Resolving Multimodal Hallucinations through semantics guided augmentation

Chaeyun Kim Seunghoon Yi Younghun Kim Jihoo Park
Seoul National University
{golddohyun,jaguar6182,yh.alex.kim, jihoopark}@snu.ac.kr

Abstract

Referring image segmentation aims to locate and segment the object within an image that a sentence refers to. With the emergence of powerful pre-trained models, recent advancements in RIS models have showcased remarkable capabilities with substantial performance improvements. However, recent improvements on this line of research is mainly done by altering the decoder architecture or fine-tuning the entire model, that incurs a significant computational cost. Additionally, fundamental concerns such as information imbalance between modality and inconsistency against prompts have received limited attention. In this research, we focus on the lack of complexity of the language side, and propose methods that leverage semantic and embedding-level augmentation. Furthermore, we introduce a novel training framework employing contrastive learning based on geometric loss to better guide the model in learning similarities from cross-modal embeddings.

1. Introduction

Referring Image Segmentation(RIS) task aims to identify the target referent in the image, given a language expression. Each RIS problem instance often requires a different level of multimodal understanding capabilities, depending on visual ambiguity as well as linguistic complexity.

Despite the advancements in state-of-the-art (SOTA) models, there remain doubts about their ability to fully capture nuanced linguistic information and its contextual relevance. One significant challenge observed is the inconsistency in outputs when conditioned on texts with identical semantics. If there are multiple sentences describing the same entity, any given text should lead to the same inferred mask. However, as seen in Fig. 2, some sentences lead to segmentations close to the correct object, while others identify entirely unrelated objects.

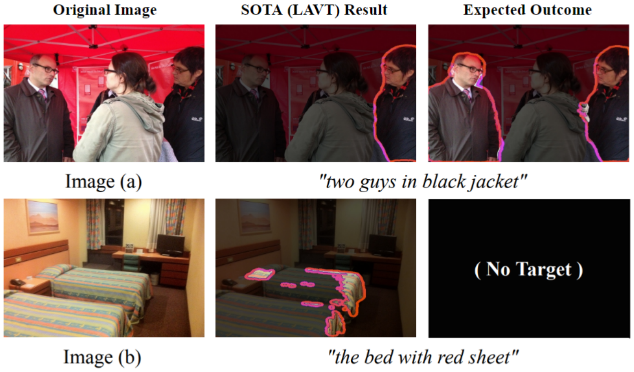

Figure 1. **Hallucination Examples.** (a) Incorrect segmentation of an unintended subject alongside ”two guys in black jacket”. (b) Erroneous segmentation of a non-existent ”bed with red sheet”, indicating the model’s failure in target absence recognition.

Additionally, these models struggle when faced with multiple similar objects or tasks requiring a deep understanding of language nuances, such as positional statements or adjectives. As can be seen from Fig. 1, when prompted with the query ’two guys in black jacket,’ the model erroneously identifies only one man instead of two, displaying a lack of comprehensive object differentiation. In Image (b), the model locates ’the bed with red sheet’ which actually does not exist. We consider such errors as an indication of **’hallucination’ in multi-modal tasks**.

Given this problem landscape, we aim to tackle the challenge of hallucination on the Referring Image Segmentation task. The critical issue here is the lack of a deep semantic understanding of the given queries of the baseline models. We first pinpoint the cause of these hallucinations to a **mismatch between visual and linguistic complexities**. An analysis of dataset statistics in Table 1, RefCOCO and RefCOCO+ reveals a severe lack of linguistic complexity compared to the visual information. Even in G-Ref, which has longer referring expressions and more objects per image, the ambiguity in linguistic descriptions still leads to a discord between visual cues and linguistic complexity.

This observation reveals that the bottleneck may not be in the models but in the training data, which has not provided sufficient difficulty to learn from. Therefore, our approach emphasizes enhancing the composite understanding of linguistic information during training. We focus on refin-

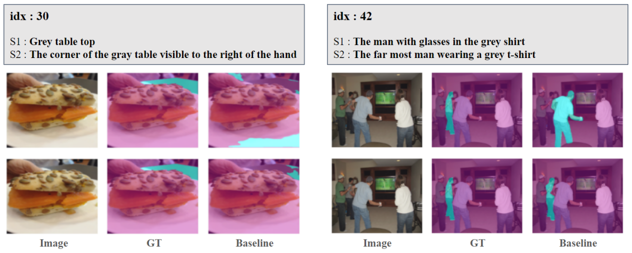
idx : 30
S1 : Grey table top
S2 : The corner of the gray table visible to the right of the hand
idx : 42
S1 : The man with glasses in the grey shirt
S2 : The far most man wearing a grey t-shirt
Image
GT
Baseline
Image
GT
Baseline

Figure 2. **Inconsistencies in Model Segmentation from Semantically Similar Descriptions.** The top row's segmentation for ”Grey table top” contrasts with the lower rows, where two descriptions of the same person yield different results, underscoring the models' inconsistent interpretations.

Table 1. Statistics of current Referring Image Datasets

Dataset	RefCOCO	RefCOCO+	G-Ref
# Images	19,994	19,992	26,711
# Ref. expressions	142,209	141,564	85,474
Avg. query length	3.61	3.53	8.43
Avg. object / query	1.76	1.67	3.03

ing language models, utilizing text data augmentation, and employing self-supervision through contrastive learning.

2. Related Work

Referring Image Segmentation RIS [5] task aims to segment a specific object in an image based on a natural language expression. The technique typically unfolds in two phases: initial feature encoding for both visual and textual inputs, then the fusion of these modalities. In the feature encoding stage, research has evolved from using RNNs and LSTMs [13, 15, 21, 26] for language processing and FCNs [5, 15, 18] for visual feature extraction to embracing Transformer-based backbones [2, 10, 11, 17].
The core stage of RIS research, visual-linguistic fusion, began with basic methods, initially merging features through concatenation [5]. Subsequent research [7, 8, 33] leverage syntax information obtained by parsing the referring expressions. The advent of self-attention [1, 6] and Transformers [3, 27, 31] brought further refinement by enhancing further multimodal alignment. Recently, CRIS [29] and ReSTR [12] maintain separate encodings for each modality before fusing them. While these developments have primarily focused on architectural enhancements, we primarily address the issue of **multi-modal hallucination and the complexity of linguistic data within datasets**.

Hallucination in Deep Learning Hallucination in deep learning refers to a model generating plausible but incorrect responses that do not accurately reflect the input data [22]. This issue has been predominantly explored within the realm of Large Language Models (LLMs), particularly during tasks like QA and summarization [28], as a part of calibration to ensure model fidelity. Notably, self-supervision techniques are utilized to automatically detect errors, using self-consistency checks [9] to infer the reliability of responses or leveraging weak-supervision approaches [34] that utilize pseudo-labels derived from multiple supervision sources for self-correction. In the multimodal field, hallucination involves generating images with objects that don't exist in the given text description. To mitigate this, recent strategies propose data refinement through self-training [23], or using language cues for better visual grounding in images [16], employing techniques like soft pseudo-labeling and adversarial learning during training.

Contrastive Learning Contrastive learning, a subset of deep metric learning, emerged in the domain of computer vision. Here, positives are defined as pairs of samples from the same category or with similar semantics, while negatives stemming from different categories. The primary objective is to reduce the distance between positives while maximizing the distance between negatives within the embedding space. Early contributions include Noise Contrastive Estimation(NCE, [4]) which estimates parameters of a statistical model and Triplet loss [25], focusing on optimizing the distances between anchors and their positives and negatives in the embedding space. Consequently, researches suggested augmenting or altering text to generate positive samples, later utilizing them as negatives if altered beyond a specified threshold. For instance, EDA [30] employes basic methods such as synonym replacement and random swapping to augment text.

3. Method

In this study, we employed two methods, prompt-and-embedding level text augmentation and contrastive learning through geometric loss. We primarily concentrate on CRIS as it stands as a pioneer in adopting large-scale pretrained models as encoders. CRIS leverages the extensive knowledge embedded in the CLIP model, making it a pivotal study in understanding the transferability and generalization capabilities of our architecture.

3.1. Setup for Contrastive Learning

We start from a pair of image $\mathbf{I}_i$ and its corresponding text $\mathbf{T}^i_{j,1,\ldots,n}$, which refer to the entity j in the image $\mathbf{I}_i$. Specifically, for an arbitrary entity E_j that exits in image $\mathbf{I}_i$, there exists a referring text $\mathbf{T}^i_{j,1}, \ldots, \mathbf{T}^i_{j,n}$, where $n \geq 1$. Next, both image and text embeddings undergo cross-attention via the decoder, and used to generate a unified **cross-modal embedding** $\mathbf{e}_{j,n}$. Specifically, we apply convolution on contextualized image embedding $\mathbf{i}_i \in \mathbb{R}^{(H \times W \times C)}$ with the contextualized sentence embedding $\mathbf{t}^i_{j,n} \in \mathbb{R}^C$. That is, $\mathbf{e}_{j,n} = \mathbf{i} * \mathbf{t}_{j,n} \in \mathbb{R}^C$, and i's are contracted for brevity.

With this setup, we perform contrastive learning on the cross-modal embeddings. First, we define **positives pairs** as $\{\mathbf{e}_{j,n}, \mathbf{e}_{j,m}\}$, since each $\mathbf{e}$'s are mapped with the exactly same segmentation mask, with $0 < m, n$ and $m \neq n$ in general. For **negative pairs**, we can choose any pairs that does not share the same entity such as $\{\mathbf{e}_{j,n}, \mathbf{e}_{j' \neq j,(\cdot)}\}$. Hereafter, following the convention, positive pairs will be denoted as $\mathbf{x}^+$ and negative pairs as $\mathbf{x}^-$.

Considering all possibilities, the number of pairs seems to be enough. However, for contrastive learning to be effective, balance between two modalities are crucial. Due to the text-sparse nature of the dataset, descriptions for each entity with similar semantics, which we define pseudo-positives are necessary. We addressed this problem by employing text augmentation, explained in the section below.

3.2. Enriching Text with Augmentation

To achieve effective results in the RIS task, the model must distinguish which entity in an image is being described by a given text. This could be enhanced by exposing the model to a wide range of textual descriptions. However, it's important not to randomly generate any positive or negative examples, as some underlying issues within the dataset remain. Often, the text descriptions are too simplistic, failing to adequately represent the complex entities in the images. Additionally, some entities are represented by a singular description. This scarcity of varied descriptions implies the necessity of text augmentation for effective metric learning.

Given these challenges, we employ two methods:

Embedding-Level Text Augmentation In Embedding-level augmentation, we introduce noise into the original text embedding to create perturbed embeddings, which are then used as pseudo-positives. This is because small perturbations does not harm the original semantic directions, but can lead to similar variants, as introduced in [24].

Prompt-Level Text Augmentation We also perform augmentations by directly adding simple prompts to the text. In our naive setting, we limit the number of positives to a minimum of two. If an entity is paired with only one sentence, we create a prompt by appending a randomly selected prefix and suffix to the original sentence. For instance, prefixes like 'In the image, ' or 'In the given image, ', and suffixes such as ' is existing.', ' is being described.' are combined with the given expression to craft these prompts.

3.3. Contrastive Learning with Geometric Loss

Various objective functions are proposed for contrastive learning, each with different degrees of constraints. In a scenario with clearly defined positive and negative pairs, the distance between embeddings within positive pairs ideally approaches zero, while the embeddings within negative pairs should be maximized. We embed such constraints with a simple **euclidean-distance based loss** as

$$\mathcal{L}_{\mathrm{E}}(\mathbf{x}^{(\cdot)}, \theta) = \alpha \cdot d(\mathbf{x}^+)^2 + (1-\alpha) \cdot \max(0, \epsilon - d(\mathbf{x}^-))^2 \quad (1)$$

where we denote $d(\mathbf{x}^{(\cdot)})$ as distance between pairs of cross-modal embeddings, θ as the model parameters, and $0 < \alpha < 1$, a hyperparameter for positive weighting. However, the modality gap observed in CLIP, as introduced in [14], poses a risk of excessive restriction, potentially leading to a dimensional collapse, which is $||\mathbf{x}||_2^2 \to 0$. This may result in additional information loss, which is undesirable. Thus, we propose a **geometric loss** that introduces constraints based on the directional discrepancy between embeddings. Concretely, we normalize embeddings as $\mathbf{z}_{j,n} \equiv \mathbf{e}_{j,n}/||\mathbf{e}_{j,n}||_2^2$, and quantify the angle between $\mathbf{z}$'s as $\phi(\mathbf{z}, \mathbf{z}') \equiv \cos^{-1}\langle \mathbf{z}, \mathbf{z}' \rangle$. Here $\langle \cdot, \cdot \rangle$ denotes the dot product. Finally, the objective function is formulated as:

$$\mathcal{L}_{\mathrm{G}}(\mathbf{z}^{(\cdot)}, \theta) = \begin{cases} \alpha \cdot \phi(\mathbf{z}, \mathbf{z}')^2 \\ (1-\alpha) \cdot \max(0, \phi_{\text{th}} - \phi(\mathbf{z}, \mathbf{z}'))^2 \end{cases} \quad (2)$$

where the upper and lower term corresponds to each cases where $\{\mathbf{z}, \mathbf{z}'\}$ form positive and negative pairs. Summations for all pairs are excluded for brevity, in both equations.

4. Experiments

4.1. Experimental Settings

Datasets. Among various benchmarks in RIS task, we utilize G-Ref [20] dataset which contains 85,474 referring expressions and 54,822 objects in 26,711 images. Notably, Different from other datasets, G-Ref provides two extra, challenging partitions, UMD [32] and Google [20]. However, due to time constraints, we initially experimented with the most challenging G-Ref dataset, and employ UMD partition for our evaluation, following CRIS.

Metrics. We adopt three evaluation metrics: overall intersection over union (oIoU), mean intersection over union (mIoU), and precision at the 0.5, 0.7, and 0.9 thresholds (Prec@ρ). The oIoU is the proportion of the total intersection area to the overall union area across all test samples. Due to its tendency to favor larger objects, mIoU is also measured. mIoU represents the average intersection between the prediction mask and the ground truth for all test samples. Prec@ρ measures the percentage of samples whose IoU with the ground truth exceeds the threshold ρ.

Model and experimental details To verify the effectiveness of our method, we initially experimented with the CRIS [29] model, which serves as a precursor to the latest RIS models. For the experiment, we initially apply noise of gaussian distribution with $\sigma = 10^{-4}$ to generate pseudo-positives, and compared results with prompt-level augmentation. For the positive weight on contrastive learning, we

Table 2. Parameter search for the contrastive learning setup. We examine the effect of ratio between weights of cross-entropy loss(W_{CE}) and the geometric loss(W_{G}). Next, we test with 3 different α's, fixing $W_{\text{G}}/W_{\text{CE}} = 0.1$. 'Best' refers to the optimal hyperparameter settings within the objective function, while 'n' denotes the number of layers in the decoder. If unspecified, the structure remains identical to the optimal setting of CRIS (decoder with 3 layers). For simplicity, we denote options as pair of ($W_{\text{G}}/W_{\text{CE}}$, α). All metrics are measured with UMD validation set.

Options	mIoU	oIoU	P@0.5	P@0.7	P@0.9
CRIS(default)	59.35	55.91	**68.93**	55.45	14.40
(1, 0.5)	56.72	53.22	64.09	50.31	12.11
(0.1, 0.5)	57.10	53.55	65.13	52.04	12.75
(0.1, 0.25) (Best)	58.99	55.17	67.85	55.09	14.87
(0.1, 0.75)	48.87	47.56	54.06	36.25	5.23
Best + **n=4**	**60.04**	**56.12**	68.81	**56.52**	**15.75**
Best + n=6	59.6	55.99	68.38	55.72	14.93

Table 3. Comparison of two text augmentation methods, along with the batch size. 'N' represents embedding-level augmentation, while 'T' denotes prompt-level augmentation, and 'B' for the batch size. Here, number of decoder layers are fixed to 4, and all experiments share the setting ($W_{\text{G}}/W_{\text{CE}}$, α) = (0.1, 0.25).

Options	mIoU	oIoU	P@0.5	P@0.7	P@0.9
CRIS(default)	59.35	55.91	**68.93**	55.45	14.40
N, B = 32	59.6	55.59	67.93	56.11	15.81
N, B = 64	**60.04**	**56.12**	68.81	**56.52**	15.75
N, B = 72	58.95	55.82	67.38	54.47	14.41
T, B = 32	59.14	54.91	67.95	55.47	15.09
T, B = 64	59.23	55.13	68.28	55.76	15.81
T, B = 72	59.45	55.20	68.03	55.58	15.47

select α within $[0.25, 0.5, 0.75]$, and choose 0.25 as default. Similarly, we select $\epsilon = 1$ and $\phi_{\text{th}} = 0.5$ for thresholds on negatives. Further, to examine the effects of the model architecture and the gradient landscape, we experiment with number of layers in $[3, 4, 6]$ and batch size in $[32, 64, 72]$. Additionally, we utilize the AdamW [19] optimizer with the learning rate of 1e-4. Training epochs remain consistent with the original paper. Unless noted above, we follow the original settings from CRIS.

4.2. Effectiveness of the Proposed Method

Combining Contrastive Loss Tab. 2 compares the performance on G-Ref validation and test split with different hyper-parameter combinations of our proposed methods on CRIS. In this experiment, we aim to figure out the basic settings of hyperparameters. First, we examine the performance using a mixture of cross-entropy loss($\mathcal{L}_{\text{CE}}$), which is the only loss adopted in CRIS, and the geometric loss($\mathcal{L}_{\text{G}}$). Hereafter, we construct our total objective function as $\mathcal{L}_{\text{tot}} = W_{\text{CE}}\mathcal{L}_{\text{CE}} + W_G\mathcal{L}_{\text{G}}$. This choice is due to the observation that even assigning small weights to the naive naive euclidean loss, most of the embeddings suffers from dimension collapse.

Additionally, we evaluate the impact of varying the positive weight parameter α, with values in range of $[0.25, 0.5, 0.75]$. Our observations indicate that beyond a certain value of α and $W_{\text{E}}/W_{\text{CE}}$, there is a noticeable degradation in performance. Hence, we conducted the remaining experiments with a fixed setting of $\alpha = 0.25$ and $W_{\text{G}}/W_{\text{CE}} = 0.1$. Further, with ablation study on the number of decoder layers, we established our default architecture as a decoder with 4 layers. This step resulted in surpassing the original CRIS model, particularly in terms of mIoU metric for over 1% increase.

Comparison between Text Augmentation Methods In cases which contrastive learning is involved, we basically used embedding-level augmentation to tackle cases where an entity has a single describing sentence. To examine the efficacy of the prompt-level augmentation, we experiment with a same range of batch sizes, based on the best setting in Tab. 2. As depicted in Tab. 3, prompt-level augmentation performed sub-optimally in every batch sizes compared to embedding-level augmentation. This might be attributed to the simplicity of the provided prefixes and suffixes.

Binding with same augmentation methods across experiments, a batch size of 64 emerged as the baseline setting. Initially, this might appear counterintuitive, as one would anticipate the model's performance to scale with the batch size. However, since we kept the learning rate and number of epochs constant for all experiments, larger batch sizes led to smaller number of steps, resulting in underfitting. In our future work, we plan to conduct additional experiments exploring setups that haven't been considered in this study.

5. Conclusion

In this paper, we propose potential strategies to bridge the complexity gap between visual and language modalities inherent in the RIS task. By employing contrastive learning alongside text augmentation methods, we attained better results compared to our baseline, CRIS. Among our experiments, utilizing embedding-level augmentation with geometric loss yielded the best performance.

Limitations and Future works. Effective utilization of contrastive learning requires essential experiments involving an extensive hyperparameter search and large batch sizes. However, due to limited computational and time resources, we could not fully explore the hyperparameter space. Also, qualitative comparisons on challenging cases along the proposed methods are missing. We set these milestones as our future line of work. Moreover, we plan to employ semi-supervised techniques by introducing a 'text to mask, mask to text' cycle consistency regularizer. This approach aims to align the mask output of augmented sentences with text embeddings derived from each mask, potentially enhancing the model's capability to accurately interpret and segment based on intricate linguistic queries.

References

[1] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7454–7463, 2019. 2

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[3] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021. 2

[4] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 2

[5] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 108–124. Springer, 2016. 2

[6] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4424–4433, 2020. 2

[7] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10488–10497, 2020. 2

[8] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 59–75. Springer, 2020. 2

[9] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating hallucination in large language models via self-reflection. *arXiv preprint arXiv:2310.06271*, 2023. 2

[10] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9858–9867, 2021. 2

[11] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 2

[12] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18145–18154, 2022. 2

[13] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2018. 2

[14] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 3

[15] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1271–1280, 2017. 2

[16] Fang Liu, Yuhao Liu, Yuqiu Kong, Ke Xu, Lihe Zhang, Baocai Yin, Gerhard Hancke, and Rynson Lau. Referring image segmentation using text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22124–22134, 2023. 2

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[18] 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*, pages 3431–3440, 2015. 2

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 4

[20] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 3

[21] Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 630–645, 2018. 2

[22] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020. 2

[23] Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. A simple recipe towards reducing hallucination in

neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679, 2019. 2

[24] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*, 2022. 3

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

[26] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 38–54, 2018. 2

[27] Jiajin Tang, Ge Zheng, Cheng Shi, and Sibei Yang. Contrastive grouping with transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23570–23580, 2023. 2

[28] Akim Tsvigun, Ivan Lysenko, Danila Sedashov, Ivan Lazichny, Eldar Damirov, Vladimir Karlov, Artemy Belousov, Leonid Sanochkin, Maxim Panov, Alexander Panchenko, et al. Active learning for abstractive text summarization. *arXiv preprint arXiv:2301.03252*, 2023. 2

[29] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 2, 3

[30] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019. 2

[31] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022. 2

[32] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 3

[33] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018. 2

[34] Theodore Zhao, Mu Wei, J Samuel Preston, and Hoifung Poon. Automatic calibration and error correction for large language models via pareto optimal self-supervision. *arXiv preprint arXiv:2306.16564*, 2023. 2