G-DAIC

Horanyi, Nora; Hou, Yuqi; Leonardis, Ales; Chang, Hyung Jin

DOI:
10.1145/3591132

License:
Other (please specify with Rights Statement)

Document Version
Peer reviewed version

Citation for published version (Harvard):
Horanyi, N, Hou, Y, Leonardis, A & Chang, HJ 2023, 'G-DAIC: A Gaze Initialized Framework for Description and Aesthetic-Based Image Cropping', *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. ETRA, 163, pp. 1-19. https://doi.org/10.1145/3591132

G-DAIC: A Gaze Initialized Framework for Description and Aesthetic-Based Image Cropping

NORA HORANYI, University of Birmingham, UK
YUQI HOU, University of Birmingham, UK
ALES LEONARDIS, University of Birmingham, UK
HYUNG JIN CHANG, University of Birmingham, UK

We propose a new gaze-initialised optimisation framework to generate aesthetically pleasing image crops based on user description. We extended the existing description-based image cropping dataset by collecting user eye movements corresponding to the image captions. To best leverage the contextual information to initialise the optimisation framework using the collected gaze data, this work proposes two gaze-based initialisation strategies, *Fixed Grid* and *Region Proposal*. In addition, we propose the adaptive *Mixed scaling method* to find the optimal output despite the size of the generated initialisation region and the described part of the image. We address the runtime limitation of the state-of-the-art method by implementing the *Early termination* strategy to reduce the number of iterations required to produce the output. Our experiments show that G-DAIC reduced the runtime by 92.11%, and the quantitative and qualitative experiments demonstrated that the proposed framework produces higher quality and more accurate image crops *w.r.t.* user intention.

CCS Concepts: • **Computing methodologies → Interest point and salient region detections**.

Additional Key Words and Phrases: Eye-tracking, Gaze-based image cropping, Aesthetics, Deep network re-purposing, Image captioning

ACM Reference Format:
Nora Horanyi, Yuqi Hou, Ales Leonardis, and Hyung Jin Chang. 2023. G-DAIC: A Gaze Initialized Framework for Description and Aesthetic-Based Image Cropping. *Proc. ACM Hum.-Comput. Interact.* 7, ETRA, Article 163 (May 2023), 19 pages. https://doi.org/10.1145/3591132

1 INTRODUCTION

Many researchers throughout the past years have demonstrated the usefulness of multimodal approaches. Regarding speech and description, the existing ambiguity and complexity of natural speech can be compensated by using other modalities, such as hand gestures [34] or gaze [21, 23, 28]. While hand gestures or other social cues are not always coupled with the scene description, observing the region of interest is inevitable to describe it [30]. With the current advancements in eye-tracking research, eye-tracking devices have become more affordable and easily accessible. Therefore we can take advantage of this extra modality free of cost by placing an eye-tracker in

Authors' addresses: Nora Horanyi, University of Birmingham, University Rd W, Birmingham, West Midlands, UK, B15 2TT, nxh840@alumni.bham.ac.uk; Yuqi Hou, University of Birmingham, University Rd W, Birmingham, West Midlands, UK, B15 2TT; Ales Leonardis, University of Birmingham, University Rd W, Birmingham, West Midlands, UK, B15 2TT; Hyung Jin Chang, h.j.chang@bham.ac.uk, University of Birmingham, University Rd W, Birmingham, West Midlands, UK, B15 2TT.

2573-0142/2023/5-ART163 $15.00
https://doi.org/10.1145/3591132

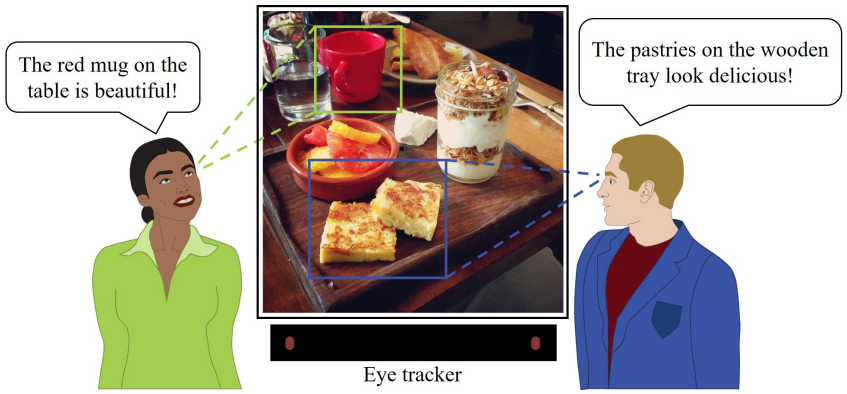

Fig. 1. **Illustration of a scenario where the users are looking at and describing the different image parts.** Our proposed framework utilises human natural language expression in combination with eye movements to localise their region of interest. This multimodal solution can effectively produce high-quality image crops corresponding to the user's intention.

front of the subject and recording their eye movements before and during describing the image region. This way, we can effortlessly collect additional, rich contextual information in addition to the recorded speech or image caption.

Despite its inarguable usefulness of gaze in multimodal communication, this non-verbal mechanism has been less studied as a communication modality [3]. Previous works used fixation points as pointers to implicitly localise the user's region of interest [4] or to detect the object of interest [37]. Recent research [18] demonstrated gaze integration's usefulness in anchoring implicit notes to digital content. Furthermore, Reinholt *et al.* [25] proposed to combine gaze information with speech to identify regions of interest in an image. This assistive system aimed to create detailed descriptions of images with minimal effort from the image creator.

In this work, we tackle the reverse problem of [25], where we obtain the image description, like in [14], and the gaze information with no additional cost from the user to crop the described part of the image. We propose directly integrating the user's gaze information into a multimodal image-cropping framework to understand their interest better. We expect this additional information to improve the accuracy of the state-of-the-art CAGIC methods and reduce their run time. This framework utilises the gaze information recorded by an eye-tracking device to initialise the iterative search to localise the described area of the image. To our knowledge, this is the first work explicitly tackling the gaze-initialised user description-based image cropping task. Our contributions are four-fold:

- We presented a novel image cropping framework that integrates the user's intentions through explicit (user caption) and implicit (user gaze) input into a multimodal framework optimised to achieve aesthetically pleasing output.
- We studied the usefulness of gaze data collected before, during, and after the caption generation and proposed the *Fixed grid* and *Region proposal* on how to leverage the correlation between them to initialise the image cropping method most efficiently.
- We propose a new multimodal framework to optimise crop parameters adaptively using the novel *Mixed scaling method* and gaze-based initialisation coupled with an *Early termination* technique.

- With the above-mentioned solution, we significantly reduced the run time compared to the state-of-the-art and improved the performance.

Utilising gaze information in a multimodal framework is a complex problem studied by many [7, 39]. We designed the gaze data collection experiment carefully, similarly [25], and performed multiple experiments to confirm the quality of the collected gaze information. We proposed the *Fixed Grid* and *Region proposal* methods to find and define the most looked-at area of the image, which was used as a start region of our optimisation framework.

For the Gaze initialised, description and aesthetics-based image cropping (G-DAIC) method, we modified the method introduced in [14] and proposed a *Mixed scaling* method, where based on the size of the initialisation area, the cropping parameter is either shrinking or expanding in every iteration. This proposed alteration is crucial to find the desired image region. Finally, we introduced *Early Termination* into the multimodal framework to reduce the number of iterations required to produce the output image crop.

2 RELATED WORKS

The primary task of the image cropping algorithms is attention-related. The description-based methods, like [8, 36], aim to find the described part of the image, keeping in mind that it might not be the main subject of the image. While aesthetics and attention-based automatic image cropping algorithms, such as [6], focus on localising the most important part and preserving the image's main subject. The attention-based methods can rely on the user's gaze information or artificial attention *e.g.* saliency maps.

2.1 Gaze and artificial attention

2.1.1 Gaze-based image cropping. Despite gaze information not being suitable for our task, gaze has been known as a crucial element of our social interactions. This non-audible signal can be interpreted and utilised in many ways; hence, its role has been studied by many during the past decades. Gaze is a well-known primary social clue that can express the subject's emotions [11]. Furthermore, it is a good guide to the subject's interest and intention [19, 22], and it also functions as a signal for turn-taking [17], and indicator of conversational roles [12, 29]. Gaze in communication is closely related to speech, meaning that in everyday settings, the user has to address speech *w.r.t.* to gaze [3].

Recent eye-tracking technology advancements enabled us to collect high-precision, and accurate eye movement data of the subjects [10]. This high-level precision and affordable prices boosted the gaze research field and gaze-based multimodal frameworks. One of the first gaze-based image cropping solutions, a semi-automatic image cropping algorithm using gaze data, has been proposed by Santella *et al.* [26]. This method uses gaze data to identify the important content of the image and generates an image crop based on a set of composition rules. While this method is useful for photo composition, it is unsuitable for the description-based image cropping task. This semi-automatic method is designed to identify the main subject of the image. Still, it is not flexible enough to take the user's intention explicitly into account to identify any part of the image.

2.1.2 Artificial attention-based image cropping. The usefulness of artificial attention, such as automatically-generated saliency maps using deep networks, has been demonstrated in attention-based image cropping methods [2, 16, 24]. While these maps are good in indicating the potentially important areas of the image based on its features, they are insufficient for the description-based image cropping task due to the lack of contextual information. Therefore in this work, we aim to extend [14] and [18] by combining gaze and description information into a multimodal system designed to crop images based on the natural language expression provided by the users.

2.2 Multimodal Image Cropping

Fang *et al.* [9] proposed an automatic image cropping method that uses composition, content preservation, and boundary simplicity clues to preserve the image's subject. This work was one of the first learning-based automatic solutions which did not hard code the cropping rules but learned it from online resources. Their study proved that combining different models in one framework can yield much better performance. Following the success of [9], Wang *et al.* proposed a novel deep network solution for attention and aesthetics-aware image cropping, and they also utilised a cascade attention box regression and aesthetic quality classification in [31, 32]. The proposed neural network consists of two branches for predicting attention bounding boxes and analysing aesthetics. This method infers the initial crop as a bounding box covering the visually important area (attention) and then selects the best crop with the highest aesthetic quality from a few cropping candidates generated around the initial crop (aesthetic). Most recently, Horanyi *et al.* [14] proposed a multimodal description and aesthetic-based image cropping framework using explicit user description information and the aesthetics assessment. They re-purposed an existing image caption and aesthetic prediction model into one framework. They could localise the described part of the image through optimisation techniques and output an aesthetically pleasing image crop. Through excessive experiments, it was shown that the proposed framework outperforms the other learning-based automatic image-cropping methods. The drawback of this solution is taking a long time to produce an image crop due to the iterative nature of the algorithm.

2.2.1 Gaze and Description-based Image Cropping. As the gaze-based semi-automatic image cropping method does not use any additional clue *w.r.t.* the image context, it is unsuitable for description-based image cropping applications. However, it is important to note that gaze has been known as a crucial element of our social interactions, and an important characteristic of gaze in communication is that it is closely connected to speech [3, 13]. Accordingly, an analysis of communication in daily settings must address speech concerning gaze. Therefore, it is natural to assume that gaze can be coupled with other modalities, such as user description, for applied computer vision tasks. Our work is the first multimodal, gaze-initialised, and description- and aesthetics-based image cropping solution.

3 METHODOLOGY

Gaze information could be used to indicate the important part of the image. The collected fixation points and the coupling temporal information tell us which part of the image caught the viewer's eye first and how its attention shifted throughout the recording. We propose to use the gaze-based user attention information to define the most attended area of the input image in addition to the user caption to enhance the performance of the description-based image cropping algorithm.

3.1 Gaze-based initialisation of the image cropping

The collected gaze points can directly define the cropping parameters [26] or be used as complementary input coupled with other modalities in a single multimodal framework. First, from the gathered gaze points, we need to define the centre point of the content and then use predefined rules to obtain the image crops. To find the centre point and the size of the bounding box ($\mathbf{I}_{init}$) around the visually relevant part of the image, we proposed the following two solutions: *Fixed grid* and *Region proposal*.

3.1.1 Fixed grid. This solution uses a generated fixed-size grid to divide the image into $N \times N$, uniform grids G_i, and then classify the gaze points ($\mathbf{g}$) into one of the N^2 classes. To select the most attended image region, we counted the number of fixation points within each grid and chose the

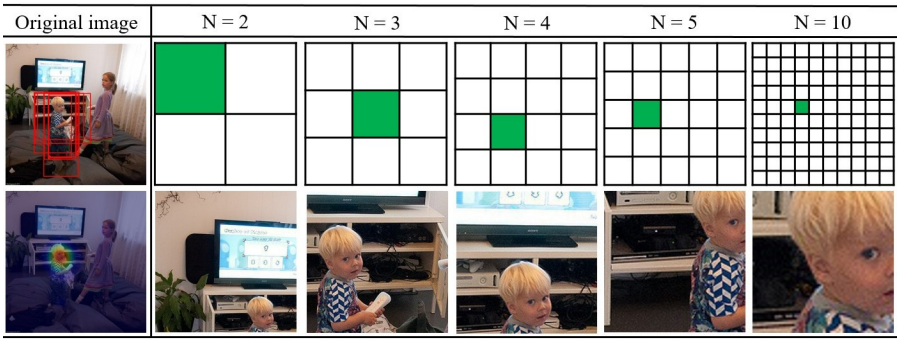

Fig. 2. **Example outputs of the fixed grid method using different scales.** In the first column, we show the original image, the ground truth bounding box annotations (top), and the heatmap generated using the recorded gaze points (bottom). We visualise the generated $N \times N$ grids, the selected section in green with the highest gaze point count, and the output initialisation region ($\mathbf{I}_{init}$) selected based on the gaze points.

one with the highest value. Denoted as

$$\mathbf{I}_{init} = argmax(\mathbf{g}(\mathbf{G}_i)), \tag{1}$$

where i=1,.., $N \times N$ corresponds to the Grid index. A generated output initialisation region ($\mathbf{I}_{init}$) example visualisation is shown in Figure 2. In the first column of the figure, at the top, we show the ground truth bounding box annotations of the image cropping dataset proposed by [14], and at the bottom, the heatmap generated from the collected gaze data. For our experiments, the grids were generated using N=2,3,4,5, and 10, shown in the first row of the figure. In the figure's bottom row, we show the gaze-based image region selection output using the *Fixed grid* method. The qualitative highlights and the quantitative evaluation (See Ablation study, Section 4.2, Table 2) of the method confirm that the gaze data is useful to initialise the search by roughly localising the centre of the described area; however, it is not flexible enough, hence not suitable to fully preserve the contextual information.

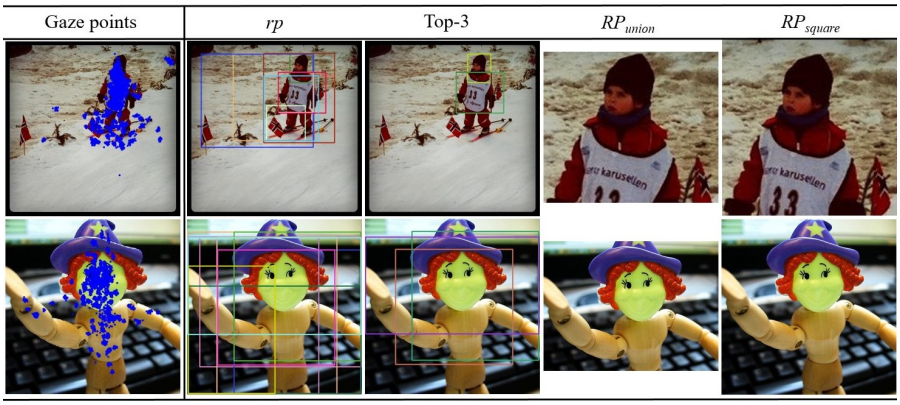

Fig. 3. **Start region generation from the collected gaze points using the region proposal method.** We show the output region proposal bounding boxes (rp) generated by [40], the selected three boxes with the highest gaze point density (Top-3), and the generated start regions RP_{union} with arbitrary aspect ratio and RP_{square} with 1:1 aspect ratio.

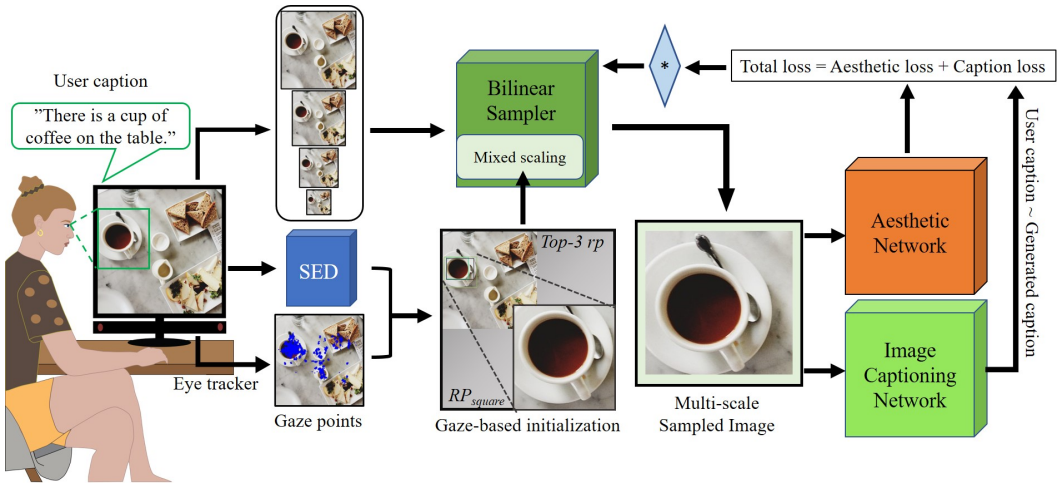

Fig. 4. **Overall framework of the proposed method: G-DAIC.** The search is initialised using the gaze points collected from the subject during the user caption generation and the Region Proposal module (SED [40]). The adaptive *Mixed Scaling* method receives the initialisation region (RP_{square}); meanwhile, the framework takes an image as input, which goes through multi-scale bilinear sampling to produce a cropped image. We then input this cropped region into the image captioning and aesthetic networks. The optimisation ends when the Total loss is below $T_{loss} = 5.23$ threshold by the proposed *Early termination* (*).

3.1.2 Region Proposal. Furthermore, the results of our ablation study, discussed in Section 4.2, show that the selection of N significantly influences the output of the *Fixed grid* method; therefore, to alleviate this problem, we propose exploiting context information by using a region proposal module. This module is implemented by Structured Edge Detector (SED) [40] to get n region bounding boxes ($rp_m, m = 1, ...n$) for each frame. Note that there are large overlaps among the generated rectangles; hence calculating the density is a better measure to identify the most attended image region than counting the number of gaze points. Therefore, for each bounding box, we calculated the gaze point density (d_m) by counting the number of gaze points inside (g_m) and dividing them by the area of the bounding box ($A(rp_m)$).

$$RP_{union} = \bigcup \left(max_3 \{ d_m = \frac{g_m}{A(rp_m)}, m = 1, .., |rp| \} \right), \tag{2}$$

where $|rp|$ is the number of bounding boxes generated by [40], d is the density function, and max_3 refers to the top-3 highest value elements of the set. We show an example of this method in Figure 3. In the first column, we visualised the collected fixation points; next, we show the generated bounding boxes using SED [40]. Due to the nature of the region proposal algorithm, we selected the three highest-density bounding boxes shown in the third column. To generate the output bounding box for the search initialisation, we merged the selected region by taking the union of the rectangles (fourth column). Then we extended this rectangle of arbitrary aspect ratio into a square (RP_{square}) (last column) initialisation region.

3.2 Proposed Framework

The proposed gaze initialised, description, and aesthetics-based image cropping framework (G-DAIC) is shown in Figure 4. Based on our experimental results, later discussed in Section 4.2, we chose the Region Proposal-based RP_{square} method for the gaze-based initialisation.

3.2.1 Mixed Scaling method. Prior methods chose the described part of the image, starting from the full image and iteratively searching for the optimal output crop. This method used a fixed scale and shrunk the sample image crop size every iteration. The gaze-based initialisation depends on the collected gaze points, which are subjective and user dependent. Therefore, the shrinking-only strategy for finding the desired output crop is not optimal. If the initialisation region is too small, the method might not be able to localise the described part of the image. Therefore, we proposed an adaptive scaling strategy based on the size of the generated initialisation area. Meaning, that based on the size of $\mathbf{I}_{init}$, the algorithm either zooms in (shrink) or zooms out (expand) every iteration with the scale. Mathematically denoted as:

$$s_{mixed} = \begin{cases} +0.98 \text{ (shrink)} & \text{if } \frac{A(\mathbf{I}_{init})}{A(\mathbf{I})} > T \\ -0.98 \text{ (expand)} & \text{otherwise} \end{cases}, \tag{3}$$

where $A(\mathbf{I}_{init})$ is the size of the gaze-based input initialisation region, $A(\mathbf{I})$ is the size of the input image, T is the threshold and s_{mixed} is the scale. The threshold $T = 0.75$ was selected empirically as part of our ablation study (See Section 4.2, Table 10). This new scaling strategy is a crucial part of the initialised search algorithm as the size of the described part of the image, and the calculated initialisation region varies based on the image content, the image caption, and the user's search behaviour.

3.2.2 Iterative optimisation. Once we calculated the RP_{square} and s_{mixed} we input these along with the original image into the *Bilinear Sampler* [15]. This module generates a multi-scale sampled image based on the input information in every iteration. The proposed sample image is chosen based on the crop parameter $\boldsymbol{\theta}$, composed of the centre coordinates of the crop x and y, and its scale s_{mixed}. A pre-trained Aesthetic Network [5] is used to generate the *Aesthetic loss* of the sample image, which reflects on the quality of the current image sample. Furthermore, we used an Image Captioning Network [35] to calculate the *Caption loss* from the user caption and the caption generated from the sampled image. The *Total loss* ($\mathcal{L}_{total}(\mathbf{I}, \mathbf{y}, \boldsymbol{\theta})$) was calculated as the sum of these two loss functions. The optimisation is performed iteratively to minimise $\mathcal{L}_{total}$ until we find the optimal output crop which best reflects the user's intention.

3.2.3 Early Termination. By using gaze-based initialisation, we better understand where the described part of the image might be. Hence, it is reasonable to assume that the method will require fewer iterations (N_{iter}) to find the optimal output image crop. To further address the runtime limitations of [14], we proposed to use an empirical threshold, $T_{loss} = 5.23$, to terminate the iterations early (*Early termination*) after n iterations ($n_{iter} < N_{iter}$) in case the total loss ($\mathcal{L}_{total}(\mathbf{I}, \mathbf{y}, \boldsymbol{\theta})$) was below a given threshold. This module was integrated into the iterative optimisation cycle.

4 EXPERIMENTS

We implement our method in Tensorflow [1]. All experiments are run on an Intel i7- CPU @3.40GHZ, 16 GB RAM, and two NVIDIA TITAN Xp GPU for fair runtime comparison with [14]. The eye-tracking data was collected using a monitor-mounted Tobii Pro Fusion Eye Tracker device and the Tobii Pro Lab v1.145 software.

4.1 Multimodal Dataset

The dataset proposed in [14] was extended with gaze data, shown in Figure 5. To analyse the correspondence between the two modalities, gaze and caption, we recorded the eye movements of the participants before and while they performed the caption-based image part localisation task.

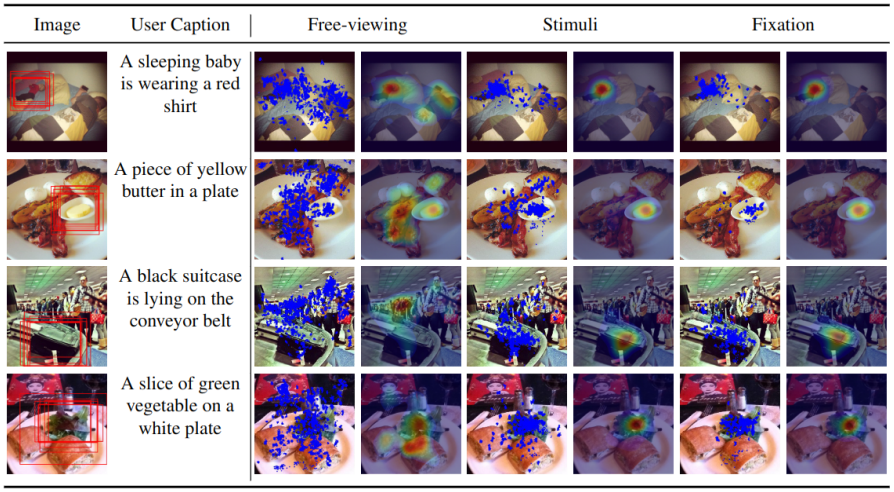

Fig. 5. **Example images of the extended, multimodal dataset.** The user-defined ground truth bounding box annotations are shown on the original images in red. We show the collected Free-viewing, Stimuli and Fixation gaze points and corresponding heatmaps.

4.1.1 Data collection. **Experimental setting.** For the data collection, we invited 14 participants to participate in our experiments. Every participant attended our experiment ten times to ensure they were not exhausted during the recording. During each session, the participant observed ten images displayed on the monitor before them. The Tobii Pro Fusion eye-tracking device was mounted to this monitor, and before every session, it was calibrated for the user. All the experiments were performed in a laboratory with controlled lighting conditions. The data collection had three stages *Free-viewing*, *Stimuli*, and *Fixation*.

Collected Gaze points. For each image, first, the users observed the image without any given instruction for 10 seconds (*Free-viewing*). We recorded the participants' eye movements during this experiment while they freely observed the previously unseen image. Without instructions, the participants naturally observe the image and spend more time on complex or interesting image parts. Following this stage, we played a recording of the corresponding image caption to the users from [14]. During the second experiment stage *Stimuli*, the participants got to know the contextual information and were asked to follow our instructions. In this phase, the participants were asked to localise the described part of the image while listening to the image caption recording. In the experiment's final *Fixation* stage, they were asked to fixate on the region of interest for 5 seconds. Using the collected gaze points, we generated heatmaps corresponding to the image caption. Note that during this stage, the participants were instructed to fixate on the described image part, which is not a single point but an image region; therefore, it is expected that the gaze points will be within a certain area but not limited to a single point.

Participant information. The participants of the eye-tracking experiment were selected to be diverse in terms of their country of origin, age, sex, and visual acuity. The participants were aged 23-29, seven males and seven females from 6 different countries. Five had perfect vision, five wore glasses, and four participants used contact lenses during the recordings. See the appendix for a detailed analysis of the collected gaze data.

4.2 Ablation Study

In this section, we compare the similarity and usefulness of human gaze data and artificial attention and the initialisation techniques introduced in Section 3.1. In addition, in the supplementary material, we present the results of additional experiments using different scaling methods (Shrink, Expand, and Mixed) and different Mixed scaling method thresholds.

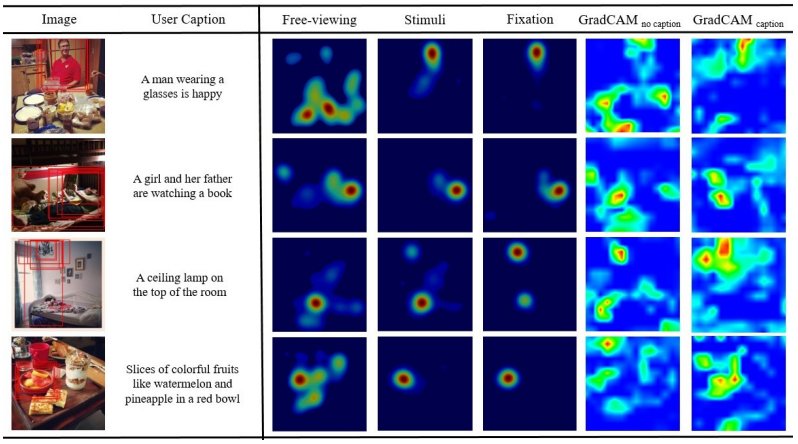

Fig. 6. **Gaze-based and artificial attention heatmap comparison.** Illustration of the heatmaps generated based on the collected gaze points from the users compared to the artificial attention heatmaps generated by with ($GradCAM_{caption}$) and without ($GradCAM_{no\ caption}$) providing the user caption to GradCAM [27].

4.2.1 Human gaze-based versus artificial attention heatmaps. In this analysis, GradCAM [27] was used in two ways to create a coarse localisation map (artificial attention heatmap) that spotlighted essential areas in the image for predicting the concept. $GradCAM_{no\ caption}$ was used without providing the user caption to the model as part of an image captioning. $GradCAM_{no\ caption}$ highlighted the image's salient features, which could be used to describe the full image. Note that this image captioning model is designed to describe the entire image, which differs from the user captions. The implementation of this model is available at [1]. $GradCAM_{caption}$ was used along with a CN to extract the parts in the image corresponding to the user caption.

Figure 6 shows example heatmaps corresponding to the collected gaze points and the generated artificial attention heatmaps. We can see that the artificial attention heatmaps are sparse and poorly aligned with the ground truth bounding box locations compared to the gaze-based heatmaps. We calculated the average AUC scores across the dataset between the heatmaps generated based on the recorded gaze points (Free-viewing, Stimuli and Fixation) to quantitatively measure the similarity between these and the artificial attention heatmaps. Table 1 summarises the average AUC scores. We found that the GradCAM heatmaps were generated without the caption, which corresponded to the free-viewing situation when the users were unaware that the caption corresponded vaguely to the human gaze-based heatmaps. $GradCAM_{caption}$ made more similar predictions to the human heatmaps. The maximum AUC achieved was 0.63, which corresponded to the fixation map.

4.2.2 Different initialisation techniques. In Table 2, we compare how the different start regions (See more in Section 3.1) correlate to the described part of the image. The start regions were generated based on the collected gaze points during Stimuli and the output heatmap of $GradCAM_{nocaption}$

[1] https://github.com/ramprs/grad-cam#image-captioning

Table 1. **Quantitative comparison of the heatmaps generated based on human gaze data and by GradCAM [27].**

AUC (Mean ± Std.)	$GradCAM_{no\ caption}$	$GradCAM_{caption}$
Free-viewing	0.556 ± 0.119	0.545 ± 0.132
Stimuli	0.572 ± 0.201	0.620 ± 0.192
Fixation	0.567 ± 0.219	**0.633 ± 0.206**

and $GradCAM_{caption}$. Without performing any optimisation or image cropping, we compare the proposed image regions with the ground truth annotations of the dataset using the Intersection over Union (IoU) measure.

The presented results show that human gaze-based initialisation results in start regions that correlate more to the desired image region than the artificial attention-based initialisation. This tendency was present regardless of the type of initialisation method. Furthermore, aligned with the results presented in Table 1, the calculated IoU score of the start regions generated by $GradCAM_{caption}$ was higher than the ones generated by $GradCAM_{nocaption}$.

The results of this experiment show that the gaze data is useful to initialise the framework; however, more is needed for the description and aesthetics-guided image-cropping task. Furthermore, the Region Proposal-based RP_{square} initialisation method was the most reliable for start region generation.

Table 2. **Quantitative comparison of the different start region generation methods using IoU measure (Mean ± Std.) on the output bounding boxes.** The gaze data used in this comparison was collected during Stimuli.

Method	$GradCAM_{nocaption}$	$GradCAM_{caption}$	Gaze data
$Fixed_{N=10}$	0.013 ± 0.003	0.013 ± 0.004	0.049 ± 0.013
$Fixed_{N=5}$	0.056 ± 0.012	0.059 ± 0.010	0.171 ± 0.007
$Fixed_{N=4}$	0.080 ± 0.011	0.094 ± 0.014	0.219 ± 0.017
$Fixed_{N=3}$	0.119 ± 0.012	0.150 ± 0.007	0.280 ± 0.005
$Fixed_{N=2}$	0.179 ± 0.009	0.223 ± 0.007	0.326 ± 0.006
RP_{union}	0.277 ± 0.009	0.290 ± 0.012	0.355 ± 0.009
RP_{square}	0.284 ± 0.010	0.294 ± 0.012	**0.361 ± 0.008**

4.2.3 Human gaze and artificial attention-based initialisation. We evaluated the performance of the G-DAIC framework using artificial attention and human gaze-based initialisation to compare their usefulness. Based on the results presented in Table 1 and 2, we used the artificial attention heatmaps generated by $GradCAM_{caption}$ to initialise our framework. Note that gaze point collection from the users during Free-viewing and Stimuli does not require additional effort from the user; therefore, it is unobtrusive.

Gaze information from different stages. In Table 3, we report the IoU scores when the Mixed scaling method was performed using different gaze data (See Section 4.1.1). Namely, we compare the differences when using Free-viewing, Stimuli, and Fixation information in the optimisation framework. In this comparison, we found that using the gaze points from the Stimuli stage results in the highest similarity with the human ground-truth annotations of the dataset. Furthermore, regardless of the chosen gaze-based initialisation technique, we can observe that the computed IoU scores are the lowest when relying on the gaze points collected during Free-viewing. We found

Table 3. **Quantitative comparison of the different attention information for start region definition using IoU measure (Mean ± Std.) on the output bounding boxes.** To obtain these results, we used the proposed Mixed scaling method for every experiment.

Attention type	$GradCAM_{caption}$	Free-viewing	Stimuli	Fixation
$Fixed_{N=10}$	0.022 ± 0.005	0.042 ± 0.016	0.075 ± 0.018	0.073 ± 0.015
$Fixed_{N=5}$	0.084 ± 0.013	0.146 ± 0.015	0.242 ± 0.007	0.237 ± 0.007
$Fixed_{N=4}$	0.140 ± 0.018	0.214 ± 0.017	0.294 ± 0.011	0.289 ± 0.013
$Fixed_{N=3}$	0.272 ± 0.020	0.271 ± 0.013	0.365 ± 0.012	0.349 ± 0.013
$Fixed_{N=2}$	0.214 ± 0.013	0.246 ± 0.015	0.328 ± 0.015	0.327 ± 0.018
RP_{union}	0.298 ± 0.013	0.332 ± 0.013	0.369 ± 0.018	0.368 ± 0.014
RP_{square}	0.307 ± 0.021	0.309 ± 0.013	**0.433 ± 0.011**	0.408 ± 0.013

the algorithm's performance using the Stimuli and Fixation information for initialisation is very similar. This is because the users knew the contextual information during both experiment stages. The difference between them might come from the fact that while the Stimuli stage is shorter and more active during the Fixation stage, the participants' attention could wander around over time, potentially influencing the initialisation by introducing Free-viewing like gaze points.

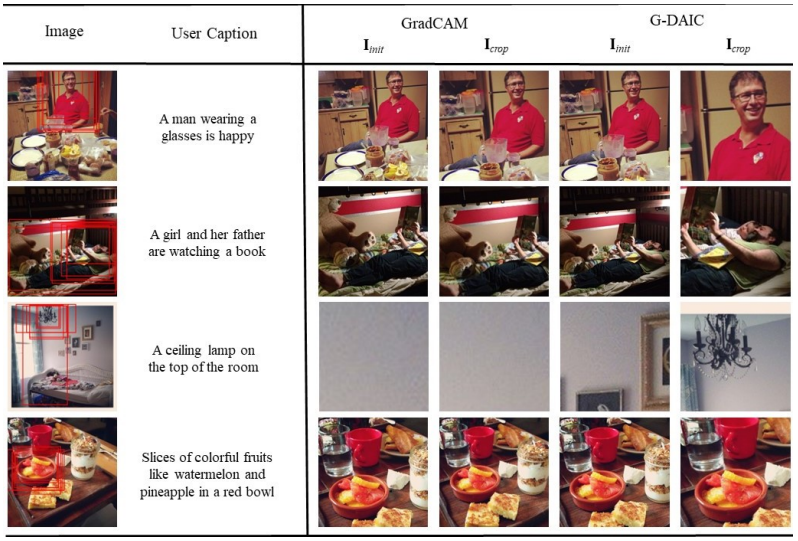

Fig. 7. **Qualitative comparison of the human gaze and artificial attention heatmap initialised output crops.** Example illustration of the generated gaze and $GradCAM_{caption}$ -based I_{init} regions and the I_{crop} output crops.

4.2.4 Human versus artificial attention. Finally, Table 3 compared also the performance of the G-DAIC framework initialised by human and artificial attention. The results invariably confirmed that the artificial attention-based initialisation ($GradCAM_{caption}$) performed worse than the gaze-based initialisation. Note that the caption-aware $GradCAM_{caption}$ initialisation yielded lower IoU scores than the Free-viewing gaze points-based output crops. This result aligns well with our qualitative (Figure 6) and quantitative (Table 1) findings regarding the artificial attention-based start regions.

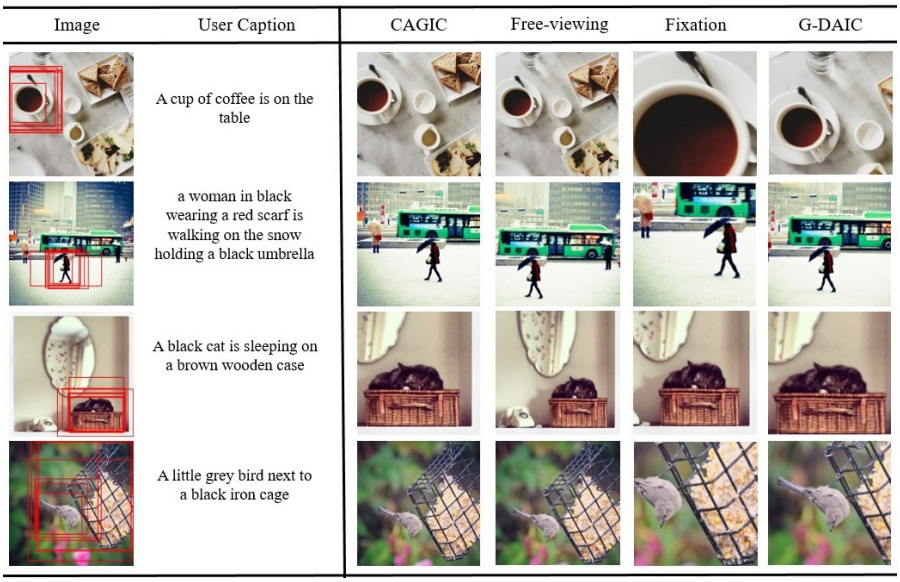

Fig. 8. **Qualitative comparison highlights.** The cropped images obtained by CAGIC, two baseline methods using Free-viewing and Fixation-based gaze initialisation and G-DAIC using Stimuli-based initialisation. The user-defined ground truth bounding box annotations are shown on the original images in red. The proposed method well crops the images as the user described.

Qualitative highlights of this experiment are included in Figure 7. This figure shows the generated RP_{square} start regions based on the collected Stimuli gaze points, the heatmap of $GradCAM_{caption}$, and the final output crops of the proposed method. The qualitative comparison shows that the human gaze data collected during Stimuli is more useful and preferable for start region generation than the artificial attention heatmaps.

4.3 Comparison with the state-of-the-art

4.3.1 Qualitative evaluation. In Figure 8, we show qualitative output examples produced by CAGIC [14], the Mixed initialisation method using Free-viewing and Fixation gaze points, and finally, the proposed method G-DAIC. In this figure, we demonstrate that using gaze data from different stages of the eye-tracking experiment in the Mixed initialisation method results in very different image crops. We can observe that initialising the optimisation framework based on the Free-viewing gaze information often returned a larger region of the image as the output image crop. Opposite to the Free-viewing output when the initialisation was based on the Fixation, the image crops tend to be tightly cropped around the subject. Overall, the crops generated by G-DAIC correspond to the captions and are aesthetically pleasing. Compared to the images of CAGIC, we found that the subjects of the caption were more centralised, and we could crop tight enough around the described image region without losing the contextual information provided by the user caption.

4.3.2 Quantitative evaluation. As part of our ablation study, we compared the quantitative performance of the proposed method using different initialisation methods, threshold levels of the proposed Mixed Scaling method, and different types of gaze information.

Overall, based on the quantitative evaluations, we found that using gaze-based initialisation is useful, and with the proposed additions, the new framework G-DAIC was able to outperform the

state-of-the-art method. Furthermore, we found that for the initialisation, it is best to use the gaze points collected during the Stimuli stage and that the proposed adaptive Mixed scaling method is better suited for our multimodal framework than the previously proposed Shrinking only method. Among the proposed initialisation methods, we found that the Region Proposal-based RP_{square} initialisation was the most successful in every experiment.

Table 4. **Comparison of user intention presence.** We ask users to caption cropped images and compare with natural language metrics how similar they are with the original desired caption.

NLP Metric	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE_L	CIDEr
GradCAM[27]	0.2728	0.1410	0.0874	0.0531	0.1182	0.2790	0.6973
MAttNet[36]	0.1718	0.0937	0.0603	0.0355	0.1132	0.2947	0.7154
CAGIC[14]	0.3424	0.1876	0.1017	0.0631	0.1702	0.2970	0.9054
G-DAIC	**0.3519**	**0.1963**	**0.1065**	**0.0654**	**0.1747**	**0.3072**	**0.9536**

4.3.3 User study. **Cycle Crop-Caption consistency.** To measure the success of the proposed image-cropping framework, we performed a quantitative comparison of how well each method preserved the contextual information of the given user caption. We asked five users who did not participate in our previous experiments, to describe the output image crops of G-DAIC. For a fair comparison, the caption similarity scores were calculated using the same natural language processing metrics as in [14]. The comparative results are presented in Table 4. This experiment demonstrated that G-DAIC was the most efficient at preserving the contextual information provided by the user according to every metric. The image captions generated from the output crops were more similar to the original user caption when we used G-DAIC compared to all the other baseline methods.

Table 5. **User study quantitative result.** Qualitative comparisons among the state-of-the-art image cropping methods with the original image and G-DAIC were compared through a human survey and evaluated by aggregation.

	Original Image	MAttNet [36]	GradCAM [27]	CAGIC [14]	G-DAIC
Aggregated percentage (%)	18.07	18.73	19.13	20.87	**23.20**

Aesthetic assessment. Due to the subjective nature of this research area, we performed a user study to compare the quality of the image crops provided by the baseline methods and G-DAIC. We performed an online user study, asking the participants to select the best-looking output image crop among the five images shown. We asked 15 users, resulting in 1500 decisions, to choose between the original image and the output image crops of MAttNet, GradCAM, CAGIC, and G-DAIC. Table 5 shows the aggregated percentages. This experiment shows that using the aesthetics information in the optimisation framework improved the quality of the output image crops. Furthermore, based on the users' votes, the most popular choice was the proposed method, G-DAIC.

4.4 Runtime

Finally, we aimed to reduce the runtime (t) of the description and aesthetics-based image cropping method compared to the state-of-the-art optimisation method [14] by adding a new modality to the framework. Therefore, beyond the quantitative and qualitative experiments, we measured the

runtime of G-DAIC and the baseline image cropping methods. The results of this experiment are shown in Table 6.

Table 6. **Runtime comparison of the baseline algorithms with the proposed method using a fixed number of iterations (G-DAIC $N_{iter} = 25$), *Early termination* (G-DAIC n_{iter}) and the $GradCAM_{caption}$-based (G-DAIC $GradCAM_{caption}$).**

Method	Runtime (sec)
A2RL [20]	0.150
VPN [33]	0.008
Anchor [38]	0.005
GradCAM [27]	0.030
MAttNet [36]	0.020
CAGIC [14]	412
G-DAIC ($GradCAM_{caption}$)	23.42
G-DAIC (N_{iter})	40.92
G-DAIC (n_{iter})	32.52

In agreement with our hypothesis, the gaze-based initialisation successfully reduced the runtime by 92.11% compared to the results presented in [14]. This runtime improvement is especially impressive, considering that both the quantitative and the qualitative evaluations confirmed that G-DAIC outperformed all the baseline methods. Therefore, the proposed method is faster and more accurate at localising the described part of the image and producing an aesthetically pleasing image crop.

4.4.1 Fixed number of iterations. This runtime improvement was achieved in three steps using two methods. Firstly, optimising the code reduced the time to run one iteration (t_{iter}) from 2.06 to 1.637 seconds. This modification resulted in an average 20.53% runtime decrease per iteration. Secondly, we maximised the number of iterations (N_{iter}) by 87.5% to 25 instead of 200, assuming that the provided gaze-based initialisation provides a better starting point for our search than starting from the original image. This resulted in a 90.07% runtime decrease, where producing a single image crop took 40.92 seconds (Table 6, G-DAIC (N_{iter})).

4.4.2 Early Termination. Finally, we used the Early Termination strategy to end the optimisation after n_{iter} instead of N_{iter} when the calculated Total loss ($\mathcal{L}_{total}(\mathbf{I}, \mathbf{y}, \boldsymbol{\theta})$) was below $T_{loss} = 5.23$. The time spent to generate a single image crop, therefore, is calculated as follows:

$$t = \begin{cases} t_{iter} \times n_{iter} & \text{if } \mathcal{L}_{total}(\mathbf{I}, \mathbf{y}, \boldsymbol{\theta}) <= T_{loss} \\ t_{iter} \times N_{iter} & \text{otherwise} \end{cases} \tag{4}$$

This solution allowed us to save 513, equal to 20.52% of the iterations across the whole dataset. The Early Termination further reduced the average runtime from 40.92 to 32.52 seconds/image (Table 6, G-DAIC (n_{iter})). Overall the average total time to produce a single output crop takes 32.52 seconds, which is 92.11% faster than the original method.

The runtime achieved by G-DAIC is significantly faster than the other iterative optimisation-based description and aesthetics-based image cropping algorithm CAGIC; however, due to the iterative nature of the proposed algorithm, it is still significantly slower but better in terms of performance than the other baselines. Finally, we evaluated the runtime of G-DAIC when initialised by $GradCAM_{caption}$. While this method achieved 34% faster runtime compared to G-DAIC (n_{iter}), the quantitative and qualitative results presented in Section 4.2 showed that its performance was

significantly lower compared to the human gaze-based initialisation. Note that the models in our framework were not fine-tuned nor specifically trained for this task. The overall runtime could be further improved by using new modalities and potentially replacing existing or new pre-trained models in the framework.

5 CONCLUSION

This work proposes a new gaze-initialised multimodal optimisation method for the description and aesthetics-based image cropping problem. The main motivation was to reduce the runtime of the state-of-the-art image cropping method [14] while preserving the accuracy of the previously introduced solution. Hence, we designed a new framework, which initialised the image cropping algorithm based on the subject's eye movements recorded during the image description generation. In this work, we proposed two solutions, *Fixed Grid* and *Region Proposal*, on integrating and utilising the gaze information into the proposed multimodal framework. Furthermore, we implemented the adaptive *Mixed method* for localisation of the described image region based on the size of the gaze-based initialisation area. Finally, we proposed the *Early termination* of the optimisation to significantly reduce the runtime. In this paper, we have shown that with the gaze-based initialisation of the description and aesthetic-based image cropping method, we were able to outperform the state-of-the-art methods significantly and in addition, we reduced the runtime by more than 92%.

ACKNOWLEDGMENTS

This work was supported by the Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korean government (MSIT) (No.2021-0-00034, Clustering technologies of fragmented data for time-based data analysis). The authors wish to acknowledge and thank Lili Tombacz for using her custom-made illustrations.

REFERENCES

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D.G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. 2016. Tensorflow: A System for Large-Scale Machine Learning. In *USENIX Conference on Operating Systems Design and Implementation*. 265–283.

[2] Edoardo Ardizzone, Alessandro Bruno, and Giuseppe Mazzola. 2013. Saliency Based Image Cropping. In *Image Analysis and Processing – ICIAP 2013*, Alfredo Petrosino (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 773–782.

[3] Ülkü Arslan Aydin, Sinan Kalkan, and Cengiz Acartürk. 2021. Speech Driven Gaze in a Face-to-Face Interaction. *Frontiers in Neurorobotics* (2021), 8.

[4] Caroline PC Chanel, Raphaëlle N Roy, Frédéric Dehais, and Nicolas Drougard. 2020. Towards mixed-initiative human–robot interaction: Assessment of discriminative physiological and behavioral features for performance prediction. *Sensors* 20, 1 (2020), 296.

[5] Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma. 2017. Learning to Compose with Professional Photographs on the Web. In *ACM International Conference on Multimedia*. ACM, 37–45.

[6] Marcella Cornia, Stefano Pini, Lorenzo Baraldi, and Rita Cucchiara. 2018. Automatic image cropping and selection using saliency: An application to historical manuscripts. In *Italian Research Conference on Digital Libraries*. Springer, 169–179.

[7] Chris Creed, Maite Frutos-Pascual, and Ian Williams. 2020. Multimodal Gaze Interaction for Creative Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[8] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. 2019. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE International Conference on Computer Vision*. 2601–2610.

[9] Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. 2014. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *Proceedings of the 22nd ACM international conference on Multimedia*. 1105–1108.

[10] Onur Ferhat, Fernando Vilarino, and Francisco Javier Sánchez. 2014. A cheap portable eye-tracker solution for common setups. *Journal of eye movement research* 7, 3 (2014).

[11] Atsushi Fukayama, Takehiko Ohno, Naoki Mukawa, Minako Sawaki, and Norihiro Hagita. 2002. Messages embedded in gaze of interface agents—impression management with agent's gaze. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 41–48.

[12] Ruth B Grossman, Erin Steinhart, Teresa Mitchell, and William McIlvane. 2015. "Look who's talking!" gaze patterns for implicit and explicit audio-visual speech synchrony detection in children with high-functioning autism. *Autism Research* 8, 3 (2015), 307–316.

[13] Simon Ho, Tom Foulsham, and Alan Kingstone. 2015. Speaking and listening with the eyes: Gaze signaling during dyadic interactions. *PloS one* 10, 8 (2015), e0136905.

[14] Nora Horanyi, Kedi Xia, Kwang Moo Yi, Abhishake Kumar Bojja, Aleš Leonardis, and Hyung Jin Chang. 2022. Repurposing existing deep networks for caption and aesthetic-guided image cropping. *Pattern Recognition* 126 (2022), 108485.

[15] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. 2015. Spatial Transformer Networks. 2017–2025.

[16] Nehal Jaiswal and Yogesh K Meghrajani. 2015. Saliency based automatic image cropping using support vector machine classifier. In *ICIIECS*. IEEE, 1–5.

[17] Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. 2013. Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 3, 2 (2013), 1–30.

[18] Anam Ahmad Khan, Joshua Newn, James Bailey, and Eduardo Velloso. 2022. Integrating Gaze and Speech for Enabling Implicit Interactions. In *CHI Conference on Human Factors in Computing Systems*. 1–14.

[19] Fatemeh Koochaki and Laleh Najafizadeh. 2018. Predicting intention through eye gaze patterns. In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 1–4.

[20] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. 2018. A2-RL: Aesthetics Aware Reinforcement Learning for Image Cropping. 8193–8201.

[21] Paul P Maglio, Teenie Matlock, Christopher S Campbell, Shumin Zhai, and Barton A Smith. 2000. Gaze and speech in attentive user interfaces. In *International Conference on Multimodal Interfaces*. Springer, 1–7.

[22] Michael Nauge, Mohamed-Chaker Larabi, and Christine Fernandez-Maloigne. 2012. A statistical study of the correlation between interest points and gaze points. In *Human Vision and Electronic Imaging XVII*, Vol. 8291. SPIE, 308–322.

[23] Matheus Vieira Portela and David Rozado. 2014. Gaze enhanced speech recognition for truly hands-free and efficient text input during hci. In *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: the Future of Design*. 426–429.

[24] Ziaur Rahman, Yi-Fei Pu, Muhammad Aamir, and Farhan Ullah. 2018. A framework for fast automatic image cropping based on deep saliency map detection and gaussian filter. *International Journal of Computers and Applications* (2018), 1–11.

[25] Kyle Reinholt, Darren Guinness, and Shaun K Kane. 2019. Eyedescribe: Combining eye gaze and speech to automatically create accessible touch screen artwork. In *Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces*. 101–112.

[26] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. 2006. Gaze-based interaction for semi-automatic photo cropping. In *SIGCHI conference on Human Factors in computing systems*. ACM, 771–780.

[27] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. (2017), 618–626.

[28] Malcolm Slaney, Rahul Rajan, Andreas Stolcke, and Partha Parthasarathy. 2014. Gaze-enhanced speech recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3236–3240.

[29] Kalin Stefanov, Jonas Beskow, and Giampiero Salvi. 2017. Vision-based active speaker detection in multiparty interaction. In *Grounding Language Understanding GLU2017 August 25, 2017, KTH Royal Institute of Technology, Stockholm, Sweden*.

[30] Ece Takmaz, Sandro Pezzelle, Lisa Beinborn, and Raquel Fernández. 2020. Generating image descriptions via sequential cross-modal alignment guided by human gaze. *arXiv preprint arXiv:2011.04592* (2020).

[31] Wenguan Wang and Jianbing Shen. 2017. Deep cropping via attention box prediction and aesthetics assessment. In *Proceedings of the IEEE international conference on computer vision*. 2186–2194.

[32] Wenguan Wang, Jianbing Shen, and Haibin Ling. 2018. A deep network solution for attention and aesthetics aware photo cropping. *IEEE transactions on pattern analysis and machine intelligence* 41, 7 (2018), 1531–1544.

[33] Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomír Mech, Minh Hoai, and Dimitris Samaras. 2018. Good view hunting: learning photo composition from dense view pairs. In *CVPR*. 5437–5446.

[34] Adam S Williams and Francisco R Ortega. 2020. Understanding gesture and speech multimodal interactions for manipulation tasks in augmented reality using unconstrained elicitation. *Proceedings of the ACM on Human-Computer Interaction* 4, ISS (2020), 1–21.

[35] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. (2015).

[36] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*. 1307–1315.

[37] Liangzhe Yuan, Christopher Reardon, Garrett Warnell, and Giuseppe Loianno. 2019. Human gaze-driven spatial tasking of an autonomous MAV. *IEEE Robotics and Automation Letters* 4, 2 (2019), 1343–1350.

[38] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. 2019. Reliable and Efficient Image Cropping: A Grid Anchor Based Approach. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[39] Hongzhi Zhu, Septimiu E Salcudean, and Robert N Rohling. 2019. A novel gaze-supported multimodal human–computer interaction for ultrasound machines. *International journal of computer assisted radiology and surgery* 14, 7 (2019), 1107–1115.

[40] C Lawrence Zitnick and Piotr Dollár. 2014. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*. Springer, 391–405.

A APPENDIX

A.1 Analysis of the collected gaze data

We performed multiple experiments to evaluate the quality of the collected gaze points. First, we analysed how well the gaze points correspond to the dataset's ground truth bounding box annotations by calculating the proportion of the points inside these bounding boxes. This is an important measure as it influences the quality of the gaze-based initialisation, hence the overall framework's accuracy. Then we investigated the dynamic nature of the collected Stimuli gaze data. We aimed to understand the participants' eye movements *w.r.t.* the words of the image caption.

Table 7. **The proportion of detected gaze points inside the ground truth bounding box during Free-viewing, Stimuli and Fixation.**

GT	Free viewing	Stimuli	Fixation	Mean ± std
1	58.78 ± 29.91	87.75 ± 17.91	92.42 ± 19.91	79.65 ± 22.58
2	56.25 ± 32.44	83.75 ± 22.77	90.31 ± 23.31	76.77 ± 26.18
3	54.50 ± 31.22	84.37 ± 21.24	90.37 ± 22.77	76.41 ± 25.07
4	57.67 ± 31.46	85.34 ± 21.17	91.48 ± 22.21	78.16 ± 24.95
5	59.12 ± 29.70	86.96 ± 18.51	92.51 ± 19.94	79.53 ± 22.72
6	50.01 ± 32.69	79.41 ± 26.50	85.26 ± 28.97	71.59 ± 29.39
7	49.04 ± 29.83	81.46 ± 22.63	88.28 ± 25.11	72.92 ± 25.86
Mean ± std	55.06 ± 31.28	84.15 ± 21.86	90.09 ± 23.47	

A.1.1 Gaze points w.r.t. bounding box annotations. We evaluated the different stages of the eye tracking recording individually and compared the recorded points *w.r.t.* the seven ground truth bounding boxes of the dataset. In this experiment, we used the recorded gaze points from every participant. In Table 7, we can see that the proportion of the gaze points inside the bounding boxes is very similar for all three stages of the recording. Furthermore, our results show that over 71% of the gaze points were within the bounding boxes for every ground truth bounding box.

The lowest percentage of gaze points inside the ground truth bounding boxes belongs to the Free-viewing stage of the experiment, which is not surprising as the participants were allowed to observe the image without any contextual constraints during this stage. During the Stimuli stage, the number of gaze points within the target area increased by more than 29%, reaching nearly 84% accuracy. Overall, this is a high percentage considering that the proportion was about 90% high during the Fixation stage. In a real-world scenario, users do not tend to fixate on the described part of the image after providing the description. Therefore, while the Fixation points are more aligned with the ground truth bounding box annotations of the dataset, we used the Stimuli points in our experiments for initialisation.

A.1.2 Gaze points w.r.t. participants. It is important to note that the participants' data had some disagreements, similar to the subjective nature of the user-annotated bounding boxes of the dataset. Therefore, we experimented to better understand the subjective nature of the gaze points. The results of this experiment are shown in Table 8. We can observe that some subjects, like User 10, had higher accuracy during all three stages of the experiment than others. However, the tendency among the stages is the same for every user, and the overall percentages are close to each other too.

A.2 Ablation study

Here, we report more ablation studies left out of the main paper due to spatial constraints. We motivate all our choices based on these studies.

Table 8. **The proportion of detected gaze points inside the ground truth bounding boxes *w.r.t.* the subjects during Free-viewing, Stimuli and Fixation.**

User	Free viewing	Stimuli	Fixation	Mean ± std
1	55.98 ± 32.63	87.73 ± 17.65	92.72 ± 19.64	78.81 ± 23.30
2	52.48 ± 28.91	89.27 ± 15.63	94.67 ± 17.05	78.81 ± 20.53
3	53.39 ± 33.01	87.97 ± 18.28	94.18 ± 19.97	78.51 ± 23.76
4	59.34 ± 34.73	76.84 ± 30.38	80.87 ± 33.11	72.35 ± 32.74
5	52.45 ± 30.42	84.70 ± 19.55	86.36 ± 25.05	74.50 ± 25.01
6	47.89 ± 28.73	87.18 ± 18.15	94.67 ± 19.58	76.58 ± 22.15
7	54.84 ± 30.53	86.84 ± 18.72	93.93 ± 18.68	78.54 ± 22.65
8	54.62 ± 30.09	83.11 ± 25.92	89.95 ± 27.36	75.89 ± 27.79
9	57.96 ± 31.37	83.57 ± 22.24	92.23 ± 22.35	77.92 ± 25.32
10	66.00 ± 33.92	86.04 ± 22.10	93.03 ± 23.61	81.69 ± 26.54
11	54.58 ± 29.22	80.87 ± 25.03	88.27 ± 23.59	74.57 ± 25.95
12	57.84 ± 31.33	81.77 ± 23.34	90.39 ± 21.61	76.66 ± 25.43
13	54.46 ± 31.04	82.70 ± 20.62	89.39 ± 23.04	75.52 ± 24.90
14	55.92 ± 31.57	78.44 ± 23.64	79.90 ± 27.02	71.42 ± 27.41
Mean ± std	55.06 ± 31.28	84.15 ± 21.86	90.09 ± 23.47	

A.2.1 Different scaling methods. In Table 9 we show the IoU scores *w.r.t.* different scaling methods. Note that we used the gaze points collected during the stimuli stage during this experiment based on our ablation study's findings. CAGIC [14] iteratively zooms into the described part of the image (Shrink) and does not have gaze-based initialisation (See more in Section 3). When the image cropping framework is initialised, we use three scaling methods to find the region of interest. Namely, we zoomed in or out in every iteration despite the size of the initialisation region. Alternatively, in Section 3.2.1, we proposed the adaptive Mixed scaling method, which flexibly decides to use Shrink or Expand *w.r.t.* the size of the initialisation area. Our results show that the Mixed scaling method using Region Proposal initialisation provides the highest IoU value among the compared methods, exceeding even the score of the state-of-the-art method, CAGIC.

Table 9. **Quantitative comparison of the different scaling methods using IoU measure (Mean ± Std.) on the output bounding boxes.** All the gaze-based initialisation methods use the Stimuli gaze points.

Scaling Method	Shrink	Expand	Mixed
$Fixed_{N=10}$	0.042 ± 0.021	0.074 ± 0.016	0.075 ± 0.018
$Fixed_{N=5}$	0.151 ± 0.015	0.241 ± 0.008	0.242 ± 0.007
$Fixed_{N=4}$	0.188 ± 0.015	0.303 ± 0.011	0.294 ± 0.011
$Fixed_{N=3}$	0.262 ± 0.010	0.365 ± 0.010	0.365 ± 0.012
$Fixed_{N=2}$	0.302 ± 0.013	0.286 ± 0.019	0.328 ± 0.015
RP_{union}	0.352 ± 0.015	0.330 ± 0.017	0.369 ± 0.018
RP_{square}	0.369 ± 0.007	0.388 ± 0.011	**0.433 ± 0.011**
CAGIC	0.416 ± 0.013	-	-

A.2.2 Different Mixed Scaling method thresholds. We analysed the performance of the Mixed scaling method using different thresholds. We used a 0.75 threshold level in our experiments, as mentioned in Section 3.2.1. This threshold level was chosen empirically based on the results of our experiments presented in Table 10. In this table, we show the performance of the fixed grid and region proposal-based gaze initialisation methods using the Mixed method with different threshold levels.

Received November 2022; revised February 2023; accepted March 2023

Table 10. **Quantitative comparison of the different thresholds (T) of the Mixed scaling method using IoU measure (Mean ± Std.) on the output bounding boxes.**

T	0.125	0.250	0.375	0.500	0.625	0.750	0.875
$Fixed_{N=10}$	0.076 ± 0.025	0.070 ± 0.016	0.075 ± 0.017	0.085 ± 0.007	0.075 ± 0.019	0.075 ± 0.018	0.073 ± 0.019
$Fixed_{N=5}$	0.152 ± 0.011	0.243 ± 0.008	0.241 ± 0.008	0.246 ± 0.004	0.240 ± 0.008	0.242 ± 0.007	0.242 ± 0.005
$Fixed_{N=4}$	0.195 ± 0.018	0.289 ± 0.015	0.301 ± 0.008	0.300 ± 0.012	0.296 ± 0.009	0.294 ± 0.011	0.289 ± 0.010
$Fixed_{N=3}$	0.259 ± 0.007	0.258 ± 0.010	0.357 ± 0.011	0.371 ± 0.010	0.369 ± 0.010	0.365 ± 0.012	0.367 ± 0.013
$Fixed_{N=2}$	0.305 ± 0.014	0.301 ± 0.009	0.298 ± 0.014	0.314 ± 0.011	0.317 ± 0.017	0.328 ± 0.015	0.311 ± 0.019
RP_{union}	0.352 ± 0.016	0.362 ± 0.014	0.377 ± 0.014	0.361 ± 0.017	0.378 ± 0.021	0.369 ± 0.018	0.363 ± 0.020
RP_{square}	0.359 ± 0.007	0.368 ± 0.009	0.382 ± 0.007	0.419 ± 0.010	0.428 ± 0.011	**0.433 ± 0.011**	0.431 ± 0.012