electronics

MDPI

Article

Directorial Editing: A Hybrid Deep-Learning Approach to Content-Aware Image Retargeting and Resizing

Elliot Dickman and Paul Diefenbach *

College of Media Arts and Design, Drexel University, Philadelphia, PA 19104, USA; ed532@drexel.edu
* Correspondence: pjdief@drexel.edu

Abstract: Image retargeting is a common computer graphics task which involves manipulating the size or aspect ratio of an image. This task often presents a challenge to the artist or user, because manipulating the size of an image necessitates some degree of data loss as pixels need to be removed to accommodate a different image size. We present an image retargeting framework which implements a confidence map generated by a segmentation model for content-aware resizing, allowing users to specify which subjects in an image to preserve using natural language prompts much like the role of an art director conversing with their artist. Using computer vision models to detect object positions also provides additional control over the composition of the retargeted image at various points in the image-processing pipeline. This object-based approach to energy map augmentation is incredibly flexible, because only minor adjustments to the processing of the energy maps can provide a significant degree of control over where seams—paths of pixels through the image—are removed, and how seam removal is prioritized in different sections of the image. It also provides additional control with techniques for object and background separation and recomposition. This research explores how several different types of deep-learning models can be integrated into this pipeline in order to easily make these decisions, and provide different retargeting results on the same image based on user input and compositional considerations. Because this is a framework based on existing machine-learning models, this approach will benefit from advancements in the rapidly developing fields of computer vision and large language models and can be extended for further natural language directorial controls over images.

Keywords: computational photography; artificial intelligence; image composition

Citation: Dickman, E.; Diefenbach, P. Directorial Editing: A Hybrid Deep-Learning Approach to Content-Aware Image Retargeting and Resizing. *Electronics* **2024**, *13*, 4459. https://doi.org/10.3390/electronics13224459

Academic Editor: Duc Thanh Nguyen

Received: 7 October 2024
Revised: 11 November 2024
Accepted: 12 November 2024
Published: 14 November 2024

1. Introduction

Image resizing, or retargeting, is a fundamental image processing task in the fields of digital and print media. Despite its importance and popularity, traditional techniques often struggle to maintain image composition or subject integrity when changing the aspect ratio of an image. This paper explores methods of integrating deep-learning models for object detection and image segmentation into the image retargeting pipeline. The proposed methodology introduces a salience map informed by object and segmentation data, providing a more conceptually nuanced understanding of images and facilitating improved control over final image composition. Additionally, we incorporate natural language processing (NLP) models into this pipeline to allow for intuitive interaction, art-direction, and control over the retargeting process.

This study predominantly focuses on seam carving as a basis for content-aware retargeting. Seam carving is a naïve pixel-processing technique that aims to remove the least-noticeable pixels from an image during the retargeting process. This method was chosen due to its flexibility and popularity, but the fundamental principles can be extended to other retargeting techniques. We also propose specific implementations of these hybrid deep-learning approaches, presenting a shift from basic pixel processing to a more comprehensive, AI-assisted image retargeting.

Accordingly, this research explores the following:

- The use of computer vision models to generate alternative salience maps for use in content-aware image retargeting;
- Different methods of generating and integrating those maps into the seam carving algorithm;
- Possible methods of using machine learning models with semantic image comprehension in conjunction with seam carving in order to retarget images with greater art-directability than existing retargeting methods;
- Use-cases and limitations for the proposed methods.

The initial results of these proposed algorithms demonstrate significant qualitative improvements over naïve pixel processing approaches to content-aware image retargeting, most notably in preserving the visual integrity of important or user-specified subjects. This integration of advanced machine learning techniques in the image retargeting workflow empowers users with greater control and reduced manual intervention, paving the way for future exploration and innovation in this field.

2. Background

Image retargeting involves resizing images to meet specific design requirements while maintaining visual quality [1]. Common methods include cropping and scaling. Cropping removes unwanted sections but may lose important content, while scaling alters the image size, often leading to artifacts like jagged edges and loss of detail [2]. Image retargeting involves resizing images to meet specific design requirements while maintaining visual quality.

One approach to addressing this problem is through content-aware image retargeting techniques, which take into consideration the visual features and salient regions of the image when performing resizing operations [3]. These methods analyze the image content to identify areas of importance, such as edges, textures, and semantic objects, and prioritize their preservation during the retargeting process. By focusing on preserving the most visually significant elements, content-aware techniques can minimize the impact of data loss on the overall visual perception of the final image.

This desired directing of data loss can be achieved by computing the areas of an image that have the least impact on the visual perception of the image and removing image data from those locations [3]. The idea that some parts of an image are more noticeable to the human eye than others is known as visual salience. This can refer to image features, objects, and compositional elements.

The basic seam carving algorithm operates under the rule that harder edges—edges in an image with higher contrast—are more salient than soft edges or areas with no edges [4]. This salience model generates an energy map in which pixel groups with higher contrast edges are assigned a higher energy value, and are considered to be of greater visual salience.

Seam carving is therefore a retargeting method that leverages a contrast-based salience model to remove the lowest-contrast pixel groups from an image. It was first proposed in the paper "Seam Carving for Content-Aware Image Resizing" [4] and is a content-aware alternative to traditional image retargeting approaches such as cropping and scaling.

Other models of computing salience are used as well. A human-attention model has been proposed, where a complex set of rules was developed that accounted for color, orientation, direction of movement, and other compositional features, in order to determine the areas of an image that are associated with heightened visual attention [5].

Another popular salience metric is depth, assuming that parts of a composition that are in the foreground are more salient than parts of an image that are in the background [6]. Metrics such as depth data have been proven useful in improving the quality of image retargeting output; however, it is important to note that the requirement of depth data, whether acquired from specialized depth-sensing cameras or manually assigned through masks [7], can pose limitations in terms of data acquisition and annotation efforts. While there is increased capacity for consumer-grade devices to record depth data, these data are

not frequently captured and recorded in images that are shared across platforms, and many users do not use depth recording functionality unless they are advanced technical users [6].

An alternative energy function implementation called forward-energy has been proposed, deviating from the conventional approach of removing the seam with the lowest energy. Instead, the forward-energy function focuses on identifying the seam that results in the minimal overall change in energy across the image [8].

The forward-energy function offers a different approach to seam removal, aiming to minimize the disruption and distortion caused by the seam removal process. By considering the cumulative energy change along the path of a seam, this alternative approach attempts to better prioritize the preservation of image content and structure. By removing seams that cause the least disturbance to the image energy, the forward-energy function seeks to achieve a resizing result with reduced distortion and enhanced visual quality in certain scenarios.

While the forward-energy function proves to be successful in reducing distortion in some use cases, it does not universally outperform the traditional energy-based approach. The effectiveness of the forward-energy function is influenced by various factors, such as the image content, the specific resizing requirements, and the characteristics of the energy distribution within the image [8].

The primary idea behind seam carving is that an image can be resized by removing or adding a seam anywhere in the image, where a seam is an 8-connected path of pixels across the width or height of the image [4]. Detecting the optimal seam to remove is a fundamental part of the seam carving algorithm. Searching for the seam with the least total change in value results in a seam that crosses the least salient parts of an image, assuming that hard edges and high contrast equate to higher salience.

An energy function computes a map of the change in pixel values across an image.

The energy function used in the originally proposed naïve seam carving algorithm [4] calculates the difference in value from a given pixel x, y and the pixel immediately below and immediately to the right, with

$$e_1(I) = \left|\frac{\partial}{\partial x}I\right| + \left|\frac{\partial}{\partial y}I\right|$$

A vertical seam is defined for image I with height n in the following set of pixels:

$$s^x = \{s_i^x\}_{i=1}^n = \{(x(i), i)\}_{i=1}^n, \forall i|x(i) - x(i-1)| \leq 1$$

making the set of pixels in image I comprising seam s

$$I_s = \{I(s_i)\}_{i=1}^n$$

Computing the optimal seam s^* that minimizes the energy of the seam across $e_1(I)$ can therefore be found with

$$s^* = \min_s \sum_{i=1}^{n} e(I(s_i)).$$

Finding the set of k lowest cost seams results in a set consisting of k optimal seams s^*, where each iteration of sampling removes s^* from both I and $e_1(I)$, and shifts all pixels that are to the right of s^* one pixel to the left, reducing the image width by one pixel.

This results in an image that is decreased in size by k pixels, but without cropping or scaling the content (Figure 1). While there are times that cropping or scaling may be acceptable retargeting methods, seam carving provides a distinctly different third option that aims to selectively preserve pixel data in areas of higher visual interest. This method of image retargeting is often effective at resizing images while maintaining the visual integrity of the subject, but often results in unwanted artifacting and distortion of important image content (Figure 2).

Figure 1. Original image (**a**), Retargeted image with seam carving (**b**), scaling (**c**), and cropping (**d**) [4].

Figure 2. Distortions from naïve seam carving, with noticeable distortion in the ibex and the diagonal lines of the vehicle.

While numerous variations to the seam carving algorithm have been proposed to resolve these issues, including methods such as the human-attention model [5] and depth-based salience [6] as described above, these proposed solutions do not take into account a semantic understanding of the image and its composition, which can allow for more reliable and art-directable image retargeting.

3. Methods

We propose a set of methodologies for integrating deep-learning frameworks for object detection and image segmentation into a content-aware retargeting pipeline, with a focus on energy map augmentation for the seam carving algorithm. Leveraging the capabilities of these models can allow for the use of high-level image information to better preserve subjects and provide additional automated control over the retargeted image. In addition, this method provides the capability of filtering out unwanted objects through natural language identification, as well as weighting the relative importance of image subjects. The level of scene comprehension that can be achieved with these advanced deep-learning computer vision models can also allow for natural-language-driven adjustments to the composition of the retargeted image.

3.1. Object Detection and Segmentation

This research utilizes the InceptionResNetV2 network for object detection [9]. This network uses the Faster R-CNN framework and was authored by researchers at Google in 2016. This architecture was chosen because it is well documented, fast, and reliable, and has many pre-trained models available. One of the key features of InceptionResNetV2 is its use of "bottleneck" layers, which help to reduce the computational complexity of the network while preserving its expressive power. This allows for faster training and inference times, while still maintaining high accuracy [10]. The model that was used in this research was trained on the Open Images v4 dataset, which contains 600 object classes across 1.74 million images [11].

CLIPseg is the model that we used for semantic segmentation. It was developed by researchers at OpenAI in San Francisco in 2021, and it has been shown to achieve impressive results on a range of image and video understanding tasks. The basic idea behind CLIPseg is to use a contrastive loss function to learn representations that capture both the semantic content and spatial layout of objects in an image. These representations can then be used to perform accurate semantic segmentation, which involves labeling each pixel in the image with the corresponding object class. This allows for more fine-grained understanding of image content [12]. The output from CLIPseg is processed to create a grayscale image mask indicating the confidence that each pixel is part of a given object class.

Three different methods were used to generate confidence maps across the image processing pipeline with varying degrees of automation and manual control.

The first method ("automated") is fully automated, using the object detection model to identify which objects are in the image, and passing those detected classes on to CLIPseg to create a confidence map for each detected class. CLIPseg then scans the entire image for each class and returns the segmentations (Figure 3).

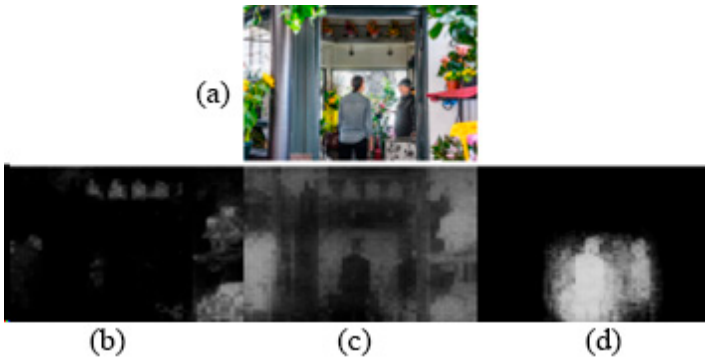

Figure 3. Original image (**a**) with confidence maps for "flower" (**b**), "houseplant" (**c**), and "man" (**d**) classes using the automated method.

The second method ("bounding box") is also fully automated, but the object detection model passes only the bounding box for each detected class to CLIPseg for segmentation instead of the entire image. This results in a higher resolution confidence map, because when a smaller image (the area of the image inside the detected bounding box) is converted to a tensor of the same size as the full image, the smaller image will result in a tensor with more granular representation. The downside to this approach is that if the detection model fails to identify all instances of an object class, the segmentation model will completely ignore it.

The third method bypasses the object detection step and relies on natural language input, but provides the most control over the result and tends to be the most accurate. With this method, the user can specify in natural language what classes to segment and how they should be weighted. A GPT language model such as ChatGPT or GPT-4 [13] is then used to convert the input to a syntax that is able to be parsed into individual classes and weights, which are passed on to CLIPseg and later used to adjust the weighting of each

confidence map (Figure 4). An example of such input would be "Identify cats and dogs in the image. Dogs should be weighted significantly higher than cats." This input results in two confidence maps, one identifying cats and given a weight of 0.2, the other identifying dogs and given a weight of 1.

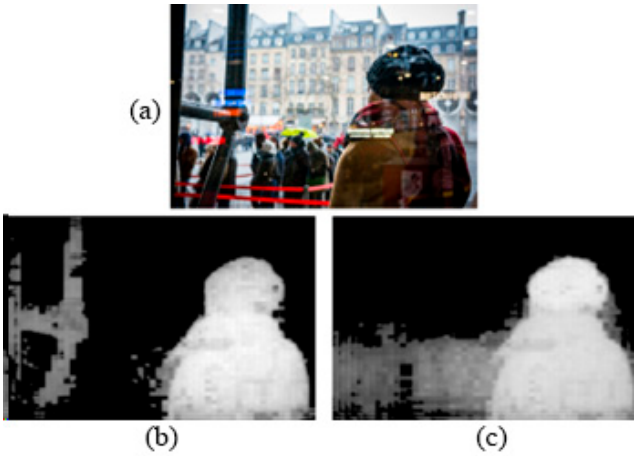

Figure 4. Original image (**a**) with NLP confidence map using prompt "Woman and metal bar. The woman is more important." (**b**), compared to the automated segmentation output map (**c**).

3.2. Processing Segmentation Output

After obtaining separate confidence maps for each detected class from the segmentation model, some post-processing steps are required to make these maps usable for energy augmentation. One crucial step is to composite these individual confidence maps into a single map that represents the overall augmentation map.

The simplest approach to compositing the confidence maps is to take the highest confidence value at each pixel location across all the individual maps. By selecting the maximum value, the resulting augmentation map will have the highest confidence value across all the segmentation classes at each pixel, indicating the presence of the most confident class in that particular location.

In addition to compositing the maps, it is often desirable to apply a threshold to the augmentation map. The thresholding process involves setting a minimum confidence value below which the pixel is considered to be of low confidence. This thresholding step helps reduce noise in areas of lower confidence, preventing them from unduly influencing the subsequent energy augmentation. Simultaneously, it enhances the relative efficacy of high-confidence areas without artificially increasing their weights.

If the size of the tensor used for segmentation is smaller than the size of the original image, grid-line artifacts may occur due to the size discrepancy. To mitigate these artifacts, a simple Gaussian blur can be applied across the entire augmentation map. The Gaussian blur smooths out the confidence values, effectively reducing the visibility of grid-line artifacts and producing a more visually coherent and seamless augmentation map.

While the described post-processing steps provide a general framework, specific implementations may vary depending on the requirements of the task and the characteristics of the segmentation model. With our models, we found that this process of compositing energy maps, applying a threshold, and then applying a gaussian blur yielded effective results.

3.3. Augmenting Energy Maps

The original e_1 energy function can be augmented with the output from the image segmentation model. Segmentation will output an image map with the confidence for each pixel mapped to how likely it is that that pixel is a part of the given input image prompt.

This information can be used to artificially increase the energy of the areas of the image that are part of an object, with a new energy function:

$$e_{aug}(I) = w_s \cdot \text{confidence}(I) + \left|\frac{\partial}{\partial x} I\right| + \left|\frac{\partial}{\partial y} I\right| = w_s \cdot \text{confidence}(I) + e_1(I)$$

where w is a weight constant. This augmentation is modeled after other successful adaptations of the e_1 energy function for seam carving [14].

Alternatively, element-wise multiplication of the confidence and e_1 matrices produces the Hadamard product of the confidence and energy maps:

$$e_{dim}(I) = \text{confidence}(I) \odot \left(\left|\frac{\partial}{\partial x} I\right| + \left|\frac{\partial}{\partial y} I\right|\right) = \text{confidence}(I) \odot e_1(I)$$

This method is less flexible as there is no control over the overall weighting of the confidence relative to the original energy, but is likely to yield slightly different results. This method effectively reduces the energy value of any pixel that is less likely to be a part of an object, and gives more weight to e_1 as compared to e_{aug} when $w_s \geq 1$ (Figure 5).

(a) (b)

Figure 5. Original image (**a**) and the combined *e_aug* energy map using the NLP segmentation output shown in Figure 4 (**b**).

3.4. Introducing Compositional Bias

Object location data can be used to maintain a degree of control over the relative positioning of objects in the output image by removing seams in quantities proportional to the space between objects.

This can be accomplished with dynamic programming by restricting the seam search areas and restricting the number of seams that the algorithm searches for in each area:

Removing k seams from image I of dimensions $w \times h$, where:

I has n detected objects, and each object O_n has a designated center point O_{np}.

The set of seams to be removed can be sampled to maintain the relative positioning of each object in the scene with $\{S\} = \bigcup_{i=1}^{n+1} S_i$, where

$$\{S_1\} = \min_s \sum_{i=1}^{h} e\left(I_{0,h;O_{1P},h} : (s_i)\right), \text{ a set of the } \frac{O_{1P}}{w} k \text{ lowestenergyseams}$$

$$\{S_n\} = \min_s \sum_{i=1}^{h} e\left(I_{O_{nP},h;O_{n+1P},h} : (s_i)\right), \text{ a set of the } \frac{O_{n+1P} - O_{nP}}{w} k \text{ lowestenergyseams}$$

$$\{S_{n+1}\} = \min_s \sum_{i=1}^{h} e\left(I_{O_{nP},h;w,h} : (s_i)\right), \text{ a set of the } \frac{w - O_{nP}}{w} k \text{ lowestenergyseams,}$$

where each iteration of sampling for seams removes s_x from I and shifts all pixels to the right of s_x one pixel to the left, reducing the image width by one pixel.

This ensures that each submatrix of image matrix I finds a number of optimal seams proportional to the size of that submatrix relative to the size of I, with each submatrix being defined by the positioning of each detected object.

This method is flexible in that it can be easily adjusted to account for different compositional considerations. One strong use case is weighing different subsets of S differently to change the position of objects relative to one another, such as by removing all the space from between two given objects (Figure 6). The same method can be used to remove objects from an image using natural language prompts, by assigning sets of seams to be removed from sections of the image where objects are detected.

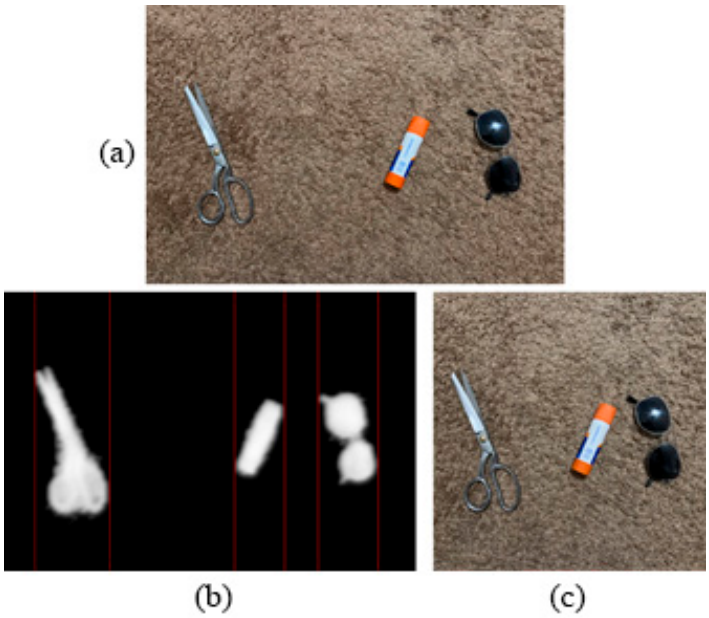

Figure 6. Retargeting an image with clearly defined object placement demonstrates this method well. Original image (**a**), the segmentation output overlaid with red lines denoting the edges of each object's predicted bounding box (**b**), and the image retargeted using the above algorithm to maintain the relative positioning of each object (**c**). While only the horizontal bounds of objects were calculated here because only vertical seams were removed, the full bounding box would be required for both vertical and horizontal retargeting.

3.5. Refined Object Manipulation Control

Another potential advancement in this retargeting workflow involves utilizing more advanced and accurate segmentation models to separate detected objects from the background. By applying this approach, each component can be retargeted separately, and the resulting layers can be composited back together. To address any discrepancies between the retargeted layers, a generative model can be employed to fill in the remaining spaces.

This approach allows for more precise retargeting by individually handling the objects and the background, thus avoiding potential distortions caused by retargeting the entire image as a whole. This segmentation-based workflow enables more localized adjustments and better preserves the integrity of the objects.

The combination of this segmentation-based retargeting approach with the NLP segmentation workflow adds another layer of automation and convenience. As outlined above, utilizing NLP models to facilitate the segmentation process provides a more streamlined and user-friendly workflow. NLP segmentation can automatically identify objects and provide initial segmentation masks, reducing the manual effort required from the user.

Running this workflow with the recently developed Segment Anything Model (SAM) yields promising results. The SAM provides accurate and robust segmentation across any image by using a zero-shot generalization framework that does not require a labeled training dataset [15]. Because SAM segmentation is not labeled, we can generate initial lower-accuracy masks with CLIPSeg and use those to isolate the high-accuracy maps from the SAM output that overlap with the desired prompted regions. This method is similar to the proof-of-concept text-to-mask method proposed in the initial SAM paper, which

describes a method of zero-shot text-to-mask segmentation that also operates by extracting the CLIP embeddings from an image. Although this method is not built into the public release of the model at this time, the hybrid CLIPSeg to SAM segmentation approach seems to be effective when used with existing text-to-image models such as DALL-E 2 [16] (Figure 7).

Figure 7. Original image (**a**), retargeted image using naïve seam carving (**b**), our proposed hybrid energy method (**c**), and retargeting the subjects separately from the rest of the image using the proposed layered method (**d**). The layered method uses SAM and CLIPSeg to segment the layers, and DALL-E is used to inpaint the missing areas of overlap. The layered method results in significantly less distortion to the background, and better preserves the primary subjects in the image.

Considering other compositional features and implementing a hierarchy of object dominance can further enhance the retargeting approach. Prioritizing different objects based on their position within an object dominance hierarchy allows for more refined control over the preservation of objects during the seam carving process.

One way to incorporate object dominance hierarchy is by assigning weights to objects based on their size and relative brightness [17]. Larger and brighter objects tend to be more visually dominant and attract greater attention. By calculating weights based on these factors, the retargeting algorithm can adjust the energy values assigned to seams passing through different objects.

The weights can be used to influence the energy computation process, making seams passing through visually dominant objects less likely to be removed compared to seams passing through less dominant objects. This hierarchical approach ensures that the more visually significant objects are preserved more effectively during retargeting.

In situations where there are too many seams to completely avoid objects, incorporating weights based on object size and relative brightness provides a mechanism to balance the preservation of visually dominant objects with the overall retargeting constraints. This approach enables the retargeting algorithm to intelligently allocate resources to preserve the most important objects while still achieving the desired resizing.

3.6. Optimizations

One of the biggest limitations to seam carving in practical applications is that the process of calculating the energy map is a fairly computationally expensive process, as it requires iterating through the entire image and applying the energy kernel at each pixel. With large images, the processing time can increase very quickly.

One approach we found that significantly reduces processing time is downscaling the image before processing. Downscaling the image decreases its resolution, thereby reducing the number of pixels that need to be processed during energy map calculation.

By operating on a smaller image, the computational load is significantly reduced. After obtaining the energy map at the downscaled resolution, it can be upscaled back to the original size, matching the dimensions of the original image.

By downscaling the image to as low as 5% of its original size before calculating the energy, and upscaling the resulting energy map, we were able to reduce the seam carving time by over 90% for 4k images with no impact to the resulting output image. Visual artifacts began to appear when downscaling below 1/24th of the original image size, and the 90% reduction in processing time after implementing downscaling was consistently present in smaller images as well.

4. Results

The proposed method for retargeting images demonstrates a notable reduction in visual distortion compared to the naïve seam carving approach. This improvement is particularly evident in images featuring subjects with low contrast or busy backgrounds, where visual distortion tends to be more pronounced using conventional seam carving techniques (Figures 8 and 9).

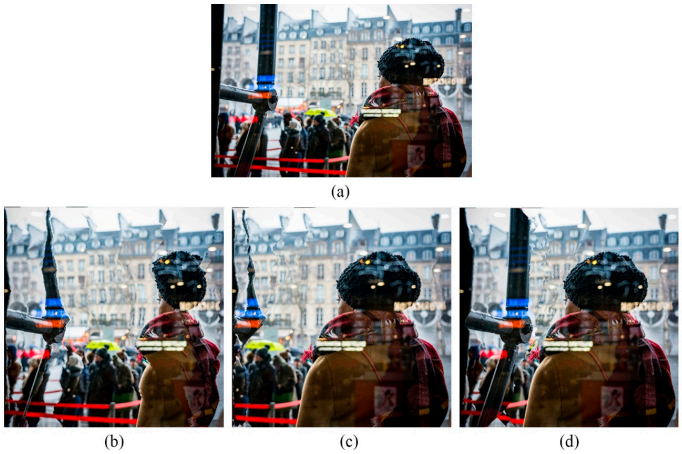

Figure 8. Original image (**a**); results of naïve seam carving (**b**); results of our method using the automatic (**c**) and manual (**d**) confidence maps as seen earlier in Figure 4.

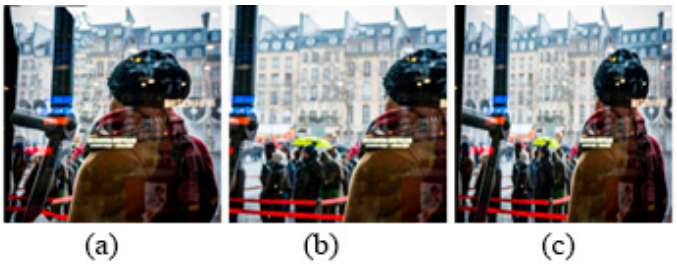

Figure 9. Our proposed method (**a**) compared with cropping (**b**) and scaling (**c**) to the same aspect ratio.

To further enhance the output quality, the inclusion of a prompt to manually specify the importance of subjects can yield even better results. By allowing users to manually indicate which subjects are more important, the retargeting process gains an additional level of customization and control (Figure 10).

Figure 10. Original image (**a**) with four variations demonstrating how different prompts can yield vastly different retargeting results. Prompts: "People and flowers" (**b**), "People and doorway" (**c**), "People and doorway, but the doorway is more important" (**d**), "Only small flowers in flowerpots" (**e**).

By considering the user-specified subject importance, the retargeting algorithm can prioritize the preservation of those subjects—or removal of unwanted subjects—during the seam removal process. This ensures that the resulting output images maintain the integrity of the important subjects while adhering to the resizing constraints.

5. Discussion and Limitations

While this approach does address many of the limitations expressed in the original seam carving proposal, it still has a number of limitations inherent in the leveraged technologies.

One significant limitation is that while the method aims to avoid distortion in primary subjects, it may introduce further distortion to the background. This trade-off is often preferable since maintaining the integrity of the subjects is a priority. However, it is important to note that this approach will not eliminate distortion in all scenarios, and could possibly be mitigated by other AI methods which separate out the foreground objects and auto-fill the background layer [18] (pp. 35–37).

In images where the subject occupies a significant portion of the width or height, there can still be unwanted distortion affecting the subject itself. This limitation aligns with the challenges identified when the naïve seam carving algorithm was initially proposed. Retargeting algorithms face inherent difficulties when resizing images that contain subjects dominating a large portion of the frame. These limitations arise due to the inherent constraints and trade-offs involved in the retargeting process. Seam carving redistributes the energy in an image to accommodate resizing, and in doing so, some level of distortion

or alteration is inevitable. Balancing the preservation of subjects with the need for overall image resizing can be a complex task, especially when subjects occupy a substantial portion of the frame.

Our approach is also subject to limitations inherent in the detection and segmentation models utilized. The accuracy and performance of these models depend on the current state of computer vision technology and may have certain constraints. While advancements in computer vision are improving the performance of these models, there is still the potential for the misclassification or incomplete identification of subjects. The effectiveness of the automated detection and segmentation method is constrained by the scope and diversity of the training data used for training the detection models. Although the Open Images v4 dataset used in this research is quite robust with 600 object classes, it does not encompass the full range of objects found in real-world images. This limitation can result in missed or misclassified subjects, impacting the accuracy of the retargeting process. Methods such as CLIPSeg which rely on the CLIP embeddings of the image for segmentation can be more robust in this regard, but are still imperfect.

Additionally, in order to be practical for real-time or even accessible consumer-level applications, the segmentation and seam carving pipeline would have to be further optimized to reduce computational expense. The computational complexity of the pipeline should be minimized to ensure efficient and timely processing, particularly for scenarios where real-time or interactive retargeting is required. Additional work would also have to be performed to reliably convert and process LLM outputs to input which can be parsed to image processing commands in a robust manner.

While the above limitations exist in our implementation, the fields of computer vision, machine learning, and computational photography are progressing at an incredibly rapid pace. Improvements to models will invariably impact the specific details of their most effective implementations in this proposed retargeting approach, but the fundamental methodology remains reliable.

6. Conclusions

By enhancing the art directability of the retargeting process, our proposed hybrid computer-vision seam-carving technique offers significant advantages over traditional methods. By integrating object detection and segmentation into seam carving, users can exercise greater control over which elements of an image are preserved or emphasized during resizing. This increased directorial control is particularly valuable in creative fields where visual composition and emphasis are critical.

As such, this technique finds its primary applications in artistically driven fields such as web design, advertising, and digital art. In these domains, creative professionals often require tools that allow them to tailor visuals to fit various design constraints without compromising key visual elements. Our method enables designers and artists to intelligently preserve or emphasize specific subjects within an image while resizing it, enhancing artistic expression, and ensuring that the visual integrity of important elements is maintained. This qualitatively enhances artistic expression and enables designers to tailor visuals to fit various design constraints without compromising the integrity of key visual elements. Practically speaking, similar natural-language-driven artist tools have already begun to be introduced in consumer products, such as the Generative Fill tool introduced in a 2023 beta release of the Photoshop desktop application [19]. A content-aware retargeting tool could be similarly implemented as a tool within existing software packages, such as Photoshop, where users would have the ability to use natural language prompts within the software to specify the importance of objects in the image.

Notably, this method may not be appropriate for applications that require completely undistorted image data. For instance, training datasets for image generation models, medical imaging analysis, or any context where pixel-level accuracy is crucial would not benefit from the distortions introduced by seam carving—even when mitigated by

advanced computer vision techniques. In such cases, maintaining the original pixel data is essential to ensure the accuracy and reliability of subsequent processing steps.

Other applications that can benefit from discarding auxiliary pixel data while maintaining the visual integrity of the image subject might find our methodology advantageous, though further research is needed. For example, deep learning-based face analysis technologies [20] could benefit from our retargeting method because it retains essential facial features while reducing extraneous background, potentially enhancing focus on relevant data and improving computational efficiency. However, it remains to be determined whether our method offers a substantial advantage over simpler techniques like cropping, especially when considering the additional computational complexity involved. The trade-off between computational expense and the quality of subject preservation is a critical factor that needs careful evaluation in such use cases.

The technologies and processes used in this research may also have applications in human-centered tasks. Using image-based identification and segmentation together with semantic control could be a powerful technique for complex robotic tasks (such as "place all mugs but not cups on the table") or self-driving cars ("avoid people and cats, but prioritize people") [21]. Semantic information can also improve techniques for re-identification of individuals in images or video where traditional models may struggle with issues such as occlusion [22] or movement in real-life settings [23]. Additional semantic information and compositional image data could potentially improve recognition reliability in those situations.

Other similar techniques for human-centered tasks can potentially be integrated into the retargeting process to improve results over the basic semantic confidence maps proposed here. The integration of spatial and temporal salience maps, as discussed in pedestrian trajectory forecasting, could enhance image retargeting by emphasizing dynamic and contextually significant regions over static ones. Such an approach would allow for more intelligent seam carving in videos or image sequences, preserving important regions based on predicted movement patterns or object intent, similarly to how trajectory prediction benefits from estimating pedestrians' destination intentions [24]. Additionally, processes used for person re-identification can be used in the semantic retargeting pipeline to estimate data about specific individuals or poses [25]. This could allow for more nuanced retargeting prompts (e.g., "Prioritize Sam and Nancy if they are walking").

This use of natural language directorial control via AI classification integrated into the image processing pipeline provides a new method of intelligent image editing where the subject(s) of the images can be easily defined. This method can be extended beyond seam-carving as a technique for directorial control over singular images or as bulk-processing on a series of images.

Author Contributions: Conceptualization, E.D.; methodology, E.D.; software, E.D.; validation, E.D.; formal analysis, E.D.; investigation, E.D.; resources, E.D.; data curation, E.D.; writing—original draft preparation, E.D.; writing—review and editing, E.D. and P.D.; visualization, E.D.; supervision, P.D.; project administration, P.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Dataset and additional result samples are available on request from the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fan, X.; Zhang, Z.; Sun, L.; Xiao, B.; Durrani, T.S. A comprehensive review of image retargeting. *Neurocomputing* **2024**, *579*, 127416. [CrossRef]
2. Pitas, I. Digital Image Processing Algorithms and Applications. In *Digital Image Processing Algorithms and Applications*; John and Wiley and Sons: Hoboken, NJ, USA, 2000.
3. Setlur, V.; Takagi, S.; Raskar, R.; Gleicher, M.; Gooch, B. Automatic image retargeting. In Proceedings of the 4th International Conference on Mobile and Ubiquitous Multimedia MUM '05, Christchurch, New Zealand, 8–10 December 2005; pp. 59–68. [CrossRef]

4. Avidan, S.; Shamir, A. Seam Carving for Content-Aware Image Resizing. *ACM Trans. Graph.* **2007**, *26*, 9. [CrossRef]
5. Koch, C.; Ullman, S. Shifts in selective visual attention: Towards the underlying neural circuitry. *Hum. Neurobiol.* **1985**, *4*, 219–227. [PubMed]
6. Shen, J.; Wang, D.; Li, X. Depth-Aware Image Seam Carving. *IEEE Trans. Cybern.* **2013**, *43*, 1453–1461. [CrossRef] [PubMed]
7. Hutchison, D.; Kanade, T.; Kittler, J.; Kleinberg, J.M.; Mattern, F.; Mitchell, J.C.; Naor, M.; Nierstrasz, O.; Pandu Rangan, C.; Steffen, B.; et al. Scene Carving: Scene Consistent Image Retargeting. In *Computer Vision—ECCV 2010*; Daniilidis, K., Maragos, P., Paragios, N., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6311, pp. 143–156. [CrossRef]
8. Rubinstein, M.; Shamir, A.; Avidan, S. Improved seam carving for video retargeting. *ACM Trans. Graph.* **2008**, *27*, 1–9. [CrossRef]
9. Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fischer, I.; Wojna, Z.; Song, Y.; Guadarrama, S.; et al. Speed/accuracy trade-offs for modern convolutional object detectors. *arXiv* **2016**. [CrossRef]
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv* **2016**. [CrossRef] [PubMed]
11. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv* **2016**. [CrossRef]
12. Lüddecke, T.; Ecker, A.S. Image segmentation using text and image prompts. *arXiv* **2022**. [CrossRef]
13. Liu, Y.; Han, T.; Ma, S.; Zhang, J.; Yang, Y.; Tian, J.; He, H.; Li, A.; He, M.; Liu, Z.; et al. Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models. *arXiv* **2023**. [CrossRef]
14. Hwang, D.-S.; Chien, S.-Y. Content-aware image resizing using perceptual seam carving with human attention model. In Proceedings of the 2008 IEEE International Conference on Multimedia and Expo, Hannover, Germany, 23 June 2008; pp. 1029–1032. [CrossRef]
15. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment Anything. *arXiv* **2023**. [CrossRef]
16. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv* **2022**. [CrossRef]
17. Obrador, P.; Schmidt-Hackenberg, L.; Oliver, N. The role of image composition in image aesthetics. In Proceedings of the 2010 IEEE International Conference on Image Processing, Hong Kong, China, 26–29 September 2010; pp. 3185–3188. [CrossRef]
18. Dickman, E. Smart Resizing: A Hybrid Deep-Learning Approach to Content-Aware and Selective Image Retargeting 2023 (Order No. 30529483). Dissertations & Theses @ Drexel University; ProQuest One Academic. (2861009977). Available online: http://ezproxy2.library.drexel.edu/login?url=https://www.proquest.com/dissertations-theses/smart-resizing-hybrid-deep-learning-approach/docview/2861009977/se-2 (accessed on 28 October 2024).
19. Adobe. Make Stunning Updates to Your Images with Text Prompts Using Generative Fill. Available online: https://web.archive.org/web/20230608020252/https://helpx.adobe.com/photoshop/using/generative-fill.html (accessed on 6 June 2023).
20. Ulrich, L.; Nonis, F.; Vezzetti, E.; Moos, S.; Caruso, G.; Shi, Y.; Marcolin, F. Can ADAS Distract Driver's Attention? An RGB-D Camera and Deep Learning-Based Analysis. *Appl. Sci.* **2021**, *11*, 11587. [CrossRef]
21. Chen, W.; Xu, X.; Jia, J.; Luo, H.; Wang, Y.; Wang, F.; Jin, R.; Sun, X. Beyond Appearance: A Semantic Controllable Self-Supervised Learning Framework for Human-Centric Visual Tasks. *arXiv* **2023**. [CrossRef]
22. Ning, E.; Wang, Y.; Wang, C.; Zhang, H.; Ning, X. Enhancement, integration, expansion: Activating representation of detailed features for occluded person re-identification. *Off. J. Int. Neural Netw. Soc.* **2024**, *169*, 532–541. [CrossRef] [PubMed]
23. Liu, Z.; Li, D.; Zhang, X.; Zhang, Z.; Zhang, P.; Shan, C.; Han, J. Pedestrian Attribute Recognition via Spatio-temporal Relationship Learning for Visual Surveillance. *ACM Trans. Multimedia Comput. Commun. Appl.* **2024**, *20*, 1–15. [CrossRef]
24. Wang, R.; Lam, S.-K.; Wu, M.; Hu, Z.; Wang, C.; Wang, J. Destination intention estimation-based convolutional encoder-decoder for pedestrian trajectory multimodality forecast. *Measurement* **2025**, *239*, 115470. [CrossRef]
25. Wang, C.; Ning, X.; Li, W.; Bai, X.; Gao, X. 3D Person Re-Identification Based on Global Semantic Guidance and Local Feature Aggregation. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 4698–4712. [CrossRef]