, Jun. 19–24, 2022.
[127] R. Cong, M. Sun, S. Zhang, X. Zhou, W. Zhang and Y. Zhao, "Frequency perception network for camouflaged object detection," *arXiv preprint arXiv:2308.08924*, 2023.
[128] Z. Wu *et al.*, "Source-free depth for object popout," presented at the IEEE/CVF Int. Conf. Comput. Vis., Paris, France, Oct. 2–6, 2023.
[129] 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*, 2022.
[130] D. Zheng, X. Zheng, L. T. Yang, Y. Gao, C. Zhu and Y. Ruan, "MFFN: Multi-view feature fusion network for camouflaged object detection," presented at the IEEE Winter Conf. Appl. Comput. Vis., Waikoloa, HI, USA, Jan. 2–7, 2023.
[131] J. Lin, Y. H. Yeung, and R. W. Lau, "Depth-aware glass surface detection with cross-modal context mining," *arXiv preprint arXiv:2206.11250*, 2022.
[132] Y. Qiao *et al.*, "Multi-view spectral polarization propagation for video glass segmentation," presented at the IEEE/CVF Int. Conf. Comput. Vis., Paris, France, Oct. 2–6, 2023, pp. 2–6.
[133] H. Lamdouar, C. Yang, W. Xie, and A. Zisserman, "Betrayed by motion: Camouflaged object discovery via motion segmentation," *arXiv preprint arXiv:2011.11630*, 2020.

[134] C. Yang, H. Lamdouar, E. Lu, A. Zisserman, and W. Xie, "Selfsupervised video object segmentation by motion grouping," presented at the IEEE Int. Conf. Comput. Vis., Montreal, QC, Canada, Oct. 11–17, 2021.
[135] H. Lamdouar, W. Xie, and A. Zisserman, "Segmenting invisible moving objects," in *Br. Mach. Vis. Conf.*, Nov. 22–25, 2021.
[136] E. Meunier, A. Badoual, and P. Bouthemy, "EM-driven unsupervised learning for efficient motion segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, pp. 4462–4473, 2022. doi: 10.1109/TPAMI.2022.3198480.
[137] X. Cheng *et al.*, "Implicit motion handling for video camouflaged object detection," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., New Orleans, LA, USA, Jun. 19–24, 2022.
[138] H. Lamdouar, W. Xie, and A. Zisserman, "The making and breaking of camouflage," *arXiv preprint arXiv:2309.03899*, 2023.
[139] A. Costanzino, P. Z. Ramirez, M. Poggi, F. Tosi, S. Mattoccia and L. Di Stefano, "Learning depth estimation for transparent and mirror surfaces," presented at the IEEE Int. Conf. Comput. Vis., Paris, France, Oct. 2–6, 2023.
[140] A. Li, J. Zhang, Y. Lv, B. Liu, T. Zhang and Y. Dai, "Uncertainty-aware joint salient object and camouflaged object detection," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Nashville, TN, USA, Jun. 20–25, 2021.
[141] A. Li *et al.*, "Joint salient object detection and camouflaged object detection via uncertainty-aware learning," *arXiv preprint arXiv:2307.04651*, 2023.
[142] W. Zhao, S. Xie, F. Zhao, Y. He, and H. Lu, "Nowhere to disguise: Spot camouflaged objects via saliency attribute transfer," *IEEE Trans. Image Process.*, vol. 32, pp. 3108–3120, 2023. doi: 10.1109/TIP.2023.3277793.
[143] J. Lin and R. W. Lau, "Self-supervised pre-training for mirror detection," presented at the IEEE/CVF Int. Conf. Comput. Vis., Paris, France, Oct. 2–6, 2023.
[144] T. -N. Lee *et al.*, "Camouflaged instance segmentation in-the-wild: Dataset, method, and benchmark suite," *IEEE Trans. Image Process.*, vol. 31, pp. 287–300, 2021.
[145] Y. Song, X. Li, and L. Qi, "Camouflaged object detection with feature grafting and distractor aware," presented at the IEEE Int. Conf. Multimedia Expo., Brisbane, Australia, Jul. 10–14, 2023.
[146] 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. doi: 10.1109/TIP.2023.3308295.
[147] T. Chen *et al.*, "SAM fails to segment anything?–SAM-adapter: Adapting SAM in underperformed scenes: Camouflage, shadow, and more," *arXiv preprint arXiv:2304.09148*, 2023.
[148] R. He, Q. Dong, J. Lin, and R. W. Lau, "Weakly-supervised camouflaged object detection with scribble annotations," presented at the AAAI Conf. Artif. Intell., Washington, DC, USA, Feb. 7–14, 2023.
[149] 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.
[150] M. Ma and B. Sun, "A cross-level interaction network based on scale-aware augmentation for camouflaged object detection," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 8, no. 1, pp. 69–81, Feb. 2024. doi: 10.1109/TETCI.2023.3299305.
[151] M. -Q. Le, M. -T. Tran, T. -N. Le, T. V. Nguyen, and T. -T. Do, "Unveiling camouflage: A learnable fourier-based augmentation for camouflaged object detection and instance segmentation," *arXiv preprint arXiv: 2308.15660*, 2023.
[152] Z. Chen, K. Sun, and X. Lin, "CamoDiffusion: Camouflaged object detection via conditional diffusion models," in presented at the AAAI Conf. Artif. Intell., Vancouver, BC, Canada, Feb. 21–28, 2024, pp. 21–28.
[153] Z. Chen, R. Gao, T. -Z. Xiang, and F. Lin, "Diffusion model for camouflaged object detection," *arXiv preprint arXiv:2308.00303*, 2023.
[154] Y. Zhang and C. Wu, "Unsupervised camouflaged object segmentation as domain adaptation," presented at the IEEE/CVF Int. Conf. Comput. Vis., Paris, France, Oct. 2–6, 2023.

[155] M. Liu and X. Di, "Extraordinary MHNet: Military high-level camouflage object detection network and dataset," *Neurocomputing*, vol. 549, 2023, Art. no. 126466. doi: 10.1016/j.neucom.2023.126466.
[156] X. Li *et al.*, "Camouflaged object detection with counterfactual intervention," *Neurocomputing*, vol. 553, 2023, Art. no. 126530. doi: 10.1016/j.neucom.2023.126530.
[157] J. Lin, X. Tan, and R. W. Lau, "Learning to detect mirrors from videos via dual correspondences," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Vancouver, BC, Canada, Jun. 18–22, 2023.
[158] D. -P. Fan, M. -M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," presented at the IEEE Int. Conf. Comput. Vis., Venice, Italy, Oct. 22–29, 2017.
[159] D. -P. Fan, C. Gong, Y. Cao, B. Ren, M. -M. Cheng and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," *arXiv preprint arXiv:1805.10421*, 2018.