Semantic Pixel Distances for Image Editing

Josh Myers-Dean Scott Wehrwein

Western Washington University
Bellingham, WA
{myersdj,scott.wehrwein}@wwu.edu

Abstract

Many image editing techniques make processing decisions based on measures of similarity between pairs of pixels. Traditionally, pixel similarity is measured using a simple L2 distance on RGB or luminance values. In this work, we explore a richer notion of similarity based on feature embeddings learned by convolutional neural networks. We propose to measure pixel similarity by combining distance in a semantically-meaningful feature embedding with traditional color difference. Using semantic features from the penultimate layer of an off-the-shelf semantic segmentation model, we evaluate our distance measure in two image editing applications. A user study shows that incorporating semantic distances into content-aware resizing via seam carving [2] produces improved results. Off-the-shelf semantic features are found to have mixed effectiveness in content-based range masking, suggesting that training better general-purpose pixel embeddings presents a promising future direction for creating semantically-meaningful feature spaces that can be used in a variety of applications.

1. Introduction

Following on rapid advances in image recognition [21], the applicability of deep convolutional neural networks has broadened to encompass a diverse range of tasks in computer vision and image processing. One reason for this wide-ranging success is the versatility of the features extracted by image classification networks. CNN architectures originally developed for image classification have been shown to produce useful image-level feature embeddings with applications in image retrieval [22], fine-grained recognition [8], and visual product search [3]. Meanwhile, CNN architectures have also been rapidly generalized to make dense predictions, for example semantic segmentation and instance segmentation [28].

In this paper, we begin to investigate the question of whether semantic feature embeddings at the *per-pixel* level can be similarly generalized from semantic prediction tasks to become useful for other applications. Where image retrieval was a natural showcase for semantically-meaningful image level features, we investigate image editing tasks as a testbed for semantic pixel features.

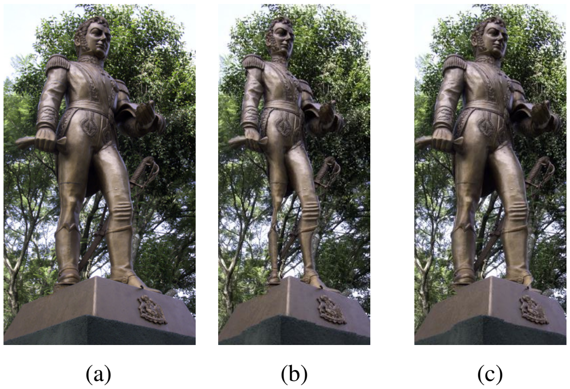


Figure 1: An input image (a) is resized to 80% of its original width by seam carving using the standard RGB energy map (b) and the proposed RGB/semantic combined energy map (c). The legs and overall shape of the statue are preserved due to the sharp difference between the statue and the background in semantic feature space.

Our key observation is that at the core of many image editing techniques is a measure of similarity between pixels. Many algorithms rely on image gradients, smoothness terms, nonlocal methods, and many other notions that boil down to measures of similarity or difference between pixels or patches. We aim to enrich these comparisons using semantic information from CNN models. Specifically, we use semantic pixel features from the second-to-last layer of an off-the-shelf semantic segmentation model to augment the notion of pixel similarity that lies at the core of several image editing applications. In this work, we try out semantically-augmented pixel distance metrics in two image editing applications: seam carving [2] and range masking. We find that off-the-shelf pixel features make noticeable improvements in seam carving results, while masking results are mixed. Figure 1 shows the effect of incorporating semantics into the seam carving application.

2. Related Work

Various prior works have begun to examine more fine-grained properties and applications of the image features learned by image recognition models. In particular, feature maps from multiple layers of image recognition networks capture meaningful notions of perceptual similarity [26] and image style [12]. These properties have been used to generate and compare images via network inversion [12] and via loss functions for feed-forward image generation models [15, 10]. Our work aims to leverage this same feature richness in image editing applications, but on an individual pixel level by using features from pixelwise prediction models (i.e., from the "decoder" of a semantic segmentation network), rather than using only features from the "encoder" recognition network.

Although per-pixel embeddings have been leveraged for traditional image and video segmentation as well as semantic segmentation [16, 13, 6, 18, 14], the utility of features beyond direct semantic applications remains less explored. In the image editing domain, Yang [24] used learned edge detectors to modulate the edge confidence in edge-preserving filtering, reasoning that edge confidences are learned from semantically-informed human judgments. Other specialized learning-based approaches have been used for editing specific image content such as faces and hair [19] or illumination estimation for object insertion [11]. Yan *et al.* [23] addressed exemplar-based image retouching with deep neural networks, but their approach does not leverage semantics from pre-trained convolutional models, opting instead to use a non-convolutional network to regress parameters of a task-specific color transformation model.

In contrast to these special-purpose techniques, we take a more generic approach: we use semantic feature embeddings alongside traditional color-based pixel distance metrics to capture a notion of pixel similarity that can be used in a wide range of applications. Recent progress has been rapid on image editing tasks where good benchmark datasets are available, such as super-resolution [17], denoising [25], and image retouching [5]. We focus our efforts on two other editing tasks that seem less amenable to learning-based approaches and have not seen as much progress: content-based image resizing using seam carving, and parametric range masking.

Prior work on content-based image resizing removes connected seams of pixels chosen using a gradient-based energy map [2]; Dong *et al.* [9] augmented this approach using image-level similarity measures. Rubinstein *et al.* [20] generalized seam carving to video and introduced "forward energy" to account for changes in the energy map upon removal of seams. The failure modes of these approaches tend to be where the low-level energy map does not fully capture the perceptual effects of removing a seam; this often occurs when semantic edges do not coincide with image edges. Our work aims to mitigate these failure cases by using a richer energy map that incorporates semantic information in addition to low-level color gradients.

Image editing programs such as Lightroom [1] and darktable [7] provide users with the ability to generate *range masks* (or *parametric* masks), including pixels in a mask based on color or luminance similarity to a selected pixel. Although we are unaware of any papers describing this simple technique, we extend the approach to use semantically-augmented similarity as a criterion for inclusion in a mask instead of only luminance or color.

3. Method

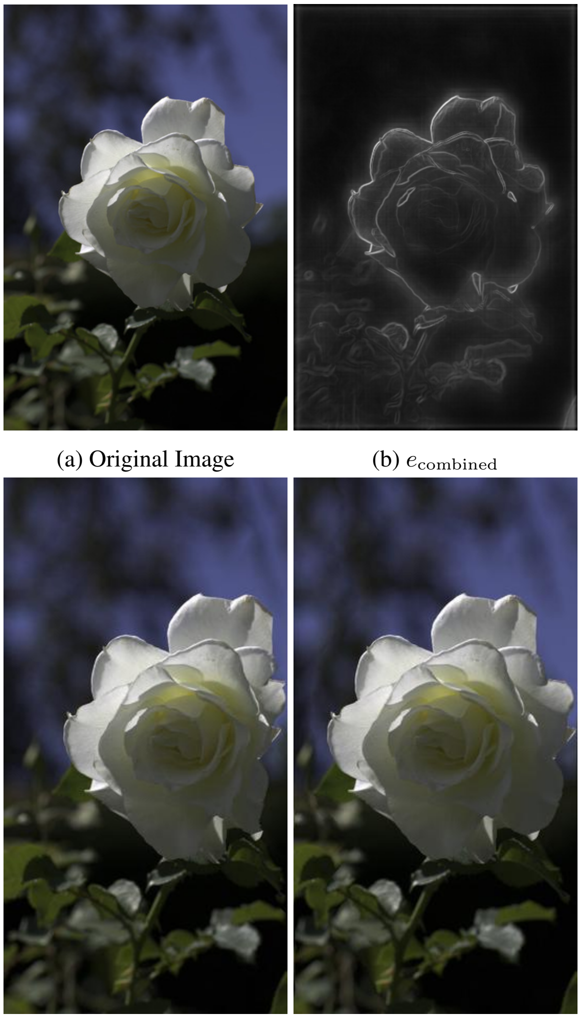
We propose to use pixel feature embedding vectors in addition to color values to compute distances between pixels in image editing applications. In this section, we describe how this general approach can be applied in two image editing applications: Seam Carving and parametric masking for image editing. The key idea in both applications is to augment a traditional pixel distance metric (i.e., RGB Euclidean distance or difference in the luminance channel) with a distance in semantic feature embedding space to incorporate a higher-level notion of similarity into low-level image editing tasks. Thus far we have focused only on using an off-the-shelf pretrained semantic segmentation model to extract per-pixel embedding feature vectors, leaving the training of purpose-built pixel embeddings for future work.

3.1. Semantic Seam Carving

The seam carving method proposed by Avidan *et al.* [2] removes "seams"—connected strings of pixels, one from each row or one from each column—in order to change the aspect ratio of the image without simply cropping or scaling the image. Their method aims to minimize disruption to the image content by choosing seams with minimal energy, according to an energy map calculated based on image gradients. Formally, the energy of a pixel is the sum of the horizontal and vertical image gradient norms

$$e_{\text{rgb}}(\mathbf{I}) = \|\frac{\partial}{\partial x}\mathbf{I}\| + \|\frac{\partial}{\partial y}\mathbf{I}\|, \tag{1}$$

where $\mathbf{I}$ is the image to be retargeted and $\|\cdot\|$ is a chosen norm (we used the L_1 norm in our experiments). A dynamic programming algorithm is used to find the lowest-energy seam for removal, where the energy of a seam is the total energy at all pixels in the seam.



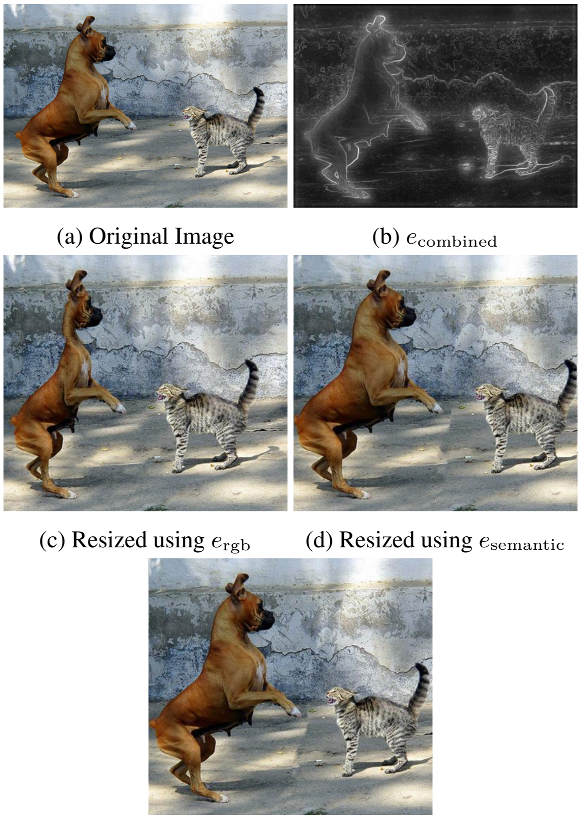
(a) Original Image

(b) e_{combined}

(c) Resized using e_{rgb}

(d) Resized using e_{combined}

Figure 2: An example input with its semantic label map, resized to 80% width using e_{rgb} and e_{combined}. Notice that in the e_{rgb} result, pixels are removed from the flower instead of from the foliage in the left side of the image. When semantics are considered, the foreground subject is better-preserved in the resized image.



(a) Original Image

(b) e_{combined}

(c) Resized using e_{rgb}

(d) Resized using e_{semantic}

(e) Resized using e_{combined}

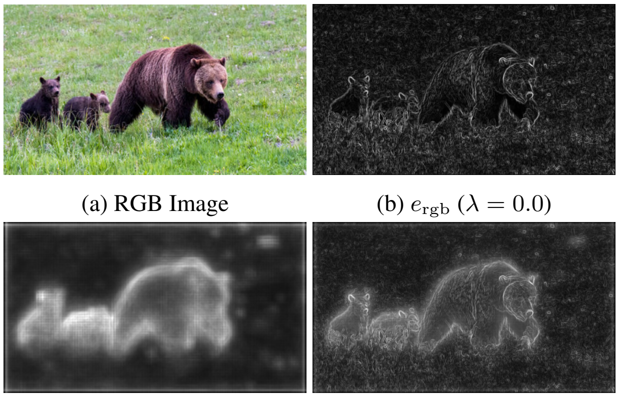
Figure 3: An example input with its semantic label map, resized to 80% width using e_{rgb}, e_{semantic}, and e_{combined}. The image resized with e_{rgb} is unaware of the semantic significance of the edges between background and foreground, so the dog's shapes is not preserved well. The image resized using e_{semantic} displays a more noticeable artifact in the continuity of the crack in the wall, because the color differences are less heavily weighted. The combined energy map achieves a balance between the two.

The intuition is that seams crossing strong image edges are most likely to be noticeable when removed, so seams with least energy should be removed first. We extend this intuition to semantics: seams that cross boundaries in semantic category are more likely to be noticeable as well. A naïve approach might use the semantic label map from a semantic segmentation network (e.g., [28]), assigning a high cost to seams that cross changes in semantic label. However, the label maps are often not pixel-accurate at object boundaries, resulting in artifacts when seam carving. This approach is also limited to semantic categories that are predicted by the network.

To incorporate semantics in a more flexible way, we propose to use semantic feature embedding vectors in place of RGB vectors. Given a semantic feature map $\mathbf{S}$, where each pixel is a d-channel semantic feature vector, the semantic energy map is

$$e_{\text{semantic}}(\mathbf{S}) = \|\frac{\partial}{\partial x}\mathbf{S}\| + \|\frac{\partial}{\partial y}\mathbf{S}\|. \tag{2}$$

Because the goal of seam carving is to minimize visual impact, we found that using only semantics gave poor results in some examples, especially where an object or background with uniform semantics exhibits distinctive structure or texture (e.g., Figure 3). We found that the best results



(a) RGB Image (b) e_{rgb} ($\lambda = 0.0$)

(c) e_{semantic} ($\lambda = 1.0$) (d) e_{combined} ($\lambda = 0.2$)

Figure 4: An input image and three example energy maps.

were achieved using a weighted combination of the two energy maps:

$$e_{\text{combined}}(\mathbf{S}, \mathbf{I}) = \lambda e_{\text{semantic}}(\mathbf{S}) + (1 - \lambda) e_{\text{rgb}}(\mathbf{I}) \quad (3)$$

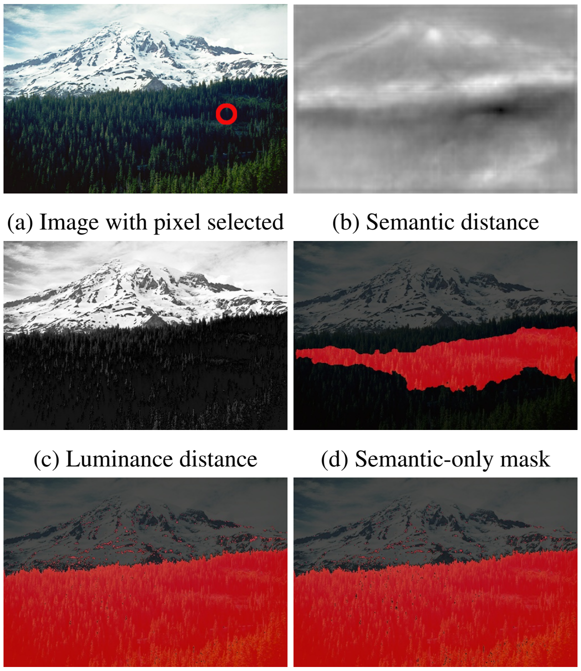
Figure 4 shows an example input image and three energy maps produced with $\lambda = 0.0, 1.0$, and 0.2. All results for e_{combined} in this paper are produced using $\lambda = 0.2$, which we found to maintain sharp details present in RGB while considering important semantic boundaries that may be less prominent in RGB.

With our combined energy map, we proceed with the seam carving algorithm as presented in [2]. We also make use of the "forward energy" approach proposed by Rubinstein *et al*. [20] to take into account changes in the energy after removal of a seam. We use forward energy in all our experiments.

3.2. Semantic Range Masking

Another use case for semantic pixel feature embeddings is an image editing technique known by various names including *Luminosity masking*, *range masking* (in Lightroom [1]) and *parametric masking* (in Darktable [7]). The basic idea of all of these techniques is to create a mask or selection containing all pixels within a user-specified distance of a given pixel or value. For example, a luminosity (or luminance) mask might allow the user to choose a pixel and create a selection of pixels whose luminance is less than 0.25 different; given such a mask, editing operations such as exposure adjustment or contrast enhancement can be selectively applied to different regions of the image. For example, Darktable allows a user to create masked adjustments based on luminance or any color channel in various color spaces [7].

As in seam carving, the most obvious approach to using semantics for masking is to directly use the label map to segment the image, but inaccuracy around edges makes semantic segmentation outputs ill-suited for photo editing purposes. Instead, our approach is to again use semantic feature embeddings to enrich the notion of pixel similarity when computing range masks to make selections that are more semantically coherent. Figure 5(a–c) shows an input image with a pixel selected and the distance of each other pixel to the selected pixel in luminance (b) and semantic feature space (c).



(a) Image with pixel selected (b) Semantic distance

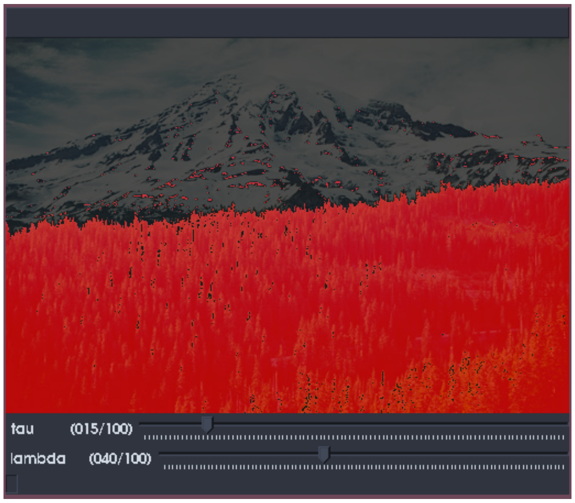
(c) Luminance distance (d) Semantic-only mask

(e) Luminance-only mask (f) Combined range mask

Figure 5: An overview of our range masking approach. All pixels are compared to the selected pixel (a) in semantic feature space (b) and luminance space (c). Range masks result from thresholded distances in semantics (d), luminance (e), and a weighted combination of the two (f).

As in the seam carving application, we found that the most effective approach was a distance metric based on a weighted combination of the traditional distance (in our experiments, we used luminance) and semantic feature embedding distance (illustrated in Fig. 5(f)). Because different images have different characteristics, we found it most useful to provide the user with control over the trade-off between luminance and semantic distance.

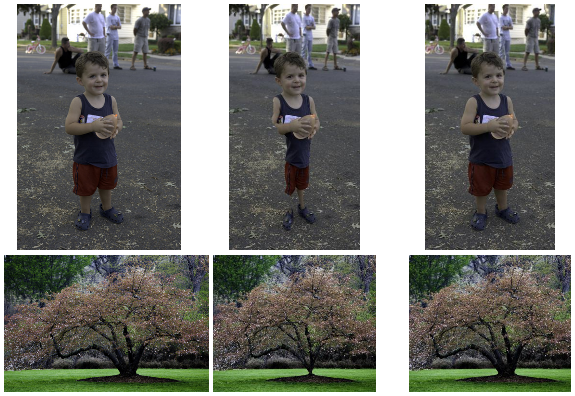
The user provides an input image $\mathbf{I}$ and selects a pixel index x to base the mask on. Sliders allow the user to choose the weight λ between luminance and semantics, and a threshold τ on distance. We start by computing a luminance map $\mathbf{L}$ and a semantic feature map $\mathbf{S}$. Then, the pixel at index p is included in the mask if the following condition



tau (015/100)

lambda (040/100)

Figure 6: Our prototype user interface for range masking. The sliders control the τ and λ parameters in Equation 4.



Input image RGB energy Combined energy

Figure 7: Example seam carving results. The combined energy map does a better job of preserving semantically significant objects in the scene.

is satisfied:

$$\lambda \|\mathbf{S}(p) - \mathbf{S}(x)\| + (1 - \lambda)\|\mathbf{L}(p) - \mathbf{L}(x)\| < \tau \qquad (4)$$

We built a prototype user interface, shown in Figure 6, that allows the user to select a pixel by clicking a point, choose a weight between luminance distance and semantic distance, and choose a threshold on the distance to be included in the mask.

3.3. Implementation Details

We extract semantic embeddings from a publicly available pre-trained semantic segmentation model [28]. We use the `ResNet101dilated` encoder and a `PPM_DeepSup` decoder trained on the ADE20k dataset[27]. We find that models trained on ADE20K work better than those trained on Pascal VOC, likely because ADE20K has more classes and therefore a richer semantic representation of a wide variety of objects and scenes.

We extract features from the second-to-last layer, yielding a (height × width × 512) feature map. For efficiency, we used random projection [4] to reduce the dimensionality of the feature vectors from 512 to 128 in the case of the masking application. We found that for the interactive masking application where distances must be recomputed with each movement of the slider, efficiency gains make for a substantially better user experience, while the quality of the results was only barely affected.

4. Results and Discussion

Evaluating image editing tasks like retargeting and range masking is challenging since the notion of ground truth is often subjective. We found that incorporating semantic feature distances into seam carving resulted in results that were often better and rarely worse than the baseline RGB approach. The range masking results were less compelling relative to the baseline, performing better on some images and worse on others. We believe that lower-dimensional feature embeddings trained on a proxy task have the potential to improve results in a wider variety of applications.

4.1. Semantic Carving

Figures 1 and 7 show a few qualitative results from our seam carving experiments. In each case, the energy map using combined RGB and semantic feature distances results in an output that shows a less-distorted representation of the subject of the image. This results from the energy map placing a higher cost on semantically-meaningful edges in the image and discouraging seams that cross those edges, such as the edge between the ground and the child's arm in Fig. 7, or between the trees and the statue's leg in Fig. 1.

To evaluate our method quantitatively, we performed a small user study to determine whether humans prefer images resized using the proposed semantic seam carving approach or the traditional RGB energy map. We chose 20 images at random from the photography-oriented MIT-Adobe FiveK dataset[5] and ran both the proposed semantic carving method and the traditional seam carving method on each image to resize it to 80% of its original width.

Each of 10 participants was shown the labeled original image; the two resized images were displayed in a random order. Tasked with selecting the image that best represented the original image, participants were given the option to select a preferred output, select both if they were equally good representations, or select neither if both results were poor representations of the original. Overall, the results produced with semantic carving received 145 votes, while RGB seam

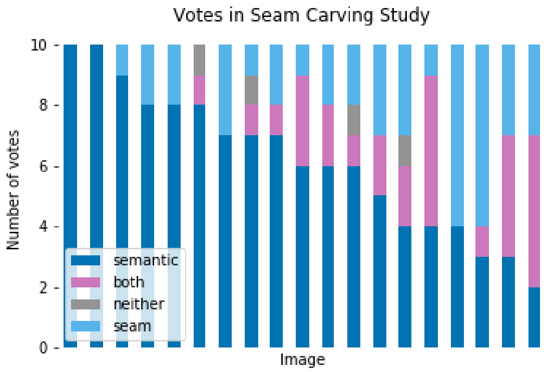


Figure 8: Results on each of the 20 images in our user study. Semantic indicates votes for the result using the combined semantic/RGB energy map; both indicates both images were considered acceptable; neither indicates that neither image was acceptable, and seam indicates that the RGB-only energy map was preferred.

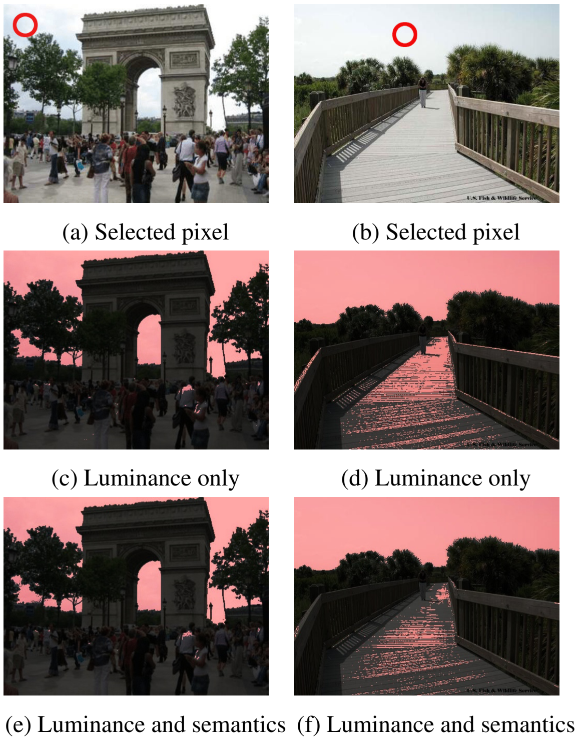
carving results received 69 votes. The vote breakdown for each individual image is shown in Figure 8.

4.2. Semantic Masking

We found that incorporating semantic features into range masking was helpful in some images, but detrimental in others. Figure 9 shows two example images with a pixel selected and example masks generated using luminance only and our combined distance. In the arch image (left column), the combined distance does a slightly better job of omitting bright pixels that are not part of the sky, though some pixels on a person's white shirt remain unmasked. The image of a boardwalk (right column) is similar: more of the non-sky pixels on the walkway are included with the sky when semantic features are included in the distance.

Landscape images with stark semantic contrasts tended to show more improvement than than images with clear subjects, even when the subjects were distinct semantic objects against a background. Even in images like the mountain from Fig. 5, the semantic features did not seem to capture the important and (from a human perspective) obvious semantic differences. This is likely due to a combination of several factors. The off-the-shelf semantic feature space used is trained for classification accuracy on a fixed set of categories; the distances in the resulting embedding space may be less meaningful than features trained with an embedding loss.

Another problem is the curse of dimensionality: in the off-the-shelf 512-dimensional feature space we used, linear classifiers perform well but pairwise distances are less meaningful. Training purpose-built lower-dimensional embeddings on a proxy task with some form of embedding loss is a promising avenue for improving the general usefulness of distances in applications like this.



(a) Selected pixel (b) Selected pixel

(c) Luminance only (d) Luminance only

(e) Luminance and semantics (f) Luminance and semantics

Figure 9: Two example images and range masks produced for each using luminance range and a range in combined luminance and semantic space. In the arch image (left column), using both luminance and semantics selects the sky while minimizing the erroneously selected pixels in other parts if the image compared to the luminance-only version. In both cases, we tuned the parameters to produce as clean a result as possible.

5. Conclusion and Future Work

This work represents early steps in the exploration of semantic pixel feature vectors as a tool for injecting semantic information into traditional image editing algorithms that typically rely on only low-level cues. Our approach worked well in seam carving, while results were less compelling in range masking.

We believe that the clear next step is to train our own general-purpose embeddings instead of using off-the-shelf feature vectors. Such embeddings could be chosen to be lower-dimensional, avoiding the problems with distances in high-dimensional space. The features could also be trained using an embedding loss that might generalize better to unfamiliar semantic content. Because the applications of these features are difficult to apply to end-to-end learning approaches, a proxy training task, such as a triplet loss derived from ground-truth semantic categories could be used.

We believe that the general notion of semantic pixel feature vectors has promise in image editing applications that are typically blind to semantics, in much the same way that feature embeddings have already proven to be useful for representing image content and style. We have shown one application where semantic information significantly improves results, but further work is required to investigate the full generality of our approach.

References

[1] Adobe, Inc. Adobe Lightroom Classic. 2, 4

[2] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. *ACM Trans. Graph.*, 26(3), July 2007. 1, 2, 4

[3] Sean Bell and Kavita Bala. Learning visual similarity for product design with convolutional neural networks. *ACM Trans. on Graphics (SIGGRAPH)*, 34(4), 2015. 1

[4] Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: Applications to image and text data. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 245–250. Association for Computing Machinery, 2001. 5

[5] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 2, 5

[6] Yuhua Chen, Jordi Pont-Tuset, Alberto Montes, and Luc Van Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[7] Darktable Open Source Contributors. darktable. 2, 4

[8] Yin Cui, Feng Zhou, Yuanqing Lin, and Serge J. Belongie. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1153–1162. IEEE Computer Society, 2016. 1

[9] Weiming Dong, Ning Zhou, Jean-Claude Paul, and Xiaopeng Zhang. Optimized image resizing using seam carving and scaling. *ACM Trans. Graph.*, 28(5):1–10, Dec. 2009. 2

[10] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems 29*, pages 658–666. 2016. 2

[11] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 9(4), 2017. 2

[12] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[13] Mohsen Ghafoorian, Cedric Nugteren, Nora Baka, Olaf Booij, and Michael Hofmann. El-gan: Embedding loss driven generative adversarial networks for lane detection. In *The European Conference on Computer Vision (ECCV) Workshops*, September 2018. 2

[14] Jianbo Jiao, Yunchao Wei, Zequn Jie, Honghui Shi, Rynson W.H. Lau, and Thomas S. Huang. Geometry-aware distillation for indoor semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. 2

[16] Abhijit Kundu, Vibhav Vineet, and Vladlen Koltun. Feature space optimization for semantic video segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[17] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[18] Yun Liu, Peng-Tao Jiang, Vahan Petrosyan, Shi-Jie Li, Jiawang Bian, Le Zhang, and Ming-Ming Cheng. Del: Deep embedding learning for efficient image segmentation. In *IJCAI*, volume 864, page 870, 2018. 2

[19] S. Qin, S. Kim, and R. Manduchi. Automatic skin and hair masking using fully convolutional networks. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 103–108, July 2017. 2

[20] Michael Rubinstein, Ariel Shamir, and Shai Avidan. Improved seam carving for video retargeting. In *ACM SIGGRAPH 2008 Papers*, SIGGRAPH '08, New York, NY, USA, 2008. Association for Computing Machinery. 2, 4

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1

[22] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[23] Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. Automatic photo adjustment using deep neural networks. *ACM Trans. Graph.*, 35(2), Feb. 2016. 2

[24] Qingxiong Yang. Semantic filtering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[25] K. Zhang, W. Zuo, and L. Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, Sep. 2018. 2

[26] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 2

[27] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5

[28] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal on Computer Vision*, 2018. 1, 3, 5