TONE MAPPING OPERATORS: PROGRESSING TOWARDS SEMANTIC-AWARENESS

Abhishek Goswami[⋆†] *Mathis Petrovich*[⋆] *Wolf Hauser*[⋆] *Frederic Dufaux*[†]

[⋆] DxO Labs
[†] Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes

ABSTRACT

A *Tone Mapping Operator* (TMO) aims at reproducing the visual perception of a scene with a high dynamic range (HDR) on low dynamic range (LDR) media. TMOs have primarily aimed to preserve global perception by employing a model of human visual system (HVS), analysing perceptual attributes of each pixel and adjusting exposure at the pixel level. Preserving semantic perception, also an essential step for HDR rendering, has never been in explicit focus. We argue that explicitly introducing semantic information to create a 'content and semantic'-aware TMO has the potential to further improve existing approaches. In this paper, we therefore propose a new local tone mapping approach by introducing semantic information using off-the-shelf semantic segmentation tools into a novel tone mapping pipeline. More specifically, we adjust pixel values to a *semantic specific target* to reproduce the real-world semantic perception.

Index Terms— Tone mapping, High Dynamic Range Imaging, Semantic aware exposure adjustment

1. INTRODUCTION

The luminance ratio between the brightest and the darkest point of a scene is called the *dynamic range*. In real world, this ratio is much higher than the dynamic range of most display devices. To render a High Dynamic Range (HDR) image on a Low Dynamic Range display or a photo print, a Tone Mapping Operator (TMO) has to be applied which maps the tonal values (pixel values) while compressing the dynamic range with the aim of preserving perceptual cues of the scene. Classical TMOs [1, 2, 3] utilize models based on pixel values only or incorporate features of Human Visual System (HVS) and tone map based on the perceptual attributes of each pixel.

The HVS can adapt to scenes of different dynamic ranges and still perceive semantic information. Therefore, preserving semantic perception is essential and it seems intuitive that a TMO should be *aware* of the semantic content of the scene. This problem has not been addressed in previous TMOs. Previous approaches have tried to preserve lightness perception, but have never used any information explicitly to emulate *Semantic Awareness* as a guide towards tonal adjustment.

In this work, we propose a new tone mapping pipeline around a probabilistic *semantic framework*, which guides the TMO with *target lightness* values for pixel specific exposure adjustment. More specifically, the proposed pipeline includes the following steps. We first decompose an image into regions (frameworks of reference) based on semantic similarity instead of only luminance distribution, using off-the-shelf semantic classifiers. We then apply tools such as *matting* to create soft segments (mattes) from the output of the semantic classifier. Finally, we propose to learn semantic specific lightness values for target tone modification and use them to compute semantic specific gains for each region. The two contributions of the work lie in the new semantic TMO pipeline and the methodology to learn semantic specific target lightness.

2. RELATED WORK

Over the past decades, TMOs have been widely studied. Functionally, TMOs are of two types: *global* and *local*. Global TMOs apply the same luminance compensation throughout the image, whereas local TMOs take into account the spatial neighborhood of each pixel. We follow a recent comparative subjective study of several classical TMOs provided by Cerda-Company et al. [4] to understand the performance of different TMOs. Kim et al. [2] (rated highly over subjective experiments [4]) propose a global TMO based on the luminance adaptation of human visual cortex. They suggest that human visual sensitivity is adapted to the average log luminance of the scene and that it follows a Gaussian distribution.

As a local approach, Krawczyk et al. [1] propose a TMO based on a probabilistic model of lightness perception (rated highly over subjective experiments [4]). They decompose an HDR image into areas of consistent luminance (lightness framework) and map each framework by adjusting the perceived 'white' point based on Gilchrist's anchoring rule [5]. They follow Durand et. al's [3] approach of smoothing the base layer and penalise local variations in the probabilistic framework to preserve local contrast. The pixel precise framework allocation based on luminance helps overcome distortions like halos.

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 765911 (RealVision)

Fig. 1. Gamma-corrected HDR content *Petroglyphs* (left) [6] and the tone mapped image using *KrawczykTMO* [7] (right).

However, the idea that the HVS breaks a scene down on the basis of consistent luminance has shortcomings. We postulate that consistent luminance should not necessarily mean allocation to same framework, especially if the luminance is encountered in a different semantic context.

Fig. 1 presents the *Petroglyphs* image from the Fairchild HDR dataset [6]. In this image, the bright *rocks* on the bottom left and the *sky* on the top right have the same luminance. After tone mapping using Banterle's [7] implementation of *KrawczykTMO* [1], the rocks are no longer underexposed, but the sky seems to be overexposed. As parts of the rocks and the sky fall in the same lightness framework, both semantics get treated uniformly. We suppose that the HVS also breaks a scene down based on the semantic consistency rather than only consistent luminance. Hence we propose to create semantic frameworks instead and adjust them towards a target.

Krawczyk et. al [1] sets the white point as a local anchor inside each framework and adjusts exposure based on this anchor. Renowned photographer Ansel Adams claimed that the global average perceived luminance of mid-grey for photography is 18% of visible light [8], which has been the anchor for many TMOs. Our hypothesis is that every semantic content has a *target lightness* (perceived luminance) which should be reproduced by the TMO. More specifically, HDR images need to be tone mapped towards target lightnesses which are a function of the semantic label of a region in the scene and each region should be adjusted accordingly.

Recent data-driven approaches use neural networks for HDR tone mapping. Unlike classical methods, learning based TMOs do not follow an explicit model based on pixel values or human vision. They tend to learn the tone curve from a high dimensional feature representation based on training images. Rana et al. [9] follow such a data-driven approach and propose a deep learning based parameter-free TMO. Such approaches might implicitly incorporate semantic attributes but do not explicitly use semantic information.

In this paper, our objective is to explicitly define a semantic-specific target based on the content observed in the image and its luminance perception in the real world. We then guide the exposure adjustment towards that target lightness for each semantic framework. To create semantic frameworks, we use available off-the-shelf semantic segmentation methods. Since 2012, the rise of neural networks and deep learning has given us multiple models and various annotated datasets to learn and predict pixel wise semantic labels [10, 11, 12, 13]. Neural networks used for semantic segmentation include PSPNet [14], DeepLab [15, 16] and FastFCN [17]. In this work, we use these available tools as a black box to learn semantic-specific target lightness values and to segment images into semantic frameworks. Our contribution is how we use such semantic frameworks to guide the exposure adjustment, thus making the TMO semantic-aware.

3. PROPOSED TONE MAPPING SYSTEM

We propose a novel tone mapping system for decomposing an HDR scene into frameworks of reference and then adjusting the exposure of each framework based on its content. Fig. 2 presents an overview of our proposal. Each module is discussed in more details hereafter.

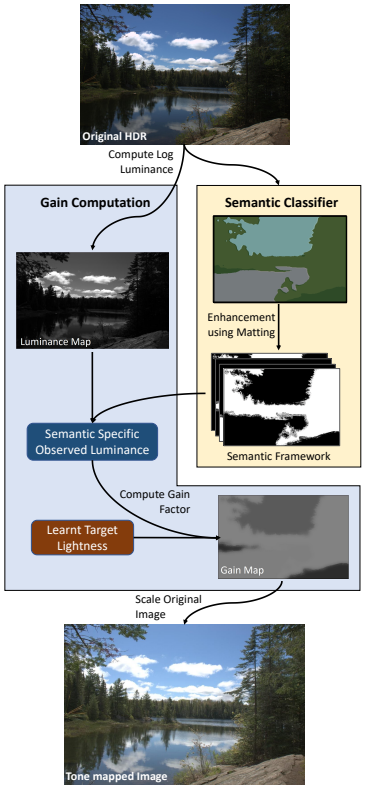

Fig. 2. Proposed tone mapping pipeline.

3.1. Semantic classifier

3.1.1. Semantic Segmentation

The first step is to create our *semantic frameworks*. We start with the original HDR image, resized by a factor of 4 on each dimension to reduce memory requirements and computational complexity. We classify the image into semantic labels,

Fig. 3. Trimaps (middle) and mattes (right) for *sky* (top) and *vegetation* (bottom) for the image *Petroglyphs* (Fig.1).

using available deep learning tools. We experimented with existing models such as PSPNet, Deeplabv3, Deeplabv3+ and FastFCN trained over annotated datasets (Pascal VOC, Cityscapes and ADE20K). We chose FastFCN architecture pretrained over ADE20K dataset due to the relatively better inference and wider range of labels (150 in total) [17].

Our hypothesis is that, the HVS analyses a scene by breaking it down into regions of *semantic similarity* and that the perceptual cues from these regions are treated similarly. It should be noted that the semantic distance between the 150 labels in ADE20K is not equal. *Dogs* and *cats*, for example, are different semantic labels, but when adjusting exposure, the HVS would probably consider them semantically similar.

Compiling an exhaustive set of semantically similar classes is a non-trivial problem and needs further research beyond the scope of this paper. Based on the contents of the images in our dataset (discussed in Sec. 3.2.1) we have heuristically defined 9 semantically similar classes, which require unique lightness adjustment. The 9 semantic classes are: *sky, mountain, vegetation, water, human subject, still-life subject, city, indoor, others*. Consequently, we grouped the 150 labels from ADE20K into these 9 predefined classes. For e.g. labels such as *sea, river, lake etc.* were mapped to the class *water*. The labels obtained for our image from FastFCN are merged according to our predefined classes to compute one binary map for each semantic class observed. The next step is to enhance the pixel precision of these binary maps.

3.1.2. Proposed enhancement using matting

Obtaining a pixel precise semantic map is a non-trivial problem. FastFCN results in imprecision in the high frequencies such as object boundaries. While making pixel specific adjustments, such imprecision can lead to distinctive distortions. Therefore, we require approaches such as soft segmentation to deal with imprecise semantic label allocation.

Hu et al. [18] proposed a two-step enhancement using the output of a semantic classifier and applying matting to it. Aksoy et al. [19] have considered the problem from a spectral segmentation viewpoint and used high and low level image features to obtain a soft segment. Learning based methods [20] also compute soft segments but they seem to associate certain semantic classes (such as *vegetation* and *sky*) systematically with the background and hence are unable to provide generalized semantic masks suitable for our purpose.

After experimenting with various matting techniques, we selected Alpha matting [21], due to the fewer artifacts leading to better quality while blending differently exposed segments. We first generate a morphologically expanded region of uncertainty called the *trimap* around the boundary of the semantic segment. Instead of binary labels, Alpha matting then provides fuzzy labels between $[0, 1]$ to each uncertain pixel denoting the probability of it belonging to that segment.

We notice that the map corresponding to semantic class *vegetation* is likely to produce segments with finer local contrast at their borders. Conversely, maps for other semantic classes such as *sky* have lower contrast borders. So a thicker trimap border for the former and a relatively thinner border for the latter yields better results (See Fig. 3). The Alpha matting returns one soft segmented matte for each observed class where each pixel has a fuzzy label i.e. a probability of belonging to that matte. The mattes may overlap at boundaries. Consequently, we normalize the label values at each pixel location over all mattes to get class-specific probability for that pixel. This collection of normalized mattes, called *semantic framework* shows, for every pixel $p_{x,y}$ in the image, its belongingness $P_{x,y,i}$ to each framework $\mathcal{F}_i$.

3.2. Gain Computation

3.2.1. Learning target lightness

The next objective is to use the semantic framework to adjust the observed luminance of the original image towards a target lightness. One of the main contributions of this work is a new methodology to learn semantic-specific perceived luminance. We propose to learn the *target lightness* (perceived luminance) for different semantic classes from a real-world dataset of *well-exposed* images. To this aim, we have created a dataset of 830 high resolution $[4000 \times 3000]$ LDR images from freely available sources [22].

More precisely, we compute a luminance histogram for each of our 9 classes over the entire dataset. Fig. 4 shows three such histograms, as well as the world histogram (which contains all pixels in our dataset). Ideally, semantics with different perceived luminance should have different histograms and hence should require different classification and exposure adjustment. This hypothesis holds for our 9 classes. Since the semantic segmentation is not pixel precise, we choose the median of the histogram to compute the class specific target lightness, for its robustness against outliers.

Table 1 shows the target lightness for our 9 classes. Intuitively, some semantics like *sky* are expected to have a brighter target, while *others*, a combined class for unrelated labels, has a target close to the global mid-grey of 18%. The class *human* has a surprisingly low target, which can be explained by the

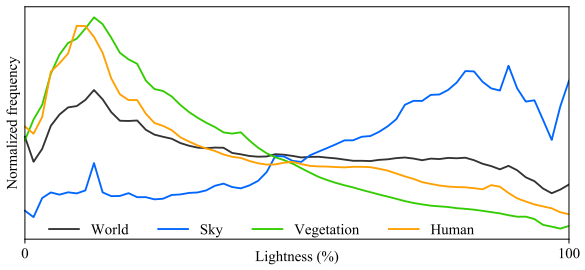

Fig. 4. Histogram of luminance values observed in our dataset, for different semantic classes. *Sky* and *vegetation* are significantly different from the world average.

Label	Target Lightness	
	sRGB (%)	Linear (%)
sky	72	48.5
mountain	36	10.5
vegetation	27	5.6
water	42	14.8
human subject	29	6.5
still-life subject	32	8.2
city	43	15.6
indoor	36.5	10.8
others	43	15.6

Table 1. Target lightness for different semantic classes.

fact that the semantic classifier does not differentiate between skin and non-skin part of humans.

3.2.2. *Computing gain factors*

Next, we compute the luminance map, $L_{x,y}$ for our test image using weighted average of the linear *RGB* values. Using the semantic framework and a probability threshold, we get a luminance distribution for each matte framework. We consider the luminance values of only those pixels which have a belongingness above a threshold T_b, empirically set to 0.8. The median of this luminance distribution gives us the observed luminance $L_{obs(i)}$ for each matte framework. The target lightness $L_{tar(i)}$ is learnt from our dataset, as detailed in Table 1. Using the two values, we compute a *class specific* gain factor, γ_i. Finally, we merge the gain factors weighted by the pixel belongingness in the semantic framework to compute a gain map of the same dimension as of the input image. Every index (x, y) in this gain map represents the *pixel specific* gain factor $\Gamma_{x,y}$ for tone mapping.

$$\gamma_i = \frac{L_{tar(i)}}{L_{obs(i)}} \tag{1}$$

$$\Gamma_{x,y} = \sum\nolimits_{\mathcal{F}_i} \gamma_i \cdot P_{x,y,i} \tag{2}$$

In order to preserve local contrast and details at transition boundaries, we use spatial information of the pixel neighborhood while computing the gain factors. More precisely, we apply a bilateral filter [23] to the mattes while creating the semantic framework to penalise for local variations. Finally, the gain map is used to scale the original image pixel by pixel to obtain our final tone mapped image.

4. RESULTS

In this section, we analyse the performance of our proposed tone mapping algorithm. Fig. 5 presents three HDR images from the Fairchild dataset and their respective tone mapped LDR images using *KrawczykTMO* [1], *KimKautzTMO* [2], and our *SemanticTMO*. We aim to analyze the results based on three factors: *Exposure compensation, Aesthetic presentation* and *Distortions*.

Exposure compensation: It is straightforward to notice the different gains achieved by *KrawczykTMO*, *KimKautzTMO* and *SemanticTMO*. *KrawczykTMO* and *KimKautzTMO* enhance shadows based on the luminance distribution only. The green bush and surrounding rocks in *Petroglyphs* are assigned positive gains but their relative distance in the original luminance histogram is not maintained after compression, leading to loss of relative contrast. Same goes for the shadows on the ground in *Jesse's Cabin*. *SemanticTMO* treats the shadows on the basis of luminance and the semantic map, thereby preserving relative contrast. Hence, images tone mapped with *SemanticTMO* are not as washed out or flat as the others.

Aesthetic representation: Aesthetic quality, though subjective, can be discussed on the basis of colour representation and preservation of photographic intent. We observe the representation of *sky* in all the images. *SemanticTMO* provides better colour and contrast representation than *KimKautzTMO* and *KrawczykTMO*. Manual photo-retouching tends to enhance primary subjects even if that requires suppressing background regions or shadows. Using *SemanticTMO*, regions of shadows are enhanced but moderately to preserve the global perceptual attributes of the primary subjects, such as the *cabin* and the *hall* dome. This is possible due to the inclusion of semantic information, as the TMO determines the gain as a function of the target lightness for each semantic framework.

Distortions: TMOs should compress the dynamic range without introducing distortions. *KrawczykTMO* and *KimKautzTMO* rate highly in this aspect. *SemanticTMO* introduces some distortions in its current implementation, due to shortcomings of the FastFCN semantic classifier. Precise semantic segmentation is an ill-posed problem. Results can be inconsistent and poor when handling translucent or complex structures. Pixel-precision is not guaranteed even with matting. The *Petroglyphs* image is an example of pixel-precise mask without halos but the *Cabin* image shows limitations due to the dense distribution of vegetation in sky region.

In order to supplement the above observations, we also score the tone mapped images using two recent *Image quality*

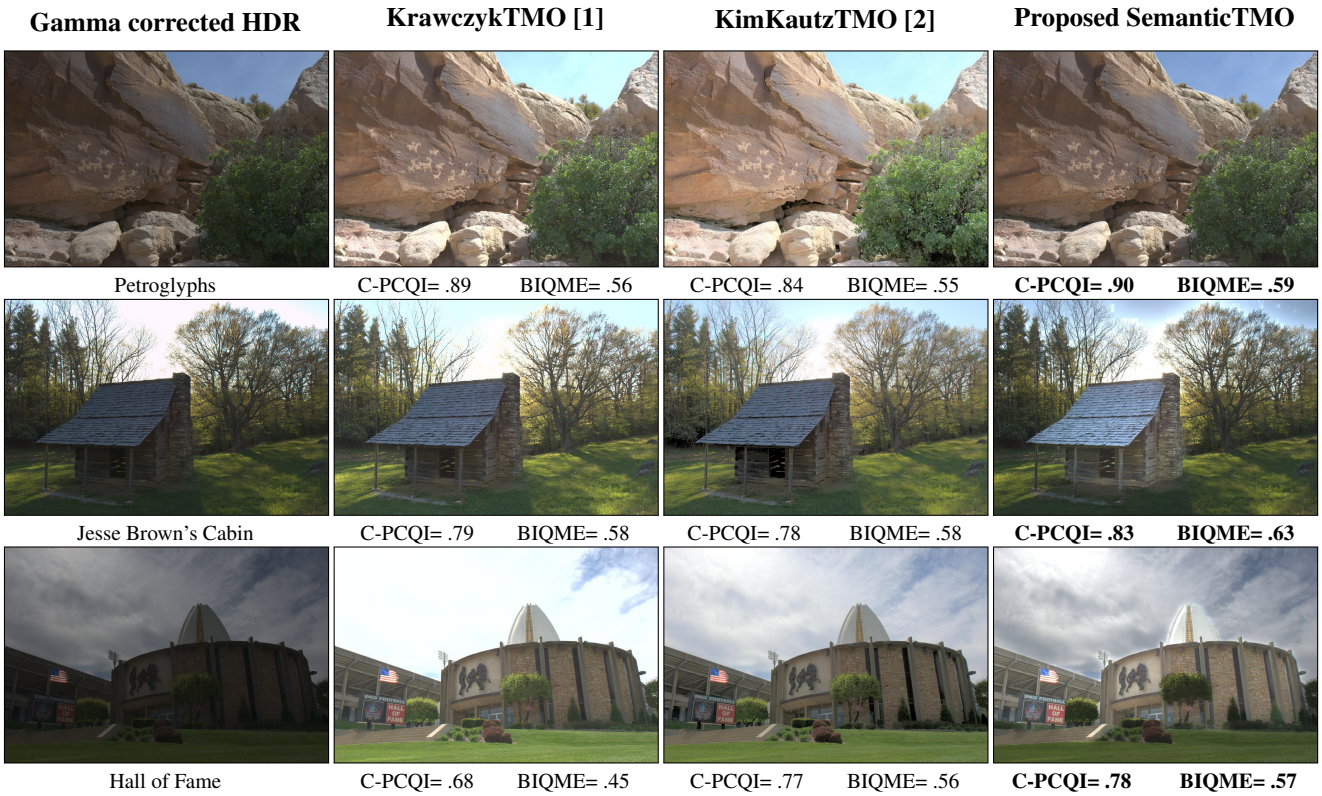

Fig. 5. Left: Gamma corrected HDR images from *Fairchild HDR Dataset* and respective tone mapped images using *KrawczykTMO* [1] & *KimKautzTMO* [2] (implemented using HDR toolkit by Banterle [7]) and the proposed *SemanticTMO*. *C-PCQI* and *BIQME* [24] scores have been computed for each tone mapped image and the best scores have been highlighted.

assessment (IQA) metrics. *C-PCQI* [24], a full reference metric, measures the quality for contrast enhanced images with importance on colourfulness. *BIQME* [24], a no reference metric, considers five influencing factors, *contrast, sharpness, brightness, colorfulness* and *naturalness of images*, contributing towards image quality and extracts a total of 17 features to assign a score to the tone mapped image. For both of these metrics, a higher score implies better image quality. We observe in Fig. 5, that our *SemanticTMO* consistently performs better than its two competitors on the test images.

5. CONCLUSION AND FUTURE WORK

In this paper, we discussed the limitations of existing TMOs and the benefits of explicitly including semantic information. We proposed a TMO, where we created a probabilistic *semantic framework* for each scene using matting and adjusted the exposure of each semantically similar region according to the semantic framework. Consequently, we created a database to learn the target lightness of different semantic classes.

The IQA scores from Fig. 5 show that our *SemanticTMO* performs better than *KimKautzTMO* and *KrawczykTMO* on the test images. This positively reinforces our hypothesis that different semantic classes require different target values and that introducing semantic awareness in the tone mapping process helps preserve the perception of the HDR image. Therefore, in this paper, we have shown that semantic awareness can improve upon existing TMOs.

Our work still has some limitations and can be further improved in several aspects. First, semantic segmentation is not as precise as luminance-based segmentation and can be refined further. Second, our dataset to learn target lightness is fairly small. A larger dataset can lead to a more robust scene classification to find semantically similar key classes for real world data. The median is not a statistically ideal choice to compute the target lightness and hence, the luminance distribution should rather be taken into account. Another important limitation is the non-availability of annotated training data for *photographically meaningful* semantic classes. The class *human subject*, for example, makes perfect sense for autonomous systems, but is almost meaningless for adjusting exposure. Skin and skin type would be more helpful.

This paper opens several avenues of future research. We aim to conduct subjective assessment of our results. We do not handle the scenario when a semantically uniform region receives non-homogeneous illumination. In future, we aim to create a hybrid approach based on both luminance and semantic similarity.

6. REFERENCES

[1] G. Krawczyk, K. Myszkowski, and H.P. Seidel, "Lightness perception in tone reproduction for high dynamic range images," in *The European Association for Computer Graphics 26th Annual Conference EUROGRAPHICS 2005*. vol. 24, Blackwell.

[2] M. H. Kim and J. Kautz, "Consistent tone reproduction," in *Proceedings of the Tenth IASTED International Conference on Computer Graphics and Imaging*. ACTA Press Anaheim, CA, USA, 2008, pp. 152–159.

[3] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," in *ACM transactions on graphics (TOG)*. ACM, 2002, vol. 21, pp. 257–266.

[4] X. Cerdá-Company, C. A. Párraga, and X. Otazu, "Which tone-mapping operator is the best? A comparative study of perceptual quality," *CoRR*, vol. abs/1601.04450, 2016.

[5] A. Gilchrist, C.s Kossyfidis, F. Bonato, T. Agostini, J. Cataliotti, X. Li, B. Spehar, V. Annan, and E. Economou, "An anchoring theory of lightness perception," *Psychological review*, vol. 106, pp. 795–834, Nov. 1999.

[6] M.D. Fairchild, "The hdr photographic survey," pp. 233–238, Jan. 2007.

[7] F. Banterle, A. Artusi, K. Debattista, and A. Chalmers, *Advanced High Dynamic Range Imaging (2nd Edition)*, AK Peters (CRC Press), Natick, MA, USA, July 2017.

[8] B. Brown, *Cinematography: Theory and Practice : Imagemaking for Cinematographers, Directors & Videographers*, Focal Press, 2002.

[9] A. Rana, P. Singh, G. Valenzise, F. Dufaux, N. Komodakis, and A. Smolic, "Deep Tone Mapping Operator for High Dynamic Range Images," *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 1285–1298, Dec. 2019.

[10] M. Everingham, S. M. Eslami, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[11] R. Mottaghi, X. Chen, X. Liu, N. Cho, S. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 891–898, IEEE Computer Society.

[12] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *CoRR*, vol. abs/1612.01105, 2016.

[15] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017.

[16] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[17] H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Yu, "Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation," 2019.

[18] G.G Hu and J.J. Clark, "Instance segmentation based semantic matting for compositing applications," 2019.

[19] Y. Aksoy, T.H. Oh, S. Paris, M. Pollefeys, and W. Matusik, "Semantic soft segmentation," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 72, 2018.

[20] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2970–2979.

[21] E. S. L. Gastal and M. M. Oliveira, "Shared sampling for real-time alpha matting," *Comput. Graph. Forum*, vol. 29, pp. 575–584, 2010.

[22] "unsample.net powered by unsplash!," https://unsplash.com/, http://unsample.net/, Accessed: 2019-09-15.

[23] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images.," in *Iccv*, 1998, vol. 98, p. 2.

[24] K. Gu, D. Tao, J.F. Qiao, and W. Lin, "Learning a no-reference quality assessment model of enhanced images with big data," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 4, pp. 1301–1313, 2017.