Rethinking Cross-Modal Interaction for Efficient Referring Image Segmentation

Rethinking Cross-Modal Interaction for Efficient Referring Image Segmentation

Claudia Cuttano, Francesca Pistilli, Fabio Cermelli and Giuseppe Averta

***Abstract*—Referring Image Segmentation, the task of finding and segmenting objects in an image conditioned on a natural language description, is crucial for human-robot collaboration. However, current RIS methods often implement visual-text alignment relying on computationally intensive Transformer-based self-attention mechanisms, which impairs deployment on robots, especially those with limited computational resources. Indeed, beyond accuracy, practical robotic applications demand efficient models with small footprints. This paper introduces ERIS, an Efficient RIS approach designed for real-world deployment. ERIS achieves effective multi-modal interaction through a novel dual-branch architecture: a Visual Text Alignment branch and a Text Visual Refinement branch. This design implements bilateral alignment between textual and visual modalities without the computational burden of self-attention. Of note, the progressive alignment in ERIS enhances interpretability, revealing how textual cues guide segmentation. For the sake of efficiency, our alignment strategy produces structured embeddings which can be directly mapped into the final segmentation mask, without the need for additional segmentation heads. Thus, ERIS footprint scales linearly with respect to the number of visual and text tokens, making it suitable for both cloud-based and edge deployment. Experimental results demonstrate that ERIS achieves competitive or superior performance compared to state-of-the-art methods while significantly reducing computational cost, proving that efficiency and accuracy are not mutually exclusive.**

***Index Terms*—AI-based methods, Deep Learning for Visual Perception, Referring Image Segmentation.**

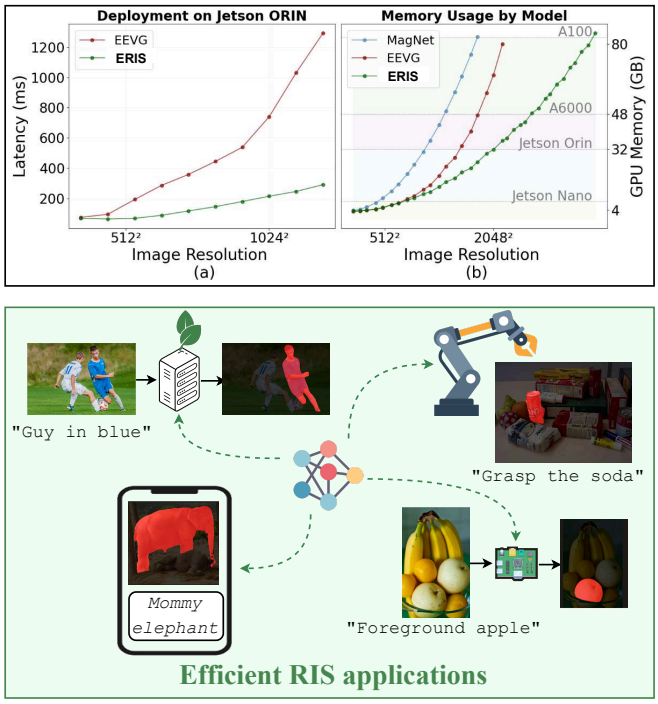

Fig. 1. ERIS delivers state of the art accuracy in Referring Image Segmentation while minimising computational demand, enabling edge applications. The plots in the top section report latency (a) and GPU memory footprint (b) with varying resolutions, comparing with previous RIS methods. The charts show ERIS better scalability, in terms of both time and space complexity.

I. Introduction

Referring Image Segmentation (RIS) models enable intelligent machines to understand natural language queries, and act accordingly by locating and segmenting an object given a textual prompt. This capability is pivotal for a large variety of applications, including human-robot interaction [1], autonomous navigation [2], [3], and robotic manipulation [4]–[6] (see also Fig. 1). To solve the task, the model is required to generate multi-modal representations by fusing textual and visual features extracted from dedicated backbones. To do this, state of the art RIS solutions [7]–[13] typically rely on heavy Transformer architectures [14], by concatenating linguistic and visual tokens - of length L and N respectively - and then applying self-attention (SA) [14] to extract cross-modal dependencies. However, as highlighted by [15], the quadratic complexity of SA, $\mathcal{O}((N+L)^2)$, makes such methods computationally expensive. Such complexity is limiting, if not unbearable, for many real-world applications, both on the cloud [16], where inference cost and energetic impact are not negligible, and on the edge, where computational resources are intrinsically limited. To unlock the potential of RIS on real-world usecases, it is therefore crucial to develop models which optimize the model accuracy-efficiency trade-off.

To this end, we propose a new paradigm for RIS that explicitly rethinks multi-modal interaction to improve efficiency without sacrificing segmentation performance. Our idea stems from the observation that the main requirement for a robust RIS method is the ability to align visual and text modalities into a shared space. Several influential works [17], [18] studied the problem of visual-text alignment for representation learning, demonstrating the effectiveness of cross-attention (CA)

Manuscript received February 12 2025; Revised May 1 2025; Accepted May 29 2025. This paper was recommended for publication by Editor Markus Vincze upon evaluation of the Associate Editor and Reviewers' comments.

This work was supported by the Sustainable Mobility Center (CNMS) which received funding from the European Union Next Generation EU (Piano Nazionale di Ripresa e Resilienza (PNRR), Missione 4 Componente 2 Investimento 1.4 "Potenziamento strutture di ricerca e creazione di "campioni nazionali di R&S" su alcune Key Enabling Technologies") with grant agreement no. CN_00000023.

The authors are with the Visual and Multimodal Applied Learning Lab, Department of Control and Computer Engineering, Politecnico di Torino, 10138 Torino, Italy (e-mail: claudia.cuttano@polito.it; francesca.pistilli@polito.it; fabio.cermelli@polito.it; giuseppe.averta@polito.it).

Digital Object Identifier: see top of this page.

in grounding vision and language representation [18], which comes with the benefit of linear scalability w.r.t. N and L. At the same time, the issue of redundant computations in SA has been also extensively investigated in the last few years, with research demonstrating an accumulation of redundancy through low-entropy representations and repetitive token patterns [19], [20]. Accordingly, we argue that full pairwise token interactions of SA in Referring Image Segmentation methods may not be necessary, and analogous results can be achieved also through a more structured and efficient use of CAs.

Building on this intuition, we propose to decompose the multi-modal interaction in two distinct yet interconnected branches: the Visual-Text Alignment and the Text-Visual Refinement branch. The two branches are designed to progressively incorporate multi-modal cues, to achieve bidirectional alignment of the two modalities without resorting to computationally heavy SA layers. With the first branch, visual features are contaminated with textual semantics via cross-attention in an early fusion module. These visual features are then processed by our novel Deformable Cross-Modal Feature Pyramid Network (FPN) module, where the sentence embedding guides the fusion of multi-resolution features by highlighting relevant pixels while suppressing background tokens. With the Text-Visual Refinement, the sentence embedding is refined by attending visual features at different resolutions, obtaining a scene-specific sentence representation. Interestingly, this intertwined alignment process improves interpretability, as we show throughout the paper that our embeddings reveal explicit correspondences between image elements and the referring query, providing insight into the model decision-making process. Lastly, since the visual and linguistic features are well-aligned, their structure can be leveraged to directly obtain the output mask via a simple dot product, eliminating the need for an additional segmentation head. Our proposed architecture, named **ERIS**, is purposefully designed for **E**fficient **R**eferring **I**mage **S**egmentation. We evaluate ERIS on RefCOCO, RefCOCO+, and RefCOCOg to assess its performance across different levels of textual query difficulty. This includes testing with simple expressions that indicate location (RefCOCO), expressions without location cues (RefCOCO+), and longer, more complex descriptions (RefCOCOg). ERIS achieves competitive and even superior performance (+0.4% on average) compared to prior state-of-the-art methods, while significantly improving efficiency in terms of GFLOPs overhead and latency. Unlike existing Referring Image Segmentation approaches that suffer from quadratic complexity, our method scales linearly with respect to the number of visual and linguistic tokens, making it a robust and scalable solution for both edge and cloud applications.

To summarize, our paper contributes with the following:

- We introduce ERIS, a novel RIS architecture that splits the multi-modal interaction in two dedicated lightweight branches, enabling bilateral alignment between visual and linguistic features without the quadratic complexity of traditional self-attention layers.
- ERIS is significantly lighter than existing solutions and showcases excellent linear scalability w.r.t. the number of visual and linguistic tokens. This enables deployment on resource-constrained devices and supports applications where high-resolution images are critical.
- Our model achieves state-of-the-art results on all benchmarks tested, with a lighter architecture, demonstrating that striving for efficiency does not necessarily require to trade-off performances.

II. Related works

Referring Image Segmentation (RIS). First defined in [21], RIS aims to segment the referred object given a natural language expression. Previous works either focus on vision and linguistic feature extraction, using CNNs [21]–[23] and recurrent neural networks [23]–[25]. The breakthrough of Transformers [14] has inspired numerous works [7]–[11], [26]–[29] to employ attention mechanisms to align visual and textual modalities. A cornerstone of modern approaches is represented by MDETR [7], which introduces a novel paradigm by concatenating visual and linguistic features and feeding them to a Transformer encoder-decoder architecture to fuse modalities. Several works follow its footsteps; [8], [11] enhance this approach by exploiting the linguistic feature to generate input-specific queries, while [9], [10] feed the concatenated features to a regression-based encoder-decoder to generate the mask as a sequence of polygon vertices. Others [12], [13] propose an encoder-only architecture to encode the two modalities. These approaches all leverage self-attention operations to align modalities, resulting in a quadratic complexity, which hinders their scalability under resource constrained scenarios. On the contrary, our focus is on efficient cross-modal alignment for RIS.

Efficient Image Segmentation. Efficient segmentation is crucial for robotic perception in various applications, including robot-assisted surgery [30], [31], segmentation from aerial vehicles such as UAVs and MAVs [32], [33], and autonomous driving systems [34], [35]. Given such request, substantial research has been dedicated to efficient segmentation architectures [36]–[39]. While most are unimodal, some methods incorporate additional modalities like depth, thermal, or 3D point clouds [40]–[42]. However, efficient segmentation with vision-language modalities remains underexplored. Recently, in RIS, [15] proposes a decoder-only framework to limit the cost of cross-modal interaction, carried out through cross-attention, which scale linearly w.r.t. the input sequence. However, it remains quadratic w.r.t. number of visual tokens, which is order of magnitude larger than the number of linguistic tokens. Differently, we propose an architecture which promotes bilateral alignment between modalities based on linear cross-attention operations.

III. Method

Overview. The core challenge in RIS lies in modeling the interactions between visual and linguistic modalities, which together define the target object in an image. To tackle this problem, recent approaches concatenate visual and textual features and pass them through transformer blocks equipped with stacked self-attention layers [14]. This approach lacks

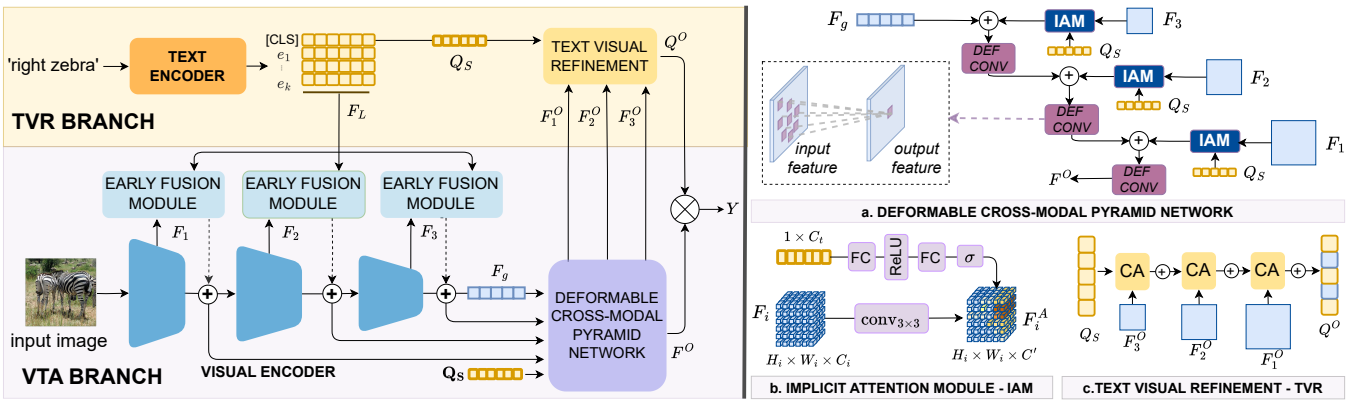

Fig. 2. **Overview of ERIS.** We rethink modality interaction by proposing a framework based on bilateral alignment. We design two complementary branches, Visual-Text Alignment (VTA) and Text-Visual Refinement (TVR), which progressively align visual and linguistic features. The VTA branch incrementally injects textual cues into the visual feature starting from the initial encoder layers, enabling early semantic focus on relevant image regions. At the end of feature extraction, this branch incorporates the novel Deformable Cross-Modal Pyramid Network (right-a), which employs deformable convolutions to upsample feature maps guided by textual information via our Implicit Attention Module (right-b). The TVR branch (right-c) refines textual features with visual context in a masked cross-attention framework, generating a scene-specific sentence representation. Finally, the output features from both branches are aligned, and can thus be combined to produce the segmentation mask through a simple dot product.

explicit guidance of the modality fusion process, delegating to the SA the task of extracting dependencies across the whole set of tokens. This redundant full-token interaction ends up being quadratic w.r.t. the number of tokens. Conversely, we rethink the cross-modal interaction paradigm as a progressive alignment process, where the visual and textual features are gradually brought into alignment. To this end, ERIS leverages two complementary branches: Visual-Text Alignment (VTA) and Text-Visual Refinement (TVR). In the former, (section III-A) visual features are progressively aligned with linguistic cues from the feature extraction process. On the other hand, TVR (section III-B) refines textual features using visual context, adapting the linguistic representation to the specific scene. As both modalities interact continuously, the final segmentation map can then be derived directly from these aligned embeddings (section III-C), eliminating the need for additional segmentation heads which bring additional computational overhead. Finally, in section III-D we discuss the time complexity of our solution. Our architecture is shown in fig. 2.

Problem setting. Given an image $x \in \mathbb{R}^{H\times W\times 3}$, where H and W denote the height and width, and a tokenized referring expression $t \in \mathbb{R}^{1\times L}$, with L being the number of tokens, our goal is to generate a segmentation mask $Y \in [0,1]^{H\times W}$. The text is tokenized and then augmented by adding a global sentence representation token `[CLS]`. The tokens are then processed using a language encoder [43] to extract the language features $F_L \in \mathbb{R}^{L\times C_t}$ where C_t is the number of channels.

A. Visual-Text Alignment (VTA)

Early Fusion Module. Given an image x, and a hierarchical visual encoder, we propose to guide the feature extraction process from the very beginning with linguistic cues. Starting from the first encoder block, we apply a cross-attention based mechanism to facilitate direct interaction between the visual and textual modalities, enabling the model to selectively attend to different parts of the sentence while processing the visual input. Given the visual feature maps $F_i \in \mathbb{R}^{H_i\times W_i\times C_i}$ from layer i, and textual embeddings F_L, we project both of them in a shared C-dimensional space to obtain, respectively, queries $Q_i \in \mathbb{R}^{H_i\times W_i\times C}$, keys and values $K, V \in \mathbb{R}^{L\times C}$. After the visual queries attend to the textual token, the output values are projected back through W_{up} to the initial C_i dimensionality, and can thus be used to reweight F_i, maintaining a balance between the original visual signal and the injected linguistic cues. Formally, we define the Early Fusion Module as:

$$F_i = F_i + F_i \odot \mathrm{W}_{\text{up}}\left(\texttt{softmax}\left(Q_i \cdot K^T\right) V\right),$$

where $\odot$ is the Hadamard product. This *language-aware* features maps are propagated through the encoder, repeating this process at each layer. As depicted in fig. 3a, the features (F_1, F_2, and F_3) progressively highlight candidate regions where the referred object is likely to be located.

Deformable Cross-Modal Pyramid Network. After extracting visual features, a common approach is to combine them with linguistic tokens in order to jointly process the two modalities with a self-attention. This approach allows visual tokens to capture both cross-modal cues (from linguistic tokens) and intra-modal information (within visual tokens) in a unified framework. However, performing self-attention over all tokens entails a significant computational cost. To overcome this limitation while preserving the dual capability of cross-modal interaction and self-modality refinement, we propose a novel Deformable Cross-Modal Pyramid Network (fig. 2a), made up of i) a cross-modal interaction mechanism that enhances visual features with linguistically relevant information and ii) intra-modal interaction through deformable convolutions, enabling adaptive feature refinement across varying resolutions.

1) Cross-modal interaction: In the early fusion stages, visual features attend to all linguistic tokens, resulting in enriched features that are structured to highlight candidate regions based on similarities with each token of the textual query. However, to identify the correct object, contextual words or descriptive elements might introduce ambiguity. To filter out unnecessary noise, and reduce the computational

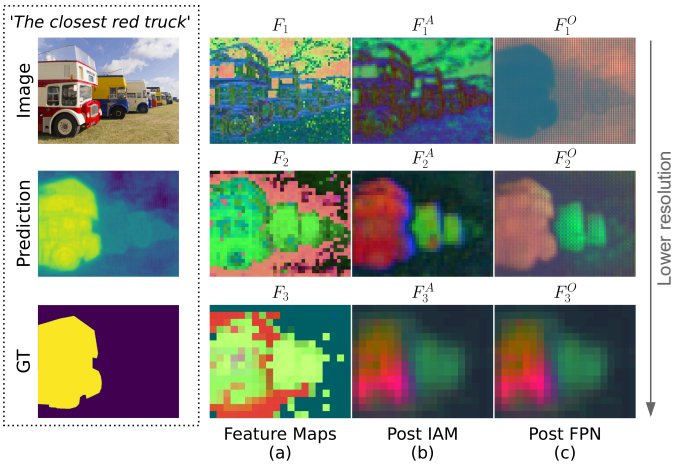

Fig. 3. Visual exemplification of **ERIS alignment process**. Given the input image and its caption in the top left, we visualize PCA of feature maps at different stages of the pipeline. After feature extraction (col. 'a'), thanks to early contamination with textual embeddings, the feature maps show a progressive bias towards candidate regions, with the shallower features retaining more spatial details. Col. 'b' shows how our IAM module promotes text-aligned features, thereby suppressing the background. Finally, the FPN (col. 'c') progressively refines the deeper features, enriching them with the spatial details from shallower layers, to obtain the final prediction.

burden as well, we employ the `[CLS]` token, which represents a summary of the query, capturing the relevant high-level semantic information about the referred object [43].

To exploit the information of the `[CLS]` token, we introduce the Implicit Attention Module (IAM), depicted in fig. 2b, which enables a channel-wise modulation of the visual features. Specifically, we derive Q'_s, F'_i by projecting respectively the `[CLS]` embedding $Q_S \in \mathbb{R}^{1\times C_t}$, and $F_i \in \mathbb{R}^{H_i\times W_i\times C_i}$ into a common low-dimensional space C'. We then apply a sigmoid function σ on Q'_s to retain only the most informative channels for the given caption, and finally perform the Hadamard product with F'_i:

$$F_i^A = F'_i \odot \sigma(Q'_s)$$

This module, inspired by [44], serves as an implicit attention mechanism by: *i)* weighting each visual channel based on the relative importance in the `[CLS]` token; and *ii)* suppressing irrelevant regions, thereby achieving clearer object-background separation. We incorporate multiple IAM at different scales, to produce aligned multi-modal features across the entire pyramid. As shown in fig. 3b, our IAM effectively guides the model focus toward the correct object.

2) Intra-modal interaction: Features after IAM represent multi-modal information at different scales. Reading fig. 3b, features at higher resolution provide fine-grained visual details but are less capable of capturing the semantic depth of the referring text. Conversely, lower resolutions features encode broader contextual information and stronger alignment with the textual representation, but they have limited spatial detail. To allow the network to capture both coarse global context and fine details, we propose to employ deformable convolutions. As the visual features have already been modulated based on the referring expression, deformable convolutions can adjust their receptive fields dynamically, refining the visual features where the referred object is likely to be. For example, if the text refers to an object located in a specific part of the image, the deformable convolution can adaptively sample more features from that region, adjusting its focus to regions of the image that are aligned with the textual query, rather than being constrained by a fixed grid as in standard convolutions.

Specifically, we start by extracting a global visual embedding $F_g \in \mathbb{R}^{C'}$, obtained by pooling the last backbone feature. The feature F_g represents a summary of the whole image. Starting from F_g, lower-resolution features are incrementally fused with higher-resolution ones. Formally, we compute the features F_i^O as:

$$F_3^O = F_3^A + F_g$$
$$F_i^O = F_i^A + \texttt{Up}(\texttt{DefConv}(F_{i+1}^O)), \quad i = 2, 1$$

where `Up` is an upsampling operation that increases the resolution of F_{i+1}^O, while `DefConv` stands for a 2D deformable convolution. Finally, the highest resolution feature map F_1^O is upsampled and projected to obtain F^O:

$$F^O = \texttt{Up}(\texttt{DefConv}(F_1^O)),$$

This bottom-up approach, as visible in fig. 3c, transfers the semantic content from lower-resolution feature maps upwards, where it is gradually merged with higher-resolution features that contribute fine-grained spatial information. The final feature map is both rich in spatial detail, enhancing segmentation quality, and aligned with the textual cues, enabling precise localization of the referred object.

B. Text-Visual Refinement (TVR)

In human perception, sentence meaning adapts dynamically to the accompanying visual scene. For instance, the interpretation of "the brown dog" shifts significantly depending on whether the image depicts an animal or a toy. Following this analogy, ERIS leverages a complementary Text-Visual Refinement (TVR) branch, shown in fig. 2c, with the goal of enhancing textual features by conditioning them on the visual information extracted from the image, enabling a bidirectional alignment between modalities. In detail, we obtain Q^O by refining Q_S, i.e. the embedding of the `[CLS]` token, with three sequential cross-attention (CA) layers, starting from the lowest-resolution (but highly semantic) VTA feature F_3^O and progressively introducing spatial details using F_2^O and F_1^O.

Formally, both the Q_S and visual features F_i^O, with $i \in 3, 2, 1$, are projected in a D-dimensional space to obtain, respectively, query $Q_0 \in \mathbb{R}^{1\times D}$, keys and values $K_i, V_i \in \mathbb{R}^{H_i\times W_i\times D}$ where H_i, W_i denote the respective feature resolutions. Each masked CA is computed as:

$$Q_j = \texttt{softmax}\left(Q_{j-1} \cdot K_i^T\right) V_i + Q_{j-1},$$

where $j \in 1, 2, 3$ and $i \in 3, 2, 1$. Note that the query obtained at step j is then employed at the subsequent step in a residual manner. The output sentence representation Q^O is then set to be equal to the last masked CA output, i.e. $Q^O = Q_3$. To provide insight into our proposed refinement scheme, we plot in fig. 4 the distribution of the cosine similarity among the visual features F^O and the refined textual representation Q_j at various stages. The plot exemplifies how with subsequent

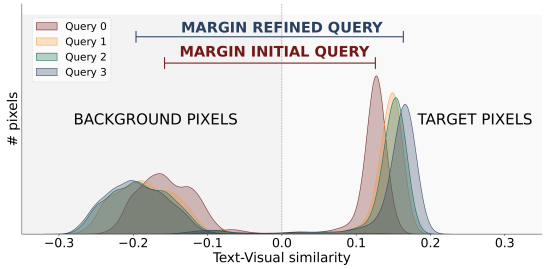

Fig. 4. Effect of **Text-Visual Refinement**. ERIS obtains segmentation masks via a dot product, i.e. pixels in the predicted masks have non-negative cosine similarity with the query. The histogram shows the distribution of target vs background pixels when computing the cosine similarity of output features F^O with the textual query at subsequent refinement steps Q_i. The plot shows how the margin between the two distributions, which can be interpreted as a confidence on the prediction (i.e. how well target pixels are aligned with the query), is increased throughout the process.

refinement steps, the query representation maximizes the margin between positive and negative pixels, *de-facto* increasing the confidence of the prediction.

C. Parameter-Free Prediction Head

ERIS leverages two distinct branches, the Text-Visual Refinement (TVR) and the Visual-Text Alignment (VTA), to create a shared multi-modal space where visual and linguistic semantics are aligned. Given these shared representation, we can compute the final segmentation mask through a simple dot product of the embeddings, eliminating the need for an additional segmentation head [11], [15], [29] or decoder [9], [10]. Formally, we compute the final prediction $Y \in [0,1]^{H \times W}$ by computing the matrix multiplication between the output of the VTA (F^O) and TVR (Q^O) branches:

$$Y = \sigma(Q^O \cdot (F^O)^T),$$

where σ and T denote the sigmoid and transpose operations. This mechanism ensures that all pixels aligned with the textual query are activated based on their semantic relevance.

D. Time Complexity Analysis

Preliminaries. Traditional RIS methods employ a standard Transformer Encoder architecture [14]. By concatenating visual and linguistic tokens as a single sequence of length $(N+L)$, where N and L represent the number of visual and textual tokens, respectively, the attention complexity is:

$$\mathcal{O}((N+L)^2 C)$$

To mitigate this quadratic cost, EEVG [15] proposes a Decoder-only framework. Complexity in the SA layer is reduced by eliminating the linguistic tokens from the query, and the cross-modal interaction is performed through CA. The total complexity, as discussed by the authors [15], is:

$$\mathcal{O}(N^2 C + NLC),$$

where the second term accounts for the CA between visual and linguistic tokens. While this structure achieves a reduction in complexity, it remains quadratic in the number of visual tokens N, which is a significant bottleneck due to the typically larger value of N compared to L.

Ours. ERIS proposes a self-attention-free paradigm for cross-modal interaction to substantially reduce computational complexity. Our method progressively integrates visual and linguistic features through bilateral cross attention operations (Early Fusion Module and TVR modules) and channel-wise operations (IAM). This leads to a complexity of:

$$\mathcal{O}(NLC + NC + NC) = \mathcal{O}(NC(L+2))$$

The first two terms refer to the complexity of Early Fusion Module and TVR, respectively. Notably, the complexity of the standard CA in the TVR is further reduced from $\mathcal{O}(NLC)$ [14] to $\mathcal{O}(NC)$ as we employ a single query token (specifically, the `[CLS]` token). The final term reflects the complexity of the IAM blocks, which consists in multiplying a feature matrix of size $N \times C$ and a C-dimensional vector with a Hadamard product. Notably, our approach achieves linear scalability with respect to both N and L. This complexity analysis considers the theoretical costs of operations that involve multi-modal interactions. In the experimental section, we provide an extensive evaluation on the impact of these design choices on real world scenarios and metrics.

IV. Experiments

Datasets. Following previous works [26]–[29], we evaluate our method on three datasets: RefCOCO [45], RefCOCO+ [45] and RefCOCOg [46]. **RefCOCO** contains 142,209 language expressions for 50,000 objects in 19,994 images. The average length of each expression is 3.6 words. **RefCOCO+** consists of 141,564 expressions for 49,856 objects in 19,992 images. RefCOCO+ presents a greater challenge as it avoids using location words for object reference in its expressions. Finally, **RefCOCOg** includes 104,560 referring expressions for 54,822 objects in 26,711 images. The expressions are longer and more complex, with an average length of 8.4 words.

Evaluation metrics. To evaluate our solution, we follow previous works [8], [11], [26]–[29] adopting mean Intersection-over-Union (mIoU). Additionally, we quantify the computational demand through Floating Point Operations (FLOPs), memory usage (GB) and latency (ms).

Implementation Details. Following [13], [15], [27]–[29], for the main comparison we adopt a Swin-B backbone. Additionally, we experiment with the lighter Swin-T. As text encoder, we adopt BERT [43] with dimension $C_t = 768$. We set $C' = 128$. The network is trained for 50k iterations with a batch size of 32 and image size 640×640. We employ Adam with learning rate $\lambda_v = 1e^{-4}$ for the vision and $\lambda_t = 1e^{-5}$ for the language encoders, with a polynomial decay strategy.

A. State-of-the-art comparison

In table I, we compare ERIS with recent state-of-the-art approaches. Notably, despite being developed with efficiency as a primary focus, ERIS not only maintains state-of-the-art performance but even outperforms previous methods, achieving an average mIoU of 70.3% on the standard benchmark.

Method	Visual Backbone	RefCOCO val	RefCOCO test A	RefCOCO test B	RefCOCO+ val	RefCOCO+ test A	RefCOCO+ test B	RefCOCOg val	RefCOCOg test	Avg (%)
Pretrain on Visual-Genome, Flickr30k										
PolyFormer [10]	Swin-B	76.0	77.1	73.2	70.7	74.5	64.6	69.4	69.9	71.9
Train on RefCOCO/+/g										
SeqTR [9]	DN53	67.3	69.8	64.1	54.1	58.9	48.2	55.7	55.6	59.2
VLT [8]	DN53	65.7	68.3	62.7	55.5	59.2	49.4	53.0	56.7	58.8
REFTR [11]	ResNet101	70.6	73.5	66.6	61.1	64.5	52.7	58.7	58.5	63.3
LAVT [29]	Swin-B	74.5	76.9	70.9	65.8	71.0	59.2	63.3	63.6	68.2
DMMI [27]	Swin-B	74.1	77.1	70.2	57.0	69.7	57.0	63.5	64.2	66.6
GRES [28]	Swin-B	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0	
MagNet [13]	Swin-B	75.2	78.2	71.1	66.2	71.3	58.1	65.4	66.0	68.9
EEVG [15]	Swin-B	**75.8**	77.9	72.8	67.6	71.5	59.1	**67.4**	67.3	69.9
ERIS (ours)	Swin-B	75.3	**78.5**	**72.9**	**67.7**	**72.9**	**60.1**	66.8	**68.0**	**70.3**
EEVG [15]	Swin-T	71.0	72.9	67.7	61.8	66.9	52.7	61.3	62.3	64.6
ERIS (ours)	Swin-T	**72.4**	**75.5**	**70.1**	**63.6**	**69.2**	**55.1**	**64.0**	**64.7**	**66.8**

TABLE I
STANDARD RIS BENCHMARK: COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THREE DATASETS IN TERMS OF MIOU [%].

Remarkably, it also remains competitive against methods that also use additional training data, such as PolyFormer [10]. This finding underscores that prioritizing efficiency does not necessarily come at the cost of accuracy. Notably, with the lighter Swin-T backbone ERIS outperforms EEVG, also designed for efficiency, by 2.1%, 2.2%, 2.6% on average on RefCOCO, RefCOCO+ and RefCOCOg. This highlights ERIS adaptability to smaller, lighter architectures, making it adaptable for deployment in resources constrained scenarios.

B. Efficiency Analysis

Computational Overhead. Our main goal is to design a framework for visual-text feature alignment with as little overhead as possible beside feature extraction. We compare in fig. 5a the amount of computation (in GFLOPs) that various RIS methods introduce on top of the backbone at 640p. ERIS adds just 25 GFLOPs, which is half the overhead of the closest competitor EEVG, while also improving in performance. The overhead of ERIS is negligible w.r.t. the feature extraction stage, whereas previous methods that do not prioritize efficiency end up doubling the total FLOPs count (as a reference, Swin-B comes with 147 GFLOPs).

Model scale. In fig. 5b and in table II we analyze the trade-off between model size and performance, showing the superior adaptability of our method to smaller backbones. With Swin-T, ERIS strikes an optimal balance that, with only 63 GFLOPs, maintains competitive performance, losing less than 4% in mIoU. Lastly, paired with a lightweight backbone, STDC [39], ERIS yields an ultra-compact model (28 GFLOPs) that still guarantees robust performance (59.7% mIoU), outperforming prior methods by a wider margin under constrained compute. Notably, its extreme compactness enables deployment at 640p on a Raspberry Pi (4GB RAM).

Hardware Deployments. In table III, we benchmark ERIS and prior methods on a Jetson Orin. With Swin-B, ERIS achieves SOTA performances while improving fps of +25% w.r.t the best competitor EEVG. Employing smaller backbones such as Swin-T and STDC further boost efficiency, achieving 12 and 24 fps respectively, enabling practical edge deployment. Finally, we provide an analysis in terms of scalability

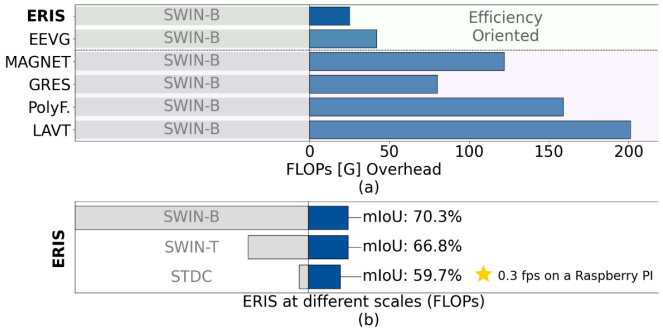

Fig. 5. Computational efficiency and scalability of ERIS. (a) Overhead (in GFLOPs) introduced by different RIS methods w.r.t the backbone. (b) Trade-off between model scale and performance, demonstrating that ERIS enables lightweight models while maintaining strong accuracy.

	Backbone	RC	RC+	RCg
ERIS	RN18	65.7	54.6	54.9
LAVT	STDC	53.6	35.0	20.5
GRES	STDC	56.4	41.2	42.0
DMMI	STDC	63.9	51.6	51.3
ERIS	STDC	67.1	56.0	55.9
EEVG	Swin-T	71.0	61.8	61.3
ERIS	Swin-T	72.4	63.6	64.0

TABLE II
EXPERIMENTS WITH EFFICIENT BACKBONES ON THE VAL SPLIT OF REFCOCO (RC) SERIES.

	Backbone	mIoU	FLOPs	FPS
GRES	Swin-B	68.3	227	2.9
MagNet	Swin-B	68.9	269	1.6
EEVG	Swin-B	69.9	189	3.9
ERIS	Swin-B	**70.3**	**172**	**4.8**
EEVG	Swin-T	64.6	80	**6.3**
ERIS	Swin-T	**66.8**	**63**	**12.2**
GRES	STDC	47.1	83	7.4
ERIS	STDC	**59.7**	**28**	**23.6**

TABLE III
COMPARISON IN TERMS OF MIOU, FLOPS, AND FPS ON NVIDIA JETSON ORIN.

to hardware requirements. In fig. 1a, we deploy ERIS and EEVG [15] on a Jetson Orin and evaluate latency w.r.t. image resolution (i.e., the number of visual tokens N). The plot validates experimentally the superior scalability of our framework, showing that ERIS can run at 12 fps at resolution of 640, whereas EEVG drops to 6 fps. This gap dramatically increases with resolution. Furthermore, in fig. 1b, we evaluate ERIS, EEVG [15] and MagNet [13] in terms of memory footprint against image size. The plot reports the maximum resolution processed by each model without exceeding device memory. We show that, because of the quadratic dependency on N, at 2048p MagNet exceeds the 80 GB of an A100, whereas ERIS can fit on a Jetson Orin for the same input. These results highlight the versatility and scalability of our

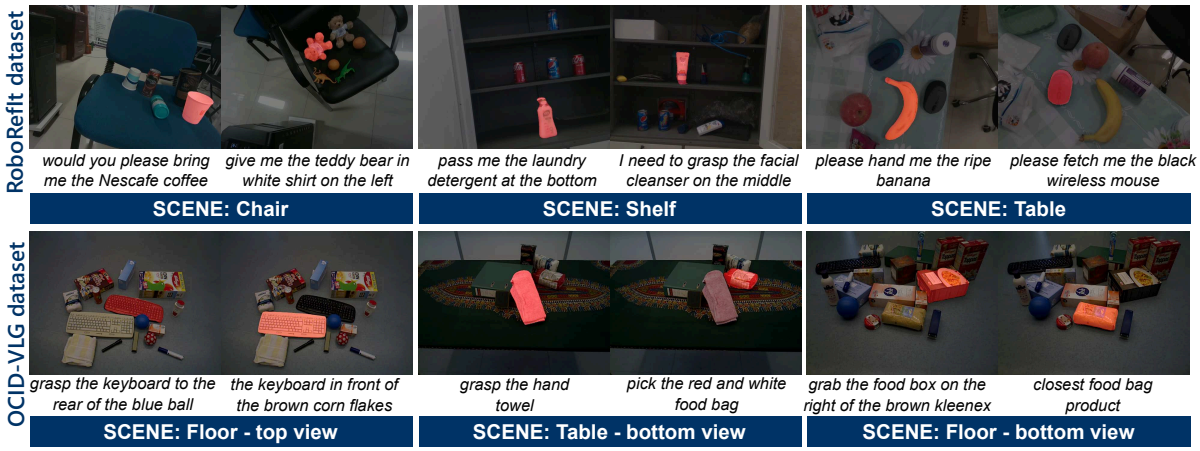

Fig. 6. **Qualitative results** of ERIS in zero-shot settings on RoboRefIt [4] and OCID-VLG [6]. Equipped with the most lightweight backbone, STDC [39], ERIS is queried with different natural language expressions to segment graspable objects across different scenes, viewpoints, and cluttered environments.

model, making it an ideal solution for edge deployments, as well as for applications in domains where high-resolution processing is critical.

C. Qualitative results and Ablation studies

Qualitative results. To further evaluate ERIS, we conduct qualitative experiments in zero-shot (i.e., without fine-tuning) on the RoboRefIt [4] and OCID-VLG [6] datasets, in fig. 6. Both datasets focus on referring grasp synthesis, where a robot must predict a grasp pose for an object specified via natural language. On RoboRefIt, the model handles diverse scenes with objects placed on chairs, shelves, or tables, while OCID-VLG involves objects on the floor or table from various perspectives, including top-down and bottom-up views. ERIS excels in disambiguating similar objects using attributes like color, location, or context, especially in cluttered environments with multiple instances of the same category, such as keyboards, teddy bears, or mouses.

Visual-Text Alignment branch. In table IV-a, we evaluate components of our Visual-Text Alignment branch on RefCOCO+ testA. Removing the Implicit Alignment Module leads to a drop to 61.3%. Replacing IAM with spatial attention via Multi-Head Cross-Attention (MHCA) yields only a modest gain (62.5%). In contrast, IAM improves performance (69.2%) by reweighting *channel-wise* the features based on their semantic relevance. We further evaluate IAM at different backbone stages (F_1^A, F_2^A, F_3^A), observing improved results with deeper features. Ablations of the Deformable Cross-Modal Network show that removing the global visual context F_g reduces performance by 1.1%, and replacing deformable convolutions with standard ones leads to a drop of 5%. In table IV-b, we examine the impact of early fusion. Introducing linguistic features early consistently outperforms the no-fusion baseline (66.1 $\rightarrow$ 69.2). Notably, gains decrease with earlier fusion points: F_3 (+1.4), F_2 (+1.2), F_1 (+0.5), suggesting that deeper features benefit more from semantic guidance.

Text-Visual Refinement branch. In table IV-c we evaluate the effectiveness of our TVR, investigating how refining the linguistic query from multi-scale visual features F_1^O, F_2^O, F_3^O affects mIoU. We first evaluate the performance of ERIS without query refinement, where the prediction is obtained by the product between F_O and the projected vector Q_0. This approach results in a 3.7% decrease in mIoU. Using a single refinement step with the lowest-resolution feature F_3^O improves results of 1.1% mIoU. Incorporating F_2^O further improves the performance (+0.9%), providing spatial visual details. Finally, using three refinements additionally improves the results (+1.7%), benefiting from both the enriched visual information and fine-grained details.

Visual-Text Alignment

		mIoU
IAM	w/o IAM	61.3
	on F_1^A	61.9
	on F_2^A	66.9
	on F_3^A	67.6
	MHCA	62.5
FPN	w/o visual context	68.1
	w/o deformable	64.3
		69.2

(a)

Visual-Text Alignment

F_3	F_2	F_1	mIoU
			66.1
✓			67.5
✓	✓		68.7
✓	✓	✓	**69.2**

(b)

Text-Visual Refinement

Q_0	Q_1	Q_2	Q_3	mIoU
✓				65.5
✓	✓			66.6
✓	✓	✓		67.5
✓	✓	✓	✓	**69.2**

(c)

TABLE IV
ABLATION STUDY OF VISUAL-TEXT ALIGNMENT (A) AND (B), AND TEXT-VISUAL REFINEMENT (C).

V. CONCLUSIONS

In this work, we reframe multi-modal interaction as a bidirectional alignment problem, reducing computational cost through efficient cross-attention. ERIS progressively aligns modalities, ensuring interpretability while enabling mask retrieval via a simple dot product, eliminating the need for a segmentation head. We demonstrate that searching for model efficiency does not necessarily come at the expenses of performance, by showing that our method achieves state-of-the-art results at a fraction of the computational cost, and that it can seamlessly deployed on low-power devices.

REFERENCES

[1] Y. Iioka, Y. Yoshida, Y. Wada, S. Hatanaka, and K. Sugiura, "Multimodal diffusion segmentation model for object segmentation from manipulation instructions," in *IEEE/RSJ Int. Conf. on Intel. Robots and Syst.* IEEE, 2023, pp. 7590–7597.

[2] N. Rufus, K. Jain, U. K. R. Nair, V. Gandhi, and K. M. Krishna, "Grounding linguistic commands to navigable regions," in *IEEE/RSJ Int. Conf. on Intel. Robots and Syst.* IEEE, 2021, pp. 8593–8600.

[3] K. Jain, V. Chhangani, A. Tiwari, K. M. Krishna, and V. Gandhi, "Ground then navigate: Language-guided navigation in dynamic scenes," in *IEEE Int. Conf. on Robot. and Autom.* IEEE, 2023.

[4] Y. Lu, Y. Fan, B. Deng, F. Liu, Y. Li, and S. Wang, "Vl-grasp: a 6-dof interactive grasp policy for language-oriented objects in cluttered indoor scenes," in *IEEE/RSJ Int. Conf. on Intel. Robots and Syst.* IEEE, 2023, pp. 976–983.

[5] V. Bhat, P. Krishnamurthy, R. Karri, and F. Khorrami, "Hifi-cs: Towards open vocabulary visual grounding for robotic grasping using vision-language models," *CoRR*, 2024.

[6] G. Tziafas, Y. XU, A. Goel, M. Kasaei, Z. Li, and H. Kasaei, "Language-guided robot grasping: CLIP-based referring grasp synthesis in clutter," in *7th Annual Conf. on Robot Learning*, 2023.

[7] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "Mdetr-modulated detection for end-to-end multi-modal understanding," in *Int. Conf. Comput. Vis.*, 2021, pp. 1780–1790.

[8] H. Ding, C. Liu, S. Wang, and X. Jiang, "Vlt: Vision-language transformer and query generation for referring segmentation," *PAMI*, 2022.

[9] C. Zhu, Y. Zhou, Y. Shen, G. Luo, X. Pan, M. Lin, C. Chen, L. Cao, X. Sun, and R. Ji, "Seqtr: A simple yet universal network for visual grounding," in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 598–615.

[10] J. Liu, H. Ding, Z. Cai, Y. Zhang, R. K. Satzoda, V. Mahadevan, and R. Manmatha, "Polyformer: Referring image segmentation as sequential polygon generation," in *Conf. on Comput. Vis. and Pattern Recog.*, 2023, pp. 18 653–18 663.

[11] M. Li and L. Sigal, "Referring transformer: A one-step approach to multi-task visual grounding," *Adv. Neural Inform. Process. Syst.*, vol. 34, pp. 19 652–19 664, 2021.

[12] N. Kim, D. Kim, C. Lan, W. Zeng, and S. Kwak, "Restr: Convolution-free referring image segmentation using transformers," in *Conf. on Comput. Vis. and Pattern Recog.*, 2022, pp. 18 145–18 154.

[13] Y. X. Chng, H. Zheng, Y. Han, X. Qiu, and G. Huang, "Mask grounding for referring image segmentation," in *Conf. on Comput. Vis. and Pattern Recog.*, June 2024, pp. 26 573–26 583.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Adv. Neural Inform. Process. Syst.*, vol. 30, 2017.

[15] W. Chen, L. Chen, and Y. Wu, "An efficient and effective transformer decoder-based framework for multi-task visual grounding," in *Eur. Conf. Comput. Vis.*, 2024, pp. 125–141.

[16] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon emissions and large neural network training," *arXiv preprint arXiv:2104.10350*, 2021.

[17] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Adv. Neural Inform. Process. Syst.*, vol. 32, 2019.

[18] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Adv. Neural Inform. Process. Syst.*, vol. 34, pp. 9694–9705, 2021.

[19] S. Lin, P. Lyu, D. Liu, T. Tang, X. Liang, A. Song, and X. Chang, "Mlp can be a good transformer learner," in *Conf. on Comput. Vis. and Pattern Recog.*, 2024, pp. 19 489–19 498.

[20] T. Chen, Z. Zhang, Y. Cheng, A. Awadallah, and Z. Wang, "The principle of diversity: Training stronger vision transformers calls for reducing all levels of redundancy," in *Conf. on Comput. Vis. and Pattern Recog.*, 2022, pp. 12 020–12 030.

[21] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Eur. Conf. Comput. Vis.* Springer, 2016, pp. 108–124.

[22] Z. Hu, G. Feng, J. Sun, L. Zhang, and H. Lu, "Bi-directional relationship inferring network for referring image segmentation," in *Conf. on Comput. Vis. and Pattern Recog.*, 2020, pp. 4424–4433.

[23] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *Conf. on Comput. Vis. and Pattern Recog.*, 2018, pp. 1307–1315.

[24] G. Luo, Y. Zhou, X. Sunf, L. Cao, C. Wu, C. Deng, and R. Ji, "Multi-task collaborative network for joint referring expression comprehension and segmentation," in *Conf. on Comput. Vis. and Pattern Recog.*, 2020, pp. 10 034–10 043.

[25] D.-J. Chen, S. Jia, Y.-C. Lo, H.-T. Chen, and T.-L. Liu, "See-through-text grouping for referring image segmentation," in *Int. Conf. Comput. Vis.*, October 2019.

[26] Z. Wang, Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, and T. Liu, "Cris: Clip-driven referring image segmentation," in *Conf. on Comput. Vis. and Pattern Recog.*, 2022, pp. 11 686–11 695.

[27] Y. Hu, Q. Wang, W. Shao, E. Xie, Z. Li, J. Han, and P. Luo, "Beyond one-to-one: Rethinking the referring image segmentation," in *Int. Conf. Comput. Vis.*, October 2023, pp. 4067–4077.

[28] C. Liu, H. Ding, and X. Jiang, "Gres: Generalized referring expression segmentation," in *Conf. on Comput. Vis. and Pattern Recog.*, June 2023, pp. 23 592–23 601.

[29] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr, "Lavt: Language-aware vision transformer for referring image segmentation," in *Conf. on Comput. Vis. and Pattern Recog.*, June 2022, pp. 18 155–18 165.

[30] Z.-L. Ni, G.-B. Bian, Z.-G. Hou, X.-H. Zhou, X.-L. Xie, and Z. Li, "Attention-guided lightweight network for real-time segmentation of robotic surgical instruments," in *IEEE Int. Conf. on Robot. and Autom.* IEEE, 2020, pp. 9939–9945.

[31] Y. Sun, B. Pan, and Y. Fu, "Lightweight deep neural network for real-time instrument semantic segmentation in robot assisted minimally invasive surgery," *IEEE Robot. and Autom. Letters*, vol. 6, no. 2, pp. 3870–3877, 2021.

[32] Q. Li, J. Cai, J. Luo, Y. Yu, J. Gu, J. Pan, and W. Liu, "Memory-constrained semantic segmentation for ultra-high resolution uav imagery," *IEEE Robot. and Autom. Letters*, 2024.

[33] T. Nguyen, S. S. Shivakumar, I. D. Miller, J. Keller, E. S. Lee, A. Zhou, T. Özaslan, G. Loianno, J. H. Harwood, J. Wozencraft, *et al.*, "Mavnet: An effective semantic segmentation micro-network for mav-based tasks," *IEEE Robot. and Autom. Letters*, vol. 4, no. 4, pp. 3908–3915, 2019.

[34] Z. Xiang, A. Bao, J. Li, and J. Su, "Boosting real-time driving scene parsing with shared semantics," *IEEE Robot. and Autom. Letters*, vol. 5, no. 2, pp. 596–603, 2020.

[35] C. Cuttano, A. Tavera, F. Cermelli, G. Averta, and B. Caputo, "Cross-domain transfer learning with corte: Consistent and reliable transfer from black-box to lightweight segmentation model," in *Int. Conf. Comput. Vis.*, 2023, pp. 1412–1422.

[36] P. Hu, F. Perazzi, F. C. Heilbron, O. Wang, Z. Lin, K. Saenko, and S. Sclaroff, "Real-time semantic segmentation with fast attention," *IEEE Robot. and Autom. Letters*, vol. 6, no. 1, pp. 263–270, 2020.

[37] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "Pidnet: A real-time semantic segmentation network inspired by pid controllers," in *Conf. on Comput. Vis. and Pattern Recog.*, 2023, pp. 19 529–19 539.

[38] G. Rosi, C. Cuttano, N. Cavagnero, G. Averta, and F. Cermelli, "The revenge of bisenet: Efficient multi-task image segmentation," in *Conf. on Comput. Vis. and Pattern Recog.*, 2024, pp. 8066–8074.

[39] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei, "Rethinking bisenet for real-time semantic segmentation," in *Conf. on Comput. Vis. and Pattern Recog.*, 2021, pp. 9716–9725.

[40] L. Sun, K. Yang, X. Hu, W. Hu, and K. Wang, "Real-time fusion network for rgb-d semantic segmentation incorporating unexpected obstacle detection for road-driving images," *IEEE Robot. and Autom. Letters*, vol. 5, no. 4, pp. 5558–5565, 2020.

[41] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *IEEE/RSJ Int. Conf. on Intel. Robots and Syst.* IEEE, 2017, pp. 5108–5115.

[42] J. Wald, K. Tateno, J. Sturm, N. Navab, and F. Tombari, "Real-time fully incremental scene understanding on mobile platforms," *IEEE Robot. and Autom. Letters*, vol. 3, no. 4, pp. 3402–3409, 2018.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[44] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Conf. on Comput. Vis. and Pattern Recog.*, 2018, pp. 7132–7141.

[45] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Eur. Conf. Comput. Vis.* Springer, 2016, pp. 69–85.

[46] V. K. Nagaraja, V. I. Morariu, and L. S. Davis, "Modeling context between objects for referring expression understanding," in *Eur. Conf. Comput. Vis.* Springer, 2016.