OBJECT SEGMENTATION IN THE WILD WITH FOUNDATION MODELS: APPLICATION TO VISION ASSISTED NEURO-PROSTHESES FOR UPPER LIMBS

Bolutife Atoki, Jenny Benois-Pineau, Fabien Baldacci
LaBRI, CNRS, Univ. Bordeaux, UMR 5800, F-33400
Talence, France
bolutife-oluwabunmi.atoki@etu.u-bordeaux.fr,
{jenny.benois-pineau, fabien.baldacci}@u-bordeaux.fr

Aymar de Rugy
Univ. Bordeaux, CNRS, INCIA, UMR 5287, F-33000
Bordeaux, France
aymar.derugy@u-bordeaux.fr

***Abstract*—In this work, we tackle the problem of semantic object segmentation with foundation models. We investigate whether foundation models, trained on a tremendous number and variety of objects, can sufficiently perform object segmentation without fine-tuning on specific images containing objects of everyday life, but in strongly cluttered visual scenes "in-the-wild" context for guiding upper limb neuro-prostheses. We adapt Segment Anything Model (SAM) to our segmentation scenario and propose strategies to guide the model with gaze fixations, and fine tune it on egocentric visual data. The results of evaluation of our approach show the improvement of IoU segmentation quality metric by up to 0.5 points on real-world data.**

***Index Terms*—Semantic Object Segmentation, Foundation Model, Gaze Fixations, Assistive Robotics, Fine-tuning.**

I. Introduction

Object detection and segmentation is a widely researched problem in image analysis, with many application areas including; surveillance [1], detecting cars, cyclists and traffic signs for traffic control using a common dense feature extraction network coupled with specialized detectors for each object class [2]. In robotic vision, [3] implemented for a small quadcopter, dynamic obstacle detection and tracking with an RBG-D camera, through an ensemble of computationally efficient but low-accuracy models to obtain precise object detection in real-time. Assistive robotics, is a large consumer of object detection, recognition and segmentation methods. One of its branches is the design and production of robotic myoelectric-prostheses. These prostheses however require more muscle than is available when high-place amputations are performed [4]. Hence, there is a need for alternative means of prosthesis control. Upper limb vision-guided prostheses are an example of such a solution [5].

In the application case of prosthesis wearers [6], object detection in video recorded by a glasses-worn camera allows for the determination of the object of interest to be grasped to ensure servoing of the prosthetic arm towards it. The exact shape of the object is required for the estimation of its 6D pose in order to control the grasping orientation, movement, and opening of the palm of the prosthetic arm. Object segmentation is a mandatory part of the whole vision-guided prosthesis control system for the above mentioned reasons: controlling the palm orientation and its opening. The authors of [7] propose one of the best 6D pose estimators nowadays, which might be used in a prosthesis control prototype. The method requires the object mask in the video frame or scene image, therefore, the accurate segmentation of the object to be grasped is necessary.

Object segmentation involves identifying objects within images to precisely obtain object boundaries at the pixel level [8]. With the introduction of Foundation models [9] such as Detectron2 [10] and Segment Anything Model (SAM) [11] trained on large amount of data and adaptable to a wide range of downstream tasks, the quality of obtained segmentation masks has drastically improved. They yield cleaner boundaries and quality in terms of human mask quality rating [11]. The contribution of this paper involves

- adapting a foundation model to utilize gaze fixation points which are available due to the vision-guided prosthesis scenario, to obtain object segmentation masks.
- evaluating if fine-tuning the foundation model is required for real-world video data in an object grasping scenario, in natural environments.

The remainder of the paper is organised as follows: in section II, we give a short overview of popular segmentation models and justify our choice of base-line model. In section III, we present our approach for the adaptation of a foundation model for object segmentation in a vision-guided neuro-prosthesis scenario. In section IV, we describe our natural data, evaluation metrics and obtained results. Section V concludes this work and outlines its perspectives.

This research is supported by I-Wrist 2023 Grant of French National Research Agency (ANR)

979-8-3503-9114-5/24/31.00 © 2024 *IEEE*

II. Literature Review

Image segmentation is a research problem with a long history. Initially, it meant partitioning the image plane into a set of non-overlapping regions, using the chrominance/colour homogeneity as a low-level criterion [12]. To extract objects, which might have different homogeneous regions inside, external prompts were included, e.g. in video scenes, and the regions animated by a homogeneous motion would represent semantic objects or the background [13].

With the advent of supervised machine learning approaches, image segmentation made a step towards semantic segmentation, with the introduction of semantic homogeneity of sets of neighboring regions in the image plane [14].

Nevertheless, the real break-through in the problem of semantic object segmentation was achieved with the advent of Deep Learning methods. Here, ground-truth segmentation masks are used for training the models that assign pixels to objects in the scene. Thus, [15] implemented a symmetric *U-shaped* encoder-decoder network, utilizing high-resolution features from the encoder network combined with up-sampled features in the decoder network, to help the model obtain precise masks. As a result, the encoder network has a high number of feature channels, thus allowing the network propagate contextual information to its high resolution layers.

Detectron2 [10] is a foundation model with training datasets including Pascal VOC, ADE20k Scene Parsing, cityscapes, coco, LVIS, and PanopticFPN that performs downstream tasks of object detection, instance, and semantic segmentation. It uses backbone feature extractor networks [16] [17] for convolutional features, fed to the region proposal network for object regions, which are pooled and passed into respective prediction heads. The Detectron2 foundation model accepts images as input for its tasks, while Segment Anything [11] additionally accepts optional prompts in the form of object points, object bounding box, low resolution object mask,... These additional prompts improve the quality of obtained masks. In the vision-guided prosthesis scenario, the gaze fixations of the amputee are measured and generally located on the object to be grasped. Therefore, it is natural to resort to the SAM segmentation model, using these gaze fixations as the object points prompt, to improve segmentation accuracy in real-world scenes.

III. Methodology

In this section, we present our segmentation scenario along with proposed methods.

A. Segmentation scenario

In our scenario, the object of interest is pointed by the gaze of the user before the reaching and grasping movement starts. Gaze fixations are a minimal information expressing a user's intention of obtaining an object. Regardless, it can be extended with the knowledge of the rules of fovea projection, into the image plane as in [18], the reasonably real-time prosthetic system has to remain light. An example of a recorded gaze fixation on the object (bright point) is given in fig. 1. Thus subjects look at the objects they wish to grasp. The objects should be segmented before the grasping is accomplished. Therefore, we use the gaze fixation prompt and adapt this noisy information for its efficient use as an input to the foundation segmentation model. In the following, we briefly review the foundation model we have chosen, and explain how we adapt gaze fixations to drive it.

Fig. 1: Illustration of our segmentation scenario, inspired by [19]

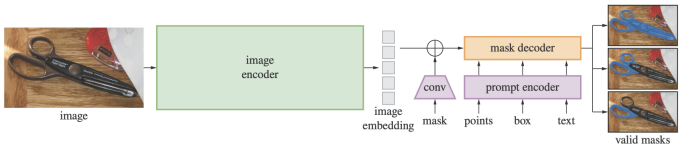


Fig. 2: Segment Anything Architecture

B. Segment Anything Model

The Segment Anything Foundation Model [11] consists of three main blocks, as depicted in fig. 2.

1) **Image Encoder:**
It accepts an input image of dimension $C \times H \times W$, and utilizes a Vision Transformer (ViT) [20] to generate a $16\times$ down-scaled feature embedding, that represents the input image, see "encoder" block in fig. 2. The generated embedding serves as part of the input to the mask decoder.
2) **Prompt Encoder:**
The prompt encoder receives optional prompt inputs in the form of image points–foreground or background–and corresponding binary labels. Other options are the bounding box coordinates of the object of interest, an initial low-resolution mask of the object, and a text description of the object (yet to be implemented by Meta AI). In our approach we use only image point prompt. With a point input, the encoder provides the sum of the positional encoding of the point and a learned embedding which varies based on the point being foreground (on the object) or not (background), see the block "prompt encoder" in fig. 2.
3) **Mask Decoder:**
The Mask Decoder is a modified Transformer Decoder Block [21] which receives all available input embedding vectors, and uses bi-directional self-attention and cross-attention to update them. The embedding vectors

are subsequently up-sampled and provided to an MLP, which predicts the probability of each pixel belonging to the foreground class, see the block "mask decoder" in fig. 2. We refer the reader to [11] for more details.

Thus, we need to adapt the information available in our scenario–the gaze fixation points–into foreground points for the model. In the following, we describe the adaptation steps from acquisition, to projection, filtering, clustering and finally labelling.

C. Getting Prompts for SAM model

The information for object segmentation is acquired with a device which provides a video sequence at 25 frames per second, and is incorporated with an eye-tracker which obtains the gaze fixation information at a rate of 50Hz. Due to the video frame rate being fixed at 25Hz, the gaze points are interpolated to match the number of frames using the spline interpolation [22]. This is done to reduce the probability of a saccade to the distractor and to reduce the data to process.

1) Gaze Fixation Points Projection: Rapid changes in the fixation point of a subject either during movement, due to the nature of the task being performed, lightning conditions and visual distractors in the scene, called visual saccades could occur [23], thereby temporarily shifting the fixation away from the object of interest. Thus, the reliance on the single gaze point for the current frame to locate the object of interest could lead to errors. Therefore, a projection of previous gaze points unto the current frame using homography estimation between consecutive frames as in [6] is implemented. In the present work, the implementation is modified to reduce the propagation of errors across frames as the homography model assumes that the scene contains planar surfaces, and may not accurately capture more complex distortions in the video scene. The proposed solution utilizes a temporal window of T frames to prevent this propagation by chaining the estimated homography matrices between frames in the temporal window through matrix multiplication. The final transformation is thus a composition of homographies as expressed in (1).

$$\tilde{P}_t = H_1 \times (H_2 \times (\ldots \times (H_{T-1} \times (H_T \times P_{t-T+k})))) \quad (1)$$

Where $\tilde{P}_t = (\tilde{x}, \tilde{y}, 1)_t$ are homogeneous gaze point coordinates projected to frame t, $H_n, n = 1, 2, ..., T$ are the homography matrices between frames n and $n+1$, T is the Temporal window size, $k = 0, ..., T-1$. Note that in the current frame t, we collect all the gaze fixations: projected ones $\tilde{P}_t$ and the current recorded gaze point P_t.

2) Visual Saccades Elimination using Density Based Clustering: The visual saccades towards distractors represent outliers in the distribution of projected gaze points and the current point set on the current frame. To eliminate outliers, we perform clustering of the points using a density based method, as proposed in [24]. The goal is to identify the population of points which are close to one another and thus situated on the object of interest. Hence the Density Based Clustering (DBSCAN) is an appropriate algorithm for this. DBSCAN [25] is a well-known non-parametric clustering algorithm that does not require defining the target number of clusters and can handle clusters of non-spherical shape. This is appropriate for our case as humans perform micro-saccades and never foveate the same spatial locus over the time. The algorithm is controlled by two parameters: the minimum number of points ($minPts$) and the radius of the point neighbourhood (ϵ value). The $minPts$ parameter specifies the high-density areas and in our case, its upper bound is the temporal window size T. The ϵ parameter determines the distance threshold for points to be considered neighbours. It depends on the magnitude of microsaccades, while the subject maintains fixation on the object. According to some sources [26], the magnitude in degrees is around 0.6 - 1.0. Hence, with the video acquisition conditions and distance from the glasses-worn eye-tracker, we parameterize it approximately as $\epsilon = 1.4$.

TABLE I: Number of Frames considered

Object Class	Object Sub-Class	No of Frames Before Grasping
Bowl	Blue	113
	Cream	113
	Red	154
	Transparent	124
sub-total		**504**
Milk Bottle	Bottle	418
	Carton	94
sub-total		**512**
Total		**1016**

IV. Results and Analysis

We describe here the natural dataset we use, the parameterization of our methods, and the analysis of the results.

A. Dataset

The Grasping-in-the-Wild dataset corresponds to our object grasping scenario [6]. It is publicly available on the NAKALA server [27]. It was captured using Tobii Glasses 2 [28] during an experiment where five subjects were grasping objects in seven natural kitchen environments. From the entire object taxonomy of sixteen different classes, we have considered only two: bowls and milk bottles. These shapes are the most common for precision grasping assessment in robotics applications. For each instance of the object classes, it contains a video file, a *json* file with gaze fixations, and foreground and random background patches. Note that for our purpose of precisely evaluating segmentation quality, we could not use this latter information and instead had to annotate object masks with pixel precision. We have only considered frames before the grasping time, as required by our object segmentation scenario. The overall number of object instances is depicted in the table I. It is important to note that the object classes exhibit strong variability in appearance.

For pixel-wise annotation of the object mask, both for model fine-tuning (training set) and evaluation (validation set), we designed a semi-automatic annotation process. Initially, objects were segmented using SAM-CLIP, trained on SA-1B [11] and LAION-400M [29] datasets. Subsequently, the masks were manually refined using the GIMP tool [30] . Regarding

the selection of foreground points as essential prompts for the model, we adopted our proposed approach, see section III-C. In the SAM fine-tuning experiment, we shuffled all object class samples and split them into 80% for training and 20% for validation.

B. Evaluation metrics

To evaluate segmentation masks, three metrics are the most popular: intersection over union (IoU), pixel accuracy and DICE coefficient [31]. In our work, we use only the IoU metric. Pixel accuracy measures the ratio of correctly classified pixels over the number of all image pixels. However, due to our small object sizes relative to frame resolution, pixel accuracy suffers from class imbalance. DICE coefficient metric is similar to IoU and is mainly used in medical image segmentation.

The IoU metric measures the overlap between the predicted and ground-truth masks. For a video frame f, it is given by:

$$IoU_f = \frac{\mid M \cap \tilde{M} \mid}{\mid M \cup \tilde{M} \mid} \tag{2}$$

Where $\tilde{M}$ is predicted binary mask, M is ground-truth binary mask, $\mid M \cap \tilde{M} \mid$ is the number of pixels in the masks' intersection area, $\mid M \cup \tilde{M} \mid$ is the number of pixels in the masks' union area. The mean IoU for the considered grasping sequence of n frames is given by:

$$mIoU = \frac{1}{n} \sum_{i=1}^{n} IoU_{f_i} \tag{3}$$

C. Experiment Configuration

The set-up of the experiment conducted is provided below.

1) **SAM Architecture:**
   The chosen architecture was the ViT-Base model [11], which employs the ViT-B model as its image encoder. It is the lightest of the three available models: ViT-Base, ViT-Large, and ViT-Heavy and contains 91M parameters.
2) **Software and Hardware Specifications:**
   The experiment was performed using Tesla P100 GPU with 16GB of graphics memory. The fine-tuning process involved freezing the parameters of both encoder models–image and prompt–thus, updating the weights of the mask decoder alone.
3) **Experiment Hyper-parameters:**
   Fixed Learning rate: $1e^{-5}$, Optimizer: Adam, Number of epochs: 10, Batch size: 4, Number of object points: 5.

D. Evaluation of pre-trained model on GITW Dataset

The results of the pre-trained model's performance (without fine-tuning) on the sub-classes of each of the considered objects are provided in fig. 3, as a function of temporal window length T. The "blue bowl" sub-class represents the most occluded object; the "milk carton" is occluded as well. Thus, projecting points from larger number of frames ($T = 5$)

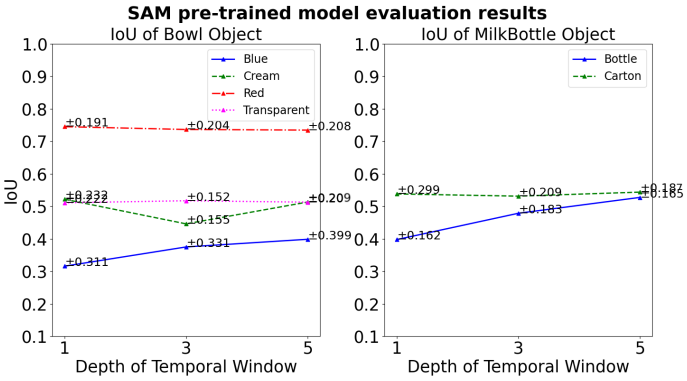


Fig. 3: Evaluation results of pre-trained model on two classes; Bowl and Milk Bottle

TABLE II: $mIoU$ Pre-trained vs Fine-tuned Model Evaluation Results

		mIoU	
Class	Sub-Class	Pre-trained Model	Fine-tuned Model
Bowl	Blue	0.41 ±0.41	**0.86 ±0.13**
	Cream	0.52 ±0.22	**0.98 ±0.01**
	Red	0.77 ±0.22	**0.91 ±0.08**
	Transparent	0.52 ±0.23	**0.76 ±0.11**
Milk Bottle	Bottle	0.54 ±0.17	**0.93 ±0.05**
	Carton	0.55 ±0.19	**0.76 ±0.14**

improves the $mIoU$ metric from 0.32 to 0.41 for the "blue bowl". For other objects, the figures remain stable: 0.52 for the transparent bowl, 0.77 for the red bowl (the least occluded), and 0.53 for the cream bowl; the milk bottle maintains 0.55. Corresponding standard deviations are indicated in fig. 3 as figures on the curves and range from 0.16 for the cream bowl to 0.41 for the blue bowl. These results show an interest to use projected gaze points from several frames, but still remain quite low in terms of the quality of segmented masks. To push the investigation to its limits, we also have evaluated the model without any prompts. The resulting $mIoU$ is 0.0.

Thus, we investigate if fine-tuning the SAM foundation model on the dataset might be of use, despite the claims that foundation models trained on such vast amount of images can segment anything.

E. Evaluation of model fine-tuned on GITW Dataset using Ground-truth Binary masks

The results of the fine-tuned model's performance on the test set of the considered objects are provided in table II.

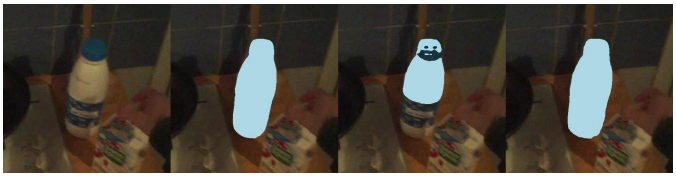
One can see that when fine-tuned on the target corpus of natural objects in strongly cluttered scenes, the model's performance improves across all objects classes considered, up to 0.5 points for the blue bowl sub-class. It is also more stable - with reduced standard deviation. Some examples are shown in fig. 4.

F. Ablation Study

The ablation means using only one point on the object and has been already reported in section IV-E. Another ablation involves the use of the pre-trained model and not a fine-tuned

Transparent Bowl

Blue Bowl



Milk Bottle

Milk Carton

Fig. 4: Examples of segmented masks. From Left to Right: orignal frame, ground truth mask, mask from pre-trained model, mask from fine-tuned model.

one. It has also been reported in table II. This shows the usefulness of our approach to adapt the foundation model and to fine-tune it.

V. Conclusion and perspectives

Hence in this paper we have proposed an adaptation of the foundation model SAM for object segmentation in real-world cluttered visual scenes by utilizing gaze fixation points available in our vision-guided prosthesis scenario, we also have investigated the need for fine-tuning the foundation model on real-world video data. Our results show that utilizing gaze-fixation points as additional input prompts to the model through temporal window gaze point projection improves the performance. The segmentation results are better when the number of gaze-fixation points on the occluded object of interest is higher. Furthermore, despite claims that foundation models can segment any object in natural images, we have shown that fine-tuning is recommended and improves the segmentation quality metric $mIoU$ by up to 0.5 points, with a better stability of results.

For future work, the resolution of input images to the model can be reduced by extracting a fuzzy mask of the object based on gaze fixation points on the frame, and building an adaptive object bounding box. This would limit the image area to the object of interest and fovea projection, resulting in lower computational time which is crucial in prostheses scenarios. A supplementary study for building such adaptive bounding box is required.

References

[1] Samet Akcay and Toby P. Breckon, "An evaluation of region based object detection strategies within x-ray baggage security imagery," in *ICIP*. 2017, pp. 1337–1341, IEEE.

[2] Yi-Nan Chen, Hang Dai, and Yong Ding, "Pseudo-stereo for monocular 3d object detection in autonomous driving," in *CVPR*. 2022, pp. 877–887, IEEE.

[3] Zhefan Xu, Xiaoyang Zhan, Yumeng Xiu, Christopher Suzuki, and Kenji Shimada, "Onboard dynamic-object detection and tracking for autonomous robot navigation with rgb-d camera," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 651–658, 2024.

[4] Sébastien Mick, Effie Segas, Lucas Dure, Christophe Halgand, Jenny Benois-Pineau, Gerald E Loeb, Daniel Cattaert, and Aymar de Rugy, "Shoulder kinematics plus contextual target information enable control of multiple distal joints of a simulated prosthetic arm and hand," *Journal of NeuroEngineering and Rehabilitation*, vol. 18, pp. 1–17, 2021.

[5] Boxuan Zhong, He Huang, and Edgar J. Lobaton, "Reliable vision-based grasping target recognition for upper limb prostheses," *IEEE Trans. Cybern.*, vol. 52, no. 3, pp. 1750–1762, 2022.

[6] Iván González-Díaz, Jenny Benois-Pineau, Jean-Philippe Domenger, and Aymar de Rugy, "Perceptually-guided understanding of egocentric video content: Recognition of objects to grasp," in *ICMR*. 2018, pp. 434–441, ACM.

[7] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martin Martin, Cewu Lu, Li Fei-Fei, and Silvio Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *CVPR*. 2019, pp. 3343–3352, Computer Vision Foundation / IEEE.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018.

[9] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, and Russ B. Altman et al., "On the opportunities and risks of foundation models," *CoRR*, vol. abs/2108.07258, 2021.

[10] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick, "Segment anything," in *ICCV*. 2023, pp. 3992–4003, IEEE.

[12] Jean-Marc Chassery and Catherine Garbay, "An iterative segmentation method based on a contextual color and shape criterion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 794–800, 1984.

[13] Ling Wu, Jenny Benois-Pineau, Philippe Delagnes, and Dominique Barba, "Spatio-temporal segmentation of image sequences for object-oriented low bit-rate image coding," *Signal Process. Image Commun.*, vol. 8, no. 6, pp. 513–543, 1996.

[14] Rémi Vieux, Jenny Benois-Pineau, Jean-Philippe Domenger, and Achille J.-P. Braquelaire, "Segmentation-based multi-class semantic object detection," *Multim. Tools Appl.*, vol. 60, no. 2, pp. 305–326, 2012.

[15] Olaf Ronneberger, "Invited talk: U-net convolutional networks for biomedical image segmentation," in *Bildverarbeitung für die Medizin*. 2017, Informatik Aktuell, p. 3, Springer.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*. 2016, pp. 770–778, IEEE Computer Society.

[17] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in *CVPR*. 2017, pp. 5987–5995, IEEE Computer Society.

[18] A. M. Obeso, J. Benois-Pineau, M. S. García-Vázquez, and A. A. Ramírez-Acosta, "Visual vs internal attention mechanisms in deep neural networks for image classification and object detection," *Pattern Recognit.*, vol. 123, pp. 108411, 2022.

[19] Bianca Lento, Effie Segas, Vincent Leconte, Emilie Doat, Frederic Danion, Renaud Péteri, Jenny Benois-Pineau, and Aymar de Rugy, "3d-arm-gaze: a public dataset of 3d arm reaching movements with gaze information in virtual reality," *bioRxiv*, 2024.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*. 2021, OpenReview.net.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[22] M. Unser, "Splines: a perfect fit for signal and image processing," *IEEE Signal Processing Magazine*, vol. 16, no. 6, pp. 22–38, 1999.

[23] Florian Kadner, Tobias Thomas, David Hoppe, and Constantin A. Rothkopf, "Improving saliency models' predictions of the next fixation with humans' intrinsic cost of gaze shifts," in *WACV*. 2023, pp. 2103–2113, IEEE.

[24] Attila Fejér, Zoltán Nagy, Jenny Benois-Pineau, Péter Szolgay, Aymar de Rugy, and Jean-Philippe Domenger, "Hybrid fpga-cpu-based architecture for object recognition in visual servoing of arm prosthesis," *J. Imaging*, vol. 8, no. 2, pp. 44, 2022.

[25] Bernard W Silverman, *Density estimation for statistics and data analysis*, Routledge, 2018.

[26] Jorge Otero-Millan, Xoana G. Troncoso, Stephen L. Macknik, Ignacio Serrano-Pedraza, and Susana Martinez-Conde, "Saccades and microsaccades during visual fixation, exploration, and search: Foundations for a common saccadic generator," *Journal of Vision*, vol. 8, no. 14, pp. 21–21, 12 2008.

[27] LaBRI, "Grasping-in-the-wild dataset," https://www.labri.fr/projet/AIV/graspinginthewild.php, Accessed: 2024-07-11.

[28] Tobii Technology, "Tobii pro glasses 2," https://www.tobii.com/products/discontinued/tobii-pro-glasses-2, Accessed: 2024-07-11.

[29] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki, "LAION-400M: open dataset of clip-filtered 400 million image-text pairs," *CoRR*, vol. abs/2111.02114, 2021.

[30] GIMP, "Gimp: Gnu image manipulation program," https://www.gimp.org/, Accessed: 2024-07-11.

[31] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Research Notes*, vol. 15, no. 1, pp. 210, Jun 2022.