POLITECNICO DI TORINO
Repository ISTITUZIONALE

Promptable image segmentation: a survey of guided input techniques

Original
Promptable image segmentation: a survey of guided input techniques / Nejabat, H., D'Asaro, F., Pecora, A.E., Monopoli, T., Bottino, A.. - In: FOUNDATIONS AND TRENDS IN COMPUTER GRAPHICS AND VISION. - ISSN 1572-2740. - ELETTRONICO. - 18:1(2026), pp. 1-139. [10.1108/FTCGV-03-2026-001]

Availability:
This version is available at: 11583/3009210 since: 2026-03-26T12:48:31Z

Publisher:
Now Publishers

Published
DOI:10.1108/FTCGV-03-2026-001

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

19 November 2026

Promptable Image Segmentation: A Survey of Guided Input Techniques

Hadi Nejabat
Politecnico di Torino
hadi.nejabat@polito.it

Federico D'Asaro
Politecnico di Torino
federico.dasaro@polito.it

Alessandro Emmanuel Pecora
Politecnico di Torino
alessandro.pecora@polito.it

Tommaso Monopoli
LINKS Foundation
tommaso.monopoli@linksfoundation.com

Andrea Bottino
Politecnico di Torino
andrea.bottino@polito.it

Abstract

Prompt-based image segmentation has revolutionized computer vision by enabling more adaptive and efficient segmentation through prompts. In the context of image segmentation, the term prompt broadly refers to any auxiliary input, such as clicks, boxes, scribbles, support sets, or free-form text that guides a model's segmentation behavior. These inputs operate as task-specific signals that enable models to adapt their segmentation behavior to different contexts and objectives. This survey categorizes Promptable Image Segmentation into five primary areas: Interactive Segmentation, Referring Segmentation, Few-Shot Semantic Segmentation, Open Vocabulary Segmentation, and Foundation Models.

We explore how different prompting strategies improve segmentation performance while enabling few-shot learning and reducing reliance on extensive labeled datasets. The discussion highlights the role of foundation models in advancing segmentation capabilities by integrating separate components of these complex models and leveraging multimodal interactions.

By synthesizing state-of-the-art techniques, this study provides a structured taxonomy, identifies key challenges in multimodal fusion and generalization, and outlines future directions for developing more intelligent and adaptable segmentation systems.

1 Introduction

Prompt-based Segmentation represents a revolutionary advancement in computer vision and image processing. This technique employs both textual and visual prompts to guide the segmentation process, enabling the identification of coherent and meaningful components within an image. By utilizing these prompts, the algorithm can logically section the image, enhancing segmentation accuracy, versatility, and performance. This method has seen increased adoption due to its broad applicability across various fields, including image analysis, medical imaging, environmental monitoring, and robotics [1, 2, 3, 4].

Recent advances, most notably the adoption of the transformer's [5] self-attention mechanism in various approaches to improve CNNs [6, 7, 8], have greatly enhanced segmentation tasks [9]. These innovations led to the development of large-scale pre-trained models, also known as Foundation Models. The need to adapt these models for specific downstream tasks gave rise to Prompts, defined as an approach to guide pre-trained models in performing downstream tasks by augmenting the model input with task-specific instructions. The key to the power and flexibility of prompt-based segmentation lies in the ability to guide the model to perform accurate and

relevant segmentation based on explicit criteria or context. The majority of previous research in this field has focused on (i) how to properly craft prompts, resulting in optimal ways to encode prompts, and (ii) designing a well-structured module to fuse the input prompt with the query, where the two may differ in modality. This approach enables the interception of relationships between the query and the prompt, favoring semantic alignment between them, reducing bias, minimizing information loss, and enhancing the overall understanding of the user's objective [10, 11]. However, the Image Segmentation Task has presented an intriguing new approach by incorporating prompts directly from the pre-training stage, enabling flexible and zero-shot segmentation inference. Integrating prompts into segmentation models has its origins in the field of **Interactive Segmentation** [12], where visual prompts provided by the user, such as *clicks, bounding boxes, scribbles,* and *polygons*, are used to select the target object in the input image, guiding the model during both training and inference.

Expanding the scope of segmentation tasks, **Referring Segmentation** introduces the ability to segment images from free-form textual prompts and focuses on how to effectively tokenize the text prompts and facilitate their interaction with cross-modal image queries. This paves the way for new extensions of the semantic segmentation task. For example, prompts may target an object in the image based on its attributes (e.g., color, shape, size, etc.) or relationships to other objects in the scene, thus requiring a certain level of semantic visual, and textual reasoning skills to address this challenge effectively.

Few-Shot Semantic Segmentation (F3S) enables models to generalize to new tasks from a minimal set of examples, reducing reliance on large labeled datasets. While few-shot methods are not always directly linked to prompts, we draw connections between the structured guiding role of examples (shots) in few-shot learning and prompts in adaptive learning frameworks. Rather than focusing on traditional fine-tuning—which, while effective, requires additional training for each adaptation and is computationally demanding—we review strategies that align more naturally with prompt-based learning. Specifically, we highlight meta-learning-inspired approaches such as episodic training and in-context learning, which enable models to generalize to novel classes without extensive parameter updates. By framing F3S within this prompt-centric perspective, we emphasize methodologies that facilitate flexibility and adaptability across different segmentation scenarios.

Due to the impressive zero-shot transferability of large-scale **Pre-trained Vision-Language Models (VLPs)** such as CLIP [13] and ALIGN [14], a growing body of research has focused on leveraging these models for Open Vocabulary Semantic Segmentation [15, 16]. Unlike traditional segmentation methods that rely on fixed label sets, this approach enables segmentation based on arbitrary categories described through textual input. By projecting both visual and textual data into a shared semantic space, VLPs allow users to guide segmentation using natural language prompts, making it possible to generate segmentation masks for previously unseen categories.

Finally, with the introduction of the **Foundation Models**, such as *Segment Anything Model (SAM)* [17], the trend has shifted toward adopting and integrating different components of such models, demonstrating how their modular design can enhance segmentation quality through various prompting mechanisms. SAM-based models, distinguished by their ability to understand and interpret prompts in a highly adaptable manner, signify a major leap forward in the utilization of VLPs. These models showcase the transformative power of prompt-based methods in achieving remarkable performance with minimal annotations. The transition towards these models encapsulates the ongoing evolution in image analysis technologies, spotlighting prompt-based segmentation as a crucial area of focus within current research and applications. This trend underscores a collective recognition of the transformative potential of this approach to redefine the paradigms of image analysis across multiple domains, facilitating more accurate, context-aware, and adaptable segmentation solutions. Furthermore, Users are now presented with a multitude of choices, from selecting appropriate prompt inputs and encoders to configuring the convolutional neural networks responsible for the segmentation tasks.

In exploring the field of Prompt in Image Segmentation, we focus on how user prompt encoding and integration have evolved over the course of the methodologies proposed in recent years, with a focus on Deep Neural Network architectures. Specifically, *Prompt Encoding* refers to how user information is encoded to be injected into the segmentation model, while *Prompt Integration* denotes the mechanisms by which the prompt codification is interleaved with the input image features. Additionally, we explore an emerging research direction in prompting segmentation models, i.e., *Compositionality*—the process of combining multiple prompts to achieve improved segmentation results. We discuss the effectiveness of various prompt integration mechanisms and highlight new encoding techniques developed for unified prompt representation.

Figure 1 presents a taxonomy organized into five main areas: Interactive Segmentation (Section 5), Referring Expression Segmentation (Section 14), Few-Shot Semantic Segmentation (Section 20), Open Vocabulary Semantic Segmentation (Section 24), and Foundation Models in Segmentation (Section 25). Across these areas, we observe many emerging possibilities, with diverse strategies for encoding task information and integrating it with query images to enhance predictions. These techniques are often applied at various stages of the segmentation pipeline—such as during prompt encoding, feature extraction, or mask refinement—making their mechanisms less explicit. This work aims to highlight, define, and formalize these strategies to better align them with existing segmentation paradigms.

Contribution: The growing interest in technologies that simplify data manipulation and reduce reliance on expert analysts underscores the importance and relevance of this study. Our research offers a more sophisticated breakdown of the state-of-the-art in prompt-based segmentation compared to existing surveys and prior works. This, in turn, enables practitioners to make informed decisions when selecting methods that best align with their domain-specific segmentation needs.

The main contributions of this survey are:

- Analyzing the progression of segmentation models, focusing on advancements in prompt encoding and integration techniques.
- Proposing a comprehensive survey of promptable segmentation models, organized across five distinct areas: *Interactive Segmentation, Referring Expression Segmentation, Few-Shot Segmentation, Open Vocabulary Segmentation, and Foundation Models.*
- Presenting an extensive review of the most performant SotA models for each segmentation approach and their corresponding application domain.
- Introducing the emerging concept of Compositionality as a new property in promptable segmentation models, which enhances segmentation outcomes.

Prior Work and Distinctions: While the concept of promptable image segmentation has recently attracted significant attention, a cohesive and focused synthesis of its core methodologies remains underdeveloped. Existing surveys have examined adjacent domains, such as vision-language models [18, 19], transformer-based segmentation [20, 21], and broader image segmentation techniques [22, 23, 24, 25, 26]. However, these works address prompting only in passing or as a secondary aspect, without recognizing it as a foundational mechanism driving segmentation performance and adaptability. In contrast, our survey delivers the first systematic and comprehensive review dedicated solely to prompt-based segmentation. We dissect how prompts are constructed, encoded, and integrated across different paradigms—Interactive, Referring, Few-Shot, Open Vocabulary, and Foundation Models. Furthermore, we introduce the novel concept of compositionality in prompting, examining how multiple prompts can be combined in a coordinated manner to enhance segmentation outcomes. Compared to recent concurrent surveys [27, 28, 29], which only briefly mention prompting, our work uniquely foregrounds it as the central axis of segmentation advancements, providing a structured taxonomy and paving the way for future developments in the field.

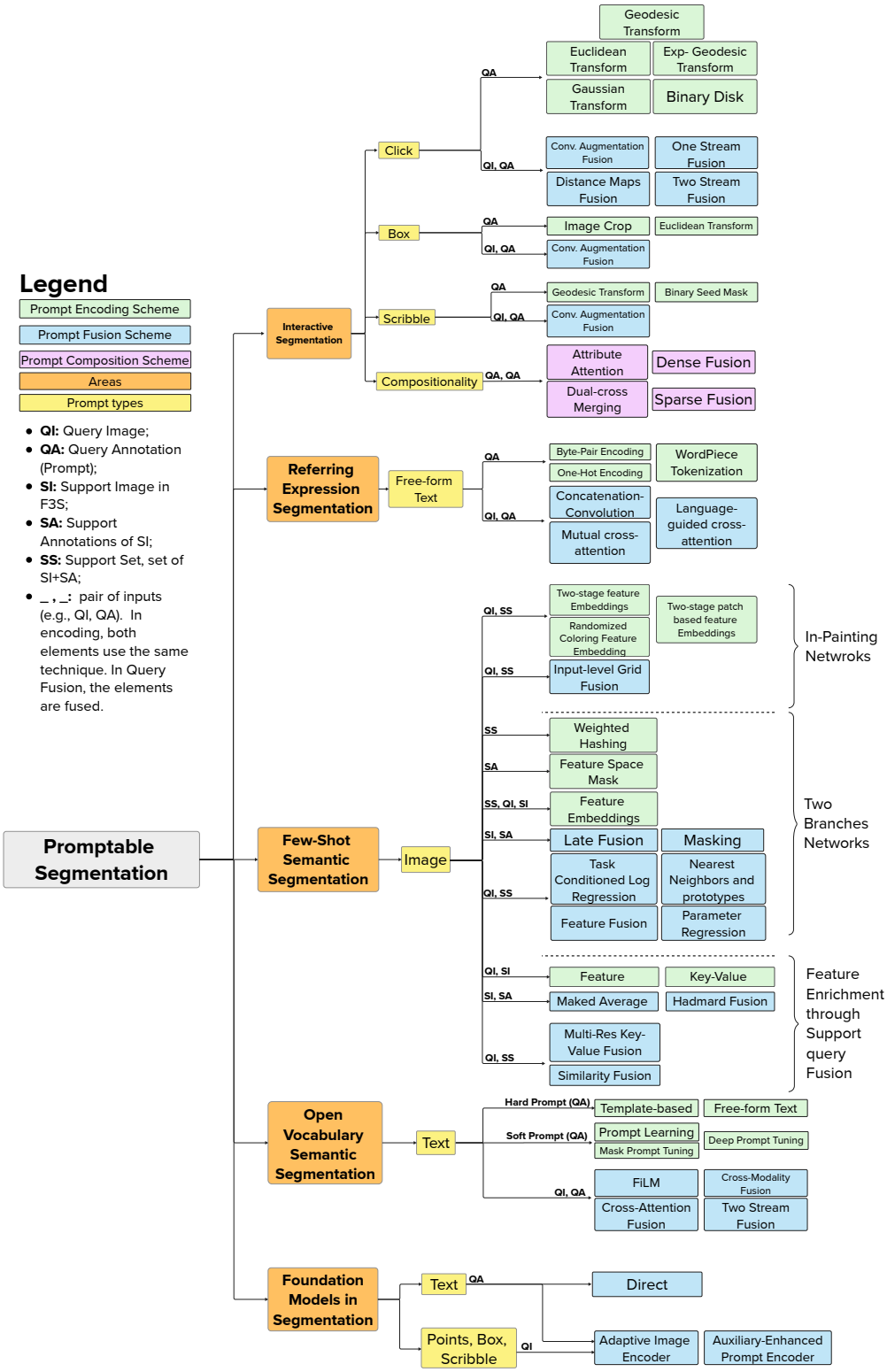


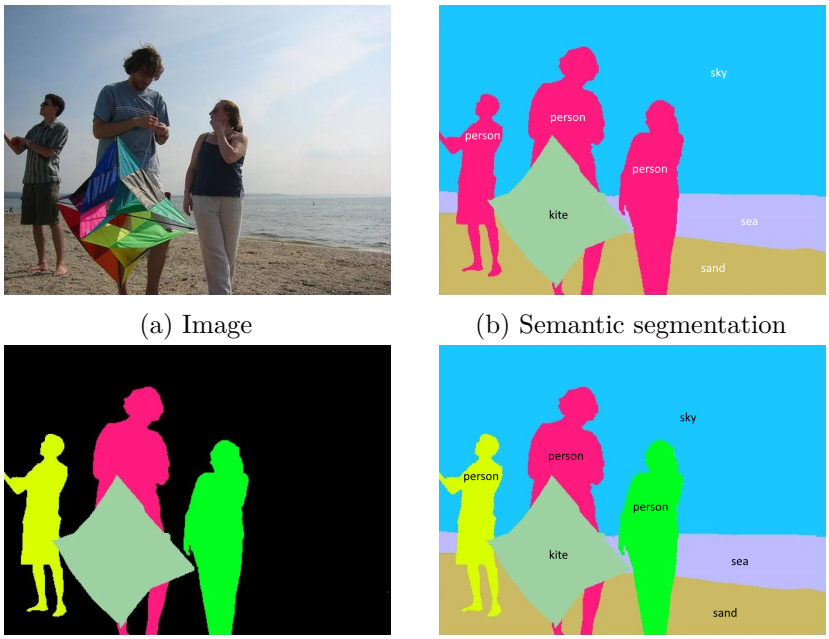
Figure 1: Taxonomy of Prompt-based Segmentation. For each prompt type (e.g., click, box, free-form text) in the corresponding area, encoding techniques are represented by green blocks, fusion techniques by blue blocks, and prompt composition techniques by pink blocks. Letters above arrows (e.g., QI, QA) indicate the inputs to the encoding, fusion, or composition techniques.

2 Background Knowledge

In this section, we provide the theoretical foundation necessary for understanding the scope of this work. We begin with an overview of the general task of Image Segmentation, followed by a detailed introduction to Promptable Image Segmentation, a paradigm that incorporates user prompts to guide segmentation outcomes. This background sets the stage for the subsequent sections, which explore diverse strategies for leveraging prompts in segmentation pipelines.

3 Image Segmentation

Image segmentation is a core computer vision task that divides an image into meaningful regions or objects [30], providing pixel-level labels for detailed content understanding beyond image classification. This process supports various applications, such as medical image analysis [31], autonomous driving [32], image editing [33], and video surveillance [34]. Segmentation can be categorized as: (i) *Semantic Segmentation*, which labels each pixel with a general class, (ii) *Instance Segmentation*, which distinguishes individual instances within each class, (iii) *Panoptic Segmentation*, combining both semantic and instance labeling. A visual example of each segmentation task is reported in fig. 2.



(a) Image (b) Semantic segmentation

(c) Instance segmentation (d) Panoptic segmentation

Figure 2: A visual example of image segmentation tasks.

4 Promptable Image Segmentation

Promptable Image Segmentation extends traditional segmentation by including user-provided prompts to guide the model toward the desired output. Formally, as described in [35], given an input image I and a set of prompts P, the input space X is defined as $I \times P$, which is mapped by a function f in to the output space Y, consisting of the segmentation mask M and class labels C:

$$f : X \to Y, \quad \text{where } X = I \times P,\ Y = M \times C. \tag{1}$$

Prompts serve as auxiliary signals that specify the segmentation intent, enhancing adaptability across contexts. These include visual cues (clicks, bounding boxes, scribbles, polygons) and linguistic inputs (textual descriptions or labels). Promptability, thus, enables few-shot segmentation by leveraging prompts to convey task-specific knowledge without requiring model parameter updates.

The concept has roots in: (i) **Interactive Segmentation (IS)**, where visual prompts guide the selection of target regions; (ii) **Referring Expression Segmentation (RES)**, where free-form text identifies objects within images. While IS and RES traditionally manage prompts per modality, Vision-Language models (VLMs) like CLIP [13], introduce a unified framework that integrates Multimodal prompts, enhancing segmentation outcomes. This advancement is largely driven by Transformer architectures trained on large-scale datasets, which have significantly improved segmentation capabilities through facilitating knowledge transfer to segmentation, enabling open vocabulary segmentation through textual prompts [36, 37, 38, 39]. Open vocabulary segmentation allows models to handle novel test categories unseen during training, supporting zero-shot transfer where textual queries specify target objects for segmentation, leading to the development of Promptable Image Segmentation models.

Promptability has driven the evolution of new segmentation pipelines, resulting in foundation models like SAM [17] and SEEM [40], as well as few-shot models such as SegGPT [41]. This work focuses on Promptable Image Segmentation, defined as guiding segmentation through pre-trained model prompts. We provide a comprehensive overview of methods in IS, RES, and foundation models, highlighting strategies for encoding and integrating prompts within segmentation networks.

Prompts in segmentation models are broadly categorized as **hard prompts**, which refer to explicit, human-interpretable inputs such as text labels or clicks, and **soft prompts**, which are continuous, learned embeddings. In natural language processing, the practice of designing effective text prompts to elicit specific model behavior is known as *prompt engineering* [42]. While prompt engineering has played a central role in language and vision-language tasks, in the context of image segmentation, its relevance is primarily limited to methods involving textual inputs, such as Referring Expression Segmentation [36] and Open Vocabulary Segmentation [37, 38, 39].

In contrast, visual prompt types such as clicks, bounding boxes, and scribbles are processed through spatial encodings and integrated via fusion mechanisms rather than relying on linguistically crafted instructions. Strategies like in-context learning [43] and prompt compositionality [44, 45] further extend the model's ability to interpret and combine multiple prompt forms. For instance, click-based Interactive Segmentation uses geometric or probabilistic encodings such as Euclidean or Geodesic transforms [17, 40], while RES leverages vision-language models such as CLIP [13] and advanced tokenization techniques to align textual descriptions with visual content.

This work surveys promptable segmentation techniques across IS, RES, and foundation models, with a focus on the encoding and integration of various prompt types into segmentation pipelines. The aim is to illustrate how these prompts guide segmentation in a flexible and modular manner, reducing the need for task-specific retraining.

5 Interactive Segmentation

Interactive Segmentation (IS) reduces the labor and expertise required to annotate large datasets at the pixel level by enabling human-guided refinement through visual prompts. As a precursor to modern promptable segmentation, IS utilizes key interaction paradigms such as Clicks [46, 47], Bounding Boxes [48, 49], Scribbles [50, 51], and Polygons [52] to guide the model in improving segmentation results. Early IS methods relied on simple image features such as color similarity and boundary detection [53, 54, 50]. With advances in deep learning, CNN and transformer-based

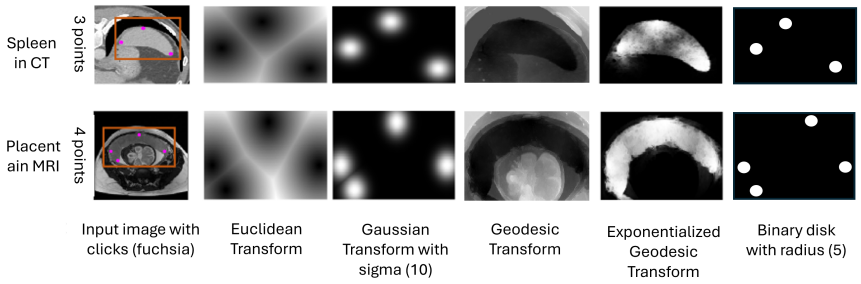


Figure 3: Example of click encodings. Image adapted from [73].

approaches have become the dominant paradigm, offering higher accuracy and robustness [12, 55, 56, 57].

IS research has focused on three main areas: (i) designing efficient backbone architectures [57, 47]; (ii) optimizing simulation of user clicks, evolving from random placement to adaptive strategies that iteratively refine segmentation [12, 58, 59, 55, 46]; and (iii) the development of refinement modules to improve segmentation accuracy [57, 56, 60].

In contrast to previous work, this section reviews interactive segmentation models from a promptability perspective, emphasizing how IS techniques contribute to promptable segmentation by exploring strategies for encoding and integrating visual prompts with input images (Table 3). A summary table of the reviewed IS works is provided in Appendix **??**.

6 Click

Among the interaction types studied—Clicks, Bounding Boxes, Scribbles, and Polygons—Clicks have received the most attention. Clicks offer a more intuitive method for specifying target objects compared to bounding boxes, which often perform poorly with complex shapes [61, 62]. As a result, significant progress in Interactive Segmentation (IS) has been driven by click-based strategies, successfully applied in video analysis [63, 64], autonomous driving [65], medical imaging [66, 67, 68, 69, 70, 71], and mobile photo editing [72].

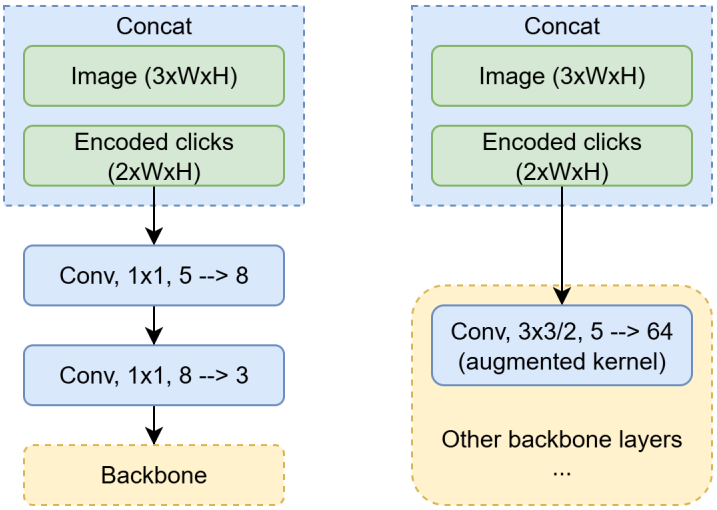
With Clicks, users specify target objects through points that are either *Positive* (indicating regions to include in the segmentation) or *Negative* (indicating regions to exclude). Formally, let C denote the set of clicks, where C_p and C_n represent Positive and Negative clicks, respectively, with $C_p \cup C_n = C$ and $C_p \cap C_n = \emptyset$.

To integrate clicks into deep learning models, they are encoded into *guidance maps* serving as additional inputs to direct the network toward the expected segmentation outcome (Figure 3). These encoding methods include Euclidean Transform, Gaussian Transform, Scale-Aware Guidance Map, Geodesic Transform, and Binary Disk Transform. The fusion of these guidance maps with the input image follows either an **Early Fusion** approach, where encoding maps are concatenated with the input image before feature extraction, or a **Late Fusion** approach, where click information is processed separately and merged later in the network.

6.1 Euclidean Transform

The Euclidean Transform encodes user interactions by computing two distance-based guidance maps: one for positive clicks and one for negative clicks [12]. Formally, for a given pixel p located at coordinates (m, n), the transformation is defined as:

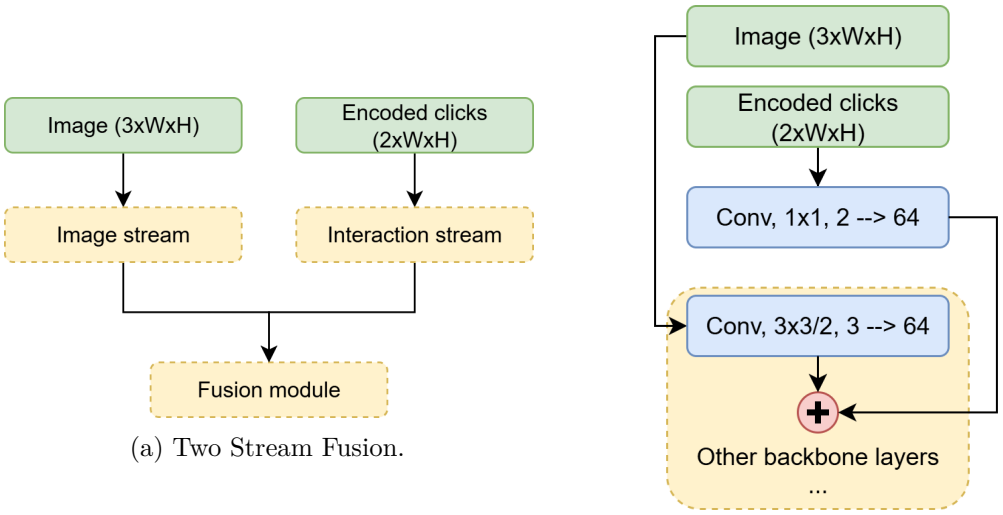
$$Y_{m,n} = \min_{(i,j)\in C} \sqrt{(m-i)^2 + (n-j)^2} \tag{2}$$



(a) Conv augmentation Fusion.

(b) Distance Maps Fusion (DMF).

Figure 4: **Early Fusion** Techniques for Click interaction maps and input image.



(a) Two Stream Fusion.

(b) One Stream Fusion.

Figure 5: **Late Fusion** Techniques for Click interaction maps and input image.

where C can be either C_p or C_n. The resulting guidance maps are denoted as Y^+ and Y^-, corresponding to the positive and negative clicks, respectively.

In the **Conv Augmentation Fusion** approach, categorized under Early Fusion strategies, these distance maps are concatenated with the input RGB image along the channel dimension. This results in a 5-channel input tensor of the form $[\text{RGB}, Y^+, Y^-]$, which is subsequently processed by the segmentation backbone [12, 74, 75, 76, 77] (see Figure 4a). Alternative approaches, such as **Distance Maps Fusion (DMF)** [58], preprocess the interaction maps through two 1x1 convolutions and a LeakyReLU activation before feeding them into the backbone (Figure 4b). The first convolution expands the input to 8 channels, while the second reduces it to 3 channels for compatibility with the segmentation backbone. This design avoids modifying the backbone's original input size or architecture, thereby allowing better utilization of pretrained backbone checkpoints. On the contrary, late Fusion strategies, such as **Two-Stream Fusion** [78], process image features and interaction maps separately before merging them, leading to improved segmentation accuracy (Figure 5a). The ad-hoc encoder enables late fusion to extract

key user interaction data, integrating it later in the network to enhance the final result.

6.2 Gaussian Transform

A probabilistic encoding method [79] models interaction maps using 2D Gaussians centered around clicks, achieving better results than Euclidean Transform [80]. The spatial map for each click c_{ij} is:

$$Y_{c_{ij}}^{\sigma}(p) = \exp\left(-\frac{d(p, c_{ij})^2}{2(\sigma \cdot L)^2}\right) \tag{3}$$

where $d(p, c_{ij})$ is the Euclidean distance between pixel p and click c_{ij}, and L is the smaller side of the image. The overall interaction map is obtained by pixel-wise max aggregation of all clicks. The parameter σ controls the spatial extent of the click.

Methods using Conv Augmentation Fusion refine masks when a single click is provided [80, 81, 56, 59]. Other techniques define guidance maps with clicks placed at object boundaries (left, right, top, and bottom), concatenated with the input image to create a 4-channel input [79, 82]. Inspired by [78], [72] implement Two Stream Fusion for high-resolution photo editing. As proposed by [75], the 2D Gaussian transform allows encoding clicks at multiple scales, balancing precision (small σ) and context (large σ).

6.3 Scale aware guidance map

Gaussian and Euclidean Transforms are considered overly simplistic for encoding clicks as they ignore structural details like object scale [83]. To address this, a scale-aware guidance map is proposed, incorporating hierarchical structures. Low-level structures are represented by superpixels [84], while high-level structures use region-based object proposals.

Let SP be the superpixel set, and $f_{SP}(p)$ map each pixel p to its superpixel. Positive and negative superpixels are derived from clicks C_p and C_n, forming sets $\{sp_p \mid sp_p = f_{SP}(c_p), c_p \in C_p\}$ and $\{sp_n \mid sp_n = f_{SP}(c_n), c_n \in C_n\}$. Two guidance maps encode superpixel distances:

$$Y_t^{sp}(p) = \min_{sp \in \{c_t\}} d_c(sp, f_{SP}(p)) \tag{4}$$

where $d_c(sp_i, sp_j)$ is the Euclidean distance between superpixel centers.

Region-based object proposals further refine the map. Let L_p be the proposal set covering pixel p, then:

$$Y^o(p) = \sum_{p' \in \{S_p\}} \sum_{L \in \{L_{p'}\}} 1[p \subset L] \tag{5}$$

where $1[p \subset L]$ is an indicator function. Finally, guidance maps $Y_p^{sp}(p)$, $Y_n^{sp}(p)$, and $Y^o(p)$ are concatenated with the RGB image using Conv Augmentation Fusion.

Experimental results indicate that using a moderate number of superpixels enhances performance compared to the Euclidean Transform, while an excessive number of superpixels leads to a decline in accuracy [83].

6.4 Exponentialized Geodesic Transform

The Euclidean Transform is scale-unaware and does not account for image context, treating all directions equally without considering the appearance of neighboring pixels. The Geodesic Transform addresses this limitation by encoding contextual information, differentiating neighboring pixels based on their appearance, and improving label consistency in homogeneous regions [85, 86].

Given an image I, the unsigned Geodesic Transform from a pixel p to the click set C ($C \in \{C_p, C_n\}$) is defined as:

$$Y^{gt}(p, C, I) = \min_{c \in C} D_{\text{geo}}(p, c, I) \tag{6}$$

$$D_{\text{geo}}(p, c, I) = \min_{\pi \in \Pi_{p,c}} \int_0^1 \|\nabla I(\pi(s)) \cdot u(s)\| \, ds \tag{7}$$

where $\Pi_{p,c}$ is the set of all paths between pixel p and click c. The path π is parameterized by $s \in [0, 1]$, and $u(s) = \frac{\pi'(s)}{\|\pi'(s)\|}$ is a unit vector tangent to the path.

Similar to other encoding techniques, the Conv Augmentation Fusion strategy is commonly employed to incorporate the geodesic guidance map. Comparative studies on medical image segmentation datasets demonstrate that the Geodesic Transform achieves performance comparable to both the Euclidean and Gaussian Transforms [73].

To further enhance geodesic encoding, the Exponentialized Geodesic Transform (EGT) was proposed as a parameter-free, context-aware metric [73]. Unlike Euclidean, Gaussian, and standard Geodesic Transforms, EGT does not require tuning hyperparameters such as thresholds or sigma values, making it more adaptable.

For a pixel p in image I and user clicks C, the unsigned EGT is defined as:

$$Y^{egt}(p, C, I) = \min_{c \in C} e^{-D_{\text{geo}}(p,c,I)} \|\nabla I(\pi(n)) \cdot v(n)\| \, dn \tag{8}$$

where $D_{\text{geo}}(p, c, I)$ is the Geodesic distance from Eq. 7. Experiments with EGT demonstrate its superior ability to distinguish foreground from background while enhancing shape, position, and contextual representation. This leads to improved segmentation performance compared to other transforms [73].

6.5 Binary Disk

Euclidean and Gaussian Transforms change significantly with new clicks, confusing networks during training and inference. The Binary Disk Transform (BDT) mitigates this by encoding clicks as small-radius binary disks, generating separate maps for positive and negative clicks [55].

For a pixel p, user clicks C, and radius r, BDT is defined as:

$$Y^{bdt}(p, C, r) = \begin{cases} 1 & \text{if } \exists c \in C \text{ such that } \|p - c\|_2 \leq r \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where $\|p - c\|_2$ is the Euclidean distance. Experiments show that small radii (3, 5, 10) outperform Euclidean and Gaussian Transforms [87, 55]. Conv Augmentation Fusion [88, 60, 47] is the primary fusion strategy, while *One Stream Fusion* [55] encodes guidance maps as a tensor matching the backbone's first layer, improving segmentation performance (see Figure 5b).

7 Bounding Box

Bounding boxes are widely used in computer vision for interactive segmentation, introduced by [90] using the graph-cut approach [50], which models segmentation as an energy minimization problem by partitioning the image into foreground and background. Methods like [91] and [92] extend bounding boxes to CNNs for medical applications. However, these methods encode bounding box information by cropping the image to the bounding box region before feeding it into the network. Formally, given an image $I \in \mathbb{R}^{H \times W \times 3}$ and a bounding box $B = [x_{\min}, y_{\min}, x_{\max}, y_{\max}]$, the cropped image region is defined as:

$$I_B = I[y_{\min} : y_{\max}, \, x_{\min} : x_{\max}] \tag{10}$$

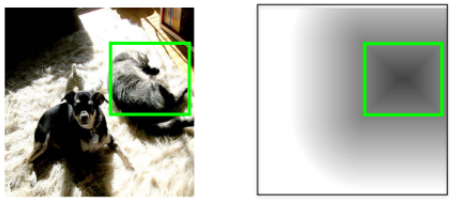


Figure 6: Example of a bounding box encoded using the Euclidean Transform. Image adapted from [89].

This operation removes all pixels outside the bounding box, thereby discarding valuable contextual information from the global image structure, which can lead to suboptimal segmentation performance. To address this, [89] proposed the Euclidean Transform, mapping loosely-placed rectangles to a guidance map fused with the image via Conv Augmentation Fusion (Figure 6). Given a rectangle B in image I, the edge pixels form set S_e, interior pixels S_i, and exterior pixels S_o. A 2-D distance map Y for each pixel p_i is computed as:

$$Y(p_i) = \begin{cases} 128 - \min_{p_j \in S_e} \|p_i - p_j\|, & \text{if } p_i \in S_i, \\ 128, & \text{if } p_i \in S_e, \\ 128 + \min_{p_j \in S_e} \|p_i - p_j\|, & \text{if } p_i \in S_o, \end{cases} \tag{11}$$

where $\|\cdot\|$ denotes the Euclidean distance. The advent of Foundation Models for image segmentation, such as the Segment Anything Model (SAM), has transformed bounding box-based segmentation paradigms. SAM encodes bounding box prompts into high-dimensional feature representations, which are then fused with image features to guide segmentation (see Section 25 for architectural details). Several recent works have explored SAM's adaptation to medical image segmentation tasks using bounding box prompts [93, 94, 95, 96, 97, 98].

Extensive evaluations demonstrate that SAM's zero-shot performance consistently favors bounding box prompts over other visual inputs and that the model performs particularly well on larger objects [93, 95, 97]. Hu, Li, and Yang [94] fine-tune SAM on dermoscopic images using simulated bounding box prompts, achieving strong performance in skin lesion segmentation. Similarly, Lei et al. [98] employ extreme points that implicitly define bounding box prompts, effectively reducing the annotation burden while maintaining segmentation accuracy.

In a comprehensive analysis, Huang et al. [96] compile 52 public datasets to evaluate SAM's "Segment Everything" mode under both click and bounding box prompts. Their study reveals substantial variability in SAM's performance across datasets and imaging modalities. Notably, bounding box prompts consistently produce higher and more stable results than click-based alternatives.

Fine-tuning has also emerged as an effective strategy for adapting foundation models to specific applications. In this context, user-provided bounding boxes are employed to localize the target region and initialize segmentation using a pre-trained model, enabling efficient online refinement through interactive annotation [97].

8 Scribble

A scribble consists of pixels inside or outside the target segmentation object, labeled as Positive or Negative. The user draws lines or shapes as an overlay mask M on the image I, represented as seed points S, where each seed $s = (p, l)$, with p as the pixel location and $l \in \{1, 0\}$ indicating Positive or Negative. The Binary Seed Mask Y_S is generated from these seeds [68, 69, 70, 71]. The common fusion strategy involves early fusion, creating a 4-channel input for the backbone

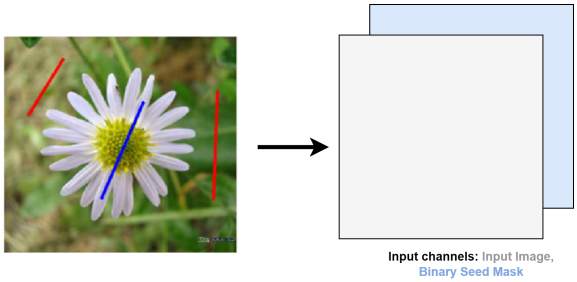


Figure 7: Example of input scribbles (positive in blue and negative in red) encoded using a Binary Seed Mask. Image adapted from [99, 68].



Figure 8: Example of polygon interaction: the user can select any incorrect control point and move it to the boundary to correct the output mask. Image from [102].

model. [86] uses the Geodesic Transform to encode scribbles similarly to the formulations in Eqs. 6 and 7.

Scribble interactions are commonly used in medical applications where automatic systems struggle due to poor image quality, patient variability, and varying clinical protocols. They also facilitate training CNNs for semantic segmentation [100], as scribbles are effective for annotating ambiguous boundaries of "stuff" (e.g., water, grass).

9 Polygon

A polygon is defined as a sequence of connected image segments that form a closed shape, aligning closely with object boundaries. This task is typically framed as a sequence prediction problem. Recent approaches employ Recurrent Neural Networks (RNNs) to predict polygon vertices sequentially from an input image [101, 52]. After constructing the polygon, these methods enable user intervention to refine vertex positions (see Figure 8).

However, due to the recurrent architecture, adjusting a single vertex may propagate changes to all subsequent points. To overcome this limitation, FIO-GCN [102] reformulates the task as a regression problem, predicting all vertices simultaneously using a Graph Convolutional Network (GCN). This design confines adjustments to the local neighborhood, ensuring that only adjacent vertices are affected when a control point is corrected.

10 Compositionality in Visual Prompting

In this subsection, we introduce the notion of **Compositionality** and emphasize its growing relevance in Interactive Segmentation. We define compositionality as the ability to integrate multiple prompts to improve segmentation outcomes. Since the discussion so far has focused solely on visual prompts, we restrict this section to the composition of visual prompts only. The integration of other modalities, such as text, will be addressed in subsequent sections.

Formally, given a set of prompts $P = \{p_1, p_2, \ldots, p_n\}$, the expected segmentation performance with the combined prompts should be at least as good as that obtained with the best individual

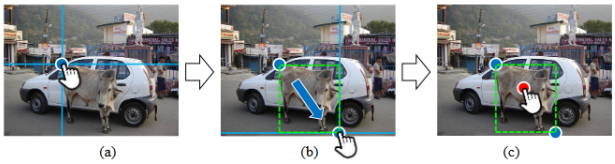


Figure 9: Two-click method to create a bounding box using horizontal and vertical guide lines [103].

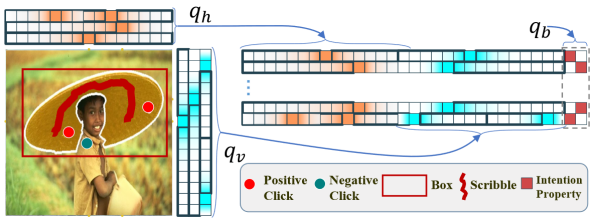


Figure 10: Prompt unified Encoder (PuE) of user input interaction. The two spatial vectors q_h and q_v are concatenated with the intention property q_b, creating a unique one-dimensional dense representation for clicks, bounding boxes, and scribbles. Image adapted from [106].

prompt. Let $S(p)$ denote the segmentation performance using prompt p, then the following inequality should hold:

$$S(P) \geq \max_{p_i \in P} S(p_i), \tag{12}$$

where the composition of prompts leverages complementary information to refine the segmentation.

Early studies explored combining bounding boxes and clicks using early fusion strategies. For example, Benenson, Popov, and Ferrari [87] concatenate the input image with a binary bounding box mask and binary disk-encoded clicks, leading to refined segmentation. Similarly, Zhang et al. [103] propose a two-click bounding box method, where an interior click helps eliminate irrelevant regions. A redesigned GUI significantly reduces the time for drawing bounding boxes from 25.5s [104] to 6.7s (see Figure 9), with clicks encoded using 2D Gaussian kernels.

Lin et al. [105] further demonstrate the compositional integration of bounding boxes, scribbles, and polygons within a multi-modal framework. This design enables dynamic user input, refining ambiguous segmentations through additional clicks or scribbles.

Recent work introduces novel encoding mechanisms to better integrate heterogeneous visual prompts. For instance, [106] propose the **Prompt unified Encoder (PuE)**, which encodes clicks, bounding boxes, and scribbles into a unified vector representation. This vector includes horizontal and vertical positional encodings along with an intention bit (see Figure 10). A late fusion strategy, termed **Dual-Cross Merging Attention**, employs cross-attention between image features and the prompt vector, enabling more effective feature modulation.

These methods collectively demonstrate the promising potential of compositionality in promptable segmentation frameworks, allowing richer and more efficient use of user input. In the subsequent chapter, we broaden this discussion by incorporating semantic interaction, demonstrating how textual cues can effectively complement visual prompts to enhance segmentation performance.

11 Datasets

Several well-established datasets are used to evaluate interactive segmentation (IS) methods. These datasets differ in scale, complexity, and annotation precision, presenting diverse challenges such as object occlusion, cluttered backgrounds, and ambiguous boundaries. In IS research,

Dataset	Scale	Objects / Instances	Annotation Type
PASCAL VOC12 [107]	7,282 images	19,694 objects (20 classes)	Pixel-wise segmentation masks with class labels.
GrabCut [90]	50 images	1 object per image	Binary foreground/background masks.
Berkeley (BSDS) [108]	96 images	100 object masks	Manually drawn contour-based object masks.
DAVIS [109]	50 videos (3,455 frames)	1 object per video	Dense per-frame binary segmentation masks.
SBD [110]	11,318 images	20,000+ instances (20 classes)	Instance-level masks with boundary annotations.
MS COCO [111]	118K images	1.2M object instances (80 classes)	Instance segmentation masks.

Table 1: Summary of commonly used datasets for interactive segmentation evaluation, listing their scale, object counts, and annotation types. All datasets rely on simulated user clicks derived from ground-truth masks.

they are typically repurposed by simulating user interactions—such as clicks or scribbles—on annotated masks to measure how efficiently a method can reach a given accuracy threshold (e.g., NoC@X). Table 1 provides a concise overview of the most common benchmarks.

PASCAL VOC [107]. The PASCAL VOC datasets serve as a standard benchmark for semantic object segmentation. The VOC2012 edition (segmentation subset) includes 7,282 images spanning 20 object classes with 19,694 annotated objects, of which 5,826 contain full pixel-level segmentations. Its wide variability in pose, illumination, and scene context makes it suitable for evaluating segmentation robustness. However, as it does not contain user interactions, synthetic click prompts are generated to simulate interactivity.

GrabCut [90]. GrabCut contains 50 single-object images with clear foreground–background separation. It is one of the earliest datasets used for evaluating interactive segmentation algorithms. While widely adopted for historical comparison, its small scale and simplified binary scenes limit its representativeness for modern click-based evaluation.

Berkeley (BSDS) [108]. The Berkeley dataset consists of 96 images with 100 manually annotated object masks, derived from the Berkeley segmentation dataset [113]. It is challenging due to low contrast and ambiguous object boundaries. Originally designed for boundary detection, it requires manual selection of a single foreground object for IS benchmarking, making evaluation less standardized across works.

DAVIS [109]. The DAVIS dataset, designed for video object segmentation, comprises 50 high-quality sequences with dense frame-by-frame annotations. Its inclusion of motion blur, occlusion, and deformation offers a strong test of model generalization under temporal variation. For interactive segmentation, static frames are sampled and evaluated individually with simulated clicks.

Semantic Boundaries Dataset (SBD) [110]. SBD builds upon the PASCAL VOC2011 dataset with detailed instance-level boundary annotations, totaling 11,318 images (8,498 for training and 2,820 for testing). Its fine-grained contours make it particularly useful for evaluating the precision of click-based refinement strategies.

MS COCO [111]. MS COCO is a large-scale dataset containing 118K training images and 1.2M object instances across 80 categories. It captures diverse, complex scenes with multiple overlapping objects, making it a preferred dataset for pretraining and large-scale generalization testing. Like most benchmarks, user interactions are simulated from ground-truth segmentations.

Overall, these datasets collectively provide a hierarchical testing framework—ranging from

Work	Encoding	Fusion	GrabCut NoC@90	Berkeley NoC@90	SBD NoC@85	DAVIS NoC@90	VOC NoC@85
DIOS [12]	Euclidean	Early CA	6.04	8.65	–	–	6.88
RIS-Net [74]	Euclidean	Early CA	5.00	6.03	–	–	5.12
FCTSFN [78]	Euclidean	Late TS	3.76	6.49	–	–	4.58
DEXTR [79]	Gaussian	Early CA	4.00	–	–	–	4.00
ITIS [80]	Gaussian	Early CA	5.60	–	–	–	3.80
BRS [76]	Euclidean	Early CA	3.60	5.08	6.59	8.24	–
MultiSeg [77]	Euclidean	Early CA	2.30	4.00	–	–	3.88
CAMLG-IIS [83]	Scale Aware	Early CA	3.58	5.60	–	–	3.62
F-BRS [58]	Euclidean	Early DM	2.72	4.57	4.81	7.41	–
FCA-Net [59]	Gaussian	Early CA	2.24	4.23	–	8.05	2.98
99A-IS [72]	Gaussian	Late TS	2.54	3.53	3.90	–	–
CAISLC [88]	Binary Disk	Early CA	3.07	4.94	–	5.16	3.18
PhraseClick [60]	Binary Disk	Early CA	2.06	3.26	–	–	3.12
IOG [81]	Gaussian	Early DM	3.00	–	–	–	3.00
CDNet [81]	Gaussian	Early DM	2.64	3.69	4.37	6.66	–
FocusCut [56]	Gaussian	Early CA	1.64	3.01	3.40	6.22	–
PseudoClick [60]	Binary Disk	Early CA	2.04	3.23	–	6.57	2.34
RITM [55]	Binary Disk	Late OS	2.04	3.22	3.39	6.71	2.51
FocalClick [57]	Binary Disk	Late OS	1.90	3.14	4.34	7.06	–
SimpleClick [46]	Binary Disk	Late OS	1.54	2.46	3.28	5.48	2.38

Table 2: Benchmark comparison of interactive segmentation methods, extended with Encoding and Fusion strategies. Lower is better for all NoC@X metrics. Abbreviations: CA = Convolutional Augmentation Fusion; DMF = Distance Maps Fusion; OS = One Stream Late Fusion; TS = Two Stream Late Fusion.

simple binary scenes (GrabCut) to complex, multi-object environments (COCO)—enabling consistent and progressive evaluation of interactive segmentation methods.

12 Metrics

The primary metrics for evaluating the performance of interactive segmentation models are Intersection over Union (IoU) and Number of Clicks (NoCs). In the following, S and G represent the segmentation result to be evaluated and the ground truth, respectively. Additionally, FG and BG denote the abbreviations for Foreground and Background.

Intersection over Union (IoU). Also known as the Jaccard index, it is the ratio of the number of correctly labeled FG/BG pixels to the number of pixels labeled as FG or BG in either G or S.

$$IoU = \frac{|S \cap G|}{|S \cup G|} \tag{13}$$

Number of Clicks (NoC@X) represents the number of clicks needed to achieve a specific Intersection over Union (IoU) value of X. Common IoU thresholds are 90% and 95%, corresponding to the metrics NoC@90 and NoC@95, respectively. A maximum number of clicks is typically

Prompt type	Encoding	Fusion Strategy	Works
Bounding Box	Euclidean Transform	Early Fusion Conv Augmentation Fusion	[91, 89]
Bounding Box + Click	Gaussian Transform	Early Fusion Conv Augmentation Fusion	[103, 105]
	Binary Disk	Early Fusion Conv Augmentation Fusion	[87]
Click	Euclidean Transform	Early Fusion Conv Augmentation Fusion	[12, 74, 75, 76, 77]
		Early Fusion Distance Maps Fusion	[58]
		Late Fusion Two Stream Fusion	[78]
	Gaussian Transform	Early Fusion Conv Augmentation Fusion	[79, 82, 80, 56, 59]
		Early Fusion Distance Maps Fusion	[81]
		Late Fusion Two Stream Fusion	[72]
	Scale Aware Guidance Map	Early Fusion Conv Augmentation Fusion	[83]
	Exponentialized Geodesic Transform	Early Fusion Conv Augmentation Fusion	[73]
	Binary Disk	Early Fusion Conv Augmentation Fusion	[88, 60, 47]
		Late Fusion One Stream Fusion	[55, 57, 46]
Click + Text	Euclidean Transform	Early Fusion Conv Augmentation Fusion + Attribute Attention Fusion	[112]
Scribble	Geodesic Transform	Early Fusion Conv Augmentation Fusion	[86]
	Binary Seed Mask	Early Fusion Conv Augmentation Fusion	[100, 68, 69, 70, 71]

Table 3: Summary of Interactive Segmentation methodologies, categorized by encoding and fusion strategies.

defined, and exceeding this limit classifies the scenario as a failure. In such cases, the **Number of Failures (NoF@X)** metric is used alongside NoC, with a common limit set at 20 clicks. For instance, NoF@90 indicates the number of cases where more than 20 clicks are necessary to reach an IoU of 90%.

$$\text{NoC@X} = \min\{n \mid \text{IoU}(n) \geq X\}_i \tag{14}$$

$$\text{mNoC@X} = \frac{1}{|D|} \sum_{i=1}^{|D|} \text{NoC@X}(i) \tag{15}$$

$$\text{mNoF@X} = \frac{1}{|D|} \sum_{i=1}^{|D|} \mathbb{I}[\text{NoC@X}(i) > 20] \tag{16}$$

where mNoC@X and mNoF@X represent the mean values of the respective metrics over the dataset D.

13 Benchmark Analysis of Encoding and Fusion Strategies

In the following benchmark analysis, we focus on methods that utilize *click-based* input prompts, as they represent the majority of existing Interactive Segmentation Approaches. Table 2 provides a summary of these methods evaluated across standard benchmarks, including GrabCut, Berkeley, SBD, DAVIS, and VOC. The table reports the NoC@X metric, where lower values indicate higher segmentation efficiency. Each method is further classified by its **Encoding technique** and **Fusion strategy**, illustrating the evolution of design choices over time.

13.1 Encoding Techniques

A clear progression is observed across encoding schemes. Early approaches using the *Euclidean Transform* (e.g., DIOS, RIS-Net) provided global spatial cues but lacked boundary precision. The introduction of the *Gaussian Transform* (e.g., DEXTR, ITIS, FocusCut) improved localization and contour accuracy through smoother spatial gradients. Recent models employing *Binary Disk* encodings (e.g., RITM, FocalClick, SimpleClick) offer discrete, localized guidance that aligns better with convolutional and transformer-based backbones. Overall, this shift from continuous to discrete encodings enhances segmentation precision and efficiency.

13.2 Fusion Strategies

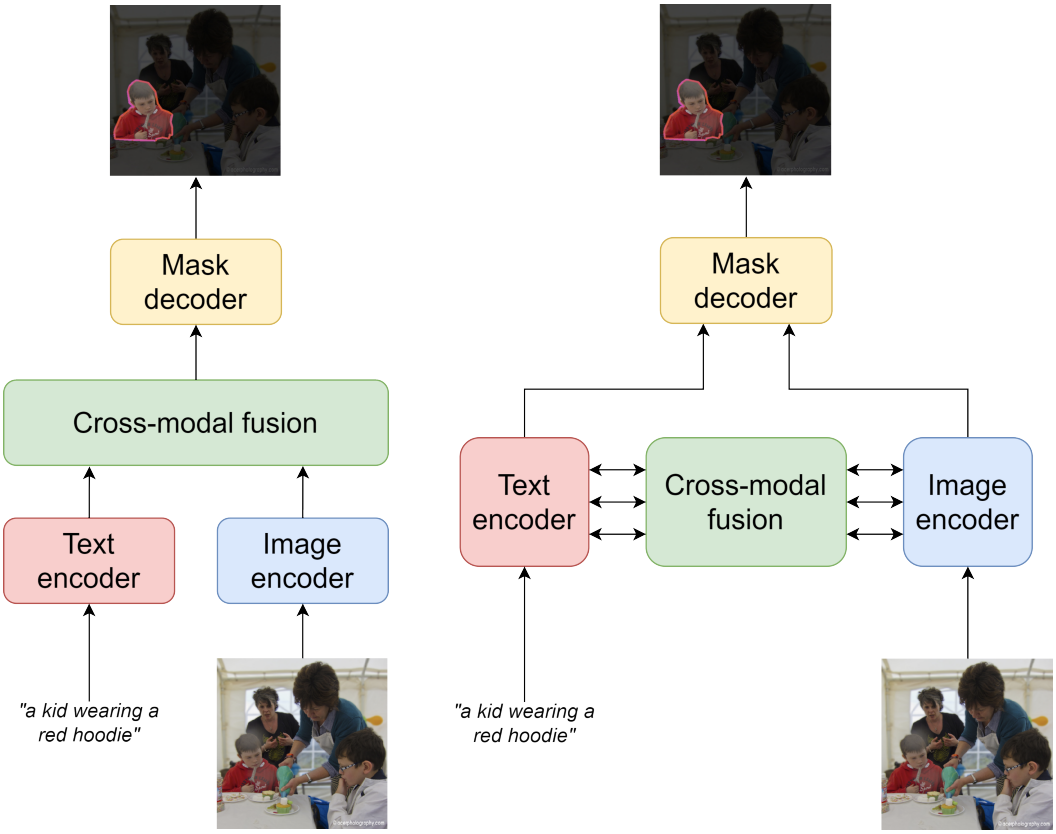
Fusion design shows a parallel evolution. *Early fusion* methods (e.g., DIOS, RIS-Net) concatenate user prompts and image inputs directly, limiting representational flexibility. *Late fusion* strategies (e.g., FCTSFN, 99A-IS) separate feature extraction before merging, improving contextual reasoning and reducing NoC. Recent *One-Stream Late fusion* architectures (e.g., RITM, FocalClick, SimpleClick) embed interaction cues within intermediate layers, enabling adaptive and efficient information propagation. Overall, progressively decoupled and context-aware fusion designs consistently outperform early concatenation approaches.

14 Referring Expression Segmentation

Referring Expression Segmentation (RES), also known as Referring Image Segmentation (RIS), is the task of segmenting an image based on a given natural language expression [114]. The output mask is expected to contain the object(s) inside the image which is *referred by* the given textual prompt. Referring expressions are free-form textual prompts, not restricted to any fixed template, e.g., “two men sitting on the right bench” or "a red cup filled with coffee". Therefore, unlike traditional semantic segmentation tasks, RES is not restricted to a predefined set of semantic classes. Research on the RES task has surged in recent years thanks to its potentially wide range of applications, including interactive image editing [115, 116, 117], human-computer interaction [118, 119, 120], assistive technologies for the visually-impaired [121], search engines and e-commerce [122, 123].

Textual prompts pose a unique set of challenges, with respect to visual prompts:

- **Ambiguity**. An expression could target a single object or multiple objects. For example, "the big dog" could refer to any large dog in the image, if there are multiple. Referring expressions can also involve subjective perceptions, as in "the cutest dog".
- **Variability**. There are endless possible different referring expressions for targeting the same object. For example, the expressions "the brown dog", "the dog on the left", "a being with four legs", "the animal on a leash" may target the same dog in the image.



(a) Late fusion (b) Early fusion

Figure 11: Fusion strategies employed by RES models

- **Complexity**. Referring expressions can be arbitrarily long and semantically complex. They can describe complex attributes of the target object and intricate spatial or semantic relationships with other objects in its surroundings. For example, "the car which was produced in Europe" or "the fifth chair on the second row". Expressions can also target an object in a more or less implicit way, and may thus require a significant reasoning effort to be understood and contextualized. For example, "the object which people usually use to bake their food".

- **Multiple targets, no targets**. Referring expressions can refer to more than one target object (multi-target referring expression), or even to none (no-target/negative referring expression). For example, for an image of an open sky with clouds, "all the clouds" may refer to multiple targets, "a dog" refers to no targets.

Another core challenge of RES is vision-language alignment. Textual and visual data exist in distinct feature spaces. Bridging these spaces requires sophisticated feature extraction methods that can capture the essence of both modalities, and multi-modal fusion strategies to align them effectively. In order to produce a correct segmentation mask, RES models need to effectively understand, interpret and integrate information from both modalities.

Recent years have seen a surge in the development of RES frameworks and model architectures. The core differences between RES models lie in the particular strategy employed to fuse together the visual and textual modalities. While earlier models usually consisted of independent vision and language encoders followed by a multi-modal fusion module (a strategy called *late fusion*), recent ones usually employ multi-modal interactions already in the early stages of the architecture (*early fusion*). The two strategies are schematized in Fig. 11a and 11b respectively.

Another point in which RES models differ is in their architectural choices. Earlier models

typically employed CNNs image encoders and RNNs textual encoders [114]. More recently, a shift has been observed towards transformer-based architectures [5, 7]. Transformers have been integrated both as language and visual encoders and as multi-modal fusion modules, leading to a significant increase in performances.

As a final note, different RES methods employ different tokenization strategies for casting free-form textual prompts to numerical embeddings, which are computer-readable. Early RES architectures leverage word-level tokenization algorithms such as simple one-hot word tokenizers or *GloVe* [124], which convert words into continuous vector representations based on co-occurrence statistics in large text corpora. On the other hand, recent RES models use more advanced sub-word tokenization strategies, such as *Byte-Pair Encoding* (BPE) [125] and *WordPiece* [126], which handle rare or out-of-vocabulary words by splitting them into smaller, more frequent sub-units, thereby enhancing robustness and reducing the vocabulary size.

This section of the survey shares similarities with past reviews on RES [127, 128]; however, differently from those, we discuss RES by placing it in the broader topic of promptable image segmentation. In the next section, we delve into the main RES methods and models, categorizing them by the particular multi-modal fusion strategy that they employ.

15 RES models

15.1 Concatenation-convolution fusion

Early research in RES primarily used a "concatenation-convolution" method for multi-modal fusion. In this approach, visual and textual embeddings are extracted independently by separate encoders, stacked, and processed by a convolution module in the mask decoder. Since encoders do not share data during encoding, the embeddings lack cross-modality awareness, and the final convolution essentially performs a weighted average along the channel dimension. Thus, concatenation-convolution is a late fusion strategy.

LSTM-CNN [114] has been the first published RES model based on this approach. It uses an LSTM network to encode the referring expression, a fully convolutional network as the image encoder, and a de-convolutional decoder for segmentation mask generation. The textual and visual features are simply concatenated and sent to a cross-modal CNN decoder.

RRN (Recurrent Refinement Network) [129] is a late fusion RES model that improves upon the LSTM-CNN architecture by adding a Recurrent Refinement (RR) module in the mask decoder. This module enhances the fused multi-modal features from the LSTM-CNN backbone by progressively incorporating hierarchical visual features from the CNN encoder, treating each feature map as an input step to a convolutional LSTM [130]. This approach refines the segmentation mask using multi-scale visual features, which are known to be critical in closed-vocabulary segmentation [131, 132]. DenseCRF post-processing is also applied to boost performance [133].

DMN (Dynamic Multimodal Network) [134] is a late fusion RES model featuring a Synthesis Module (SM) for multimodal fusion. DMN uses an LSTM-based language encoder to generate a feature vector and a "language filter" for each of the T tokens in the referring expression. In the SM, these T filters are convolved with the feature map from the image encoder, yielding an image response map for each token. Then, multiple concatenation-convolution operations integrate the T response maps into a multimodal feature stack, as shown in Fig. 12. A Simple Recurrent Unit (SRU) [135] finally aggregates these maps into a single response map, which is upsampled to form the segmentation mask.

RMI (Recurrent Multimodal Interaction Network) [136] is the first RES model to introduce recurrent interaction between textual and visual modalities. Based on the LSTM-CNN architecture, RMI incorporates a multimodal convolutional LSTM (mLSTM). After each step t of the LSTM encoder, the mLSTM takes as input a concatenation of the image feature vector

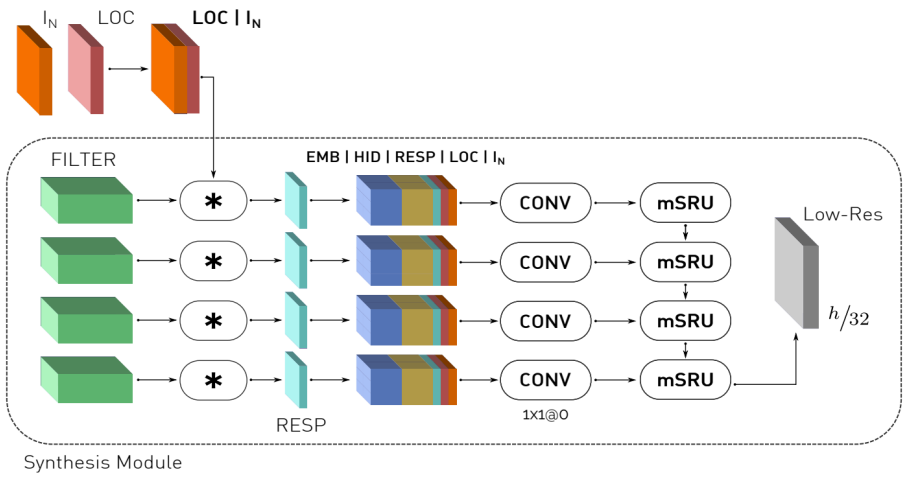


Figure 12: Synthesis Module (SM) of the DMN model. It takes the visual feature map (I_N) produced by the image encoder, stacked with its spatial coordinate representation (LOC), and convolves it with dynamic language filters (FILTER) to produce a response map (RESP) for each token. Then, a multimodal stack is created by concatenating the hidden state of the language encoder (HID), word embedding (EMB), RESP, I_N, and LOC, and a single response map is output through a 1×1 convolution followed by a multimodal Simple Recurrent Unit (SRU). Taken from [134].

($\mathbf{v}$), spatial coordinates, the t-token embedding (l_t), and the t-th hidden state of the LSTM text encoder (h_t). The mLSTM's hidden states serve as multimodal feature vectors, enabling recurrent word-image interaction. At the final step, the textual, visual, and multimodal feature maps are concatenated and processed by a cross-modal CNN decoder. Despite the recurrent interaction, the text encoder remains agnostic to the multimodal fusion, so RMI is still considered a late fusion method. RMI set a new state-of-the-art in RES, demonstrating the importance of recurrent early textual-visual interaction and influencing subsequent research in the field.

15.2 Language-guided cross-attention fusion

Recent works have found that late fusion is a suboptimal multimodal fusion strategy. In fact, delayed interaction between modalities has been shown to prevent models from learning the complex, nuanced relationships between text and images, which could be critical for RES [137]. Experimental results show that jointly encoding images and text from the beginning of the encoding process (i.e., early fusion) increases performances, as it enables modeling joint contextual information more effectively. The introduction of the early fusion mechanism and the advent of the Transformer architecture [5] marked a turning point in RES, as models employing multi-modal attention-based early fusion strategies set a new state-of-the-art. Specifically, such models rely on the vision-language cross-attention operation, where the visual features iteratively attend to the referring expression features being encoded, generating language-informed visual features.

LAVT [138] has been the first early fusion RES architecture proposed. In LAVT, the image encoder attends to language features at each encoding stage through a Pixel-Word Attention Module (PWAM), depicted in Fig. 13. After each stage i of the image encoder, PWAM receives the i-th visual feature map V_i and token embeddings L, projects them through 1×1 convolutions to generate key, query, and value matrices, and performs cross-modal attention to produce a language-aware visual feature map F_i. This map is then fused back with the original visual features via a language pathway. A lightweight decoder processes the final language-aware visual feature map to produce the segmentation mask.

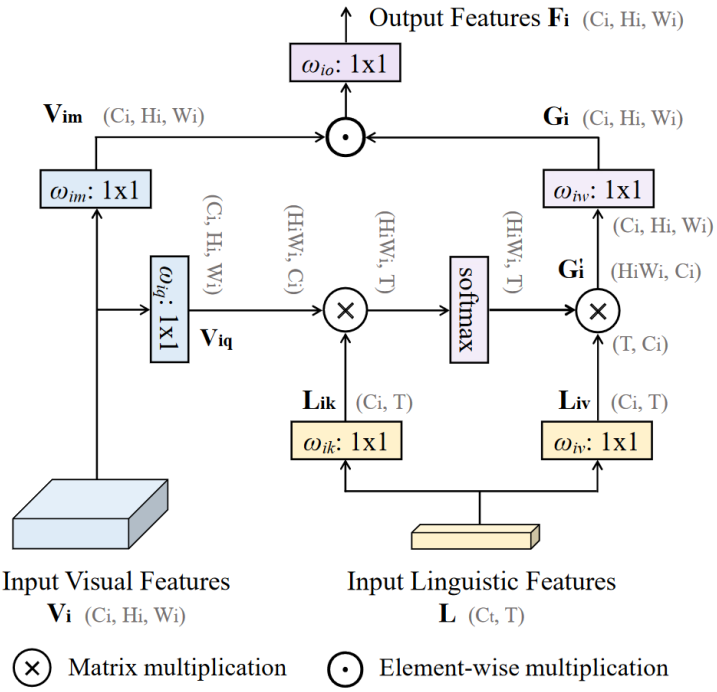


Figure 13: Pixel-Word Attention Module (PWAM) of the LAVT model. Text in grey refers to the shape of the tensor, and ω blocks to a convolution operation. Taken from [138].

RefSegformer [139] is a RES model designed for Robust RES (a generalization of RES which also allows negative referring expressions, see Sec. 16). At each stage i of the transformer vision encoder, the vision features V_i are fused with the language features L (extracted by an independent language encoder) through the Vision-Language Token Fusion (VLTF) module (Fig. 14). The resulting multimodal features serve as input for the next encoding stage. The VLTF module performs early fusion using three cross-attention operations: first, vision-aware language features are generated via vision-guided cross-attention; then, conditional tokens containing multimodal information are created through attention with learnable memory tokens; finally, language-enriched visual features are produced via attention between visual features and conditional tokens. Unlike models such as LAVT that rely solely on language-guided cross-attention, RefSegformer also uses memory tokens to dynamically select relevant language cues. Conditional and blank tokens are decoupled, with conditional tokens aligning modalities and blank tokens attending to image areas unrelated to the referring expression. Finally, a multimodal binary classifier processes the tokens and the decoder output mask to predict if the referring expression is negative.

15.3 Mutual cross-attention fusion

Vision-language cross-attention enables one modality to guide or inform the processing of the other. In a cross-attention layer, the key and value vectors are derived from the first modality, while the query vectors come from the second modality. The output is a set of features of the first modality informed (i.e., re-weighted) by those of the second modality. Therefore, vision-language cross-attention allows for two possible interactions between modalities: language-informed visual features (as seen in the previous section), and vision-informed linguistic features.

A recent finding by Liu et al. [140] shows that relying solely on language-informed visual features (as done by models in the previous paragraph) limits performances. In fact, such features tend to be dominated only by visual information over linguistic one. To address this issue, they introduced **M^{3}Net**, a RES model featuring a mutual multi-modal cross-attention

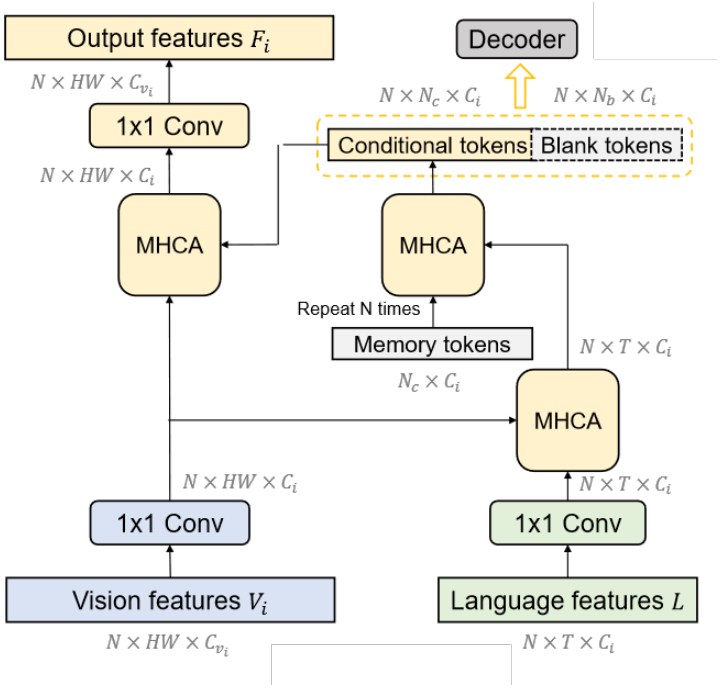


Figure 14: Vision-Language Token Fusion (VLTF) module employed in the RefSegformer model. Taken from [139].

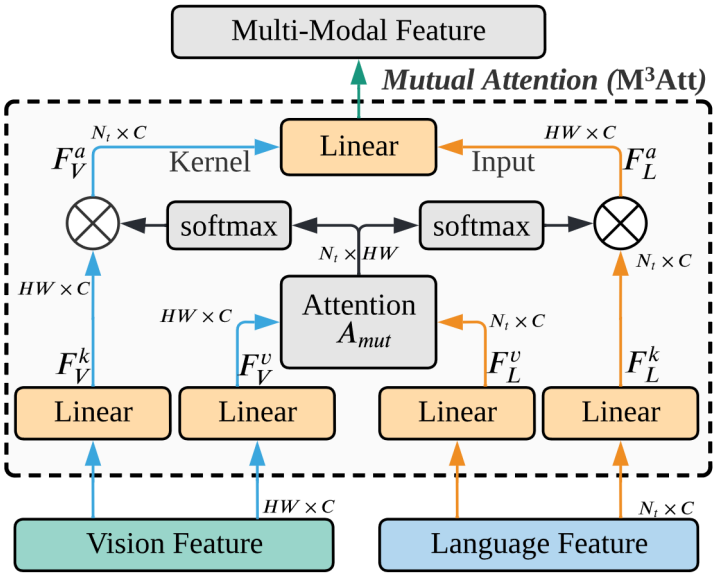


Figure 15: The M^{3}Att (Mutual Multi-Modal Attention) module, which integrates both language-guided and vision-guided cross-attention for multi-modal feature fusion. Taken from [140].

fusion module (M^{3}Att) for balanced interaction between language and vision (Fig. 15). This module calculates a mutual attention matrix A_{mut} to generate both language-attended visual features and vision-attended linguistic features. An Interactive Multi-Modal Interaction (IMI) module is placed between M^{3}Att stages in the mask decoder, enriching multi-modal outputs with linguistic context at each stage. Additionally, a Language Feature Reconstruction (LFR) module encourages retention of language information by reconstructing the language feature vector from the decoder's final multi-modal vector. This multi-task learning approach trains the model jointly on text reconstruction and referring expression segmentation.

Taking into account M^{3}Net's paper findings, several recently published models incorporate

mutual cross-modal mechanisms. **DMMI** (Dual Multi-Modal Interaction) [141] is a RES model employing a Multi-scale Bidirectional Cross-Attention (MBA) module in the encoding phase and a dual-branch transformer decoder, facilitating bidirectional information flow between linguistic and visual streams. The MBA module aligns text and vision bidirectionally, accounting for pixel-token, region-token, and pixel-(token sequence) interactions, thus enabling multi-scale cross-attention views of the inputs. The mask decoder comprises two branches: text-to-image and image-to-text. In the text-to-image branch, text features guide visual decoding to generate the segmentation mask via cross-attention. The image-to-text branch reconstructs a partial linguistic feature vector from language-informed visual features through a Context Clue Recovery (CCR) module, encouraging visual features to retain semantic information of the referring expression. The model is trained with a multi-task learning approach combining pixel-level supervision and contrastive learning for text reconstruction.

15.4 Weakly-supervised RES

To train a RES model, a large dataset of images, referring expressions and pixel-level segmentation masks triples is required. However, producing such datasets, i.e. generating referring expressions and segmentation masks for each image, is a labor-intensive task. In the context of RES, weakly-supervised learning reduces the reliance on these detailed annotations by learning from coarser or less expensive forms of supervision, such as image-level labels, bounding boxes, image-only and text-only datasets. Weak supervision drastically expands the amount of data available for training. It may even lead the model to stronger generalization capabilities and reduced overfitting, due to the model being forced to learn more meaningful correspondences between image regions and referring expressions [142].

The first example of weakly-supervised RES learning was published by the creators of LSTM-CNN, the first RES model [143]. They expanded the LSTM-CNN model by training it on large-scale language-only datasets and closed-vocabulary semantic segmentation datasets. In particular, they pre-trained a word embedding matrix using the GloVe method [124] on a large-scale language-only dataset, which is then kept fixed at training time. Moreover, they leverage closed-vocabulary semantic segmentation datasets by directly using the category label as a referring expression.

TSEG [144] (Text-grounded semantic SEGmentation) is a weakly-supervised learning framework to learn RES models from image-level referring expressions, without pixel-level ground truth masks. Similarly to CLIP [**radford2021learning**], TSEG includes independent encoders for each modality and a cross-modal late-fusion mechanism which computes a patch-text similarity matrix. Specifically, an input image and a set of input textual expressions are encoded independently, and a patch-text similarity matrix is computed. The matrix is then summarized by a text-wise pooling operation, obtaining an "image-text" similarity vector. Hence, the model is trained by learning to classify which textual expressions effectively refer to the image, without needing ground-truth segmentation masks.

Similarly, **Shatter and Gather** [145] is another framework for learning a RES model from image-level text supervision. It consists of two parts: a bottom-up attention module (based on slot attention [146]), which identifies individual object entities in the input image, and a top-down attention module, which fuses the visual features of the identified entities with features extracted from the referring expression to infer semantic relationships between the identified entities and the referring expression, and combine them based on an estimated mutual relevance score. At inference time, a segmentation mask is predicted by considering the attention matrices of the bottom-up and top-down module through an interpolation and binary thresholding process.

Omni-RES [147] is a weakly-supervised training framework which aims to make full use of unlabeled, weakly-labeled and fully-labeled data to efficiently train RES models. Namely, it can learn RES models from text-only supervision, bounding boxes of the referred object or even a single point placed on the object. Such "omni-labels" are not directly transformed into

supervision signals, but they are instead used to produce high-quality pseudo-masks to train on (pseudo-learning framework). Specifically, a teacher model is responsible for producing pseudo segmentation masks. The raw masks returned by the teacher are refined through an Active Pseudo-Label Refinement Process (APLR): if a point annotation is available, the mask is kept only if it contains the ground truth point, whereas if a bounding box is given, it is kept only if the intersection between the ground truth box and the mask is above a certain threshold. Then, a student model is trained on the refined pseudo masks, and updates the teacher model after each training step through an exponential moving average (EMA) operation.

Partial-RES [148] is a partially-supervised training paradigm for RES, which aims to learn RES models from Referring Expression Comprehension (REC) datasets, i.e. datasets in which ground-truth annotations are bounding boxes of the referred objects, and a limited fraction (as low as 1%) of RES pairs. It revisits the SeqTR model [149], which approaches RES as a point sequence prediction task, and extends it to the weakly-supervised setting. Namely, SeqTR produces a segmentation mask as a sequence of contour points, rather than a binary map. Similarly to Omni-RES [147], Partial-RES consists in generating pseudo segmentation masks from just the bounding boxes, through a Resampling Pseudo Point (RPP) strategy. SeqTR is then finetuned both on annotated samples and pseudo-annotated ones.

15.5 Transfer Learning from large pretrained models

The major challenge of RES resides in the effective alignment of visual and linguistic features. A large amount of image-expression data pairs are required to capture the huge heterogeneity within each modality and the complex relationships between them, in order to train a sufficiently general and robust RES model. However, as already mentioned in the previous section, a limited number of RES datasets are available in the literature. One of the main strategies to make up for the limited availability of training data is using transfer learning methodologies. By leveraging large pretrained models through transfer learning, RES models can achieve better generalization with reduced training data requirements. In the context of RES, the transfering of vision/language knowledge can happen from pretrained vision, language or vision-language models.

Many of the RES models discussed in previous sections leverage large vision models (e.g. Swin Transformer [150]) and language models (e.g. BERT [151]) pretrained on large-scale unimodal datasets as encoders to speed up training and facilitate convergence. Moreover, the availability of large vision-language alignment models such as CLIP [**radford2021learning**] has led to their integration into RES models even without any finetuning required (zero-shot RES). CLIP is trained on a dataset of over 400 million image-text pairs collected from the internet, and is shown to transfer well to a wide variety of language-vision tasks, thanks to its learned strong cross-modal alignment capability. As CLIP is trained to predict if an expression and an image are paired together, it excels in image-level downstream tasks such as image classification and retrieval, but adapting it to be reused in pixel-level tasks such as RES is not straightforward.

For example, **CRIS** [152] (CLIP-driven Referring Image Segmentation) is a RES model leveraging on CLIP. To transfer the multi-modal knowledge of CLIP, CRIS uses a vision-language decoder and a text-to-pixel contrastive learning framework to finetune CLIP and achieving cross-modal alignment at the pixel-level. Specifically, the vision-language decoder is designed to propagate semantic linguistic information to the pixel-level visual feature map, promoting consistency between the two modalities. In addition, the text-to-pixel contrastive learning explicitly enforces linguistic features to be similar to referred pixel-level visual features and dissimilar to the irrelevant ones, thus achieving pixel-text alignment needed for the RES task.

ETRIS [153] is another RES model which uses CLIP as its vision and language encoders. This model eases computational requirements by freezing CLIP encoders and training through a parameter-efficient finetuning strategy, using a newly introduced Bridger module (which is basically a multi-modal version of the Adapter [154]). This facilitates efficient transfer of CLIP

image-text pretrained knowledge to the RES task.

Global-Local CLIP [155] is a zero-shot RES model which leverages pretrained CLIP to perform RES, without any finetuning needed. It features an unsupervised instance segmentation method called FreeSOLO [156], which produces mask proposals for an input image. CLIP visual encoder is used to encode both the full-sized image (global-context features) and each crop corresponding to the proposed masks (local-context features), which are separately processed and finally summed together to produce a global-local visual feature representation for each proposed mask. In parallel, CLIP text encoder is used to encode both the full referring expression (global-sentence features) and just the category of the referred object (local target noun, e.g. "cat" or "person"), parsed by the spaCy tool [157]. Since CLIP is already trained on the task of aligning images with their text description, the method returns the mask proposal which maximizes the similarity between the textual feature and the visual features among all the mask proposals.

LISA [158] (Language-Instructed Segmentation Assistant) is a multi-modal Large Language Model (LLM) capable of performing RES. LISA introduces the reasoning segmentation task, a RES task in which a complex and implicit referring expression is given, as it often occurs in conversations. Similar to large conversational AI models, LISA excels in complex reasoning, world knowledge, explanatory answers and multi-turn conversation, while also performing RES. Under the hood, LISA leverages pretrained LLaVa [159] as a multi-modal encoder and SAM [17] as a visual encoder. Both LLaVa and SAM are kept frozen at training time. Transfer learning is facilitated by using a single LoRA module [160] on top of LLaVa to perform parameter-efficient finetuning. To integrate segmentation capabilities into LISA, an "embedding-as-mask" paradigm is used: when a special <SEG> token is included in the output text of the model, its hidden representation is decoded into the corresponding segmentation mask by a mask decoder.

16 Extensions of RES

Over time, researchers have expanded the RES task by broadening the range of possible referring expressions and trying to make RES models more robust to them. As a result, several tasks closely related to RES have emerged.

A core assumption of traditional RES is that the referred object(s) always exist in the input image, which may limit model robustness and interpretability. A more general task, **Robust RES** (R-RES) [139], removes this assumption by handling both "positive" and "negative" referring expressions, with the latter referring to objects absent in the image. R-RES models are trained to output an empty segmentation mask for negative expressions.

An even more general task, **General RES** (G-RES) [161], extends RES to include multi-target and no-target expressions in addition to single-target ones. Multi-target expressions refer to multiple objects in the image (e.g., "All the kids wearing shorts"), while no-target (or negative) expressions do not match any objects in the image. G-RES is a flexible yet challenging task with applications in a wide range of practical scenarios.

Multi-granularity RES (M-RES) [162] specifically focuses on a RES setting in which expressions may refer both to objects (e.g., a person) and to parts of it, with varying degrees of granularity (e.g., a person's arm, hand, index finger or nail).

Theoretically, any RES model can perform any of the tasks presented here, as they are generally able to output empty or multi-target masks, and seamlessly handle referring expressions of different granularity. While some RES model architectures are explicitly designed to address an extended RES task (e.g., RefSegformer [139] is tailored for R-RES), the key to achieving good performances is the choice of a training dataset which is suited for the task at hand. The main datasets available for RES and its extensions are discussed in Sec. 19.

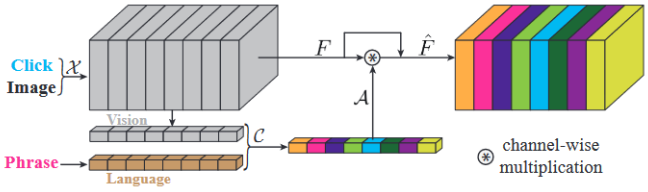


Figure 16: Attribute Attention [112]: It emphasizes feature channels that have larger response in semantic attribute learning, which is based on the phrase interaction, click interaction and visual patterns.

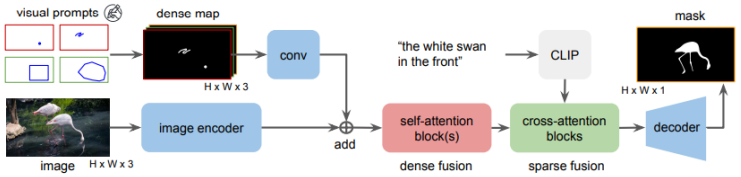


Figure 17: Decoupling visual and textual prompts through separate encoding and Late Fusion strategies, using self-attention for visual prompts and cross-attention for textual prompts. Image from [163].

17 Compositionality with Visual and Text Prompts

In the context of Interactive Segmentation (sec. 5), we have seen how the compositionality of visual prompts has emerged as a promising direction to improve segmentation performance by combining multiple forms of user input. With the advancement of referring segmentation techniques, new opportunities arise that integrate both visual and textual prompts in a unified framework.

Early efforts in this direction aim to leverage spatial clicks to localize the target object and semantic phrases to describe its attributes [112]. This approach is motivated by the complementary roles of the two modalities: while phrases convey *what* the object is, spatial clicks indicate *where* it is located. Combining both forms of interaction facilitates more accurate segmentation with fewer user interactions. Specifically, **Attribute Attention** is introduced (see Figure 16) to integrate semantic attributes into the spatial encoding pipeline. Experimental results demonstrate that combining phrase-based and click-based inputs outperforms individual approaches, and incorporating speech transcription further improves usability and segmentation quality. Further developments investigate the optimal stage for fusing visual and textual prompts with the image input. In particular, [163] propose to decouple the encoding processes of visual and textual prompts and apply two distinct late fusion mechanisms. Visual prompts are processed independently through an image encoder and a convolutional module, and then merged using self-attention—referred to as *Dense Fusion*. When textual input is available, it is encoded by the CLIP [13] text encoder and fused with the visual feature representation via cross-attention—termed *Sparse Fusion* (see Figure 17). This dual-fusion strategy leads to substantial improvements in segmentation performance, underscoring the value of prompt compositionality across modalities.

18 Metrics

Most of the evaluation metrics for RES models are based on the already described IoU metric (Eq. 13). By convention, the IoU is 1 if the union is empty: this happens only in the edge case of a negative referring expression, for which the model correctly predicts an empty mask. Throughout this section, the following notation is used: $\mathcal{D}$ is a RES dataset, $\mathcal{D}_p$ and $\mathcal{D}_n$ are the subsets of $\mathcal{D}$

containing respectively the target (positive) and no-target (negative) image-expression pairs, $\mathbf{g_i}$ and $\mathbf{m_i}$ are respectively the ground-truth and predicted segmentation mask for the i-th image, $|\cdot|$ is the cardinality of a set, $\mathbb{1}_{[\cdot]}$ is the indicator function.

mIoU. Mean Intersection over Union (mIoU) is the average per-image IoU. It quantifies how close the predicted segmentation masks match the ground truth masks on average.

$$mIoU = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|} IoU_i = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\frac{|\mathbf{m_i}\cap\mathbf{g_i}|}{|\mathbf{m_i}\cup\mathbf{g_i}|} \tag{17}$$

oIoU. Overall Intersection over Union (oIoU) is the ratio between the total number of intersection pixels and the total number of union pixels in a dataset (i.e., it considers all the samples in a dataset as just one large sample and computes the IoU on it).

$$oIoU = \frac{\sum_{i=1}^{|\mathcal{D}|}|\mathbf{m_i}\cap\mathbf{g_i}|}{\sum_{i=1}^{|\mathcal{D}|}|\mathbf{m_i}\cup\mathbf{g_i}|} \tag{18}$$

Due to its formulation, the oIoU favors large objects (i.e., it tends to be higher for images with a large foreground area), whereas the mIoU treats tiny and large objects equally [138]. Therefore mIoU is usually preferred over oIoU when comparing RES performances among models, especially when referred objects have very different sizes.

Precision@X. It is an accuracy metric of "correctly" segmented regions. Specifically, it is the fraction of samples which achieve an IoU higher than X.

$$Prec@X = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\mathbb{1}_{[IoU_i\geq X]} \tag{19}$$

gIoU. Generalized IoU (gIoU) [161] is an extension of the mIoU metric which can handle negative referring expressions.

$$gIoU = \frac{1}{|\mathcal{D}|}\left(\sum_{i=1}^{|\mathcal{D}_p|}\frac{|\mathbf{m_i}\cap\mathbf{g_i}|}{|\mathbf{m_i}\cup\mathbf{g_i}|} + \sum_{i=1}^{|\mathcal{D}_n|}\mathbb{1}_{[|\mathbf{m_i}|=0]}\right) \tag{20}$$

The $\mathbb{1}_{[|\mathbf{m_i}|=0]}$ term penalizes the model for outputting a non-empty segmentation mask for a negative referring expressions.

PointM. PointM [164] is a simple accuracy metric for segmentation masks. Given a confidence score map $\mathbf{s}$ (which may be thresholded to obtain a segmentation mask $\mathbf{m}$) and its corresponding ground-truth mask $\mathbf{g}$, let $(i,j) = \arg\max_{i,j}(\mathbf{s})$ the pixel achieving the maximum confidence score. If $\mathbf{g}_{i,j} = 1$, then PointM = 1 ("Hit"), otherwise 0 ("Miss"). For a set of images, the mean PointM metric is considered. Overall, the mean PointM metric can be formulated as

$$PointM = \frac{Hits}{Hits + Misses} \tag{21}$$

This metric is rarely used in literature, as it is a measure of localization accuracy based on just one pixel, i.e. the most "confident" one.

19 Datasets

In this section, several existing RES datasets are briefly overviewed. Relevant statistics about them are summarized in Tab. 4

RefCLEF. The RefCLEF (aka ReferIt) dataset [165] was the first published large-scale RES dataset. It was generated from a simple two-player game in which player A writes an expression

Dataset	N. images	N. categories	N. targets	N. ref. expressions			Avg. len.	Dict. size
				single-target	no-target	multi-target		
RefCLEF [165]	19,997	238	99,296	130,364	–	–	3.5	9,320
RefCOCO [166]	19,994	80	50,000	142,210	–	–	3.5	10,341
RefCOCO+ [166]	19,992	80	49,856	141,564	–	–	3.5	12,226
RefCOCOg [167]	25,799	80	49,822	95,010	–	–	8.5	12,728
VGPhraseCut [168]	77,262	3,103	360,615	284,649	–	60,837	2.0	8,456
R-RefCOCO [139]	19,994	80	50,000	142,210	199,869	–	3.7	10,344
R-RefCOCO+ [139]	19,992	80	49,856	141,564	196,783	–	3.9	12,230
R-RefCOCOg [139]	25,799	80	49,822	95,010	159,806	–	8.4	13,114
gRefCOCO [161]	19,994	80	60,287	166,008	32,202	80,022	4.6	16,529
Ref-ZOM [141]	55,078	80	74,942	56,972	11,937	21,290	6.8	11,714
ReasonSeg [158]	1,218	≤ 620	1,218	2.667	–	2,411	24.0	4,543
GRD [169]	10,578	-	18,514	9,323	22,201	-	5.9	614
MRES-32M [162] (objects)	1M	365	15.3M	15.3M	–	–	4.6*	?*
MRES-32M [162] (parts)	1M	2,299	16.9M	16.9M	–	–	4.6*	?*
RefCOCOm [162] (objects)	3,000	80	7,596	21,586	–	–	3.5	3,511
RefCOCOm [162] (parts)	3,000	391	26,476	70,324	–	–	4.9	3,074

Table 4: Overview of existing RES datasets. *MRES-32M dataset has not been published, and average length of referring expressions was only reported for the objects' and parts' referring expressions as a whole; dictionary size was not reported.

referring to a given object in an image and player B tries to find and click on the referred object. The RefCLEF dataset is built on top of the ImageCLEF IAPR TC-12 image dataset [170], which depicts scenes of everyday life (people, animals, cities, sports, landscapes), and its expansion SAIAPR TC-12 [171], which enriches the dataset with segmentation masks of various objects in each image.

RefCOCO, RefCOCO+, RefCOCOg. The RefCOCO dataset (aka UNC-Ref-COCO or UNC-Ref for short) [166] was generated from a ReferIt-like two-player game based on the MS-COCO image dataset [111]. RefCOCO expressions use a concise and simple language.

The RefCOCO+ dataset is a variant of RefCOCO which prohibits the use of absolute location words. Instead, its referring expressions focus purely on the appearance of objects (e.g., "the second giraffe on the right" is not an allowed expression, but "the giraffe eating leaves from a tree" is), making this dataset harder than RefCOCO.

The RefCOCOg dataset (aka G-Ref, short for Google Refexp) [167] is also based on RefCOCO images, but features longer and semantically richer referring expressions than RefCOCO's ones.

Sample referring expressions from the RefCOCO, RefCOCO+ and RefCOCOg datasets are reported in Fig. 18.

VGPhraseCut. VGPhraseCut (or simply PhraseCut) [168] was collected on top of the Visual Genome semantic segmentation dataset [172]. It describes a larger number of object and

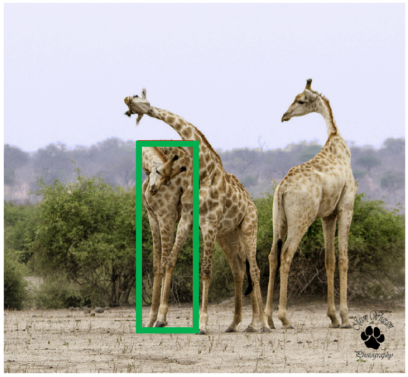


Figure 18: Sample image and referring expressions from the RefCOCO, RefCOCO+ and RefCOCOg datasets.

stuff categories, with widely varying frequency and object sizes. Moreover, referring expressions are not limited to targeting a single object instance per image, as in ReferIt and RefCOCO/+/g datasets, as this dataset also contains multiple-target samples. Referring expressions focus on objects' attributes such as colors, shapes, parts, and relationships with other entities in the image. Each referring expression combines concepts following a common template "category + attribute + relationship", e.g. "large cake on plate", "wipers on train" or "black shirt".

R-RefCOCO, R-RefCOCO+, R-RefCOCOg. These datasets extend RefCOCO, RefCOCO+ and RefCOCOg respectively by introducing negative referring expressions. They were introduced in [139] along with the new task of Robust RES (Sec. 16). Negative referring expressions for each object are generated automatically by several different methods, such as selecting referring expressions from other images or replacing relevant words such as target's category name, position, color, relationship to other objects.

gRefCOCO. This dataset extends RefCOCO by introducing multi-target and no-target referring expressions. It was introduced in [161] along with the new task of Generalized RES (Sec. 16). It is a far more challenging and general dataset than previous ones, and provides enhanced flexibility and robustness to RES models in practical applications.

Ref-ZOM. The Ref-ZOM (**Z**ero/**O**ne/**M**any) dataset [141] is another extension of the RefCOCO/+/g datasets to the multi-target and no-target RES setting.

ReasonSeg. The ReasonSeg dataset [158] is a small collection of image-instruction pairs in which the provided referring expression is complex and implicit. It is based on OpenImages [173] and ScanNetv2 [174] image datasets, annotated with implicit text instructions and high-quality target masks. Instruction-like referring expressions included in this dataset are of two types: short phrases (e.g. "the camera lens that is more suitable for photographing nearby objects") and long sentences (e.g. "toddlers are curious and often enjoy exploring their surroundings: what object in the picture can provide a safe and enclosed space for a toddler to play in?"). Given its small size, ReasonSeg is mainly used as a benchmark dataset for validation and testing purposes.

GRD. The Group Referring Dataset (GRD) [169] is designed for Group-wise RES (GW-RES, Sec. 16). This dataset consists of images acquired from the Internet, grouped in 106 scenes (depicting real indoor and outdoor environments). Each group contains around 100 images and 3 textual expressions referring to one or multiple objects appearing in a subset of the images.

MRES-32M. The MRES-32M dataset [162] is designed for Multi-granularity RES (M-RES, Sec. 16). With around one million images taken from the Object365 dataset [175], it is the largest RES dataset to date. It includes 32.2M objects and parts, and a single referring expression for each.

RefCOCOm. The RefCOCOm dataset [162] is another dataset for Multi-granularity RES (M-RES, Sec. 16). It builds on a subset of RefCOCO and extends its original object-level data samples with part-level masks and corresponding referring expressions. Sample masks from



Figure 19: Object-level and part-level masks associated to the same image in the RefCOCOm dataset. Taken from [162].

Work	Fusion	ReferIt	RefCOCO		RefCOCO+		RefCOCOg	
		test	test A	test B	test A	test B	test(U)	test(G)
LSTM-CNN [114]	CC	48.03	–	–	–	–	–	28.14
RRN [129]	CC	63.63	57.26	53.95	42.15	36.11	–	36.45
DMN* [134]	CC	52.81	54.84	45.20	44.25	32.49	–	37.64
RMI [136]	CC	58.73	45.69	45.57	30.48	29.50	–	34.52
LAVT [138]	LG-XA	–	75.82	68.79	68.38	55.10	62.09	60.50
RefSegformer [139]	LG-XA	–	79.31	74.25	69.24	55.91	63.07	58.48
M^{3}Net [140]	M-XA	72.97	76.23	70.36	70.50	56.98	67.37	63.90
DMMI [141]	M-XA	–	77.13	70.16	69.73	57.03	64.19	61.98

Table 5: Benchmark comparison of referring segmentation methods. The reported metric is the overall IoU (%), except for DMN (*) for which only mIoU was published. Abbreviations: CC = Concatenation-Convolution; LG-XA = Language-Guided Cross-Attention; M-XA = Mutual Cross-Attention. We report results on two slightly different test splits of the RefCOCOg dataset, created by UMD (U) [176] and Google (G) [167].

RefCOCOm are depicted in Fig. 19.

Table 5 summarizes the performances of the RES models described in Sec. 15 on the most commonly RES benchmark datasets (RefCOCO/+/g).

20 Few-Shot Learning in Promptable Segmentation

Few-Shot Learning (FSL) is a paradigm that enables a pre-trained model to generalize to new tasks with minimal examples. This capability extends to semantic segmentation, known as *Few-Shot Semantic Segmentation (F3S)*, where models learn to segment novel classes with limited supervision. The concept of F3S was first formalized by [177], which defines the task as follows: given a support set $S = \{(I_i^s, M_i^s)\}_{i=1}^k$, composed of k annotated images I^s with their corresponding binary masks M^s, and a query image I_q, the goal is to predict a segmentation mask $\hat{M}_q$ that identifies the target class in I_q. The parameter k specifies the number of annotated examples (shots), with each pair (I_i^s, M_i^s) providing guidance for the segmentation task. The objective is to enable the model to generalize to new classes not seen during pretraining.

Within this context, *Promptable Segmentation* refers to segmentation methods that leverage prompts—structured inputs that provide guidance to the model—to enhance its ability to generalize to unseen tasks or classes. These prompts can take various forms, such as images annotated with masks, points, boxes, or even textual descriptions, and are designed to convey task-specific information that guides the segmentation process at inference time. Building on these foundations, we introduce the notion of **Promptable Few-Shot Semantic Segmentation (PF3S)**, a formulation that consolidates and formalizes the conceptual bridge between few-shot

learning and prompt-based paradigms in semantic segmentation. In PF3S, the support set and the query image together form a structured prompt that provides task-specific guidance at inference time, enabling the model to generalize to unseen classes without the need for additional training.

Recent literature has already highlighted this convergence: [178] discuss how different forms of guidance—such as images, masks, and textual cues—can be modularly integrated to inform segmentation models, implicitly reflecting prompt-based mechanisms; [179] explicitly introduce the concept of visual prompting in the context of few-shot segmentation; and [35] emphasize that support sets are increasingly used as structured prompts within foundation model frameworks.

While traditional few-shot approaches typically rely on fine-tuning pre-trained models to adapt to new tasks—a process that involves computationally expensive retraining for each novel class—PF3S distinguishes itself by focusing on methods that achieve adaptation through conditioning on prompts alone. This avoids the need for additional task-specific training and aligns with the broader principles of prompt-based learning.

To systematically review PF3S methodologies, we categorize existing methods into two main classes: **episodic training**, where support and query are processed separately and fused during inference, and **in-context learning and tuning**, where support and query are integrated jointly within the model input to directly guide predictions. This taxonomy is further structured by architectural typologies distinctive of each class and by the encoding and fusion strategies adopted in individual works, offering a clear framework to analyze how prompt-based principles are embedded in F3S approaches.

21 Episodic Training

Episodic training is the first strategy used to implement PF3S. In this setting, training is organized into episodes, where each episode simulates a few-shot segmentation task by providing a prompt composed of a support set S and a query image I_q.

The model's goal is to predict the segmentation mask $\hat{M}_q$ for the target class in I_q, guided by the information contained in S. The training objective is to maximize the probability of producing the correct mask conditioned on the prompt:

$$P_\theta(\hat{M}_q \mid I_q, S), \tag{22}$$

where θ are the model parameters. This formulation teaches the model to extract task-specific information from the support set and apply it to segment the query image, enabling generalization to new classes not seen during pretraining.

Episodic training thus functions as a mechanism for encoding task structure: the support set provides contextual cues that describe the segmentation task, and the model learns to adapt dynamically based on this input. A central challenge is how to effectively fuse the support and query information—whether through feature-level, parameter-level, or hybrid strategies—to achieve robust performance on unseen classes.

In the following sections, we review key architectures that implement episodic training for PF3S, analyzing how different encoding and fusion strategies translate the prompt into actionable knowledge for segmentation.

21.1 Two-Branches Network

One of the foundational architectural paradigms in PF3S is the **two-branches network**, in which the support set and the query image are processed through two distinct branches. The support branch is responsible for encoding task-specific knowledge derived from the annotated examples, while the query branch focuses on extracting a representation of the query image that will be fused with the task information from the support branch. Introduced in early

meta-learning [180], this architectural separation naturally maps onto the prompting framework: the support set plays the role of the prompt, specifying what to segment, and the query image is the input to be interpreted accordingly after fusion.

Two representative works in this category are **One-Shot Learning for Semantic Segmentation (OSLSM)** [177] and **Few-Shot Segmentation Propagation with Guided Networks (FSSPGNet)** [181]. While both adopt the two-branches framework, they differ substantially across multiple components.

21.1.1 Support Branch: Task Representation Construction.

The first point of divergence lies in how the support branch constructs the task representation, combining encoding and early fusion of support images and annotations.

In OSLSM, the authors apply a **Masking** operation to fuse the image with its mask: the RGB image is element-wise multiplied by the binary mask to retain only the target object, effectively suppressing background content. This masked image is then encoded by a VGG-16 network (unaltered in its first layer), producing a **Feature Embedding**, i.e., a 1000-dimensional feature vector. The encoded support set is subsequently transformed into task-specific parameters (w, b) through a **Weight Hashing** mechanism, which serves as a task encoding technique, by using a fixed-weight projection to avoid introducing excessive learnable parameters, thereby reducing the risk of overfitting. These parameters define a linear decision boundary that guides segmentation over the query image.

In contrast, FSSPGNet constructs the support branch by combining sparse supervision and visual features through a multi-stage process involving distinct encoding and fusion strategies. The support image is first encoded by a convolutional encoder to obtain a Feature Embedding $F_s = \Omega(x)$, capturing high-level visual representations. The support annotations consist of a small number of positive and negative points indicating the presence or absence of the target class. These annotations are directly projected onto the spatial coordinates of the feature map F_s, forming two binary masks $m(+)$ and $m(-)$, which respectively highlight locations marked as positive or negative examples. This mapping procedure, which associates annotated pixels with their corresponding feature locations, defines the encoding strategy we refer to as **Feature Space Mask**. To obtain a compact representation of the task, the masked regions of the support feature map are aggregated separately for positive and negative evidence. These masked features are pooled, via averaging, and subsequently combined into a single vector z that summarizes the segmentation objective. This stage is referred to as **Late Fusion**, since the fusion between the encoded support image and encoded support annotations occurs post-embedding—unlike in OSLSM, where support annotations modulate the input image prior to encoding.

21.1.2 Query Branch and Task Fusion.

Turning to the query branch, both models encode the query image through standard Feature Embedding. However, they differ markedly in how they fuse task information with the embedded query features.

In OSLSM, the task parameters (w, b) are applied through a **Task-Conditioned Logistic Regression**, a fusion strategy that injects information from the support set into the query prediction by modulating each spatial location (i, j) of the query feature map. Specifically, for each feature vector of the query image $F_q^{(i,j)}$, the model outputs $M_q^{(i,j)} = \sigma(w^\top F_q^{(i,j)} + b)$, with σ denoting the sigmoid activation. This technique interprets the task as a fixed linear classifier operating across the query feature space, making it efficient but structurally constrained.

For FSSPGNet, different alternatives are explored to fuse the task information—captured by the support-derived vector z—with the query representation. Among these, the most effective is **Feature Fusion (Concatenation)**: the task vector z is spatially broadcast and concatenated to each position of the query feature map, producing a joint representation that is then decoded

into a segmentation mask. Two additional fusion strategies are evaluated: in **Parameter Regression**, z is used to dynamically generate the weights of a convolutional decoder, enabling adaptation to the current task; in **Nearest Neighbors and Prototypes**, segmentation is achieved by comparing query pixels to class prototypes derived from the support. While both approaches are theoretically appealing, they perform worse than direct concatenation in practice. Unlike OSLSM, where fusion relies on a fixed linear projection, FSSPGNet's more expressive mechanisms allow it to better accommodate sparse supervision and intra-class variability.

21.2 Feature Enrichment through Support-Query Fusion

A further evolution of PF3S architectures shifts from compressing the support into fixed task vectors or parameters toward mechanisms that explicitly intertwine support and query representations across spatial and semantic dimensions. **Cyclic Memory Network (CMN)** [182], illustrated in Figure 20, and **Prior Guided Feature Enrichment Network (PFENet)** [183], shown in Figure 21, exemplify this paradigm by designing feature enrichment pipelines that allow task-relevant information to flow directly into the query representation throughout the inference process.

21.2.1 Task Cue Preparation.

Both models begin by encoding the support and query images through a shared convolutional backbone, applying standard Feature Embedding to extract multi-level representations as part of the overall task cue preparation.

In PFENet, this process yields two levels of embeddings: high-level features, which capture semantic content, and middle-level features, which retain finer spatial details.

In CMN, the extracted features are passed to a Multi-Resolution Shared Encoder, which processes intermediate outputs from block2 and block3 of a ResNet backbone. These features are compressed via 1×1 convolutions into structured key-value pairs—a technique we refer to as **Key-Value Encoding**—where keys encode semantic identity and values retain detailed appearance information.

A crucial step in both models is the fusion of the support image with its annotation.

In PFENet, the fusion process exploits a **Hadamard Fusion** module, which performs an element-wise multiplication between the high-level support features and the binary support mask, producing a masked support feature that selectively retains the semantic signature of the target object.

In CMN, this is implemented via **Masked Average Pooling**, which filters out background activations by averaging only the foreground regions defined by the binary mask. This operation produces two compact global descriptors: a semantic key vector f_K^s and a content-rich value vector f_V^s.

Unlike two-branch networks, CMN and PFENet do not rely solely on the support set to construct the task representation. Instead, they allow the query image to actively contribute to this process, shaping the task encoding prior to mask prediction.

In CMN, this is achieved through **Multi-Resolution Key-Value Fusion**. The query key-value features are expanded via Pyramid Average Pooling to obtain multi-scale representations, which are then aligned with the global support descriptors f_K^s and f_V^s through spatial interpolation. At each resolution i, these elements are fused to form enriched key-value pairs (k_i, v_i) that combine support-derived cues with query-specific context. These multi-resolution pairs serve as input to the memory refinement stage, where they are iteratively updated via cross-attention with other representations.

In PFENet, a similar effect is obtained via **Similarity Fusion**. The masked high-level support feature is compared against the query features using pixel-wise cosine similarity, generating a

prior mask Y_Q that reflects the likely spatial extent of the target class within the query. This prior acts as a query-aware task representation that informs the following decoding stages.

21.2.2 Support-Query Fusion and Enrichment.

Another substantial distinction from two-branch architectures lies in the introduction of a dedicated feature enrichment phase before segmentation. This phase represents the core innovation of both CMN and PFENet: rather than directly decoding a fused support-query representation, these models refine the query features using task-derived information prior to prediction.

In PFENet, the Feature Enrichment Module (FEM) refines the query feature map using three sequential steps. During inter-source enrichment, the middle-level query features are concatenated with both the masked high-level support features—produced via Hadamard Fusion—and the prior mask Y_Q, enabling the injection of semantic and spatial cues from the support. In the inter-scale interaction phase, feature maps at different resolutions exchange information, allowing the integration of fine details and global context. Finally, in the information concentration step, all enriched multi-scale features are resized and aggregated into a single refined representation. This enhanced query feature is passed to a classification head with convolutional layers and a softmax function to predict the segmentation mask.

In CMN, the enriched key-value representations $\{(k_i, v_i)\}$, obtained via Multi-Resolution Key-Value Fusion, are used within a cyclic memory reading mechanism. At each step, one (k_i, v_i) pair is treated as the active query, while the others serve as external memory. The model computes similarities between the query key and memory keys, retrieves relevant content from the memory values, and updates the query features through a recursive module based on ConvGRU. This iterative refinement captures long-range and cross-resolution dependencies. The final segmentation is generated by the episode decoder, which integrates both the original and memory-enhanced features through convolutional and pooling operations.

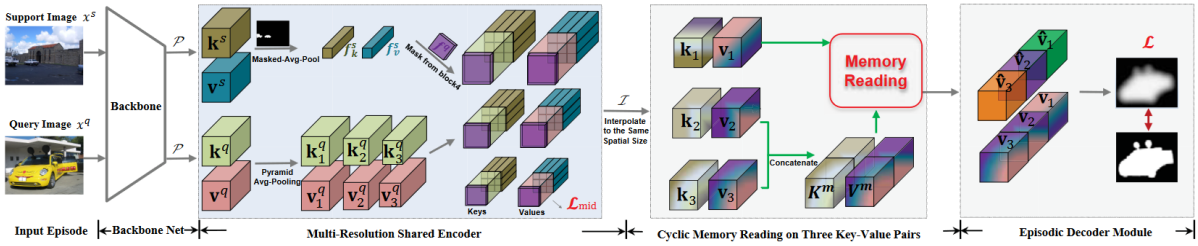


Figure 20: The Cyclic Memory Network (CMN) architecture consists of four main modules: a shared ResNet backbone, a Multi-Resolution Shared Encoder, a Cyclic Memory Reading module, and an Episode Decoder. Support and query images are both encoded into multi-resolution key-value representations via the encoder. The support mask is used to extract compact descriptors via Masked Average Pooling. These are fused with the query features using Multi-Resolution Key-Value Fusion to produce enriched representations (k_i, v_i), which are progressively refined via cyclic memory reading based on ConvGRU before being decoded into the final segmentation mask. Image from [182].

22 In-context learning & tuning

In-context learning has emerged as a transformative paradigm in artificial intelligence, enabling models to generalize to new tasks by conditioning directly on a structured prompt. Originally introduced in Natural Language Processing [43], this approach has gained traction in vision through the use of visual prompts—combinations of support examples and query inputs that define the task without requiring model retraining.

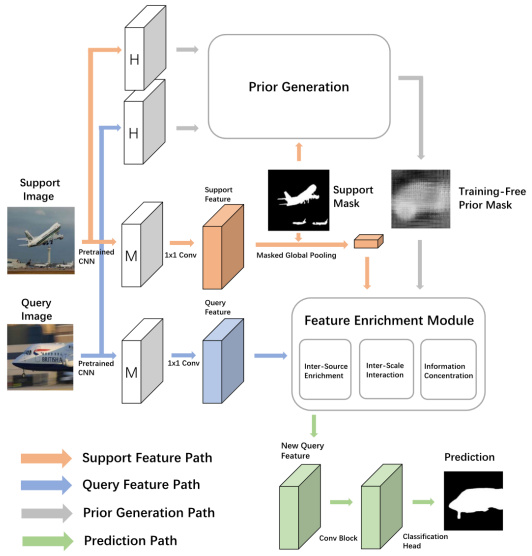


Figure 21: The Prior Guided Feature Enrichment Network (PFENet) processes support and query images through a shared CNN backbone to extract high-level features H and middle-level features M. The high-level support features are fused with the binary mask using Hadamard Fusion to produce a masked support feature, which is compared with query features to generate a prior mask Y_Q. The Feature Enrichment Module (FEM) receives as input the middle-level query features, the masked support feature, and the prior Y_Q, and progressively enriches the query representation through inter-source enrichment, inter-scale interaction, and information concentration. The resulting enhanced features are used to produce the final segmentation mask. Image adapted from [183].

In the context of segmentation, a visual prompt typically consists of one or more support images annotated with their expected outputs (e.g., segmentation masks), together with a query image whose output the model must infer. Crucially, in-context learning does not rely on encoding the support into intermediate task vectors or parameters. Instead, it provides the entire support-query structure as input, enabling the model to interpret the task directly through its internal representations.

Formally, given a query input x and a visual prompt $C = \{(x_1, y_1), \ldots, (x_k, y_k)\}$, where each pair represents a support image and its associated output, a pretrained model M estimates the conditional probability of a candidate output y_j as:

$$P(y_j|x, C) \equiv f_M(y_j, C, x) \tag{23}$$

and produces the final prediction as:

$$\hat{y} = \underset{y_j \in Y}{\operatorname{argmax}}\ P(y_j|x, C) \tag{24}$$

This formulation highlights the core principle of in-context learning: the task is defined entirely by the visual prompt, and the model must infer the output for the query by reasoning over the examples it sees—without any further training or parameter adaptation.

22.1 Inpainting networks

A compelling application of in-context learning in vision emerges in generalist inpainting networks. Originally designed to complete missing parts of images, these models have evolved into universal solvers capable of handling tasks such as segmentation, edge detection, and colorization. They do so by learning to complete an image given partial information—a capability that aligns naturally with the structure of visual prompts.

In this setting, support and query examples are composed into a single large image where each row or block consists of a task input (e.g., an image) and its corresponding output (e.g., mask). These are arranged sequentially, forming a visual demonstration of the task. The query image is appended at the end, left without its output, prompting the model to complete it based on the prior examples.

This layout forms the basis of a simple yet powerful strategy we refer to as **Input-Level Grid Fusion**. Rather than fusing features or tokens at intermediate stages, the model receives a single unified input where both the task context (support examples) and the query image are spatially juxtaposed in a grid-like structure. The inpainting model is thus tasked with "filling in" the missing output for the query, just as it would complete a masked region in a corrupted image.

This fusion of structure and objective is both intuitive and elegant: the prompt is simply constructed by placing support examples and the query image side by side in a grid, and the model completes the missing output based on what it has seen. In doing so, inpainting networks become inherently promptable and task-adaptive, operating across domains with minimal architectural changes.

Works such as [184, 185, 41] have leveraged this principle to construct flexible models capable of performing segmentation and other tasks via pure input-level prompting. In what follows, we review these architectures, illustrating how Input-Level Grid Fusion enables inpainting models to function as generalist segmenters within the PF3S framework.

The progression of inpainting-based models for in-context segmentation tasks reveals a clear path of innovation through increasingly refined encoding strategies and task formulations.

22.1.1 Visual prompting via Image Inpainting.

This evolution begins with **Visual Prompting via Image Inpainting (VPIP)** [184], which first formalizes the Input-Level Grid Fusion paradigm. The input is processed using a **Two-Stage Feature Embedding**, an encoding strategy designed to capture both global semantics and local details. (1) In the first stage, a Masked Autoencoder (MAE) randomly masks parts of the prompt during training and trains a Vision Transformer to reconstruct the missing regions. This self-supervised pretraining step encourages the model to learn long-range dependencies and structural relationships across the input, enabling it to reason over task demonstrations holistically. Crucially, masking is applied only during training; at inference time, the full prompt—containing both support and query—is passed to the model unaltered. (2) In the second stage, a Vector Quantized GAN (VQGAN) encodes the reconstructed image into discrete tokens via a learned codebook—a finite set of representative visual embeddings. Each image region is mapped to its closest prototype in this codebook, effectively compressing the information while preserving fine visual details. The combination of these two stages equips the model to interpret both the overall layout of the task and the visual characteristics of its components. See Figure 22 for a visual example.

22.1.2 Generalist Painter

Generalist Painter [185] extends the VPIP formulation by adopting a **Two-Stage Patch-Based Feature Embedding** scheme. Rather than encoding the visual prompt as a whole, Generalist Painter decomposes input-output pairs into patches and applies selective masking during training through Masked Image Modeling (MIM). This shift to patch-level encoding increases granularity and flexibility, allowing the model to capture finer context across dense prediction tasks such as segmentation, depth estimation, and keypoint detection. Painter also simplifies task representation by translating outputs into RGB-formatted images, enabling all tasks to be treated uniformly as image-to-image reconstruction problems. As shown in Figure 23, training involves stitching multiple input-output examples into a large prompt image, followed by block-wise masking applied only to output regions. A ViT-based encoder-decoder then reconstructs the missing outputs using a smooth-ℓ_1 loss, promoting accurate spatial recovery. During inference, the model operates in a fully prompt-based fashion: new examples are appended to the visual prompt, and the model predicts the missing query output in-context.

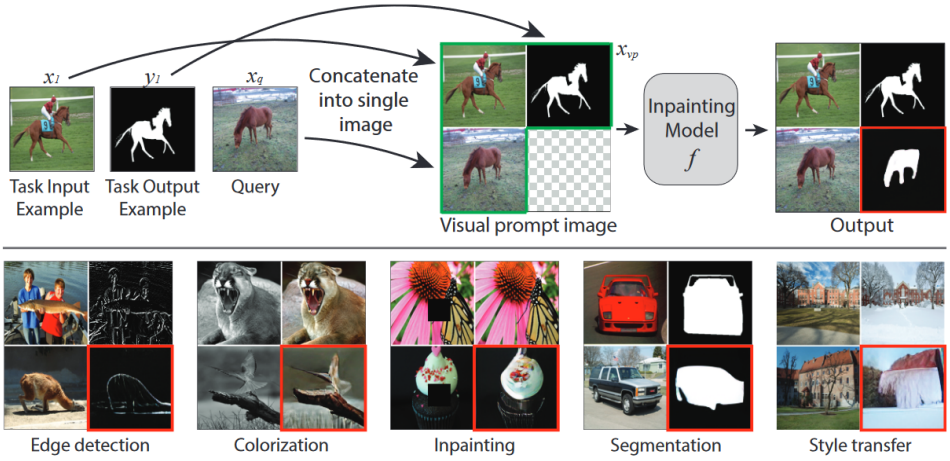


Figure 22: Demonstrating Visual Prompting via Image Inpainting across Multiple Tasks. **Top:** Construction of a visual prompt x_{vp} by concatenating task input-output examples (x_1, y_1) with a query image (x_q), highlighted in green. This grid-like single image serves as the visual prompt for the inpainting model, which is tasked with predicting the masked region (shown in red). **Bottom:** The model's versatility is showcased across various tasks such as edge detection, colorization, inpainting, segmentation, and style transfer. Particularly for segmentation, the model demonstrates its capacity to discern and outline object boundaries accurately, a capability extended to other vision tasks, demonstrating the model's generalist yet adaptable nature. Each task output is annotated in red to illustrate the model's predictive accuracy based on the provided visual prompts.

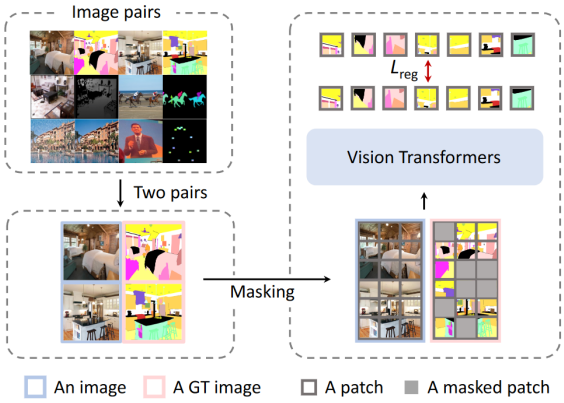


Figure 23: Training Pipeline of the Masked Image Modeling (MIM) Framework. Pairs of input and output images are processed via a Vision Transformer (ViT), with selective masking applied to the output images. Image from [185].

22.1.3 SegGPT: Segmenting Everything In Context

SegGPT [41] further advances the architecture by introducing a specialized encoding scheme called **Randomized Coloring Feature Embedding**, specifically optimized for semantic segmentation. In this strategy, instead of using fixed colors for segmentation labels, each category is assigned a random color in every support example. This prevents the model from associating colors with fixed meanings and forces it to learn to segment based on spatial and structural cues alone, thereby improving generalization to unseen classes and configurations. Additionally, SegGPT employs Mix-Context Training, where prompts are constructed from

multiple tasks or domains, pushing the model to reason over heterogeneous examples within a single input. During inference, SegGPT applies ensemble strategies (Figure 24) to enhance prediction quality: (1) spatial ensembles concatenate multiple prompts to expand the reference context; and (2) feature ensembles aggregate intermediate representations across attention layers, improving consistency and robustness. Finally, the model supports In-Context Tuning (Figure 25), a lightweight adaptation procedure where a learnable image tensor is appended to the prompt and fine-tuned on specific datasets. Only this tensor is updated, leaving the pretrained model weights untouched—allowing for efficient adaptation to new domains.

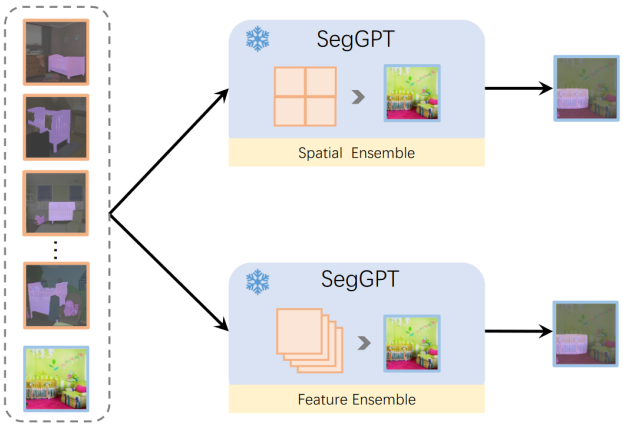


Figure 24: SegGPT's context ensemble methods for multi-example inference. The top section shows the *Spatial Ensemble* strategy, where stitched images are resized to match the input resolution, allowing unified processing. The bottom section displays the *Feature Ensemble* strategy, where query image features are averaged after each attention layer, aggregating contextual information from reference examples to enhance segmentation accuracy. Image from [41].

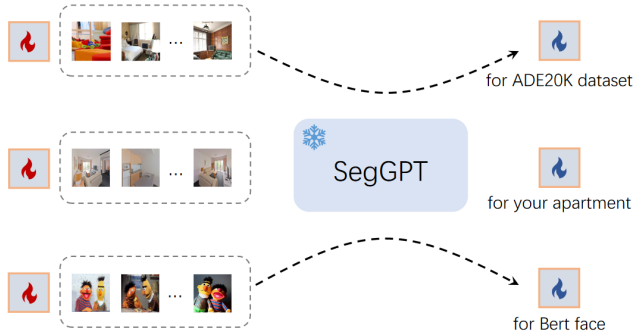


Figure 25: Diagram of the in-context tuning process used by SegGPT, tailored for specific tasks. The model's pre-trained architecture is frozen except for a learnable image tensor, which serves as the adaptive input context. This tensor is finely tuned to the dataset or scene, enabling SegGPT to perform precise segmentation without full retraining. Image from [41].

22.2 PF3S Benchmarks and Concluding Remarks

This section first outlines the datasets that serve as standard evaluation grounds for Prompt-based Few-Shot Segmentation (PF3S), highlighting their differences in class partitioning and generalization objectives. We then provide a concise synthesis of current PF3S methods, tracing the shift from explicit support–query conditioning toward prompt-driven inference, and discussing the performance trends observed across benchmarks.

22.2.1 Benchmark Datasets

COCO-20^i is constructed from the MS COCO dataset [111] and comprises 80 semantic categories evenly divided into four folds, each containing 20 classes. This partitioning ensures that, for every fold, the model is evaluated on novel categories that are not observed during training.
PASCAL-5^i extends the PASCAL VOC 2012 dataset [107] with additional annotations from SDS [110]. It comprises 20 semantic categories evenly divided into four folds, each containing five

classes. Following the standard few-shot or open-vocabulary evaluation protocol, 15 classes are used as *seen* categories for training, while the remaining five are held out as *unseen* categories for testing in each fold.

FSS-1000 [186] contains 1,000 object categories, each represented by 10 annotated images. In contrast to PASCAL-5^i and COCO-20^i, which rely on predefined class splits, FSS-1000 evaluates generalization by sampling support–query episodes over disjoint labels at test time, making it a purely class-agnostic benchmark for cross-category segmentation.

22.2.2 Concluding Remarks on PF3S

The progression of PF3S research reflects a continuous rethinking of how support and query information interact, moving from explicit task conditioning toward implicit prompt-based reasoning. Table 6 outlines this evolution across learning paradigms and network typologies, while Table 7 complements it with a quantitative comparison on standard benchmarks.

Overall, the trajectory traced by these methods reveals a clear trend: early architectures like OSLSM rely on strict masking supervision, intermediate models such as PFENet and CMN exploit multi-level feature fusion to strengthen the support–query link, and recent inpainting-based architectures unify both inputs within the prompt itself. Among them, SegGPT stands out for achieving the best results through in-context learning and input-level grid fusion, demonstrating how visual prompting can serve as a general and highly adaptable strategy for segmentation under minimal supervision.

Learning Method	Network Typology	Support Set (prompt component)	Work
Episodic Training	Two-branches Network	Masked Images	OSLSM [177]
		Points annotated Images	FSSPGNet [181]
	Feature Enrichment through Support-Query Fusion	Masked Images	CMN [182]
			PFENet [183]
In-Context Learning & tuning	In-Painting Network	Grid-like Masked Images	VPIP [184]
			Generalist Painter [185]
			SegGPT [41]

Table 6: Summary of Few-Shot Learning Models, grouped by Learning Method, Network Typology, and Support Set, which form a key component of the prompt alongside the query image.

Work	mIoU (1-shot)	mIoU (5-shot)	FB-IoU (1-shot)	FB-IoU (5-shot)
		PASCAL-5^i		
OSLSM ([177])	40.8	43.9	–	–
FSSPGNet (2 annotated points) ([181])	67.2	–	–	–
FSSPGNet (10 annotated points) ([181])	73.2	–	–	–
PFENet ([183])	60.8	61.9	73.3	73.9
CMN ([182])	62.8	63.7	**72.3**	**72.8**
Painter ([185])	64.5	64.6	–	–
SegGPT ([41])	**83.2**	**89.8**	–	–
		COCO-20^i		
PFENet (VGG-16) ([183])	34.1	37.7	60.0	61.6
PFENet (ResNet-50) ([183])	32.4	37.4	58.6	61.9
CMN ([182])	39.3	43.1	**61.7**	**63.3**
Painter ([185])	32.8	56.1	–	–
SegGPT ([41])	**56.1**	**67.9**	–	–
		FSS-1000		
Painter ([185])	61.7	62.3	–	–
SegGPT ([41])	**85.6**	**89.3**	–	–

Table 7: Comparison of PF3S methods across standard benchmarks. Metrics: mean Intersection-over-Union (mIoU, %) and Foreground–Background IoU (FB-IoU, %). PASCAL-5^i and COCO-20^i adopt class-split protocols for episodic evaluation, while FSS-1000 measures cross-class generalization on unseen categories. FSSPGNet is evaluated with two supervision levels (2 and 10 annotated points), and PFENet with two backbones (VGG-16 and ResNet-50). All values correspond to available data reported in the original studies.

23 Foundation Models for Promptable Image Segmentation

Foundation Models, trained on large-scale datasets, have attracted significant attention in both Natural Language Processing and Computer Vision due to their strong generalization capabilities across a wide range of downstream tasks [187]. Among these tasks, image segmentation has emerged as a domain where prompting is increasingly leveraged in innovative ways.

The current body of work can be broadly divided into two major directions. The first focuses on adapting Vision-Language Models (VLMs) for open-vocabulary segmentation, where the goal of textual or visual prompts is to segment novel categories in an open-set manner by leveraging the semantic knowledge captured by text encoders such as CLIP [13]. The second direction involves the development of dedicated foundation models for image segmentation, such as SAM [17], which are designed with promptability as a core feature from the training phase.

Both research directions offer valuable insights into the evolving role of prompts in guiding segmentation and demonstrate the growing importance of prompt-based interaction in image understanding tasks.

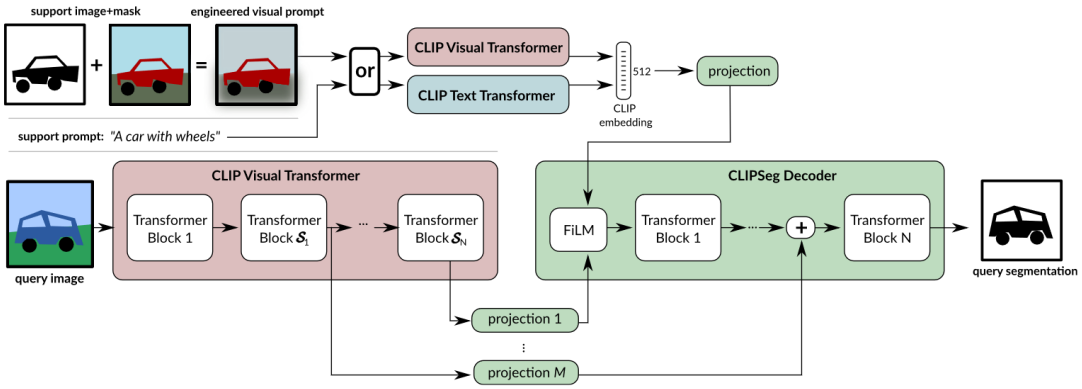


Figure 26: Example of VLM used for Open-Vocabulary Semantic Segmentation. Image from [22].

24 Pre-trained VLMs for Open Vocabulary Segmentation

With large-scale visual and language datasets, Vision Language Models (VLMs) have gained attention [188]. Trained on image-text pairs with weak supervision, VLMs are categorized as Dual Encoder, Fusion Encoder, or Unified Encoder. Dual Encoders like CLIP [13] and ALIGN [14] use separate encoders and contrastive learning for retrieval and zero-shot classification. CLIP learns a shared image-text space via contrastive learning on 400M pairs. The image encoder (ResNet or ViT) embeds visual inputs and uses a [CLS] token for global representation. The text encoder (Transformer) tokenizes input via BPE [125]; the [EOS] token encodes the entire sentence. Training maximizes similarity for aligned pairs, enabling zero-shot vision-language tasks. Fusion Encoders, such as SimVLM [189], jointly process image and text, excelling in tasks like Visual Question Answering. Unified Encoders, including Meta-Transformer [190], use a shared encoder for cross-modal tasks. This survey focuses on Dual Encoders, as VLM zero-shot classification has extended to Image Semantic Segmentation. Unlike CNN-based [191, 192, 193, 194] and Transformer-based segmentation [132, 195, 196], which operate in closed sets, VLMs enable Open Vocabulary Semantic Segmentation [197, 198, 39], broadening category recognition.

In this section, we examine how VLMs like CLIP facilitate Open-Vocabulary Semantic Segmentation by leveraging the broad knowledge embedded in their text encoders. Specifically, these solutions share a common overall structure, including a Text Encoder $\mathcal{E}_{\mathcal{T}}$, an Image Encoder $\mathcal{E}_{\mathcal{I}}$, and optionally a Decoder $\mathcal{D}$, as illustrated in Figure 26. The CLIP pre-trained Text Encoder is used to encode input textual prompts P consisting of class labels [CLASS] and surrounding context words as $[t_1] \dots [t_m]$ [CLASS] $[t_{m+1}] \dots [t_M]$, where M is the number of prompt tokens. The text encoder is kept frozen during training, and its prompt features are used as weights to classify pixel-, patch-, or mask-proposal-level image features. The Image Encoder is usually a CLIP-based or DenseCLIP [199] image encoder modified to output features at different granularities depending on the specific approach.

We focus on two main aspects: *Prompt Design* and *Prompt Integration*. In the first, we discuss hard and soft prompt engineering techniques, and in the second, we examine prompt-image fusion approaches. Both aim to enhance prompt effectiveness to guide the segmentation model in obtaining valid output masks.

An overview table summarizing the examined works can be found in Appendix **??**.

24.1 Prompt Design

Prompts can be classified into two categories: *Hard prompts* and *Soft prompts*, based on whether they are discrete or continuous.

24.1.1 Hard Prompts

Hard prompts consist of fixed templates such as "This is a photo containing a [CLASS]," where [CLASS] represents the segmentation class to be specified at input time. Early approaches use one of CLIP's templates, like $P =$ "a photo of a [CLASS] in the scene", to extract text features from a special token [SEP] appended as the last token of the sequence [200, 201], resulting in $T = \mathcal{E}_{\mathcal{T}}(P)$. However, CLIP templates are not specifically designed for dense prediction tasks and often fail to guide the model effectively. Inspired by *Prompt Ensembling* [13], other works average prompt features extracted from more than 80 templates, resulting in $T = \text{mean}(\mathcal{E}_{\mathcal{I}}(P_i))$ [36, 37, 38, 202, 39, 198]. Compared to single prompt strategies, ensembling has shown superior performance [36, 37].

Further refinements have been made to input prompts. Since only a small number of the many classes provided as input appear in a single image, the remaining classes act as distractors, reducing performance. Thus, authors in [38] introduced prompt denoising to remove classes with confidence scores lower than a certain threshold at pixel, patch, region, or location levels. Target objects can be referred to by multiple terms, such as "person" including terms like "girl," "woman," "boy," and "man." Using a list of synonyms, subcategories, and plurals as [CLASS] labels has proven effective [203]. Additionally, the same authors suggest providing [CLASS] with additional contextual words to account for polysemy, such as "fan" being expanded to "ceiling fan" or "floor fan." Other works argue that prompt ensembling emphasizes the prominent class in multi-label images, suppressing the scores of other target classes [204]. Therefore, they introduced a Sharpness-based Prompt Selection technique inspired by the Coefficient of Variation to select the best prompt using only image-level annotations. It is defined as:

$$\text{sharpness}(P) = \frac{\sum_{i=1}^{n} \text{var}(s_{i1}, \ldots, s_{ik})}{\sum_{i=1}^{n} \text{mean}(s_{i1}, \ldots, s_{ik})} \tag{25}$$

where s_{ij} represents scores for the j-th class after softmax in the i-th image. This metric has been observed to negatively correlate with segmentation performance, allowing the selection of the most effective prompt in advance. Other works utilize prompts in the form of natural language queries, allowing a certain degree of flexibility in the choice of surrounding words $[t]$ for the class label [CLASS] at input time [22]. Techniques such as noun and adjective filtering, context augmentation for polysemy, and synonym assembly have proven effective in preprocessing input prompts in natural language [203].

24.1.2 Soft Prompt

Hard prompt performance heavily depends on the choice of surrounding words $[t]$ and their positioning, leading to fluctuating results [205]. Prompt learning in the continuous space has been proposed as an effective technique to learn domain-specific prompts [206]. Formally, consider a prompt of the form $[t_1] \ldots [t_M]$[CLASS], where each $[t_i]$ represents a learnable token. Rather than projecting these tokens onto discrete vocabulary entries, it has been shown that utilizing their continuous embeddings directly leads to improved performance. Experimental results have shown that this technique outperforms Hard Prompt Ensembling by a large margin, proving more effective in adapting VLM performance from Zero-Shot Classification to Zero-Shot Segmentation [197]. Authors from [207] note that simply allowing the image decoder and text prompts to be the only learnable components, while keeping the VLM image encoder and text encoder frozen, results in sub-optimal outcomes for dense prediction tasks. This phenomenon arises because the learnable text prompts and image embeddings lack proper semantic alignment, given that VLMs are pretrained primarily on global alignment between image-sentence pairs. To favor text alignment with pixel-level features, multiple text learnable prompts are matched with frozen image embeddings through Optimal Transport. Formally, $P = \{p_i \mid i = 1, \ldots, N\}$, where N is the number of text prompts and $p_i = [t_{i,1}] \ldots [t_{i,M}][CLASS]$. This technique enables each of the

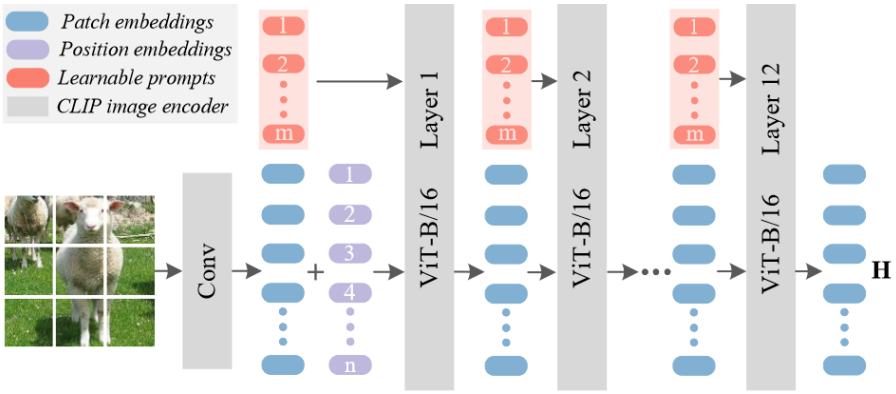


Figure 27: Deep Prompt Tuning (DPT). Image from [37].

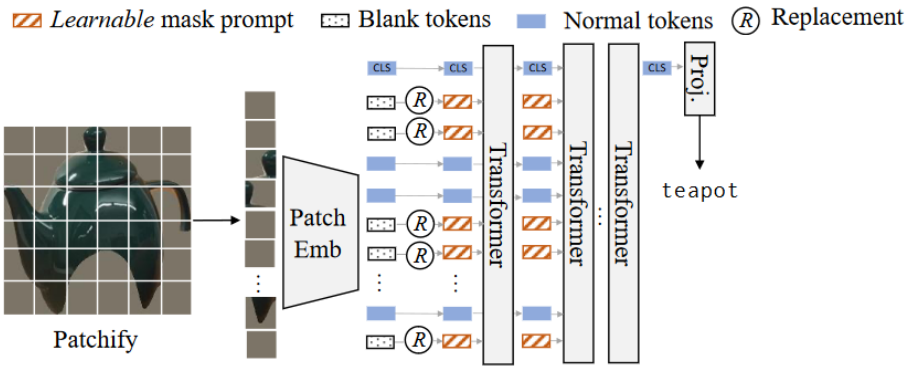


Figure 28: Mask Prompt Tuning (MPT). Image from [198].

multiple text prompts to effectively focus on distinct visual semantic attributes and has proven to be more effective than single prompt learning.

To alleviate the Image Encoder $\mathcal{E}_\mathcal{I}$ from overfitting on training data while making the training more efficient, *Deep Prompt Tuning* (DPT) has been adopted to adapt the Image Encoder for the Semantic Segmentation task [37]. Formally, given the input features $H^l = h_1^l, \ldots, h_N^l$ to the l-th transformer layer $\mathcal{L}^l$, DPT appends learnable tokens $P^l = p_1^l, \ldots, p_M^l$ to the input, resulting in

$$[_, H^l] = \mathcal{L}^l([P^{l-1}, H^{l-1}]) \tag{26}$$

The output features corresponding to the learnable tokens are discarded and will not proceed to the next transformer layer (see Figure 27).

When masked images are given as input, the image encoder can suffer from domain shift caused by zero tokens corresponding to background areas. To mitigate this and help the model learn from masked images, the authors from [198] propose *Mask Prompt Tuning*, a visual prompt tuning technique specifically designed to adapt the image encoder to the segmentation task. Given N_p patch image features $T \in \mathbb{R}^{N_p \times D}$ and the corresponding binary mask $M_p \in \{0,1\}^{N_p}$ assigning 0 to background patches and 1 to foreground ones, a learnable tensor $P \in \mathbb{R}^{N_p \times D}$ is substituted in correspondence with masked tokens as shown in Figure 28.

24.2 Prompt Integration

Integrating the input prompt into the segmentation pipeline requires effectively fusing the textual features with the visual ones. Approaches have evolved from *Late Fusion* to *Early Fusion*, allowing a deeper interaction between input image and prompt features.

Late Fusion generally consists of encoding the input image $I \in \mathbb{R}^{C \times H \times W}$ and prompt P with

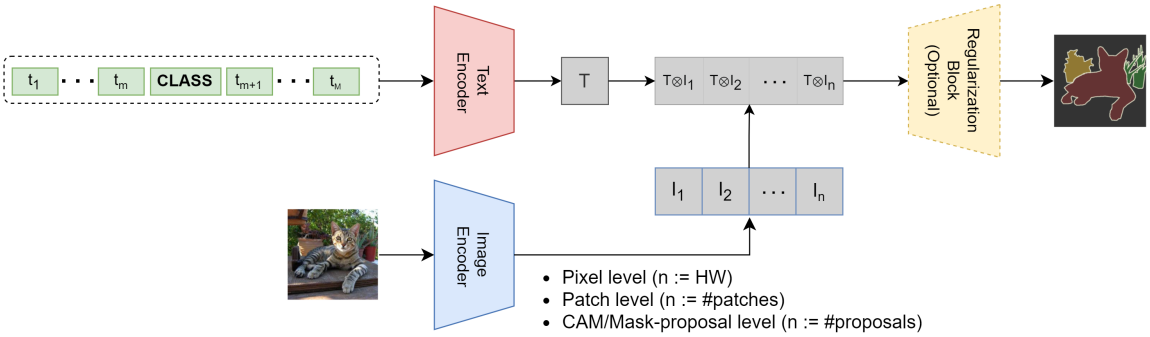
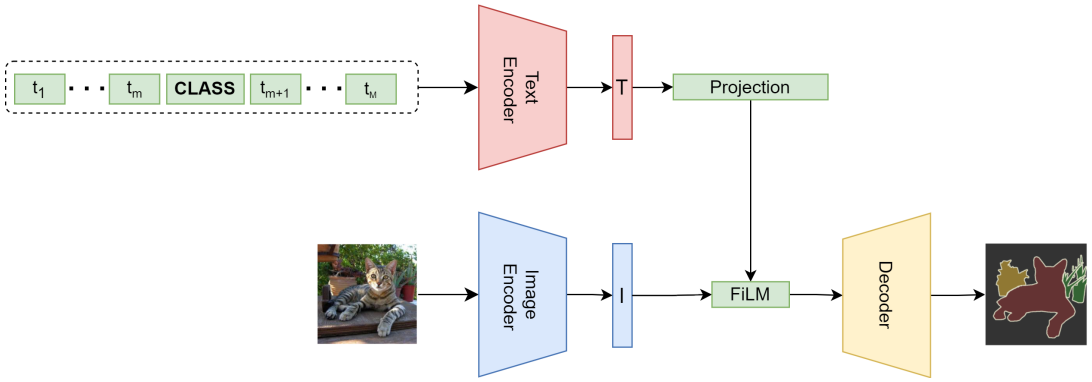
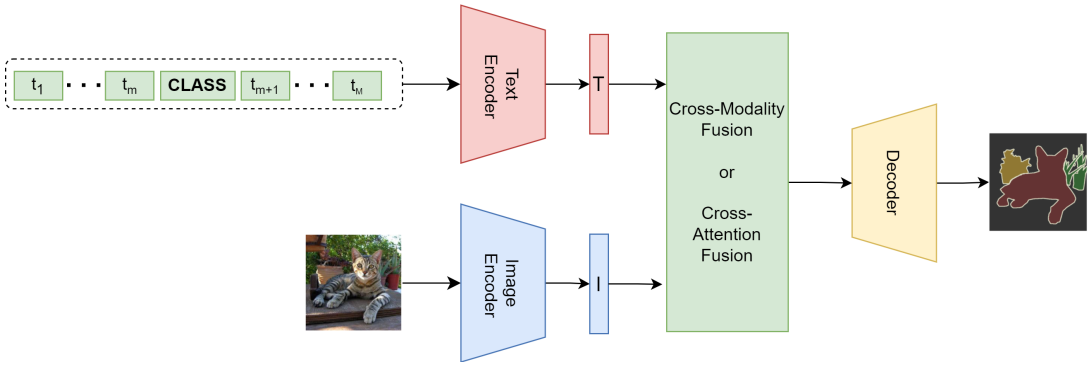


Figure 29: The Late Fusion strategy integrates visual and textual information in Open Vocabulary Segmentation.



(a) Early Fusion using a FiLM module.



(b) Early Fusion using attention modules.

Figure 30: Early Fusion strategies combine visual and textual information in Open Vocabulary Segmentation. (a) Utilizes a FiLM module to condition the input segmentation based on the textual input. (b) Employs either a cross-attention or self-attention module to enhance the information exchange between the input image and text features.

two separate encoding streams, resulting in features $f_I \in \mathbb{R}^{n \times d}$ and $f_T \in \mathbb{R}^d$, respectively. Here, d indicates the feature dimensionality, and n represents the number of features extracted from the input image. Depending on the approach used, this can be at the pixel-level ($n = HW$) [208, 209, 38, 202], patch-level ($n =$ number of patches) exploiting ViT feature generation by patch region [200], mask proposal level ($n =$ number of proposals) if a mask proposal generator like MaskFormer is used on the input image, followed by the image encoder $\mathcal{E}_I$ applied to each proposal [197], or CAM proposal level if such proposals come from class activation maps used in weakly supervised approaches [201]. Prompt features $f_T \in \mathbb{R}^d$ are then used as classification weights with respect to the image features by comparing them with a similarity metric such as

Work	COCO-Stuff			PASCAL-VOC			PASCAL Context		
	S	U	Mean	S	U	Mean	S	U	Mean
ZegFormer [36]	36.6	33.2	34.8	86.4	63.6	73.3	–	–	–
OpenSeg [203]	–	–	–	–	–	72.2	–	–	48.2
ZSSeg [197]	39.3	36.3	37.8	83.5	72.5	77.5	–	–	–
MaskCLIP+ [38]	39.6	54.7	45.0	88.1	86.1	87.4	48.1	66.7	53.3
ZegCLIP [37]	40.2	41.4	40.8	91.9	77.8	84.3	46.0	54.6	49.9
SegCLIP [200]	–	–	26.5	–	–	52.6	–	–	24.7
ReCo+ [202]	–	–	32.6	–	–	27.2	–	–	–
OVSeg [198]	–	–	–	–	–	94.5	–	–	55.7
ZegOT [207]	38.7	57.5	46.2	91.9	90.9	91.4	50.5	72.5	59.5

Table 8: Benchmark comparison of zero-/open-vocabulary segmentation methods across COCO-Stuff, PASCAL-VOC, and PASCAL Context datasets. Values are mIoU (%). Higher is better.

Work	COCO-20^i					PASCAL-5^i				
	20^0	20^1	20^2	20^3	mean	5^0	5^1	5^2	5^3	mean
SPNet [16]	–	–	–	–	–	23.8	17.0	14.1	18.3	18.3
ZS3Net [15]	18.8	20.1	24.8	20.5	21.1	40.8	39.4	39.3	33.6	38.3
LSeg [209]	28.1	27.5	30.0	23.2	27.2	61.3	63.6	43.1	41.0	52.3
Fusioner [39]	31.7	35.7	34.9	31.8	33.5	61.9	70.0	51.2	52.7	59.0
CLIPSeg [22]	–	–	–	–	33.2	–	–	–	–	59.5

Table 9: Comparison of zero-/open-vocabulary segmentation methods on COCO-20i and PASCAL-5i benchmarks. Values are mIoU (%). Higher is better.

cosine similarity. The resulting guidance map is then optionally fed to a Regularization Block that is resized to the input image resolution.

More recent approaches encourage interaction between prompt and image features by fusing the two in the early stages of the network. Authors from [22] propose conditioning the input image features f_I of the segmentation decoder $\mathcal{D}$ with prompt features f_T using *FiLM* conditioning (see Figure 29). FiLM applies a feature-wise affine transformation to image features, resulting in:

$$\text{FiLM}(f_I) = \gamma(f_T) \odot f_I + \beta(f_T), \tag{27}$$

where γ and β are scaling and shifting vectors depending on f_T (see Figure 30a). Other approaches leverage attention mechanisms to encourage interaction between image and prompt features. Considering word-level prompt features $f_T \in \mathbb{R}^{|W| \times d}$, where $|W|$ is the number of prompt tokens, a first approach is to concatenate them with image features f_I and feed them through self-attention transformer layers [39], resulting in:

$$[F_I', F_T'] = \Phi_{\text{FUSE}}([F_I, F_T]). \tag{28}$$

In this way, all image features as well as all prompt ones attend to each other, acquiring information from the other modality (see Figure 30b). Alternative methods [37] employ a cross-attention mechanism, where the prompt serves as the query and the image features act as keys and values. Experimental results show that attention mechanisms allow the input prompt to learn to attend to correlated image locations, thereby enhancing segmentation results compared to late fusion strategies [39].

24.3 Datasets and Metrics

Open-Vocabulary Semantic Segmentation models are commonly evaluated on **PASCAL VOC**, **MSCOCO** (described in Section 11), **PASCAL Context**, and **ADE20K-Full**. To better assess the open-vocabulary capabilities of the proposed models, these datasets are divided into seen classes C_s in the training set and unseen classes C_u in the test set, where the intersection $C_s \cap C_u = \emptyset$. For example, PASCAL VOC is split into 15 seen classes and 5 unseen classes, while MSCOCO has 156 seen classes and 15 unseen.

COCO-Stuff extends the MS COCO [111] dataset by densely labeling "stuff" regions in every image, yielding 171 classes in total (the 80 original COCO "thing" classes plus 91 "stuff" classes). The publicly released annotations cover roughly 118,287 training images and 5K validation images from COCO, all densely annotated. In open-vocabulary segmentation benchmarks, COCO-Stuff's 171 classes are commonly split into 156 seen classes (used for training) and 15 unseen classes (held out at test time).

PASCAL Context [210] extends PASCAL VOC 2010 with additional annotations that provide extra labeling for the entire scene in each image. It includes 60 classes, with 4,996 images for training and 5,104 for testing. The classes are split into 50 seen and 10 unseen categories.

ADE20K-Full [211] contains 25,000 images for training and 2,000 for validation. Different works split the classes in various proportions; for instance, [36] separates them based on their frequency, with seen classes appearing in more than 10 images, while unseen classes appear in fewer than 10. This results in 572 seen classes and 275 unseen.

To evaluate the proposed models' ability to generalize to new classes while preserving knowledge of the training classes, the mean Intersection over Union (mIoU) metric, introduced in Section 12, is typically divided into $mIoU_s$ and $mIoU_u$, which denote the mIoU computed for seen and unseen classes, respectively. As a unified metric, the harmonic mean between the two is sometimes used:

$$hIoU = \frac{2 \times mIoU_s \times mIoU_u}{mIoU_s + mIoU_u} \tag{29}$$

Tables 8 and 9 summarize benchmark results for Open-Vocabulary Segmentation methods. The first table reports performance on the COCO-Stuff, PASCAL VOC, and PASCAL Context datasets, while the second presents results on COCO-20^i and PASCAL-5^i (described in Section 22.2.1), thereby encompassing a comprehensive range of open-vocabulary segmentation benchmarks.

25 Foundation Models in Image Segmentation

Traditional approaches in image segmentation have evolved into Promptable Image Segmentation by incorporating prompt learning strategies, which leverage model promptability during pre-training. This shift, combined with multi-task learning, has led to the emergence of Foundation Models specifically designed for segmentation. Pre-trained on a massive dataset of 1 billion masks from 11 million images, these models function as zero-shot segmenters, generating versatile and detailed representations across various segmentation tasks.

What sets these models apart is their ability to handle multiple levels of granularity, from high-level object categories to fine-grained details, capturing rich semantic information. This flexibility makes them highly effective for tackling complex segmentation challenges.

Leveraging CNNs and ViTs, these models learn to recognize patterns, shapes, and textures across diverse image sets. As a result, they can accurately segment previously unseen images without extensive fine-tuning, achieving high accuracy and strong generalization [212]. This significantly enhances their reliability and adaptability across different applications.

Although several surveys have reviewed foundation models in the context of image segmentation [27, 212, 24, 213], they often treat prompting as a secondary feature. In contrast, our work centers on prompting as the primary mechanism for segmentation. A pivotal development in this space is the Segment Anything (SA) project [17], which exemplifies the shift toward prompt-driven segmentation by enabling flexible visual prompting at scale. To contextualize the models and techniques discussed in later sections, we briefly review SA's core components—such as its promptable architecture and generalization capabilities—as a foundation for understanding how subsequent methods build upon or adapt its framework.

25.1 Segment Anything

SA project introduces a novel task, model, and dataset for general image segmentation, aiming to create a foundation model for segmentation by being capable of generating high-quality segmentation masks for any object in an image or video, based on various types of input prompts. SA Model (SAM) combines three components, as illustrated in Fig. 31: an image encoder, a flexible Transformer-based prompt encoder, and a fast mask decoder that forms the versatility found in NLP foundation models along with zero-shot generalization.

25.1.1 SA Task

The promptable segmentation task is designed to handle ambiguous prompts by generating valid masks for any input, mirroring language models' ability to respond to vague prompts. This method involves training a foundation model for segmentation that can respond to a variety of prompts, from foreground/background points to rough boxes, masks, or even free-form text, indicating what needs to be segmented in an image. This approach facilitates a natural pre-training algorithm where the model is trained against a series of simulated prompts and their corresponding ground truth masks. This is important for automatic annotation applications because it allows the model to adapt to a variety of segmentation tasks through zero-shot transfer, which is the process of applying the model to a wide range of segmentation tasks by just giving the right prompts. This eliminates the need for retraining or specific tuning of the model for every new task or domain.

25.1.2 SA-1B Dataset and Data Engine

The SA-1B dataset, developed using the SA Data Engine, addresses the scarcity of high-quality segmentation masks by generating over 1.1 billion masks from 11 million images. Its annotation process progresses in three stages: manual mask creation using interactive tools, semi-automated refinement through annotator feedback, and fully automated generation using a grid-based approach to produce around 100 masks per image, enhancing scalability and diversity along with significantly streamlining the production. The SA-1B prioritizes high-quality, general-purpose annotations, covering a wide range of objects, domains, and scenarios. This emphasis is crucial for enhancing the complex pattern and representation learning of more targets across all categories, particularly from underrepresented regions.

25.1.3 SA Model

As mentioned in Section 24, a significant development in VL models is the adoption of the language head as their main source of supervision. By leveraging the learned relationships between textual descriptions and visual content, CLIP can make accurate predictions or find relevant matches without the need for extensive training on specific datasets. While SAM shares the foundational concept of aligning visual and textual representations with models like CLIP, it diverges in its specialized approach to image encoding and multi-modal integration.

The **Image Encoder** facilitates a Masked Autoencoder (MAE) pre-trained ViT [214]. This encoder handles images as sequences of fixed-size patches. These patches are linearly embedded and supplemented with positional encodings to retain spatial information. Processing of these embeddings is done through multiple transformer layers, each comprising multi-head self-attention mechanisms and feed-forward networks. This configuration allows SAM to capture long-range dependencies within the image, resulting in a high-dimensional feature map that is crucial for precise image segmentation [215].

Additionally, SAM's text encoder, similar to CLIP's, integrates text prompts specifically for segmentation tasks. The text encoder tokenizes the input text, converts it into embeddings, and processes these embeddings through multiple transformer layers. Positional encodings are

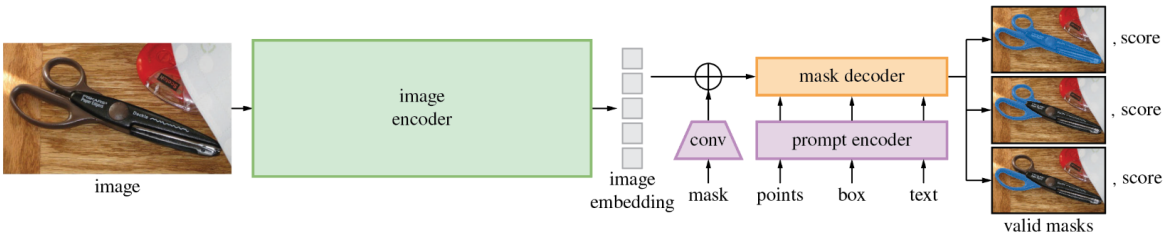


Figure 31: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at real-time speed. [17]

added to maintain the order of tokens, and the resulting feature vector encapsulates the semantic meaning of the text prompt. Next, SAM introduces a unique component called the **Prompt encoder**, which integrates multi-modal inputs, including image features, text features, and additional user-provided prompts like points, boxes**(sparse prompt)** or masks**(dense prompt)**. This prompt encoder uses a cross-attention mechanism to align and synthesize these features, enabling the model to focus on relevant parts of the image as guided by the prompts. The integrated representation is then processed through the segmentation head and lightweight mask decoder to produce the final segmentation mask as shown in Figure 31.

Finally, the **Mask Decoder** structure is an optimized design that uses a modified Transformer decoder block and a dynamic mask prediction head. The procedure involves prompt self-attention and cross-attention between the prompt and the key/value weight image embeddings, which allows effective updating of all embeddings. This encoder does not only directly map the image embedding to the prompt embedding to produce the segmentation masks. However, the connection between classification and mask head includes novel components like dynamic linear classifiers that return probabilities of mask foregrounds at every image location. It is also possible to support ambiguous prompts by producing multiple mask outputs and selecting the one with the highest confidence while training via a combination of focal and dice loss [216]. Real-time operation is evident in the fact that the model runs in a web browser very efficiently, usually taking around 50 milliseconds, which confirms the suitability for interactive applications.

As noted, a wide range of novel methods have emerged that aim to improve image segmentation through innovative prompting strategies, many of which build upon components introduced by the SA project. In this section, we examine recent **SAM-based approaches** by analyzing the prompting techniques they introduce and the segmentation challenges they address. Our goal is to highlight how these methods extend the capabilities of promptable segmentation and their most successful applications.

25.2 Prompt Techniques

As explained in Section 24, the main prompt techniques used in this scope are Hard-Text, Free-Text, and Visual prompt. However, the distinct needs and capabilities of the foundation models compared to VLMs lead to some differences in the usage of these techniques, which are presented in the following sections.

Hard-Text Prompts, includes **Defined Prompts** and **Template-Based Prompts**. Defined Prompts are used in models such as SEEM [40], which leverage pre-defined textual cues to direct specific tasks, like segmenting all water bodies in a landscape image, that manifests as Queries and Prompt Interaction during training (See Figure 33). This creates a direct mapping between the specific cue and the label that ensures the model focuses on relevant elements without deviation. Conversely, Template-Based Prompts provide a structured yet flexible approach. In this method, employed by models like RSPrompter [217], a fixed syntactic template is used where certain keywords can be swapped with other entities e.g., here the [object] keyword in the

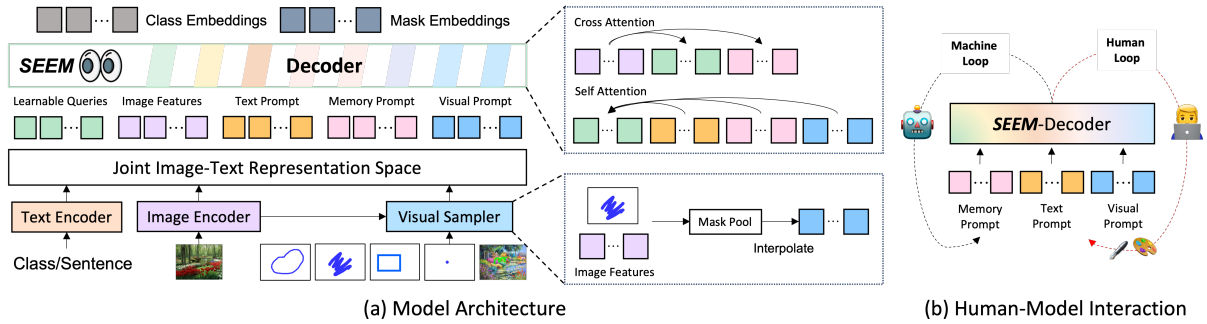


Figure 32: SEEM leverages predefined prompts to guide tasks via text queries and prompt interactions during training. It additionally supports user-provided inputs such as clicks or scribbles for segmentation and generates semantic labels for segmented masks, capabilities that distinguish it from SAM. [40]

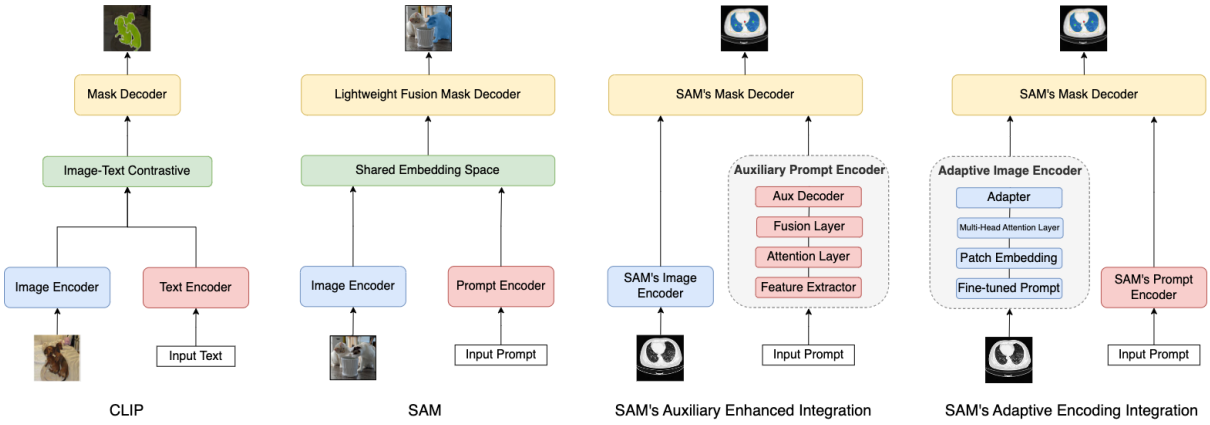


Figure 33: Overview of prompt-based VLM architectures and their Corresponding Encoding Methods. Left to right: **CLIP**'s Dual-encoder designs use a parallel visual and language encoder with aligned representations. SAM process both image and text representations via a decoder, where the image encoder can process visual prompts such as points and boxes. **SAM Auxiliary-Enhanced Integration** introduces an auxiliary encoder comprising a feature extractor, attention mechanism, fusion layer, and auxiliary decoder. The feature extractor processes auxiliary data into a feature map, which is refined by the attention mechanism. The fusion layer combines auxiliary and primary features, which are then refined by the auxiliary decoder to enhance segmentation accuracy, particularly in specialized or out-of-distribution tasks. **SAM Adaptive Encoding Integration** features an adaptive architecture that processes input images at different levels of granularity. This architecture adapts to diverse object forms and complexities, improving segmentation performance across varied domains, especially in scenarios requiring high precision.

"Segment [object] in this image" can be replaced with "trees" or "buildings, allowing the model to adapt to different segmentation tasks while maintaining an underlying structure.

Free-Text or Natural Language Prompts, includes Adaptable and Conversational approach. Models like All-in-SAM [218], and SegGPT [41] can handle complex requests, such as recognizing areas in a city with key tree coverage. In a similar vein, Conversational Prompts in models like as Say-Anything [219] provide continuous conversations with users to improve comprehension and segmentation tasks. These models' precision and versatility are increased by their ability to modify their output in response to user feedback because of their interactivity.

Visual Prompts introduce a different dimension to model interactions by incorporating

visual cues or sample images to guide segmentation tasks. Models such as Grounded-SAM [220], for example, segment sick crop leaves inference from an example image by using **Image-Based Cues**. Additionally, **Visual-Textual Prompts** also incorporate text and images to improve contextual relevance and target concentration, for example, by helping users identify particular objects in complicated settings like when a model could use a picture of a crowded street with a prompt: "Identify all red cars" to focus on specific elements. This approach is evident in models like CLIP-Surgery, Caption-Anything, Label-Anything, and Semantic-SAM [221, 222, 223, 224] which include semantic reasoning as well.

25.3 Innovative Integration of SAM Components

As mentioned above, SAM is composed of several modular components, each contributing to its versatile segmentation capabilities. These components are being integrated into various models through innovative approaches based on different prompt types and specialized domains, enhancing the performance of segmentation for different applications.

25.3.1 Direct Integration

In these models, SAM functionality is directly incorporated within the core framework of segmentation networks, increasing their capabilities to handle a broader spectrum of segmentation tasks efficiently. Notably, the integration of SAM's encoders into models like Mask R-CNN, as discussed by Kirillov et al. [17], enables these models to dynamically segment a wide variety of objects, showcasing SAM's versatility in handling previously unseen object classes and complex segmentation scenarios. Models like SAMed, Text2Seg, and Grounded-SAM [225, 226, 220] are highlights of this direct integration. This integration is particularly effective in scene understanding, as demonstrated by Osco et al. [227], who incorporated SAM's components into aerial image analysis pipelines. Through direct integration of SAM's encoder into segmentation models, their approach enabled precise delineation and classification of urban and natural features in satellite imagery, supporting informed decision-making in urban planning and infrastructure development.

25.3.2 Auxiliary-Enhanced Prompt Encoder Integration

SAM's performance is limited when applied to Out-of-Distribution (OOD) domains [106]. Auxiliary-enhanced models use the main model as a supplementary mechanism to refine the outputs from primary segmentation networks making the final segmentation results more accurate or detailed. The process is done by utilizing an auxiliary encoder in charge of generating surrogate prompts directly from the input images, thereby eliminating the need for manual prompt generation and enabling automatic segmentation

In the integration of SAM, Auxiliary Prompt Encoders (APEs) are often used to enhance segmentation performance in specialized tasks, especially when dealing with rare object classes or complex scenes not well-represented in typical training datasets [224]. These encoders introduce additional semantic context to guide segmentation. APEs generally include a feature extractor, an attention module, a fusion layer, and an auxiliary decoder. The feature extractor, built with transformer layers, processes auxiliary data such as textual descriptions, scene attributes, or spatial cues. It converts these inputs into a feature map that captures task-relevant information. The attention mechanism, typically based on cross-attention, then refines this feature map by highlighting the most important elements and suppressing irrelevant noise, helping the model better focus on critical context during segmentation.

Integrating the auxiliary features with the primary prompt features from SAM is facilitated by the fusion layer. This layer can employ various techniques such as concatenation, addition, or more sophisticated methods like bilinear pooling to combine the features effectively. The resultant

fused feature map then feeds into the auxiliary decoder, which plays a critical role in refining the segmentation output. The auxiliary decoder includes additional layers for upsampling and fine-tuning, ensuring that the combined features yield a precise and well-defined segmentation mask. This meticulous integration process allows the model to leverage auxiliary data efficiently, significantly enhancing its segmentation performance. This architecture, first proposed by AutoSAM [228], leverages the strengths of the pre-trained SAM while adapting it to the domain of medical imaging.

Other significant examples of medical domain adaptation involve the use of auxiliary prompt encoders. For instance, MaskSAM [229] introduces a prompt generator integrated with SAM's image encoder to produce auxiliary classifier tokens, binary masks, and bounding boxes, enabling the model to operate without explicit prompts. Building on this, research efforts such as [225, 230, 231] demonstrate how such integrations significantly enhance segmentation accuracy in medical imagery, improving the detection of subtle abnormalities like malignant cells in MRI and CT scans, and supporting both diagnosis and treatment planning. These adaptable architectures are also being extended beyond medical applications into domains like precision manufacturing, where they assist in industrial quality control by accurately classifying components and detecting defects [232, 233, 234].

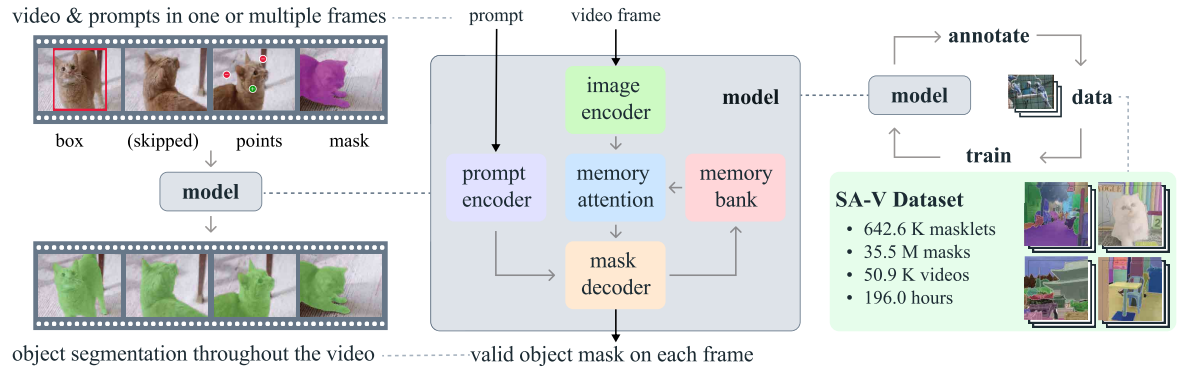
25.3.3 Adaptive Image Encoder Integration

Adaptive encoding in segmentation models refers to mechanisms that adaptively process input images to produce segmentation masks at different levels of granularity and detailed shapes. Particularly for integrating models such as SAM, these techniques are essential for managing the intrinsic variation in item forms, sizes, and complexity present in various datasets, as shown by AdaptiveSAM, SAMUS, WebSAM-Adapter [235, 236, 237]. To elaborate, as illustrated in Figure 35c, the use of pre-defined prompt methods is a significant advancement in adaptive encoding since it allows a model to produce segmentation masks at different levels of detail from a single input.

The Semantic-SAM model [224], for instance, presents a multi-choice learning scheme where each click point on an image can generate masks at different levels, which correspond to several ground-truth masks. It utilizes the shared vision-text encoder to encode both objects and parts separately, allowing for distinct segmentation processes while adapting the loss function according to the type of input data. This strategy enables the model to handle different levels of semantic annotations, from object-level to part-level, which helps in producing detailed and accurate segmentation masks without relying on an auxiliary prompt encoder. This enhancement typically functions to address specialized segmentation tasks of uncommon objects that are not well-represented in typical training datasets by integrating more context or understanding of the scene's semantics. Applications needing in-depth analysis of image regions at multiple scales can benefit from this feature [238]. Another novel development toward adaptive encoding is SegPrompt's [239] method to enhance open-world segmentation by leveraging category-level prompts during training. These prompts serve as secondary supervision, helping the model maintain its generalization capabilities while also utilizing category information to improve the segmentation performance of both seen and unseen categories. This approach of adaptive encoding helps to optimize the model's performance across varied domains, particularly in medical and surgical scenes where precision is crucial.

25.4 Segment Anything Model 2

SAM 2 [240] builds upon the original SAM architecture with key innovations aimed at improving efficiency and adaptability. One major addition is a streaming memory mechanism, which maintains contextual information across multiple prompts by storing intermediate representations in a structured memory bank and retrieving them through attention-based queries. This design



(a) Task: promptable visual segmentation **(b) Model:** Segment Anything Model 2 **(c) Data:** data engine and dataset

Figure 34: We introduce the Segment Anything Model 2 (SAM 2), towards solving the promptable visual segmentation task (a) with our foundation model (b), trained on our large-scale SA-V dataset collected through our data engine (c). SAM 2 is capable of interactively segmenting regions through prompts (clicks, boxes, or masks) on one or multiple video frames by utilizing a streaming memory that stores previous prompts and predictions. [240]

allows SAM 2 to refine segmentation results without recomputing from scratch, significantly reducing latency.

Another critical advancement is the integration of Hiera, a hierarchical Vision Transformer that processes images at multiple scales [241]. Unlike standard ViTs, which operate at a fixed resolution, Hiera progressively reduces spatial resolution while increasing channel depth across stages. This hierarchical design captures both fine-grained details and global context efficiently, without relying on complex components like shifted windows. Combined with Masked Autoencoder pretraining, Hiera achieves state-of-the-art accuracy with lower computational cost, making SAM 2 faster and more robust for diverse segmentation tasks.

SAM 2 achieves 6× faster inference and higher accuracy than SAM for images, while reducing user interactions by 3× in video segmentation tasks. These advancements position SAM 2 as a versatile foundation model for interactive and automated segmentation across dynamic visual environments. Additionally, SAM 2 introduces a new large-scale dataset that includes both images and videos, further enhancing its generalization capabilities, though its architecture remains optimized for image segmentation.

25.5 Performance Analysis

25.5.1 Medical Domain Evaluation

When evaluating SAM-based models in the medical domain, researchers establish direct performance comparisons with vanilla SAM to quantify the effectiveness of domain-specific adaptations.As shown in table 11, On the BTCV benchmark [244], vanilla SAM achieves approximately 54.8% Dice Score (ViT-B), demonstrating a significant domain gap between its SA-1B natural image training data and medical imaging characteristics. Specialized medical adaptations show substantial improvements: MedSAM achieves 84.6% Dice (+46.0% improvement) through extensive fine-tuning on 1.57M medical images across diverse modalities, SAM-Med2D reaches 84.7% (+54.6%) via 2D medical-specific optimization, SAM-Med3D achieves 84.68% through efficient 3D volumetric processing with minimal prompts, and AutoSAM establishes state-of-the-art performance at 87.15% Dice on BTCV and 88.65% on AMOS CT benchmark [245] (+59.0% and +79.8% improvements over SAM, respectively). For specialized pathology tasks, All-in-SAM achieves 82.54% Dice on the MoNuSeg benchmark [246], outperforming vanilla SAM through molecular-oriented corrective learning and weak label tolerance. Notably, SAM 2's video segmentation capabilities achieve 71.99% on BTCV, demonstrating improved 3D

Prompt Integration Mechanism	Prompt Type	Architecture Notes	Works
Direct Integration	Text, Points, Boxes	SAM architecture with domain-specific optimizations	[242]
		SAM architecture refined for enhanced ground-truth integration	[220]
	Text Queries	SAM integrated with language processing units for better query handling	[226]
Auxiliary-Enhanced	Click-Based queries	SAM architecture utilizing a query-based mask decoder	[230], [231]
		Zero-shot segmentation enhanced through SAM optimization	[243]
	Text, Points, Boxes, Mask	Domain-specific fine-tuning of SAM for specialized tasks	[225]
Adaptive Encoding	Text, Points, Boxes	SAM with adaptive vision decoder for varied processing tasks	[235], [237], [236]
		Task-specific adaptations within the SAM framework	[218], [239]
	Text, Points, Boxes, Scribbles, Clicks	SAM coupled with an assistive text encoder for versatile use	[40]
	Click-Based queries	Query-based mask decoder integrated with SAM for responsive feedback	[224]

Table 10: Summary of SAM integration mechanisms categorized by integration type, architecture, and works.

propagation through temporal memory mechanisms, yet remains inferior to specialized medical adaptations, showing that general capability expansion does not necessarily translate to medical domain effectiveness. This evaluation methodology follows a standardized four-step process: (1) establishing baseline performance using vanilla SAM with identical prompt strategies, (2) implementing domain-specific adaptations through fine-tuning or architectural modifications, (3) evaluating both models on identical medical datasets using domain-appropriate metrics (Dice Score, Hausdorff Distance), and (4) quantifying improvements as percentage gains over baseline. This demonstrates that medical domain specialization through fine-tuning, 2D medical optimization, 3D architectural redesign, or task-specific adaptation is essential for clinical deployment, with typical improvements ranging from 40–80% over vanilla SAM depending on task complexity and domain mismatch degree.

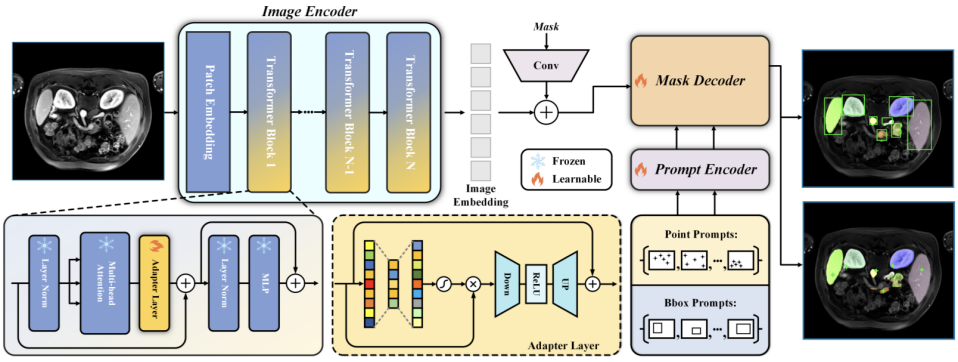
25.5.2 General Domain Evaluation

In contrast to medical domain evaluation, SAM-based models in general domains are assessed for functional capability enhancement rather than domain transfer effectiveness. SAM was designed for interactive, class-agnostic instance segmentation producing binary masks from spatial prompts, making direct performance comparisons on semantic segmentation benchmarks (COCO-Stuff, Pascal-VOC, ADE20K) fundamentally inapplicable to vanilla SAM. Grounded-SAM combines Grounding DINO's open-set object detection with SAM's segmentation capabilities, achieving 48.7 mean AP on the SegInW benchmark for zero-shot open-vocabulary segmentation. SEEM achieves 1.51 NoC@85 (Number of Clicks at 85% IoU) compared to SAM's 1.82–2.47 across

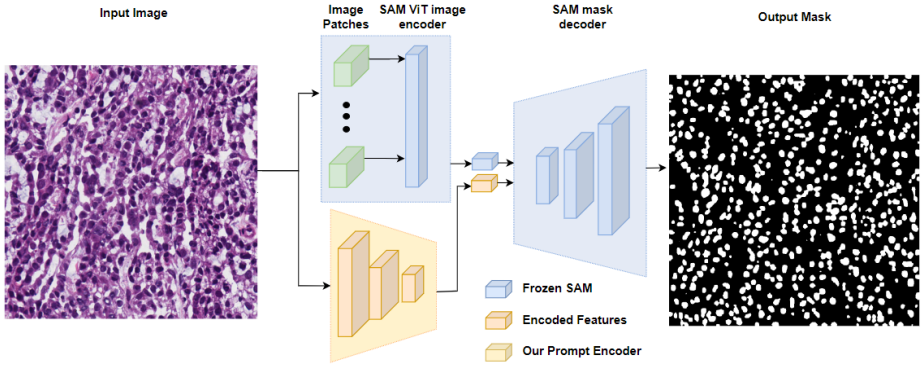
Work	BTCV	AMOS CT	MoNuSeg
SAM [17]	54.80	49.30	71.00
SAM 2 [240]	71.99	–	–
MedSAM [242]	84.60	83.10	–
SAM-Med2D [225]	84.70	–	–
SAM-Med3D [247]	84.68	–	–
All-in-SAM [218]	–	–	82.54
AutoSAM [228]	87.15	88.65	–
AdaptiveSAM [235]	–	–	–

Table 11: Comparison of SAM and SAM-based models on medical image segmentation benchmarks using Dice Score (%). "–" indicates the model was not evaluated on that benchmark. Best performance highlighted: AutoSAM achieves state-of-the-art on BTCV (87.15%) and AMOS CT (88.65%).

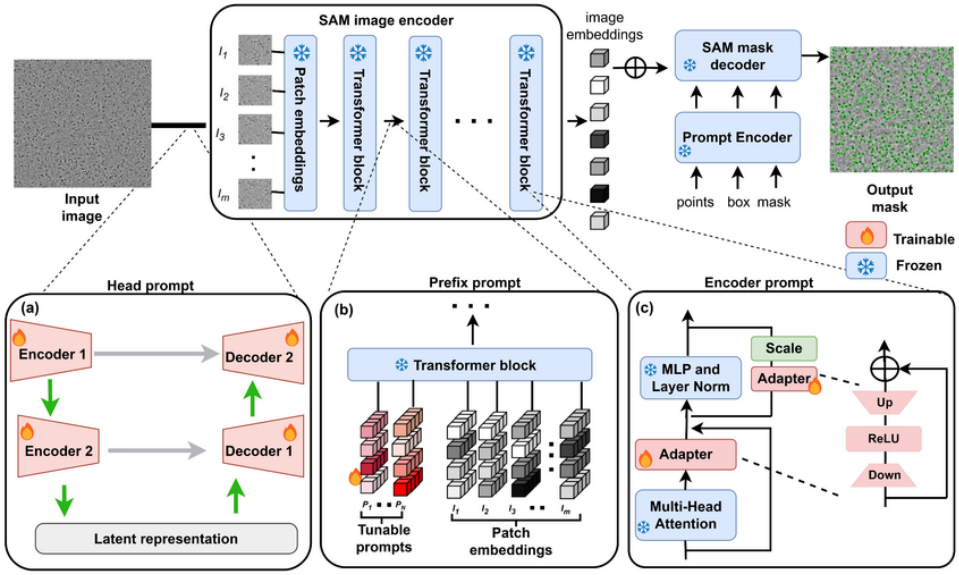
different model sizes, representing 17–39% fewer user interactions required, while demonstrating superior multi-prompt robustness with consistent 76.5–84.6% performance across diverse input types (points, strokes, scribbles, polygons, boxes, text), whereas vanilla SAM exhibits high inconsistency ranging from 22.8% to 65.0%. On generic segmentation where SAM cannot be evaluated, SEEM achieves 57.5% PQ (Panoptic Quality), 47.7% mAP (instance), and 67.6% mIoU (semantic) on COCO, establishing a universal segmentation interface. Semantic-SAM addresses multi-granularity segmentation by generating six meaningful semantic hierarchy levels from a single point click with 89.0% 1-IoU@All Granularity, compared to SAM's three-level output requiring multiple interactions. SAM-CLIP [234] achieves state-of-the-art zero-shot semantic segmentation through model weight interpolation: 60.6% mIoU on Pascal-VOC (+6.8% over previous best SegCLIP at 52.6%), 31.5% mIoU on COCO-Stuff (+5.9% improvement), and 29.2% mIoU on Pascal Context. When comparing SAM 2 to SAM on general natural images, SAM 2 shows mixed results with 2–3% improvements on certain medical imaging modalities (dermoscopy, light microscopy) through enhanced encoder and larger training data, but improvements are incremental rather than transformative. Critically, general domain evaluation metrics (NoC, 1-IoU@All, PQ) differ fundamentally from medical domain metrics, reflecting distinct research objectives: medical models are assessed for domain adaptation effectiveness in bridging natural-to-medical image gaps, while general domain models are assessed for functional capability enhancement, extending SAM's instance-centric design toward semantic understanding, multi-prompt flexibility, and hierarchical granularity. This paradigmatic difference explains why SAM-based model evaluation requires contextual interpretation within each model's intended application scope.



(a) Pipeline of SAM-Med2D showing direct integration of SAM functionality within the segmentation network. The image encoder is frozen, and learnable adapter layers in each Transformer block are used to acquire domain-specific knowledge in medical imaging. The prompt encoder is fine-tuned using point, Bbox, and mask information, while the mask decoder parameters are updated through interactive training.



(b) AutoSAM's framework integrates SAM as an auxiliary mechanism while the image encoder and mask decoder are frozen. This segmentation model processes various types of segmentation data, including generic, part, and class-agnostic segmentation, thus improving segmentation accuracy and detail.



(c) Overview of adaptive encoding strategies in SAM-based models. (a) Head Prompt: A trainable U-Net enhances input micrographs for SAM adaptation, with only the U-Net being refined during training. (b) Prefix Prompt: Tunable prompts prepended to image embeddings before entering transformer blocks in the image encoder, with only these prompts updated during training. (c) Encoder Prompt: Adjustable autoencoder-like adapters integrated into transformer blocks, with only the adapter parameters being updated during training, allowing SAM parameters to remain unchanged. These methods enable the model to manage variations in object shapes, sizes, and complexities, optimizing segmentation performance across diverse domains.

Figure 35: Illustration of SAM-based segmentation strategies.

26 Discussion and Future Directions

Prompt-based segmentation has introduced a new emerging approach in computer vision, offering flexibility and adaptability across various domains. Across all segmentation paradigms, the use of prompts, whether visual, textual, or both, marks a crucial development in making segmentation models more dynamic, adaptable, and context-aware. It also improves image segmentation use in various sectors, including medical imaging, autonomous driving, environmental monitoring, and industrial automation. Each method offers solutions to the key challenges posed by traditional segmentation techniques: reducing dependence on large labeled datasets, improving generalization, and incorporating user feedback into the segmentation process. Nevertheless, several challenges persist. In this section, we offer an overview and discussion on key findings of the research work, helping to direct the reader toward the most relevant areas for further exploration. We also provide a comprehensive look at the challenges faced by these approaches and highlight research directions that could help address these challenges.

27 Visual Interaction and Compositionality

Section 5 reviewed encoding and fusion methods for visual prompts in Promptable Segmentation, focusing on Click prompts in Interactive Segmentation. Various approaches to encoding visual prompts were discussed, with interaction maps fused with input images through Early or Late Fusion. Early Fusion integrates interaction maps with images at the channel level, while Late Fusion enhances information flow by merging image and interaction maps at the feature level in later network stages, improving segmentation performance.

Binary Disk Transform with Late Fusion is among the most efficient methods for interactive segmentation, offering notable computational benefits compared to complex encodings. However, this efficiency introduces a trade-off, as it compromises the ability to process rich contextual and scale-aware information, ultimately affecting segmentation accuracy. In contrast, advanced encoding strategies such as **Scale-Aware Guidance Maps** and the **Geodesic Transform** are capable of effectively capturing both object scale and contextual cues, substantially outperforming simpler Euclidean-Gaussian encodings. These methods, however, are mostly limited to Early Fusion, restricting compatibility with Late Fusion approaches that excel at feature-level integration and reuse of pre-computed features [248].

We introduce the concept of **Compositionality** in visual prompting, which seeks to integrate multiple spatial prompts—such as clicks, bounding boxes, and scribbles—to improve segmentation performance beyond single-prompt approaches. Recent techniques like the **Prompt Unified Encoder** encode these heterogeneous inputs into a unified representation. Furthermore, late fusion strategies, including **Dual-Cross Merging Attention**, have proven effective at combining multimodal prompts with image features.

These developments in Interactive Segmentation are beginning to influence foundation models [248], highlighting Compositionality as a promising avenue for future promptable segmentation systems. Nevertheless, current models like SAM lack compositional capabilities, restricting their ability to jointly leverage visual and textual cues. By employing multimodal fusion techniques, future models could achieve more adaptive and precise segmentation by dynamically integrating spatial information across a wide range of tasks.

28 Prompt understanding, visual domain shift and data augmentation in Referring Expression Segmentation

Referring Expression Segmentation (RES) entails a greater challenge than Interactive Segmentation. While interactive prompts can be encoded directly as images (i.e., visual guidance maps)

and can thus directly interact with query images and inform their segmentation process, this is not true for referring expressions. In fact, they require sophisticated encoding strategies and cross-modal fusion operations in order to allow natural language to guide the segmentation process. Recent advancements demonstrate that early fusion cross-modal strategies are more effective than late fusion ones, as early-on interaction during the encoding process allow models to extract more nuanced correlations between individual words in the expression and pixel-level features in the image, thus enhancing the precision of segmentation outputs. Moreover, the same results highlight that attention-based approaches in cross-modal fusion are more effective than other methods (e.g., concatenation-convolution). Indeed, the attention mechanism is effective in capturing long-range semantic dependencies within words and linking them to relevant visual regions in the query image.

Another unique challenge of RES lies in the variability and ambiguity of natural language prompts. Human expressions can be arbitrarily vague, context-dependent, or semantically complex, making it difficult for models to unambiguously interpret textual prompts and mapping them to relevant visual regions. This challenge is addressed mainly by training models on datasets which reflect the heterogeneity of possible referring expressions, such as the ones including multi-target and no-target text prompts, large word vocabulary, diverse referred categories, varying degree of implicitness and ambiguity.

Based on these observations as well as on the recent RES works analyzed in this survey, promising research directions emerge. To tackle the challenge of prompt understanding, the use of large pretrained vision-language models (LVLMs) is worth exploring. Models like CLIP [13], which are trained on massive vision-language datasets, have shown remarkable generalization capabilities over a variety of vision-language tasks and could be transfered to the RES task faster and more effectively than training a model from scratch on a RES dataset. Additionally, the design of novel architectures for multimodal fusion remains a wide area for exploration. One example are graph-based neural networks, which are already widely applied to the similar task of Referring Expression Comprehension [249] but still underexplored in RES.

Aside from improvements over RES model architectures, the creation of new RES datasets is another promising research direction for training more robust models. A limitation of currently available RES datasets (sec. 19) is that they include only natural images (i.e., images typically taken with a handheld camera). Visual domain shift is a well-known challenge for segmentation models [250, 251] which often requires the creation of new datasets for specific target domains. A few recent works extend RES to the field of aerial images [252, 253] and medical images [254] by proposing new domain-specific datasets, though the availability of non-natural image RES datasets is still scarce. Therefore, this is an area of open research for RES.

Moreover, as creating curated large-scale datasets remains a resource-intensive task, researchers may resort to data augmentation strategies to virtually enhance the size and diversity of available data. However, a recent study [255] showed that conventional image augmentation techniques are often incompatible with RES, since they could alter object spatial relationships (e.g., horizontal flip) or colors (e.g., color jitter), and would thus require a semantically consistent modification of the referring expression too. Therefore, an open issue in RES is developing RES-compatible augmentations for images, text, or both. Few attempts at this have been published, such as joint image and text masking [255] and negative-mined mosaic generation [256], though this research direction remains underexplored.

As a final note, an emerging and intriguing research direction concerns the idea of prompt compositionality within the realm of RES. Recent methods, including Attribute Attention [112] and Sparse Fusion [163], have begun to investigate joint visual-textual prompting. However, the study of visual-textual compositionality remains in its early stages, and there is still limited understanding of how to effectively integrate semantic information into large-scale models such as SAM [248].

29 Scaling Few-Shot Semantic Segmentation (F3S) with Promptable F3S

Few-Shot Semantic Segmentation (F3S) enables models to generalize to unseen classes with minimal supervision, reducing reliance on retraining. Traditional fine-tuning adapts pre-trained models by updating parameters, but this is costly and limits scalability. In contrast, Promptable F3S (PF3S) uses structured prompts to guide models at inference, enabling task adaptation without parameter updates.

This review has highlighted how episodic training and in-context learning represent two strategies for avoiding costly retraining. Episodic training constructs task representations from support sets, enhancing generalization through explicit encoding and fusion strategies that condition predictions on task-specific cues. In-context learning, on the other hand, embeds the entire prompt directly into the model's input, enabling seamless adaptation across tasks within a single forward pass. The latter approach pushes toward highly generalist segmentation models capable of handling diverse scenarios without fine-tuning.

Despite these advantages, PF3S methods also face important limitations. A central challenge is their strong dependence on prompt quality and design: variations in prompt construction, annotation sparsity, or user guidance can lead to inconsistent results, making robust validation more complex. Automating prompt selection or generation, such as choosing optimal support examples based on query similarity that represents a promising research direction to mitigate this sensitivity. Domain shift between support and query images further threatens generalization, though in-context tuning techniques offer partial mitigation by allowing lightweight adaptation at test time. Moreover, the size and complexity of prompts introduce practical constraints, especially for in-context learning approaches that concatenate many examples into grid-like inputs and risk exceeding model context limits. Finally, PF3S typically incurs higher computational costs at inference time compared to fine-tuned models, due to the need to encode, fuse, and condition on support information dynamically.

Overall, while PF3S offers a flexible and efficient alternative to traditional fine-tuning, its scalability will depend on continued advances in prompt design, context-aware learning mechanisms, and the integration of lightweight fine-tuning strategies—such as in-context tuning—that complement prompt-based conditioning to balance adaptability, accuracy, and computational efficiency in real-world segmentation tasks.

30 Adapting VLPs and Foundation Models for Image Segmentation

VLPs have significantly advanced the field of image segmentation by enabling open-set segmentation capabilities through their richly learned textual and visual representations from the extensive knowledge embedded in their text encoders. Prompt engineering plays a central role in adapting these models to semantic segmentation tasks. Early efforts focused on Hard Prompts, evolving from simple **Single Prompts** to **Prompt Ensembling**, as seen in CLIP. However, due to their limitations in dense prediction tasks, learnable **Soft Prompts** tailored to the segmentation objective have emerged as a more effective solution.

For adapting image encoders, methods such as **deep prompt tuning** and **mask prompt tuning** have been successful in maintaining efficiency while mitigating overfitting. Moreover, prompt integration has evolved from **Late Fusion**, where image and prompt features are merged post-encoding, to **Early Fusion**, where prompt tokens are embedded alongside image tokens. These are processed jointly via self-attention and cross-attention, facilitating deeper multimodal interactionand have demonstrated strong potential.

These developments culminate in the emergence of Foundation Models, which represent the

pinnacle of VLP adaptability. Trained on vast datasets, these models are inherently generalizable, supporting a wide array of segmentation tasks with minimal fine-tuning. Such models excel in **multi-prompt adaptability and generalization**, proving effective in addressing a wide range of segmentation challenges. They offer remarkable flexibility by leveraging large-scale data and transformer-based architectures to process various input prompts, ranging from textual descriptions to visual cues such as points or scribbles.

To further enhance their adaptability, various integration techniques have been explored. Direct prompt engineering refines input prompts for improved guidance, while auxiliary-enhanced prompt integration incorporates additional metadata for domain-specific segmentation. Adaptive image encoders modify feature extraction to improve performance in challenging environments, and few-shot learning with prompt tuning enables adaptation with minimal labeled data. Additionally, compositional multi-prompt integration enhances segmentation granularity by leveraging multiple input types.

Despite these advances, several challenges persist. One key issue is the **Modality Gap** [257], which refers to the separation between image and text features clustered in distinct subregions of the shared feature space. This gap can impede fine-grained alignment, especially at the pixel level, thereby limiting the effective integration of semantic information in promptable segmentation. Addressing this challenge is crucial for advancing both Referring Segmentation and broader prompt Compositionality tasks that involve textual inputs.

31 Prompt-Based Models Efficiency

As prompt-based segmentation models increase in complexity and scope, concerns about efficiency and scalability are becoming more prominent. These models offer strong generalization and flexibility, particularly through integration with VLPs, but also introduce practical limitations related to computational cost, memory usage, and deployment feasibility.

One major challenge is integrating multimodal prompts such as text, points, scribbles, or bounding boxes, which require processing diverse input types simultaneously. Foundation models often use early fusion strategies, embedding prompt tokens with image features and processing them together through attention mechanisms. While this improves multimodal interaction, it also increases computational load during inference due to the high-dimensional token space and quadratic complexity of transformer attention.

Another bottleneck is handling large prompt contexts. In in-context learning setups, models are conditioned on full support-query sets or prompt grids. For example, in PF3S, concatenating multiple support examples as visual prompts can exceed the context window of transformer models, resulting in slower inference and higher memory demands. Although foundation segmentation models support zero-shot capabilities, they are not inherently efficient. Techniques like deep prompt tuning, adaptive encoders, and auxiliary-enhanced prompts can improve performance but also introduce complex data flows that complicate real-time use in constrained environments.

The modality gap, or disconnect between image and text feature spaces, is another source of inefficiency. Bridging this gap often requires deeper fusion modules or alignment layers, which increase model depth and computational requirements. If not addressed during pretraining, models may use significant capacity on basic alignment rather than task-specific reasoning.

To improve real-world applicability, recent research has focused on strategies that reduce computational overhead without sacrificing segmentation accuracy. One promising direction for future research is **prompt compression**, which minimizes redundant information in multimodal prompts. Techniques such as token pruning and learned prompt embeddings have been extended to vision tasks, where prompt-guided token pruning selectively removes low-relevance tokens, reducing computation by up to 35–55% while maintaining accuracy [258]. Similarly, **dynamic prompt selection** mechanisms adaptively choose the most relevant prompts during inference,

avoiding unnecessary processing and improving latency [259]. Another line of work explores **lightweight transformer architectures** to address the complexity of cross-modal fusion. Methods like Performer [260] and MobileViT [261] employ linear attention or hybrid CNN-transformer designs to reduce quadratic attention costs. These approaches align with parameter-efficient fine-tuning, combining adapters and prompt tuning to reduce trainable parameters and memory usage [259]. Collectively, these strategies aim to balance accuracy with efficiency, essential for deploying prompt-based models in real-time and resource-constrained environments.

32 Conclusion

In conclusion, prompt-based semantic segmentation holds vast potential for revolutionizing computer vision tasks by providing more effective, dynamic, and context-aware models. However, addressing the discussed challenges will be necessary to achieve the full potential of these systems, particularly in complex, real-world scenarios. This survey reviewed the development of various types of promptable segmentation research areas, including interactive, referring, few-shot, and open-vocabulary segmentation. Thus, by categorizing encoding, integration, and learning strategies, we highlighted key methodologies driving progress in this field, providing readers with comprehensive yet practical knowledge of reflective developments. As the capability of incorporating user-guided prompts increases, the field is set for innovative developments in the future, particularly in reducing annotation loads and improving cross-modal understanding. We hope this survey sparks further exploration into the design and application of promptable segmentation systems, addressing current challenges such as improving model generalization across diverse visual domains, ensuring effective cross-modal alignment, and developing robust prompt encoding strategies that can handle ambiguity, variability, and compositionality in user inputs.

A Interactive Segmentation Works

Work	Year	Task	Prompt Type	Network Type	Network Detail	Datasets	Highlight
Interactive Segmentation							
DIOS	2016	IS	Click	CNN	Based on FCN-8s [191]	Grabcut Berkeley PASCAL VOC MSCOCO	First IS architecture based on FCN. Introduced Euclidean distance transform of click interaction.
DeepCut	2016	Medical-IS	Bounding Box	CNN	CNN	Database from [262]	This approach extends previous solutions, incorporating a neural network to segment the fetal brain and lungs in medical images, starting from bounding box annotations.
ScribbleSup	2016	SS	Scribble	CNN	DeepLab-MSc-CRF-LargeFOV	PASCAL Context	Proposes a scribble-supervised training by optimizing a graphical model that propagates the information from the scribbles to the unmarked pixels.
RIS-Net	2017	IS	Click	CNN	Based on DeepLab-LargeFOV	Grabcut Berkeley Pascal VOC MSCOCO	Uses multiscale global contextual information to augment each local region for improving feature representation.
IMISDL	2017	Medical-IS	Bounding Box	CNN	P-Net	-	We propose a method that performs image-specific fine-tuning to adapt a CNN model to a particular test image in the domain of medical image segmentation.
CEDN	2017	IS	Bounding Box	CNN	Convolutional encoder-decoder architecture (FCN with VGG16)	Grabcut SBD MSCOCO	Proposes a novel way to use rectangles for selection that produces more accurate results given a tight bounding box while also giving similar results for loosely-placed rectangles. Rectangles are transformed into Euclidean distance maps.
UI-Net	2017	Medical-IS	Scribble	CNN	U-Net	DYNA-CT	Proposed UI-Net architecture to semi-automatic image segmentation in medical applications where only few datasets need to be processed and only a small database of fully annotated images is available for training.
Polygon-RNN	2017	IS	Polygon	CNN+RNN	The CNN is based on VGG-16. ConvLSTM as RNN	Cityscapes KITTI	Allows a human annotator to modify the polygon vertices if needed, producing a more accurate segmentation.
IBPOS	2018	IS	Click	CNN	Custom convolutional encoder-decoder architecture	GrabCut	Concatenates multiple 2D Gaussian interaction maps with three different standard deviation values.
FCTSFN	2018	IS	Click	CNN	FCN (VGG16)	Grabcut Berkeley PASCAL VOC MSCOCO	A two-stream late fusion network that processes separately the input image and the user interactions. Then the features are fused by a convolutional fusion net.

Work	Year	Task	Prompt Type	Network Type	Network Detail	Datasets	Highlight
Interactive Segmentation							
DEXTR	2018	IS	Click	CNN	Based on Deeplab-V2 (ResNet-101)	COCO PASCAL VOC Grabcut DAVIS	Explores the use of extreme points in an object (left-most, right-most, top, bottom pixels) as input and encodes them as 2D Gaussian in order to create a single heatmap.
ITIS	2018	IS	Click	CNN	Based on Deeplab-V3+	Grabcut KITTI PASCAL VOC DAVIS	Proposes a novel iterative training strategy and compares different design choices for representing click and mask inputs.
InterCNN	2018	Medical-IS/SS	Scribble	CNN	U-Net	NCI-ISBI	Assumes the editing CNN is an auxiliary tool that supports a base segmentation algorithm and is optimized to take into account user edits and improve segmentation accuracy.
Polygon-RNN++	2018	IS	Polygon	CNN +RNN	ResNet-50+ConvLSTM	Cityscapes KITTI	Increases the output resolution of Polygon-RNN using a graph neural network.
BRS	2019	IS, SS	Click	CNN	Encoder based on DenseNet, CNN w/ deconvolutions	Grabcut Berkeley DAVIS SBD	Introduces the backpropagating refinement strategy, which corrects mislabeled pixels.
MultiSeg	2019	IS	Click	CNN	Based on DeepLabV3+ (ResNet-101)	Grabcut Berkeley PASCAL VOC	Introduces the concept of scale diversity and incorporates it to generate a set of scale-varying proposals conditioned on the user input.
CAMLG-IIS	2019	IS	Click	CNN	Based on FCN-8s	Grabcut Berkeley PASCAL VOC MSCOCO	Proposes a scale aware guidance map generated using hierarchical image information.
DeepIGeoS	2019	Medical-IS	Scribble	CNN	Based on P-Net	BraTS	Proposes to combine user interactions with CNNs through geodesic distance transforms.
FIO-GCN	2019	IS	Polygon	GCN	–	Cityscapes KITTI ADE20K Cardiac MR SsTEM	Frames object annotation as a regression problem, where the locations of all vertices of the polygon are predicted simultaneously.
LIOS	2019	IS	BBox + Click	CNN	Deeplabv2 ResNet101	MSCOCO	Combines bounding box and click interactions to enhance segmentation result.
F-BRS	2020	IS, SS	Click	CNN	Based on DeepLabV3+ (ResNet-101)	GrabCut Berkeley DAVIS SBD	Optimizes BRS and introduces the Distance Maps Fusion (DMF) module which transforms an image concatenated with additional user click channels into 3-channel input.

Work	Year	Task	Prompt Type	Network Type	Network Detail	Datasets	Highlight
Interactive Segmentation							
FCA-Net	2020	IS	Click	CNN	DeeplabV3+ with ResNet-101 Backbone	Grabcut Berkeley PASCAL VOC MSCOCO	Demonstrates the critical role of the first click integrating an effective module for utilizing the guidance information of the first click.
99A-IS	2020	IS	Click	CNN	U-Net style with ResNet-50 Backbone	Grabcut Berkeley SBD	A single architecture which allows for high resolution processing. Separate image stream from interaction stream to improve the propagation of user interaction.
CAISLC	2020	IS	Click	CNN	DeepLabV3+	Berkeley YouTube-VOS MSCOCO	Proposes a practical online adaptation method to update the parameters of an interactive segmentation model at test time. User corrections are encoded as binary disks.
PhraseClick	2020	IS, SS	Click + Text	CNN+RNN	ResNet-101 based on DeepLabv3 for image and Bi-LSTM for text	PASCAL VOC GrabCut Berkeley	First work to leverage both clicks and phrases for interactive segmentation. Proposes an attribute-guided feature attention module to bring clicks and phrases information together.
IOG	2020	IS	BBox + Click	CNN	Based on FPN	Grabcut ImageNet SsTEM PASCAL Context Cityscapes MSCOCO	Introduces modifications to the annotation interface, such as using a horizontal and a vertical guide line to assist user bounding box interaction.
TOS-Net	2021	IS	Click	CNN	Based on Deeplab-V3+ (ResNet-50 and Atrous Spatial Pyramid Pooling)	ThinObject-5K HRSOD COIFT	Introduces an edge-guided segmentation baseline to tackle thin object selection taking four extreme points (top, bottom, leftmost, and rightmost pixels) as inputs and converts them to Gaussian heatmaps before concatenating with the image.
CDNet	2021	IS	Click	CNN	DeeplabV3+ with ResNet-50 Backbone	Grabcut Berkeley DAVIS SBD	Formulates click-based interactive segmentation as a process of information diffusion and proposes Conditional Diffusion Network.
MIDeepSeg	2021	Medical-IS	Click	CNN	U-Net	Placenta in MRI Spleen in CT	Presents a new way to encode user interactions based on exponentialized geodesic distance transform, which is context-aware and parameter-free.
FocusCut	2022	IS	Click	CNN	DeeplabV3+ with ResNet-101 Backbone	Grabcut Berkeley DAVIS SBD	Introduces the focus view to grasp user intentions by considering the local segmentation from clicks' eyes.

Work	Year	Task	Prompt Type	Network Type	Network Detail	Datasets	Highlight
Interactive Segmentation							
PseudoClick	2022	IS	Click	Transformer	HRFormer or SegFormer	Grabcut Berkeley PASCAL VOC SBD MSCOCO	Generic framework that can be built upon different types of segmentation backbones, including both CNNs and transformers. Imitates human clicks and refines the segmentation with automatically generated pseudo clicks.
RITM	2022	IS	Click	CNN	HRNet	GrabCut Berkeley SBD DAVIS PASCAL VOC	Conducts ablation study to compare the distance transform encoding with the disks encoding and finds that the latter outperforms the former.
FocalClick	2022	IS	Click	CNN	HRNet	GrabCut Berkeley SBD DAVIS	Proposes progressive merge to perform local mask correction of the previous masks and the current predictions to decide where to update and preserve.
ISegformer	2022	Medical-IS	Click	Hierarchical Transformer	Swin Transformer with a lightweight MLP decoder	SsTEM BraTS OAI-ZIB	First transformer-based approach for interactive medical image segmentation.
ECONet	2022	Medical-IS	Scribble	CNN	Lightweight FCN	UESTC-COVID-19	Proposes an efficient convolutional neural network that can be learned online while the annotator provides scribble-based interaction.
MM-IIS	2022	Medical-IS	BBox /Polygon /Scribble + Click	CNN	–	COVID19-CT COVID19Xray BraTS-T1	First attempt to systematically integrate multiple interaction modes into a unified network. Users can freely choose and jointly use the appropriate interaction modes to deal with different ambiguities in segmentation.
SimpleClick	2023	IS	Click	Transformer	ViT	GrabCut Berkeley SBD DAVIS PASCAL VOC SsTEM BraTS OAI-ZIB	First interactive segmentation method that leverages non-hierarchical vision transformer (ViT) backbone.
Scribble-Prompt	2023	Medical-IS	Scribble	CNN or Transformer	–	MegaMedical	Devises a new training strategy to simulate diverse and realistic scribbles.

B Referring Segmentation Works

Work	Year	Task	Prompt Type	Network Type	Network Detail	Datasets	Highlight
Referring Segmentation							
VE: Visual Encoder, TE: Text Encoder, MF: Multimodal Fusion, MD: Mask Decoder							
LSTM-CNN	2016	RES	Free Text	CNN + RNN	VE: CNN TE: LSTM MF: concat + conv (late fusion) MD: CNN	ReferIt	Introduced RES task and first RES architecture
RRN	2018	RES	Free Text	CNN + RNN	VE: CNN TE: LSTM MF: concat + conv (late fusion) MD: CNN	ReferIt RefCOCO RefCOCO+ RefCOCOg	Refines multi-modal feature stack by recurrently incorporating multi-scale visual features
DMN	2018	RES	Free Text	CNN + RNN	VE: CNN TE: LSTM MF: concat + conv + mSRU (late fusion) MD: CNN	ReferIt RefCOCO RefCOCO+ RefCOCOg	Introduces the Synthesis Module (SM) for multi-modal feature fusion
RMI	2017	RES	Free Text	CNN + RNN	VE: CNN TE: LSTM MF: concat + conv (late fusion) MD: CNN	ReferIt RefCOCO RefCOCO+ RefCOCOg	Fuses text and image modalities recurrently as a sequential interaction
LAVT	2022	RES	Free Text	CNN + Transformer	VE: Swin TE: BERT MF: language-guided cross-attention (early fusion) MD: CNN	RefCOCO RefCOCO+ RefCOCOg	First work to leverage multimodal early fusion

Work	Year	Task	Prompt Type	Network Type	Network Detail	Datasets	Highlight
Referring Segmentation							
VE: Visual Encoder, TE: Text Encoder, MF: Multimodal Fusion, MD: Mask Decoder							
RefSegformer	2022	Robust RES	Free Text	CNN + Transformer	VE: Swin TE: BERT MF: language-guided cross-attention (early fusion) MD: CNN	RefCOCO RefCOCO+ RefCOCOg R-RefCOCO R-RefCOCO+ R-RefCOCOg	The first model tackling the Robust RES task
GRES (method), ReLA (multi-modal Fusion Module)	2023	Generalized RES	Free Text	Transformer	VE: Swin transformer TE: BERT MF: cross-attention MD: Transformer	GRefCOCO	Introduces GRES task, new dataset GRefCOCO, new ReLA multi-modal fusion method
M^{3}Net	2023	RES	Free Text	CNN + Transformer	VE: Swin TE: BERT MF: mutual cross-attention (early fusion) MD: custom transformer	RefCOCO RefCOCO+ RefCOCOg	First model to employ both cross-attention mechanisms for multi-modal fusion (language-guided visual features and vision-guided linguistic features).
DMMI	2023	Generalized RES	Free Text	CNN + Transformer	VE: Swin TE: BERT MF: mutual cross-attention (early fusion) MD: CNN	RefCOCO RefCOCO+ RefCOCOg RefZOM	Introduces RefZOM dataset for generalized RES; employs mutual cross-attention fusion; employs multitask learning framework (segmentation + ref. expr. completion)

C Few-Shot Segmentation Works

Work	Year	Task	Prompt Type	Network Type	Network Detail	Datasets	Highlight
Few-Shot Semantic Segmentation							
OSLSM	2017	FSSS	Masked images	CNN	CNN/VGG - FCNN/FCN32-S	PASCAL VOC	Introduces support set of example images to predict the task i.e. segment a never seen class
FSSPGNet	2018	FSSS	Points annotated images	CNN	VGG - Custom CNN	PASCAL VOC	Extends the support set with points annotations
CMN	2021	FSSS	Masked images	CNN	ResNet/VGG16 - ConvGRU - Custom CNN	PASCAL-5i COCO20i	Introduces Cyclic Memory Network to directly learn to read abundant support information at different resolutions
PFENet	2022	FSSS	Masked images	CNN	ResNet/VGG - Custom CNN	PASCAL-5i COCO20i	Introduces autogenerated, class-insensitive prior mask and exploits adaptive integration of support and query features for enhanced generalization
VPIP	2022	Multiple CV Tasks	Grid like Masked images	Transformer	MAE-VQGAN	CV Figures	Introduces painter model as generalist model for all computer vision tasks
Generalist Painter	2023	Multiple CV Tasks	Grid-like masked images	CNN +Trans-former	ViT -CNN Head	ADE20K COCO DAVID MOSE FSS-1000	Extends painter approach with patch-based processing for enhanced granularity and efficiency.
SegGPT	2023	Multiple CV Tasks	Grid-like masked images	CNN +Trans-former	ViT - CNN Head	ADE20K COCO DAVID MOSE FSS-1000	Extends patch-based processing with random coloring scheme, introduces in-context tuning

D Vision-Language Pretrained Segmentation Works

Work	Year	Task	Prompt Type	Network Type	Network Detail	Datasets	Highlight
Pre-trained VLMs for Open Vocabulary Segmentation							
SSIW	2021	SS	Hard Prompt (Template)	Transformer	CLIP ViT-B/32	PASCAL VOC ImageNet Segmentation	Demonstrates that vision-language models can be used to transform image-level guidance into high accuracy relevance maps.
CLIMS	2022	SS	Hard Prompt (Template)	CNN	Based on DeepLabV2 with ResNet-101	Pascal VOC	Proposes a text-driven learning framework to introduce image-text matching model based supervision in an open-world setting.
MaskCLIP	2022	SS	Hard Prompt (Template)	CNN or Transformer	CLIP ResNets/ViTs	PASCAL VOC PASCAL Context COCO Stuff	Shows that features learned via large-scale visual-language pre-training can be readily used to facilitate open vocabulary dense prediction.
LSeg	2022	Z3S, F3S	Hard Prompt (Template)	Transformer	Dense Prediction Transformer with a ViT-L/16	Pascal VOC MSCOCO	Proposes a method for zero- few-shot semantic segmentation which also performs competitively in fixed label set scenario.
ZSSeg	2022	SS	Soft Prompt	Transformer	CLIP ViT-B/16	MSCOCO PASCAL VOC PASCAL Context ADE20k Cityscapes	Shows that learnable prompts outperform manually searched prompts both on seen and unseen classes of few-shot segmentation scenario.
ZegFormer	2022	SS	Hard Prompt (Template)	CNN	Based on Detectron2 with ResNet50 and FPN	MSCOCO PASCAL VOC ADE20k	Proposes a new formulation for the task of zero-shot semantic segmentation, by decoupling it into two sub-tasks, a class-agnostic grouping and a segment-level zero-shot classification.
ReCo	2022	SS	Hard Prompt (Template)	Transformer	CLIP ViT-L/14@336px	MSCOCO Cityscapes	A vision backbone is used to co-segment concepts gathered in a curated collection.
Fusioner	2022	Z3S, F3S	Hard Prompt (Template)	Transformer	CLIP ViT-L/14@336px	MSCOCO Pascal VOC	Shows that early cross-modal fusion of vision and language features outperforms late fusion.

Work	Year	Task	Prompt Type	Network Type	Network Detail	Datasets	Highlight
Pre-trained VLMs for Open Vocabulary Segmentation							
OpenSeg	2022	SS	Hard Prompt (Free Text)	CNN	Based on EfficientNet-B7 [263]	PASCAL VOC PASCAL Context ADE20k	Trained using a weakly supervised approach on image-caption pairs, it employs synonym ensembling and context augmentation to account for polysemy in natural language prompts.
CLIPSeg	2022	SS	Hard Prompt (Free Text)	Transformer	CLIP ViT-B/16	PASCAL VOC	CLIPSeg utilizes Late Fusion with FiLM conditioning to efficiently merge the user prompt, whether textual or visual, with the input image.
CLIP-ES	2023	SS	Hard Prompt (Template)	Transformer	CLIP ViT-B/16	PASCAL VOC MSCOCO	Explores the potential of CLIP to localize different categories with image-level labels.
ZegCLIP	2023	SS	Soft Prompt	Transformer	CLIP ViT-B/16	MSCOCO Pascal VOC PASCAL Context	Performs a comparison among single and multiple text templates observing that the latter leads to reasonable improvements.
OVSeg	2023	SS	Soft Prompt	Transformer	MaskFormer	PASCAL VOC PASCAL Context ADE20k	It proposes mask prompt tuning as a new visual tuning strategy to effectively adapt CLIP for mask region classification.
SegCLIP	2023	SS	Hard Prompt (Template)	Transformer	CLIP ViT-B/16	Conceptual Captions MSCOCO	Proposes a CLIP-based model for weakly-supervised semantic segmentation.
ZegOT	2023	SS	Soft Prompt	Transformer	CLIP	PASCAL VOC PASCAL Context COCO-Stuff164K	It proposes a combination of multiple learnable text prompts and Optimal Transport to match image and prompt embeddings.

E Foundation Segmentation Works

Work	Year	Task	Prompt Type	Network Type	Network Detail	Datasets	Highlight
Foundation Models for Segmentation tasks							
SAM	2022	SS	Points, Box, Text	Transformer	CLIP ViTs	SA-1B	First foundation model for segmentation
SEEM	2023	Panoptic-RS-IS	Points, Box, Scribbles, Masks, Text, Image	Transformer	Transformer + extra text encoder for queries	COCO COCO+ RefCOCO	Enhances interactive AI in visual understanding, supports versatility and compositionality
SAM-Med2D	2023	Medical-SS	Points, Box	Transformer	Fine-tuned SAM architecture for specific domain	Custom (4.6M images, 19.7M masks (medical))	Superior in medical image segmentation benchmarks
All-In-SAM	2023	SS	Points, Box	Transformer	Adapted SAM architecture for specific tasks	MICCAI 2018 Monuseg	From weak annotation to detailed segmentation
Semantic-SAM	2023	SS	Click-based queries	CNN Transformer	Multi-granularity learning + query-based mask decoder	MSCOCO Objects365 ADE20k	Semantic awareness into SAM
PromptSA	2023	ZSS	Text queries	Transformer	SAM-based architecture tailored for zero-shot learning	Open Images Dataset	Robust zero-shot segmentation capabilities
CAT	2023	Caption+SS	Text, Points, Box, Mask	Transformer	Integration of GPT4 and SAM (Mask Decoder + LM Block)	COCO Visual Genome	Versatile image processing tool that combines the capabilities of SAM, visual captioning, and ChatGPT
LA	2023	SS	Text, Points, Image	Transformer	Combination of SAM for visual segmentation and GPT4 for NLP	COCO ADE20k	Enhances detailed annotation capabilities. Annotation in visual tasks and all-in-one pipeline with GPT4 and SAM

Work	Year	Task	Prompt Type	Network Type	Network Detail	Datasets	Highlight
Foundation Models for Segmentation tasks							
MedSAM	2023	Medical-SS	Points, Boxes	Transformer	Modified SAM architecture to handle medical images	MICCAI FLARE2022	Specialized for precision in medical image segmentation
AutoSAM	2023	SS	Learnable prompt	Transformer	Lightweight SAM configuration	LVIS	Provides robustness against OOD samples. Lightweight segmentation solution for out-of-distribution (OOD) domains
VisionLLM	2023	ZSS	Text	Transformer	ViTs integrated with language models	COCO	Facilitates flexible, user-driven segmentation tasks. Zero-shot capacity for user-tailored tasks
AdaptiveSA	2023	SS	Text, Points, Visual Cues	Transformer	SAM architecture with enhanced adaptability features	Dynamic medical datasets with changing scenarios	Ensures SAM's effectiveness under dynamic conditions. Adapts SAM for dynamic learning environments
GroundedSA	2023	SS	Text, Points, Box	Transformer	SAM architecture optimized for ground-truth integration	Datasets with detailed annotations like parts of COCO	Comprehensive prompts for detailed segmentation. Improves precision and reliability by leveraging ground-truth data. Integrates ground-truth data into SAM training
Text2Seg	2023	SS	Text	Transformer	Language processing units + SAM (ViT-H, mask decoder + LM Block)	RefCOCO	Converts text descriptions into segmentation tasks, Textual descriptions specifying segmentation tasks. Translates textual instructions into precise visual segmentation
SAM-CLIP	2023	ZSS	Text	Transformer	Unified model combining spatial (SAM) + semantic (CLIP) capabilities	zero-shot semantic segmentation benchmarks (e.g. PASCAL-VOC, COCO-Stuff)	Establishes new SOTA in zero-shot semantic segmentation

References

[1] Yizhe Zhang et al. “Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 129–139.

[2] B Basavaprasad and S Hegadi Ravindra. “A Survey on Traditional and Graph Theoretical Techniques for Image Segmentation”. In: *International Journal of Computer Applications* NC-RAIT2014.1 (2014). Special Issue on Recent Advances in Information Technology, pp. 1–6. URL: https://research.ijcaonline.org/ncrait/number1/ncrait1408.pdf.

[3] Arti Taneja, Priya Ranjan, and Amit Ujjlayan. “A Performance Study of Image Segmentation Techniques”. In: *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*. IEEE. 2015, pp. 1–6. DOI: 10.1109/ICRITO.2015.7359305.

[4] Vicente García Díaz et al. “Robot based Transurethral Bladder Tumor Resection with automatic detection of tumor cells”. In: *Measurement* 206 (2023), p. 112079.

[5] Ashish Vaswani et al. “Attention is All You Need”. In: *Advances in Neural Information Processing Systems* 30 (2017). URL: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[6] Xiaolong Wang et al. “Non-local Neural Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7794–7803. URL: https://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Non-local_Neural_Networks_CVPR_2018_paper.html.

[7] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[8] Nicolas Carion et al. “End-to-end object detection with transformers”. In: *European conference on computer vision*. Springer. 2020, pp. 213–229.

[9] Xiangtai Li et al. *Transformer-Based Visual Segmentation: A Survey*. arXiv:2304.09854 [cs]. Dec. 2023. URL: http://arxiv.org/abs/2304.09854 (visited on 01/16/2024).

[10] Jiaqi Wang et al. “Review of large vision models and visual prompt engineering”. In: *Meta-Radiology* 1.3 (2023), p. 100047. DOI: 10.1016/j.metrad.2023.100047. URL: https://www.sciencedirect.com/science/article/pii/S2950162823000474.

[11] Banghao Chen et al. “Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review”. In: *arXiv preprint arXiv:2310.14735* (2023).

[12] Ning Xu et al. “Deep interactive object selection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 373–381.

[13] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV]. URL: https://arxiv.org/abs/2103.00020.

[14] Chao Jia et al. “Scaling up visual and vision-language representation learning with noisy text supervision”. In: *International conference on machine learning*. PMLR. 2021, pp. 4904–4916.

[15] Maxime Bucher et al. “Zero-shot semantic segmentation”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[16] Yongqin Xian et al. “Semantic projection network for zero-and few-label semantic segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 8256–8265.

[17] Alexander Kirillov et al. “Segment Anything”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 14784–14794. URL: https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html.

[18] Jingyi Zhang et al. *Vision-Language Models for Vision Tasks: A Survey*. en. arXiv:2304.00685 [cs]. Feb. 2024. URL: http://arxiv.org/abs/2304.00685 (visited on 07/16/2024).

[19] Florian Bordes et al. *An Introduction to Vision-Language Modeling*. en. arXiv:2405.17247 [cs]. May 2024. URL: http://arxiv.org/abs/2405.17247 (visited on 07/16/2024).

[20] Xiangtai Li et al. *Transformer-Based Visual Segmentation: A Survey*. arXiv:2304.09854 [cs]. Dec. 2023. URL: http://arxiv.org/abs/2304.09854 (visited on 01/16/2024).

[21] Yanxin Long et al. “Fine-Grained Visual–Text Prompt-Driven Self-Training for Open-Vocabulary Object Detection”. en. In: *IEEE Transactions on Neural Networks and Learning Systems* (2023), pp. 1–11. ISSN: 2162-237X, 2162-2388. DOI: 10.1109/TNNLS.2023.3293484. URL: https://ieeexplore.ieee.org/document/10197240/ (visited on 12/12/2023).

[22] Timo Lüddecke and Alexander Ecker. “Image segmentation using text and image prompts. In 2022 IEEE”. In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 7076–7086.

[23] Jiaqi Wang et al. *Review of Large Vision Models and Visual Prompt Engineering*. en. arXiv:2307.00855 [cs]. July 2023. URL: http://arxiv.org/abs/2307.00855 (visited on 02/07/2024).

[24] JiaLu Xing et al. “A survey of efficient fine-tuning methods for Vision-Language Models — Prompt and Adapter”. en. In: *Computers & Graphics* (Jan. 2024), S0097849324000128. ISSN: 00978493. DOI: 10.1016/j.cag.2024.01.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S0097849324000128 (visited on 02/08/2024).

[25] Lingfeng Yang et al. “Fine-Grained Visual Prompting”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47.3 (Mar. 2025), pp. 1594–1609. DOI: 10.1109/TPAMI.2024.3504568. URL: https://ieeexplore.ieee.org/document/10763465/.

[26] Chaoning Zhang et al. *A Survey on Segment Anything Model (SAM): Vision Foundation Model Meets Prompt Engineering*. en. arXiv:2306.06211 [cs]. July 2023. URL: http://arxiv.org/abs/2306.06211 (visited on 02/07/2024).

[27] Muhammad Awais et al. “Foundational Models Defining a New Era in Vision: A Survey and Outlook”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47.4 (2025), pp. 1234–1256. DOI: 10.1109/TPAMI.2025.1234567. URL: https://www.computer.org/csdl/journal/tp/2025/04/10834497/23mYUeDuDja.

[28] Jindong Gu et al. *A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models*. arXiv:2307.12980 [cs]. July 2023. URL: http://arxiv.org/abs/2307.12980 (visited on 03/06/2024).

[29] Chunhui Zhang et al. *A Comprehensive Survey on Segment Anything Model for Vision and Beyond*. en. arXiv:2305.08196 [cs]. May 2023. URL: http://arxiv.org/abs/2305.08196 (visited on 02/07/2024).

[30] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022.

[31] KKD Ramesh et al. “A review of medical image segmentation algorithms”. In: *EAI Endorsed Transactions on Pervasive Health and Technology* 7.27 (2021), e6–e6.

[32] Bike Chen, Chen Gong, and Jian Yang. “Importance-aware semantic segmentation for autonomous vehicles”. In: *IEEE Transactions on Intelligent Transportation Systems* 20.1 (2018), pp. 137–148.

[33] Saikat Roy et al. “Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model”. In: *arXiv preprint arXiv:2304.05396* (2023).

[34] Monica Gruosso, Nicola Capece, and Ugo Erra. “Human segmentation in surveillance video with deep learning”. In: *Multimedia Tools and Applications* 80.1 (2021), pp. 1175–1199.

[35] Tianfei Zhou et al. “Image Segmentation in Foundation Model Era: A Survey”. In: *arXiv preprint arXiv:2408.12957* (2024).

[36] Jian Ding et al. “Decoupling zero-shot semantic segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 11583–11592.

[37] Ziqin Zhou et al. “Zegclip: Towards adapting clip for zero-shot semantic segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 11175–11185.

[38] Chong Zhou, Chen Change Loy, and Bo Dai. “Extract free dense labels from clip”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 696–712.

[39] Chaofan Ma et al. “Open-vocabulary semantic segmentation with frozen vision-language models”. In: *arXiv preprint arXiv:2210.15138* (2022).

[40] Xueyan Zou et al. “Segment Everything Everywhere All at Once”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 19769–19782. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/3ef61f7e4afacf9a2c5b71c726172b86-Paper-Conference.pdf.

[41] Xinlong Wang et al. “SegGPT: Segmenting Everything In Context”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 1130–1140. DOI: 10.1109/ICCV51070.2023.00122. URL: https://doi.org/10.1109/ICCV51070.2023.00122.

[42] Shubham Vatsal and Harsh Dubey. “A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks”. In: *arXiv preprint arXiv:2407.12994* (2024).

[43] Qingxiu Dong et al. *A Survey on In-context Learning*. 2023. arXiv: 2301.00234 [cs.CL].

[44] Zhengyuan Yang et al. “An empirical study of gpt-3 for few-shot knowledge-based vqa”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 36. 3. 2022, pp. 3081–3089.

[45] Jason Wei et al. “Chain-of-thought prompting elicits reasoning in large language models”. In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.

[46] Qin Liu et al. “Simpleclick: Interactive image segmentation with simple vision transformers”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 22290–22300.

[47] Qin Liu et al. “iSegFormer: interactive segmentation via transformers with application to 3D knee MR images”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 464–474.

[48] Jiajun Wu et al. “Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 256–263.

[49] Ming-Ming Cheng et al. “Densecut: Densely connected crfs for realtime grabcut”. In: *Computer Graphics Forum*. Vol. 34. 7. Wiley Online Library. 2015, pp. 193–201.

[50] Yuri Y Boykov and M-P Jolly. “Interactive graph cuts for optimal boundary & region segmentation of objects in ND images”. In: *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*. Vol. 1. IEEE. 2001, pp. 105–112.

[51] Leo Grady. “Random walks for image segmentation”. In: *IEEE transactions on pattern analysis and machine intelligence* 28.11 (2006), pp. 1768–1783.

[52] David Acuna et al. “Efficient interactive annotation of segmentation datasets with polygon-rnn++”. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018, pp. 859–868.

[53] Eric N Mortensen and William A Barrett. “Intelligent scissors for image composition”. In: *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. 1995, pp. 191–198.

[54] Yin Li et al. “Lazy snapping”. In: *ACM Transactions on Graphics (ToG)* 23.3 (2004), pp. 303–308.

[55] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. “Reviving iterative training with mask guidance for interactive segmentation”. In: *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2022, pp. 3141–3145.

[56] Zheng Lin et al. “Focuscut: Diving into a focus view in interactive segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 2637–2646.

[57] Xi Chen et al. “Focalclick: Towards practical interactive image segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 1300–1309.

[58] Konstantin Sofiiuk et al. “f-brs: Rethinking backpropagating refinement for interactive segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8623–8632.

[59] Zheng Lin et al. “Interactive image segmentation with first click attention”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 13339–13348.

[60] Qin Liu et al. “Pseudoclick: Interactive image segmentation with click imitation”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 728–745.

[61] Junjie Bai and Xiaodong Wu. “Error-tolerant scribbles based interactive image segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 392–399.

[62] Daniel Freedman and Tao Zhang. “Interactive graph cut based segmentation with shape priors”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*. Vol. 1. IEEE. 2005, pp. 755–762.

[63] Gedas Bertasius and Lorenzo Torresani. “Classifying, segmenting, and tracking object instances in video with mask propagation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9739–9748.

[64] Ning Xu et al. “Youtube-vos: A large-scale video object segmentation benchmark”. In: *arXiv preprint arXiv:1809.03327* (2018).

[65] Holger Caesar et al. “nuscenes: A multimodal dataset for autonomous driving”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 11621–11631.

[66] Geert Litjens et al. “A survey on deep learning in medical image analysis”. In: *Medical image analysis* 42 (2017), pp. 60–88.

[67] Dinggang Shen, Guorong Wu, and Heung-Il Suk. “Deep learning in medical image analysis”. In: *Annual review of biomedical engineering* 19 (2017), pp. 221–248.

[68] Mario Amrehn et al. “UI-Net: Interactive artificial neural networks for iterative image segmentation based on a user model”. In: *arXiv preprint arXiv:1709.03450* (2017).

[69] Gustav Bredell, Christine Tanner, and Ender Konukoglu. “Iterative interaction training for segmentation editing networks”. In: *Machine Learning in Medical Imaging: 9th International Workshop, MLMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 9*. Springer. 2018, pp. 363–370.

[70] Muhammad Asad, Lucas Fidon, and Tom Vercauteren. “ECONet: Efficient convolutional online likelihood network for scribble-based interactive segmentation”. In: *International Conference on Medical Imaging with Deep Learning*. PMLR. 2022, pp. 35–47.

[71] Hallee E Wong et al. “Scribbleprompt: Fast and flexible interactive segmentation for any medical image”. In: *arXiv preprint arXiv:2312.07381* (2023).

[72] Marco Forte et al. “Getting to 99% accuracy in interactive segmentation”. In: *arXiv preprint arXiv:2003.07932* (2020).

[73] Xiangde Luo et al. “MIDeepSeg: Minimally interactive segmentation of unseen objects from medical images using deep learning”. In: *Medical image analysis* 72 (2021), p. 102102.

[74] JunHao Liew et al. “Regional interactive image segmentation networks”. In: *2017 IEEE international conference on computer vision (ICCV)*. IEEE. 2017, pp. 2746–2754.

[75] Hoang Le et al. “Interactive boundary prediction for object selection”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 18–33.

[76] Won-Dong Jang and Chang-Su Kim. “Interactive image segmentation via backpropagating refinement scheme”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5297–5306.

[77] Jun Hao Liew et al. “Multiseg: Semantically meaningful, scale-diverse segmentations from minimal user input”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 662–670.

[78] Yang Hu et al. “A fully convolutional two-stream fusion network for interactive image segmentation”. In: *Neural Networks* 109 (2019), pp. 31–42.

[79] Kevis-Kokitsi Maninis et al. “Deep extreme cut: From extreme points to object segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 616–625.

[80] Sabarinath Mahadevan, Paul Voigtlaender, and Bastian Leibe. “Iteratively trained interactive segmentation”. In: *arXiv preprint arXiv:1805.04398* (2018).

[81] Xi Chen et al. “Conditional diffusion for interactive segmentation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 7345–7354.

[82] Jun Hao Liew et al. “Deep interactive thin object selection”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 305–314.

[83] Soumajit Majumder and Angela Yao. “Content-aware multi-level guidance for interactive instance segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 11602–11611.

[84] Radhakrishna Achanta et al. “SLIC superpixels compared to state-of-the-art superpixel methods”. In: *IEEE transactions on pattern analysis and machine intelligence* 34.11 (2012), pp. 2274–2282.

[85] Antonio Criminisi, Toby Sharp, and Andrew Blake. “Geos: Geodesic image segmentation”. In: *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10*. Springer. 2008, pp. 99–112.

[86] Guotai Wang et al. “DeepIGeoS: a deep interactive geodesic framework for medical image segmentation”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.7 (2018), pp. 1559–1572.

[87] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. “Large-scale interactive object segmentation with human annotators”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 11700–11709.

[88] Theodora Kontogianni et al. “Continuous adaptation for interactive object segmentation by learning from corrections”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer. 2020, pp. 579–596.

[89] Ning Xu et al. “Deep grabcut for object selection”. In: *arXiv preprint arXiv:1707.00243* (2017).

[90] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. “"GrabCut" interactive foreground extraction using iterated graph cuts”. In: *ACM transactions on graphics (TOG)* 23.3 (2004), pp. 309–314.

[91] Guotai Wang et al. “Interactive medical image segmentation using deep learning with image-specific fine tuning”. In: *IEEE transactions on medical imaging* 37.7 (2018), pp. 1562–1573.

[92] Martin Rajchl et al. *DeepCut: Object Segmentation from Bounding Box Annotations using Convolutional Neural Networks*. 2016. arXiv: 1605.07866 [cs.CV].

[93] Maciej A Mazurowski et al. “Segment anything model for medical image analysis: an experimental study”. In: *Medical Image Analysis* 89 (2023), p. 102918.

[94] Mingzhe Hu, Yuheng Li, and Xiaofeng Yang. “Skinsam: Empowering skin cancer segmentation with segment anything model”. In: *arXiv preprint arXiv:2304.13973* (2023).

[95] An Wang et al. “Sam meets robotic surgery: an empirical study on generalization, robustness and adaptation”. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2023, pp. 234–244.

[96] Yuhao Huang et al. “Segment anything model for medical images?” In: *Medical Image Analysis* 92 (2024), p. 103061.

[97] Jun Ma et al. “Segment anything in medical images”. In: *Nature Communications* 15.1 (2024), p. 654.

[98] Wenhui Lei et al. “MedLSAM: Localize and segment anything model for 3D CT images”. In: *Medical Image Analysis* 99 (2025), p. 103370.

[99] Hiba Ramadan, Chaymae Lachqar, and Hamid Tairi. “A survey of recent interactive image segmentation methods”. In: *Computational visual media* 6.4 (2020), pp. 355–384.

[100] Di Lin et al. “Scribblesup: Scribble-supervised convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3159–3167.

[101] Lluis Castrejon et al. “Annotating object instances with a polygon-rnn”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 5230–5238.

[102] Huan Ling et al. “Fast interactive object annotation with curve-gcn”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 5257–5266.

[103] Shiyin Zhang et al. “Interactive object segmentation with inside-outside guidance”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 12234–12244.

[104] Hao Su, Jia Deng, and Li Fei-Fei. “Crowdsourcing annotations for visual object detection”. In: *Workshops at the twenty-sixth AAAI conference on artificial intelligence*. 2012.

[105] Zheng Lin et al. “Multi-mode interactive image segmentation”. In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 905–914.

[106] Xu Zhang et al. *VPUFormer: Visual Prompt Unified Transformer for Interactive Image Segmentation*. 2023. arXiv: 2306.06656 [cs.CV].

[107] Mark Everingham et al. “The pascal visual object classes (voc) challenge”. In: *International journal of computer vision* 88 (2010), pp. 303–338.

[108] Kevin McGuinness and Noel E O’connor. “A comparative evaluation of interactive segmentation algorithms”. In: *Pattern Recognition* 43.2 (2010), pp. 434–444.

[109] Federico Perazzi et al. “A benchmark dataset and evaluation methodology for video object segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 724–732.

[110] Bharath Hariharan et al. “Semantic contours from inverse detectors”. In: *2011 international conference on computer vision*. IEEE. 2011, pp. 991–998.

[111] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context”. In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Cham: Springer International Publishing, 2014, pp. 740–755. ISBN: 978-3-319-10602-1.

[112] Henghui Ding et al. “Phraseclick: toward achieving flexible interactive segmentation by phrase and click”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer. 2020, pp. 417–435.

[113] David Martin et al. “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics”. In: *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*. Vol. 2. IEEE. 2001, pp. 416–423.

[114] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. “Segmentation from Natural Language Expressions”. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, 2016, pp. 108–124. ISBN: 978-3-319-46448-0.

[115] Chang Liu, Xiangtai Li, and Henghui Ding. “Referring Image Editing: Object-level Image Editing via Referring Expressions”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 13128–13138.

[116] Jianbo Chen et al. “Language-Based Image Editing With Recurrent Attentive Models”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[117] Myungsub Choi. “Referring Object Manipulation of Natural Images with Conditional Classifier-Free Guidance”. In: *Computer Vision – ECCV 2022*. Ed. by Shai Avidan et al. Cham: Springer Nature Switzerland, 2022, pp. 627–643. ISBN: 978-3-031-20059-5.

[118] Mohit Shridhar and David Hsu. “Interactive Visual Grounding of Referring Expressions for Human-Robot Interaction”. In: *Proceedings of Robotics: Science and Systems*. Pittsburgh, Pennsylvania, June 2018. DOI: 10.15607/RSS.2018.XIV.028.

[119] Chen Jiang, Yuchen Yang, and Martin Jagersand. “CLIPUNetr: Assisting Human-robot Interface for Uncalibrated Visual Servoing Control with CLIP-driven Referring Expression Segmentation”. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. 2024, pp. 6620–6626. DOI: 10.1109/ICRA57147.2024.10611647.

[120] Xin Wang et al. “Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[121] Kun Qian et al. “GVGNet: Gaze-Directed Visual Grounding for Learning Under-Specified Object Referring Intention”. In: *IEEE Robotics and Automation Letters* 8.9 (2023), pp. 5990–5997. DOI: 10.1109/LRA.2023.3301294.

[122] Xiaoyang Zheng et al. “MAKE: Vision-Language Pre-training based Product Retrieval in Taobao Search”. In: *Companion Proceedings of the ACM Web Conference 2023*. WWW ’23. ACM, Apr. 2023. DOI: 10.1145/3543873.3584627. URL: http://dx.doi.org/10.1145/3543873.3584627.

[123] Mingchen Zhuge et al. “Kaleido-BERT: Vision-Language Pre-Training on Fashion Domain”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 12647–12657.

[124] Jeffrey Pennington, Richard Socher, and Christopher Manning. “GloVe: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162.

[125] Rico Sennrich, Barry Haddow, and Alexandra Birch. “Neural Machine Translation of Rare Words with Subword Units”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: https://aclanthology.org/P16-1162/.

[126] Yonghui Wu et al. *Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. 2016. arXiv: 1609.08144 [cs.CL]. URL: https://arxiv.org/abs/1609.08144.

[127] Lixia Ji et al. “A survey of methods for addressing the challenges of referring image segmentation”. In: *Neurocomputing* 583 (2024), p. 127599. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2024.127599. URL: https://www.sciencedirect.com/science/article/pii/S0925231224003709.

[128] Honglin Wang. “A Survey on Referring Image Segmentation”. In: *Transactions on Computer Science and Intelligent Systems Research* 5 (Aug. 2024), pp. 538–546. ISSN: 2960-1800. DOI: 10.62051/a2t2ec16. URL: http://dx.doi.org/10.62051/a2t2ec16.

[129] Ruiyu Li et al. “Referring Image Segmentation via Recurrent Refinement Networks”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 5745–5753. DOI: 10.1109/CVPR.2018.00602.

[130] Xingjian Shi et al. “Convolutional LSTM network: A machine learning approach for precipitation nowcasting”. In: *Advances in neural information processing systems* 28 (2015).

[131] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[132] Enze Xie et al. “SegFormer: Simple and efficient design for semantic segmentation with transformers”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 12077–12090.

[133] Philipp Krähenbühl and Vladlen Koltun. “Efficient inference in fully connected CRFs with Gaussian edge potentials”. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems*. NIPS’11. Granada, Spain: Curran Associates Inc., 2011, pp. 109–117. ISBN: 9781618395993.

[134] Edgar Margffoy-Tuay et al. “Dynamic Multimodal Instance Segmentation Guided by Natural Language Queries”. In: *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*. Munich, Germany: Springer-Verlag, 2018, pp. 656–672. ISBN: 978-3-030-01251-9. DOI: 10.1007/978-3-030-01252-6_39. URL: https://doi.org/10.1007/978-3-030-01252-6_39.

[135] Tao Lei, Yu Zhang, and Yoav Artzi. “Training RNNs as Fast as CNNs”. In: *CoRR* abs/1709.02755 (2017). arXiv: 1709.02755. URL: http://arxiv.org/abs/1709.02755.

[136] Chenxi Liu et al. “Recurrent Multimodal Interaction for Referring Image Segmentation”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017.

[137] Guang Feng et al. “Encoder Fusion Network With Co-Attention Embedding for Referring Image Segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 15506–15515.

[138] Zhao Yang et al. “LAVT: Language-Aware Vision Transformer for Referring Image Segmentation”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 18134–18144. DOI: 10.1109/CVPR52688.2022.01762.

[139] Jianzong Wu et al. “Toward Robust Referring Image Segmentation”. In: *IEEE Transactions on Image Processing* 33 (2024), pp. 1782–1794. DOI: 10.1109/TIP.2024.3371348.

[140] Chang Liu et al. “Multi-Modal Mutual Attention and Iterative Interaction for Referring Image Segmentation”. In: *IEEE Transactions on Image Processing* 32 (2023), pp. 3054–3065. DOI: 10.1109/TIP.2023.3277791.

[141] Yutao Hu et al. “Beyond One-to-One: Rethinking the Referring Image Segmentation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 4067–4077.

[142] Hwanjun Song et al. *Learning from Noisy Labels with Deep Neural Networks: A Survey*. 2022. arXiv: 2007.08199 [cs.LG]. URL: https://arxiv.org/abs/2007.08199.

[143] Ronghang Hu et al. “Utilizing Large Scale Vision and Text Datasets for Image Segmentation from Referring Expressions”. In: *CoRR* abs/1608.08305 (2016). arXiv: 1608.08305. URL: http://arxiv.org/abs/1608.08305.

[144] Robin Strudel, Ivan Laptev, and Cordelia Schmid. *Weakly-supervised segmentation of referring expressions*. 2022. arXiv: 2205.04725 [cs.CV].

[145] Dongwon Kim et al. “Shatter and Gather: Learning Referring Image Segmentation with Text Supervision”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 15547–15557.

[146] Francesco Locatello et al. *Object-Centric Learning with Slot Attention*. 2020. arXiv: 2006.15055 [cs.LG]. URL: https://arxiv.org/abs/2006.15055.

[147] Minglang Huang et al. *Towards Omni-supervised Referring Expression Segmentation*. 2023. arXiv: 2311.00397 [cs.CV]. URL: https://arxiv.org/abs/2311.00397.

[148] Mengxue Qu et al. “Learning to Segment Every Referring Object Point by Point”. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 3021–3030. DOI: 10.1109/CVPR52729.2023.00295.

[149] Chaoyang Zhu et al. “SeqTR: A Simple Yet Universal Network for Visual Grounding”. In: *Computer Vision – ECCV 2022*. Ed. by Shai Avidan et al. Cham: Springer Nature Switzerland, 2022, pp. 598–615. ISBN: 978-3-031-19833-5.

[150] Ze Liu et al. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. 2021. arXiv: 2103.14030 [cs.CV]. URL: https://arxiv.org/abs/2103.14030.

[151] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[152] Zhaoqing Wang et al. “CRIS: CLIP-Driven Referring Image Segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 11686–11695.

[153] Zunnan Xu et al. “Bridging Vision and Language Encoders: Parameter-Efficient Tuning for Referring Image Segmentation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 17503–17512.

[154] Neil Houlsby et al. “Parameter-Efficient Transfer Learning for NLP”. In: *CoRR* abs/1902.00751 (2019). arXiv: 1902.00751. URL: http://arxiv.org/abs/1902.00751.

[155] Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. “Zero-Shot Referring Image Segmentation With Global-Local Context Features”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 19456–19465.

[156] Xinlong Wang et al. “FreeSOLO: Learning To Segment Objects Without Annotations”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 14176–14186.

[157] Matthew Honnibal et al. “spaCy: Industrial-strength Natural Language Processing in Python”. In: (2020). DOI: 10.5281/zenodo.1212303.

[158] Xin Lai et al. “LISA: Reasoning Segmentation via Large Language Model”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 9579–9589.

[159] Haotian Liu et al. *Visual Instruction Tuning*. 2023. arXiv: 2304.08485 [cs.CV]. URL: https://arxiv.org/abs/2304.08485.

[160] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL]. URL: https://arxiv.org/abs/2106.09685.

[161] Chang Liu, Henghui Ding, and Xudong Jiang. “GRES: Generalized Referring Expression Segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 23592–23601.

[162] Wenxuan Wang et al. “Unveiling Parts Beyond Objects: Towards Finer-Granularity Referring Expression Segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 12998–13008.

[163] Qin Liu et al. “Rethinking Interactive Image Segmentation with Low Latency High Quality and Diverse Prompts”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 3773–3782.

[164] Fang Liu et al. “Referring Image Segmentation Using Text Supervision”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 22124–22134.

[165] Sahar Kazemzadeh et al. “ReferIt Game: Referring to Objects in Photographs of Natural Scenes”. In: *EMNLP*. 2014.

[166] Licheng Yu et al. “Modeling Context in Referring Expressions”. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, 2016, pp. 69–85. ISBN: 978-3-319-46475-6.

[167] Junhua Mao et al. “Generation and Comprehension of Unambiguous Object Descriptions”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 11–20. DOI: 10.1109/CVPR.2016.9.

[168] Chenyun Wu et al. “PhraseCut: Language-Based Image Segmentation in the Wild”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020.

[169] Yixuan Wu et al. “Advancing Referring Expression Segmentation Beyond Single Image”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 2628–2638.

[170] Michael Grubinger et al. “The IAPR TC12 Benchmark: A New Evaluation Resource for Visual Information Systems”. In: *OntoImage 2006 Workshop* (Oct. 2006).

[171] Hugo Jair Escalante et al. “The segmented and annotated IAPR TC-12 benchmark”. In: *Computer Vision and Image Understanding* 114.4 (2010). Special issue on Image and Video Retrieval Evaluation, pp. 419–428. ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2009.03.008. URL: https://www.sciencedirect.com/science/article/pii/S1077314209000575.

[172] Ranjay Krishna et al. “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations”. In: *International Journal of Computer Vision* 123.1 (May 2017), pp. 32–73. ISSN: 1573-1405. DOI: 10.1007/s11263-016-0981-7. URL: https://doi.org/10.1007/s11263-016-0981-7.

[173] Alina Kuznetsova et al. “The Open Images Dataset V4”. In: *International Journal of Computer Vision* 128.7 (July 2020), pp. 1956–1981. ISSN: 1573-1405. DOI: 10.1007/s11263-020-01316-z. URL: https://doi.org/10.1007/s11263-020-01316-z.

[174] Angela Dai et al. “ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017.

[175] Shuai Shao et al. “Objects365: A Large-Scale, High-Quality Dataset for Object Detection”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.

[176] Varun K. Nagaraja, Vlad I. Morariu, and Larry S. Davis. “Modeling Context Between Objects for Referring Expression Understanding”. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, 2016, pp. 792–807. ISBN: 978-3-319-46493-0.

[177] Amirreza Shaban et al. *One-Shot Learning for Semantic Segmentation*. 2017. arXiv: 1709.03410 [cs.CV].

[178] Shijie Chang et al. “Beyond mask: Rethinking guidance types in few-shot segmentation”. In: *Pattern Recognition* (2025), p. 111635.

[179] Mir Rayat Imtiaz Hossain et al. “Visual prompting for generalized few-shot segmentation: A multi-scale approach”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 23470–23480.

[180] Luca Bertinetto et al. “Learning feed-forward one-shot learners”. In: *Advances in neural information processing systems* 29 (2016).

[181] Kate Rakelly et al. *Few-Shot Segmentation Propagation with Guided Networks*. 2018. arXiv: 1806.07373 [cs.CV].

[182] Guo-Sen Xie et al. “Few-Shot Semantic Segmentation with Cyclic Memory Network”. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 7273–7282. DOI: 10.1109/ICCV48922.2021.00720.

[183] Zhuotao Tian et al. “Prior Guided Feature Enrichment Network for Few-Shot Segmentation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.2 (2022), pp. 1050–1065. DOI: 10.1109/TPAMI.2020.3013717.

[184] Amir Bar et al. “Visual Prompting via Image Inpainting”. In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 28913–28925. URL: https://papers.nips.cc/paper_files/paper/2022/file/9f09f316a3eaf59d9ced5ffaefe97e0f-Paper-Conference.pdf.

[185] Xinlong Wang et al. *Images Speak in Images: A Generalist Painter for In-Context Visual Learning*. 2023. arXiv: 2212.02499 [cs.CV].

[186] Xiang Li et al. “Fss-1000: A 1000-class dataset for few-shot segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 2869–2878.

[187] Gerhard Paaß and Sven Giesselbach. *Foundation models for natural language processing: Pre-trained language models integrating media*. Springer Nature, 2023.

[188] Zhe Gan et al. “Vision-language pre-training: Basics, recent advances, and future trends”. In: *Foundations and Trends® in Computer Graphics and Vision* 14.3–4 (2022), pp. 163–352.

[189] Zirui Wang et al. “Simvlm: Simple visual language model pretraining with weak supervision”. In: *arXiv preprint arXiv:2108.10904* (2021).

[190] Yiyuan Zhang et al. “Meta-transformer: A unified framework for multimodal learning”. In: *arXiv preprint arXiv:2307.10802* (2023).

[191] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[192] Hengshuang Zhao et al. “Pyramid scene parsing network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2881–2890.

[193] Liang-Chieh Chen et al. “Encoder-decoder with atrous separable convolution for semantic image segmentation”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 801–818.

[194] Xin Wen et al. “Self-supervised visual representation learning with semantic grouping”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 16423–16438.

[195] Bowen Cheng et al. “Masked-attention mask transformer for universal image segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 1290–1299.

[196] Jitesh Jain et al. “Oneformer: One transformer to rule universal image segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 2989–2998.

[197] Mengde Xu et al. “A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 736–753.

[198] Feng Liang et al. “Open-vocabulary semantic segmentation with mask-adapted clip”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 7061–7070.

[199] Yongming Rao et al. “Denseclip: Language-guided dense prediction with context-aware prompting”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 18082–18091.

[200] Huaishao Luo et al. “Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 23033–23044.

[201] Jinheng Xie et al. “Clims: Cross language image matching for weakly supervised semantic segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 4483–4492.

[202] Gyungin Shin, Weidi Xie, and Samuel Albanie. “Reco: Retrieve and co-segment for zero-shot transfer”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 33754–33767.

[203] Golnaz Ghiasi et al. “Scaling open-vocabulary image segmentation with image-level labels”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 540–557.

[204] Yuqi Lin et al. “Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 15305–15314.

[205] Jialu Xing et al. “A survey of efficient fine-tuning methods for Vision-Language Models—Prompt and Adapter”. In: *Computers & Graphics* 119 (2024), p. 103885.

[206] Zexuan Zhong, Dan Friedman, and Danqi Chen. “Factual probing is [mask]: Learning vs. learning to recall”. In: *arXiv preprint arXiv:2104.05240* (2021).

[207] Kwanyoung Kim, Yujin Oh, and Jong Chul Ye. “Zegot: Zero-shot segmentation through optimal transport of text prompts”. In: *arXiv preprint arXiv:2301.12171* (2023).

[208] Nir Zabari and Yedid Hoshen. “Semantic segmentation in-the-wild without seeing any segmentation examples”. In: *arXiv preprint arXiv:2112.03185* (2021).

[209] Boyi Li et al. “Language-driven Semantic Segmentation”. In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=RriDjddCLN.

[210] Bolei Zhou et al. “Semantic understanding of scenes through the ade20k dataset”. In: *International Journal of Computer Vision* 127 (2019), pp. 302–321.

[211] Roozbeh Mottaghi et al. “The Role of Context for Object Detection and Semantic Segmentation in the Wild”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.

[212] Jindong Gu et al. *A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models*. 2023. arXiv: 2307.12980 [cs.CV]. URL: https://arxiv.org/abs/2307.12980.

[213] Ho Hin Lee et al. *Foundation Models for Biomedical Image Segmentation: A Survey*. 2024. arXiv: 2401.07654 [cs.CV].

[214] Kaiming He et al. “Masked Autoencoders Are Scalable Vision Learners”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 16000–16009. URL: https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.

[215] Yanghao Li et al. “Exploring Plain Vision Transformer Backbones for Object Detection”. In: *Computer Vision – ECCV 2022*. Vol. 13669. Lecture Notes in Computer Science. Springer, 2022, pp. 280–296. DOI: 10.1007/978-3-031-20077-9_17. URL: https://link.springer.com/chapter/10.1007/978-3-031-20077-9_17.

[216] Tsung-Yi Lin et al. “Focal Loss for Dense Object Detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.2 (2020), pp. 318–327. DOI: 10.1109/TPAMI.2018.2858826. URL: https://doi.org/10.1109/TPAMI.2018.2858826.

[217] Keyan Chen et al. “RSPrompter: Learning to Prompt for Remote Sensing Instance Segmentation Based on Visual Foundation Model”. In: *IEEE Transactions on Geoscience and Remote Sensing* PP (Jan. 2024), pp. 1–1. DOI: 10.1109/TGRS.2024.3356074.

[218] Can Cui et al. “All-in-SAM: from Weak Annotation to Pixel-wise Nuclei Segmentation with Prompt-based Finetuning”. In: *Journal of Physics: Conference Series*. Vol. 2722. Conference version of arXiv:2307.00290. 2024, p. 012012. DOI: 10.1088/1742-6596/2722/1/012012.

[219] Reid Swanson and Andrew S Gordon. “Say anything: A massively collaborative open domain story writing companion”. In: *Interactive Storytelling: First Joint International Conference on Interactive Digital Storytelling, ICIDS 2008 Erfurt, Germany, November 26-29, 2008 Proceedings 1*. Springer. 2008, pp. 32–40.

[220] Tianhe Ren et al. “Grounded sam: Assembling open-world models for diverse visual tasks”. In: *arXiv preprint arXiv:2401.14159* (2024).

[221] Yi Li et al. “A closer look at the explainability of Contrastive language-image pre-training”. In: *Pattern Recognition* 162 (2025), p. 111409. DOI: 10.1016/j.patcog.2025.111409. URL: https://www.sciencedirect.com/science/article/pii/S003132032500069X.

[222] Teng Wang et al. *Caption Anything: Interactive Image Description with Diverse Multimodal Controls*. 2023. arXiv: 2305.02677 [cs.CV]. URL: https://arxiv.org/abs/2305.02677.

[223] Wei Wang. *Advanced Auto Labeling Solution with Added Features*. https://github.com/CVHub520/X-AnyLabeling. CVHub, 2023.

[224] Feng Li et al. “Segment and Recognize Anything at Any Granularity”. In: *Computer Vision – ECCV 2024*. Vol. 15106. Lecture Notes in Computer Science. Springer, 2024, pp. 467–484. DOI: 10.1007/978-3-031-73195-2_27. URL: https://link.springer.com/chapter/10.1007/978-3-031-73195-2_27.

[225] Junlong Cheng et al. “Sam-med2d”. In: *arXiv preprint arXiv:2308.16184* (2023).

[226] Jielu Zhang et al. “Text2Seg: Zero-shot Remote Sensing Image Semantic Segmentation via Text-Guided Visual Foundation Models”. In: *Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GeoAI'24)*. Published version of arXiv:2304.10597. 2024, pp. 63–66. DOI: 10.1145/3687123.3698287.

[227] Lucas Prado Osco et al. “The Segment Anything Model (SAM) for remote sensing applications: From zero to one shot”. In: *International Journal of Applied Earth Observation and Geoinformation* 124 (2023), p. 103540. DOI: 10.1016/j.jag.2023.103540.

[228] Tal Shaharbany et al. “AutoSAM: Adapting SAM to Medical Images by Overloading the Prompt Encoder”. In: *34th British Machine Vision Conference (BMVC) 2023*. Peer-reviewed conference version of arXiv:2306.06370. BMVA Press, 2023. URL: https://papers.bmvc2023.org/0530.pdf.

[229] Bin Xie et al. *MaskSAM: Towards Auto-prompt SAM with Mask Classification for Medical Image Segmentation*. 2024. arXiv: 2403.14103 [cs.CV]. URL: https://arxiv.org/abs/2403.14103.

[230] Junde Wu et al. *One-Prompt to Segment All Medical Images*. en. Dec. 2023. URL: http://arxiv.org/abs/2305.10300 (visited on 02/07/2024).

[231] Kaidong Zhang and Dong Liu. *Customized Segment Anything Model for Medical Image Segmentation*. Oct. 2023. URL: http://arxiv.org/abs/2304.13785 (visited on 02/26/2024).

[232] Qiurui Ma et al. “SAM-Industry: Foundation Vision Model for Industrial Defect Detection”. In: *SSRN Electronic Journal* (2024). Preprint; available at SSRN. DOI: 10.2139/ssrn.4944325.

[233] Israt Zarin Era et al. “An Unsupervised Approach Towards Promptable Defect Segmentation in Laser-Based Additive Manufacturing by Segment Anything”. In: *arXiv preprint arXiv:2312.04063* (2023). Unsupervised clustering prompts for porosity segmentation using SAM.

[234] Haoxiang Wang et al. “SAM-CLIP: Merging Vision Foundation Models Towards Semantic and Spatial Understanding”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2024, pp. 3635–3647.

[235] Jay N. Paranjape et al. “AdaptiveSAM: Towards Efficient Tuning of SAM for Surgical Scene Segmentation”. In: *Medical Image Understanding and Analysis*. Vol. 14532. Lecture Notes in Computer Science. Formal version of arXiv:2308.03726. Springer, 2024, pp. 187–201. DOI: 10.1007/978-3-031-66958-3_14.

[236] Xian Lin et al. “Beyond Adapting SAM: Towards End-to-End Ultrasound Image Segmentation via Auto Prompting”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024*. Vol. 15008. Lecture Notes in Computer Science. Springer, 2024, pp. 24–34. DOI: 10.1007/978-3-031-72111-3_3. URL: https://link.springer.com/chapter/10.1007/978-3-031-72111-3_3.

[237] Bowen Ren et al. “WebSAM-Adapter: Adapting Segment Anything Model for Web Page Segmentation”. In: *Advances in Information Retrieval*. Springer Nature Switzerland, 2024, pp. 439–454. ISBN: 9783031560279. DOI: 10.1007/978-3-031-56027-9_27. URL: http://dx.doi.org/10.1007/978-3-031-56027-9_27.

[238] Runnan Chen et al. “Towards label-free scene understanding by vision foundation models”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[239] Muzhi Zhu et al. “SegPrompt: Boosting Open-World Segmentation via Category-Level Prompt Learning”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Conference version of arXiv:2308.06531. 2023, pp. 999–1008. DOI: 10.1109/ICCV51070.2023.00098.

[240] Nikhila Ravi et al. *SAM 2: Segment Anything in Images and Videos*. 2024. arXiv: 2408.00714 [cs.CV]. URL: https://arxiv.org/abs/2408.00714.

[241] Chaitanya Ryali et al. “Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles”. In: *Proceedings of the 40th International Conference on Machine Learning (ICML '23)*. Oral presentation. PMLR, 2023, —. URL: https://proceedings.mlr.press/v202/ryali23a/ryali23a.pdf.

[242] Jun Ma et al. “Segment anything in medical images”. In: *Nature Communications* 15.1 (Jan. 2024). ISSN: 2041-1723. DOI: 10.1038/s41467-024-44824-z. URL: http://dx.doi.org/10.1038/s41467-024-44824-z.

[243] Jiaxing Huang et al. *Learning to Prompt Segment Anything Models*. en. arXiv:2401.04651 [cs]. Jan. 2024. URL: http://arxiv.org/abs/2401.04651 (visited on 02/07/2024).

[244] Bennett A. Landman et al. “MICCAI Multi-Atlas Labeling Beyond the Cranial Vault — Workshop and Challenge”. In: *Proceedings of MICCAI (Multi-Atlas Labeling Beyond the Cranial Vault) Workshop Challenge*. 2015.

[245] Yuanfeng Ji et al. “AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 36722–36732.

[246] Neeraj Kumar et al. “A Multi-Organ Nucleus Segmentation Challenge”. In: *IEEE Transactions on Medical Imaging* 39.5 (2019), pp. 1380–1391.

[247] Haoyu Wang et al. *SAM-Med3D: Towards General-purpose Segmentation Models for Volumetric Medical Images*. Accepted for ECCV BIC 2024 Oral. 2024. arXiv: arXiv:2310.15161.

[248] Chongkai Yu et al. “SAM-REF: Introducing Image-Prompt Synergy during Interaction for Detail Enhancement in the Segment Anything Model”. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 19356–19365.

[249] Yanyuan Qiao, Chaorui Deng, and Qi Wu. “Referring Expression Comprehension: A Survey of Methods and Datasets”. In: *IEEE Transactions on Multimedia* 23 (2021), pp. 4426–4440. DOI: 10.1109/TMM.2020.3042066.

[250] Swami Sankaranarayanan et al. “Learning From Synthetic Data: Addressing Domain Shift for Semantic Segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[251] Yawei Luo et al. “Taking a Closer Look at Domain Shift: Category-Level Adversaries for Semantics Consistent Domain Adaptation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[252] Zhenghang Yuan et al. “RRSIS: Referring Remote Sensing Image Segmentation”. In: *IEEE Transactions on Geoscience and Remote Sensing* (2024).

[253] Sihan Liu et al. “Rotated Multi-Scale Interaction Network for Referring Remote Sensing Image Segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 26658–26668.

[254] Shuyi Ouyang et al. *LSMS: Language-guided Scale-aware MedSegmentor for Medical Image Referring Segmentation*. 2024. arXiv: 2408.17347 [cs.CV]. URL: https://arxiv.org/abs/2408.17347.

[255] Minhyun Lee et al. “MaskRIS: Semantic Distortion-aware Data Augmentation for Referring Image Segmentation”. In: *arXiv preprint arXiv:2411.19067* (2024).

[256] Seongsu Ha et al. “Finding NeMo: Negative-Mined Mosaic Augmentation for Referring Image Segmentation”. In: *Computer Vision – ECCV 2024*. Ed. by Aleš Leonardis et al. Cham: Springer Nature Switzerland, 2025, pp. 121–137. ISBN: 978-3-031-72890-7.

[257] Victor Weixin Liang et al. “Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 17612–17625.

[258] Pallabi Dutta, Anubhab Maity, and Sushmita Mitra. “Prompt-based Dynamic Token Pruning to Guide Transformer Attention in Efficient Segmentation”. In: *arXiv preprint arXiv:2506.16369* (2025). URL: https://arxiv.org/abs/2506.16369.

[259] Xin Zhou et al. “Dynamic Adapter Meets Prompt Tuning: Parameter-Efficient Transfer Learning for Point Cloud Analysis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 14707–14717. URL: https://openaccess.thecvf.com/content/CVPR2024/html/Zhou_Dynamic_Adapter_Meets_Prompt_Tuning_Parameter-Efficient_Transfer_Learning_for_Point_CVPR_2024_paper.html.

[260] Krzysztof Choromanski et al. “Rethinking Attention with Performers”. In: *International Conference on Learning Representations (ICLR)*. 2021. URL: https://arxiv.org/abs/2009.14794.

[261] Sachin Mehta and Mohammad Rastegari. “MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer”. In: *International Conference on Learning Representations (ICLR)*. 2022. URL: https://arxiv.org/abs/2110.02178.

[262] Mellisa S Damodaram et al. “Foetal volumetry using magnetic resonance imaging in intrauterine growth restriction”. In: *Early human development* 88 (2012), S35–S40.

[263] Mingxing Tan. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *arXiv preprint arXiv:1905.11946* (2019).