Image quality benchmark of computational bokeh

Wolf Hauser, Balthazar Neveu, Jean-Benoit Jourdain, Clément Viard, Frédéric Guichard
DxOMark Image Labs, 3 rue Nationale, 92100 Boulogne-Billancourt FRANCE

Abstract

Smartphone cameras have progressed a lot during recent years and even caught up with entry-level DSLR cameras in many standard situations. One domain where the difference remained obvious was portrait photography. Now smartphone manufacturers equip their flagship models with special modes where they computationally simulate shallow depth of field.

We propose a method to quantitatively evaluate the quality of such computational bokeh in a reproducible way, focusing on both the quality of the bokeh (depth of field, shape), as well as on artifacts brought by the challenge to accurately differentiate the face of a subject from the background, especially on complex transitions such as curly hairs. Depth of field simulation is a complex topic and standard metrics for out-of-focus blur do not currently exist. The proposed method is based on perceptual, systematic analysis of pictures shot in our lab.

We show that the depth of field of the best mobile devices is as shallow as that of DSLRs, but also reveal processing artifacts that are inexistent on DSLRs. Our primary goal is to help customers comparing smartphone cameras among each other and to DSLRs. We also hope that our method will guide smartphone makers in their developments and will thereby contribute to advancing mobile portrait photography.

Keywords: *Image quality evaluation, benchmark, portrait photography, depth of field, bokeh, smartphone.*

Introduction

Even some professional photographers and photojournalists use smartphone cameras for documentary, landscape and street photography. In these disciplines, the quality of recent high-end smartphones is—sometimes—indistinguishable from that of DSLRs. In portrait photography, however, even novices can immediately distinguish a smartphone picture and one taken with a DSLR. Figure 1 shows such a comparison.

There are two major differences between the smartphone and the DSLR used for taking the pictures in figure 1: perspective and depth of field. We will look at these phenomenons in more detail in the following subsection and recall how they are related to the difference in size between the smartphone camera and the DSLR. We will then have a short look at how smartphone manufacturers attempt to overcome these limits by narrowing the depth of field computationally.

Given user's interest in image quality and the strong competition among manufacturers, it seems obvious that smartphone image quality evaluation should benchmark these attempts. In the rest of this paper we propose a method and laboratory setup for evaluating the quality of shallow depth of field simulations. The proposed method is used as part of our DxOMark Mobile test protocol. We start by presenting related work and go on to describing our proposal and the results we obtain.

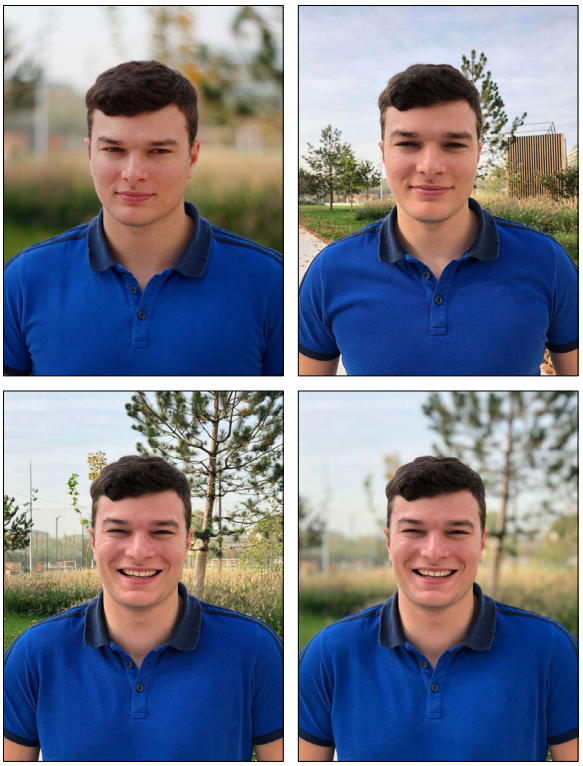


Figure 1. *Portraits taken with a consumer DSLR (a) and a smartphone (b–d). Top: large field of view and depth of field cause the background to appear very detailed in the smartphone picture, drawing attention away from the subject. Bottom: the telephoto lens (c) improves perspective, but the background remains sharp. Computational bokeh (d) allows the smartphone to deliver result rather close to that of the DSLR.*

Perspective, depth of field and why size matters

Perspective is how three-dimensional objects appear in two-dimensional images and it depends on the viewing point. In figure 1, the photographer framed the subject so that its face covers ⅓ of the image width in portrait orientation. For this, he had to stand at approximately 6 ft from the subject for the DSLR and at only 2 ft for the smartphone.

The relationship between subject distance and focal length is shown in figure 2. From intersecting lines, it follows $s = h\frac{s'}{h'}$ for the subject distance, where s' is the image distance and h and h' are the object and image heights, respectively.

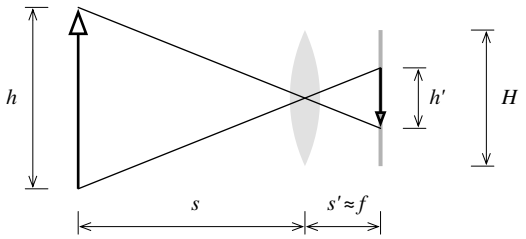


Figure 2. *Focal length f and sensor height H define the subject distance s required for a certain framing, i.e. for a certain ratio between image height h' and sensor height H.*

For portrait distances, the image distance is close to the focal length f and we can suppose $s' \approx f$. In our example, we wanted h' to be ⅓ of the sensor height H. It follows that $s \approx 3h\frac{f}{H}$. The object height is fix and the same for both cameras. The ratio between focal length and sensor height, however, is different: 50 mm : 15 mm for the DSLR, only 4 mm : 3.6 mm for the smartphone. Photographers are used to converting all focal lengths to their equivalent in traditional 35 mm format ($H = 24$ mm)[1]. This yields 80 mm for the DSLR and 27 mm for the smartphone—which corresponds precisely to the factor of three between the two subject distances.

There are several reasons why smartphones use moderate wide-angle lenses. Since they do not provide zoom, they have to provide a versatile prime lens that fits the majority of situations. Moderate wide-angle is also a sweet spot for lens designers and allows the best trade-off between sensor size and smartphone thickness. Longer equivalent focal lengths either require f to increase (which makes the phone thicker) or H to decrease (which makes the camera capture less light).

For portrait photography, however, moderate wide-angle is not ideal: it causes the face to be a bit distorted ("big nose effect") and scales down the surroundings compared to the subject, often leading to agitated backgrounds that draw attention away from the subject.

Depth of field is the distance between the closest and farthest objects that appear sharp in the image. Supposing that the subject's eyes are in focus, we obtain the configuration shown in figure 3.

Light rays coming from points in the subject plane converge right in the sensor plane. Light rays coming from points in front of or behind the subject converge behind or in front of the sensor, respectively, and form blur spots on the image—also called circles of confusion c. For a point to appear sharp in the image, the size of its blur spot must not exceed C, the *maximum acceptable circle of confusion* [1]. According to figure 3, these points lie between s_F and s_N. From similar triangles, it follows that

$$\frac{s'_N - s'}{s'_N} = \frac{s' - s'_F}{s'_F} = \frac{C}{D}$$

[1]Not only the sensor size changes between DSLRs and smartphones, but also the aspect ratio, which is 3:2 for most DSLRs and 4:3 for most smartphones. Most textbooks use the sensor diagonal for calculating the equivalent focal length. For the sake of symmetry, we decided to crop the DSLR image in our example in figure 1 so that it becomes 4:3. As a consequence, we only consider the sensor height H (i.e. the shorter sensor dimension) for both framing and computing the equivalent focal lengths.

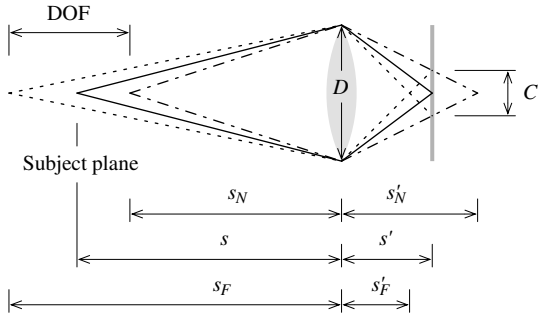


Figure 3. *The maximum acceptable circle of confusion C defines the nearest and farthest points in the scene that appear sharp on the image.*

where D is the aperture diameter. Utilizing the thin lens equation $1/s + 1/s' = 1/f$ and solving for s_N and s_F, respectively, we obtain

$$s_N = \frac{sDf}{f(D-C)+sC} \quad \text{and} \quad s_F = \frac{sDf}{f(D+C)-sC}. \tag{1}$$

The circle of confusion being smaller than the aperture by several orders of magnitude, $D - C \approx D + C \approx D$ and we obtain for the depth of field

$$DOF = s_F - s_N = \frac{2CDfs^2}{D^2f^2 - C^2s^2}.$$

For portrait photography, where the subject is much closer than infinity, we can make a second approximation and suppose that C^2s^2 is negligible against D^2f^2. This finally yields

$$DOF^p \approx \frac{2Cs^2}{Df} = \frac{2Cs^2N}{f^2}$$

where $N = f/D$ is the f-number of the lens.

We still have to determine C, which, depending on the camera and the viewing conditions, can be limited either by diffraction, sensor resolution, display resolution or the resolution of the human eye.

The influence of diffraction can be observed by computing the radius of the Airy disk, which is approximately $r = 1.22\,\lambda N$ [1]. Putting $\lambda = 550$ nm (green light) and $N = f/1.8$, we obtain $r = 1.2\,\mu$m. This is on par with the pixel size of the smartphone and far below that of the DSLR. A typical smartphone image sensor has a resolution of 12 Mpx and can therefore resolve details down to 1/5000 of its diagonal. Given that the resolution of the human eye is approximately 1 arcmin [2], this corresponds to looking at a $8 \times 12''$ print from a distance of only 10 inches. We assume that the majority of people do not look at their images so closely. As a consequence, in practice, C is neither limited by diffraction nor by sensor resolution, but by the viewing conditions. According to [3], the camera industry uses $C = 1/1500$ of the sensor diagonal for computing the depth of field scales imprinted on lenses. This corresponds to looking at a $4 \times 6''$ print from a distance of approximately 16 inches. Whatever the exact value, the important point is that diffraction and sensor resolution can be neglected. And supposing that the viewing conditions are the same for both smartphone and DSLR, C is always the same fraction of the respective sensor height.

The depth of field indicates whether or not a point in the scene appears sharp in the image. But it tells little about "how blurry" the background is, as this not only depends on the depth of field, but also on the distance between the subject and the background b. And this distance typically is fix and—unlike the subject distance—not adjusted in function of the equivalent focal length. A point at distance b behind the subject will appear as blur spot in the image. The ratio ρ between the diameter of its circle of confusion c_b and the image height H is a measure for the blurriness of the background.

According to figure 3 and by solving equation (1) for the circle of confusion, we obtain

$$\rho = \frac{c_b}{H} = \frac{f^2 b}{NH\,(s-f)(s+b)} \approx \frac{f^2 b}{NHs\,(s+b)}$$

because $f \ll s$ for portrait photography. For the particular case where $b \to \infty$, this simplifies to

$$\rho^{\infty} = \frac{f^2}{NHs}.$$

Back to the comparison between DSLR and smartphone. Suppose that, compared to the former, the latter has an image sensor that is α times smaller and an equivalent focal length that is β times shorter. Then,

$$C_{\text{phone}} = \frac{C_{\text{DSLR}}}{\alpha}$$
$$s_{\text{phone}} = \frac{s_{\text{DSLR}}}{\beta}$$
$$f_{\text{phone}} = \frac{f_{\text{DSLR}}}{\alpha\beta}$$

and finally

$$DOF^{p}_{\text{phone}} = \alpha\,\frac{N_{\text{phone}}}{N_{\text{DSLR}}}\,DOF^{p}_{\text{DSLR}} \tag{2}$$

$$\rho^{\infty}_{\text{phone}} = \frac{1}{\alpha\beta}\,\frac{N_{\text{DSLR}}}{N_{\text{phone}}}\,\rho^{\infty}_{\text{DSLR}}. \tag{3}$$

Note that β has disappeared from the depth of field comparison—equivalent focal length has no impact on the depth of field, provided we adapt the subject distance s to obtain the same framing. It does, however, have an impact on the blur spot at infinity: the shorter the equivalent focal length, the smaller the blur spots.

The sensor size impacts both and is the main reason for the difference between the two systems. In figure 1 (a) and (b), both cameras use the same f-number $f/1.8$. But because the smartphone sensor is only 3.6 mm high, compared to 15 mm for the DSLR ($\alpha \approx 4$), the smartphone has four times the depth of field of the DSLR. Combined with the difference in equivalent focal length, its background blur spots are, compared to the image height, approximately twelve times smaller than those of the DSLR. This is why the background appears so sharp, drawing even more attention away from the subject.

Just as the equivalent focal length $\tilde{f} = 24\,\text{mm}/H \times f$ allows to compare the field of view independently of the sensor size, one can define the equivalent aperture $\tilde{N} = 24\,\text{mm}/H \times N$ to compare the depth of field independently of the sensor size. The smartphone used to shoot figure 1 (b) has $\tilde{f} = 27$ mm and $\tilde{N} = f/12$. Seeing these figures, any photographer familiar with the 35 mm format understands immediately that this lens is meant for landscape or street photography and not for portrait.

Bokeh originally refers to the aesthetic quality of blur in the out-of-focus parts of an image, especially out-of-focus points of light. Unlike depth of field, it has no precise definition. Aperture blades, optical vignetting, spherical and chromatic aberrations all have influence on the character of out-of-focus points of light [4]. In recent years, the term has become synonym for out-of-focus blur and commonly includes the size of out-of-focus points of light. Larger ρ means stronger bokeh.

Overcoming physics by image processing

Some of the latest smartphones feature a second backside camera with a longer equivalent focal length. Figure 1(c) shows the same scene as in figure 1(b), shot with the telephoto camera of the smartphone. As in figure 1 (a) and (b), the subject distance s was chosen so that the face fills ⅓ of the image width. While this is a significant improvement in terms of perspective, it can make the depth of field issue even worse. To squeeze a longer equivalent focal length into the same small form factor, smartphone manufacturers have to use very small sensors and apertures ($\beta = 1.5$; $\alpha = 5$; $N = f/2.8$ for the smartphone used in figure 1 (c)). As a consequence, according to equation (2), the depth of field of the telephoto lens is twice that of the wide-angle lens! And according to equation (3), despite the longer equivalent focal length of the telephoto lens, a background at infinity is only slightly blurrier

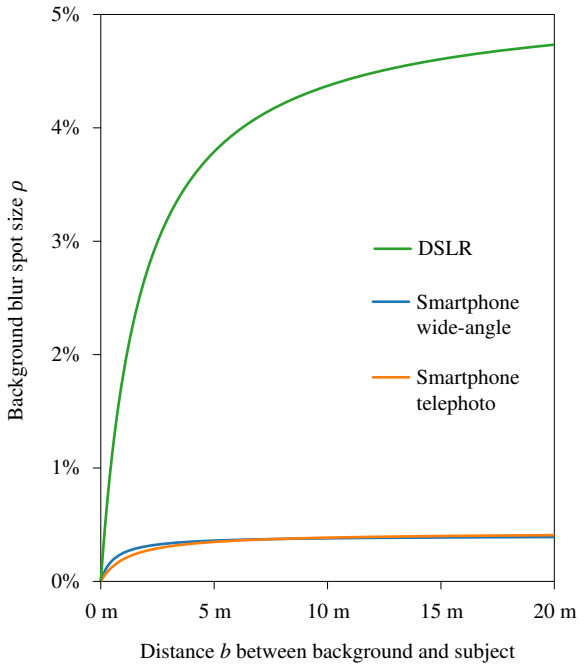


Figure 4. *Points on the background appear as blur spots of size ρ on the image. The DSLR obtains much stronger bokeh than both of the smartphone cameras. In particular, due to its even smaller sensor and its larger f-number, the telephoto camera performs worse than the wide-angle camera for backgrounds up to 7 m (23 ft) behind the subject.*

than for the wide-angle lens. Figure 4 shows how the blurriness ρ increases with b. Only for far-away backgrounds, the telephoto lens does slightly better than the wide-angle lens—and both mobile lenses are far behind the DSLR.

This is where image processing comes to help. The idea is simple: blur out-of-focus areas computationally, so that they no longer draw attention away from the subject. Figure 1(d) shows a result of such processing—and the first impression is rather convincing. There are however several challenges to this approach. Instead of a binary segmentation between subject and background—which may work on simple cases like figure 1—a depth map is required to handle the general case where the scene consists of more than two planes. The correctness and precision of this depth map are crucial for obtaining convincing, artifact-free computational bokeh. Many approaches exist for obtaining such a depth map, involving more or less sophisticated hardware and algorithms: through focus, structure from motion, dual camera (stereo vision), dual pixels and dedicated depth sensors. They all have their strengths and weaknesses and every company claims that their approach works best.

Even when the depth map is known, bokeh simulation is more challenging than just applying gaussian blur and making its radius a parameter of depth. Strictly speaking, it is impossible to simulate true optical bokeh using one single image from a single viewpoint, because even points *behind* the subject contribute to the image, as illustrated in figure 5. The smartphone has no information about these points. Fortunately, these points are not strictly necessary for obtaining a convincing illusion of bokeh—the observer usually will not guess that they should be there, either. The example in figure 5 might even look more natural when it showed only the pencil and a blurry dark background. It is, however, very important to keep a sharp transition between the sharp subject and the blurry background. Therefore, the blur computation in out-of-focus areas must not use pixels from in-focus areas.

Moreover, as can be observed in figure 5, in optical bokeh, light spots are not totally blurry—instead they appear approximately in the shape of the lens aperture [4], typically of circular shape. While this is not ideal to melt away unpleasant backgrounds [5], people are used to this particular look and associate it with professional photographs and movies. If the goal is to recall this look, out-of-focus points of light should appear as disks rather than being melted away in gaussian blur.

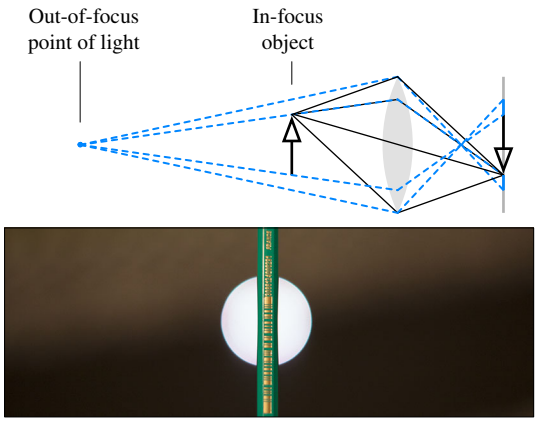


Figure 5. *Optical bokeh: points can contribute to bokeh, even if they are occluded by a sharp object/subject. The above example shows a pencil taken in front of a single point of light. Without optical bokeh, such a point of light is invisible (occluded by the pencil) and the smartphone has no chance to produce the same image as the DSLR.*

Finally, in optical bokeh, specular highlights in out-of-focus areas hit the sensor as large blur spots distributed over many pixels. In the smartphone case, however, they are concentrated on fairly few pixels, which tend to saturate. The true intensity of the highlight is therefore lost.

Although the shallow depth of field modes help to narrow the gap between smartphones and DSLRs, currently they are often plagued by processing artifacts, causing unnatural results.

Related work

Smartphone users and manufacturers care a lot about image quality. Given the number of portraits people take with their smartphones, it seems obvious that both users and manufacturers are in need of a method for evaluating and comparing the shallow depth of field modes.

Indeed, most reviews include sections on these new modes [6, 7, 8]. Some of them are particularly in-depth and reveal many of the issues that we observed ourselves when playing around with the phones. The shortcoming is that these tests are hardly reproducible and their results are hardly comparable. While [6] is very detailed and points out the capacities and limits of the shallow depth of field modes of the compared smartphones, it provides limited information in that it compares only two devices. On the other hand, [8] provides reviews of all smartphones, but the portrait scenes vary so much that a fair comparison between the devices is at least difficult.

Our ambition is to propose a method that allows comparisons between all evaluated smartphones—among each other and to DSLRs. In particular, when a new device is evaluated, the new results should be comparable to all past results without re-testing the other devices.

The traditional approach to measuring image quality in a reproducible and comparable manner is utilizing test charts. The test chart consists of a well-defined pattern, typically very different from real world scenes, designed to evaluate one specific criteria, such as resolution [10, 11] or noise and color [9]. Test charts were convenient back in the days when cameras were “dumb”, but as cameras get smarter and become aware of scene content, charts are less suitable. The highly non-linear image processing tends to behave differently on a test chart and in the real world.

For instance, measuring noise in homogeneous areas and sharpness (as indicator for detail) on edges does not provide reliable information about real-world image quality any more. This is why, already since the launch of our first DxOMark Mobile test protocol in 2012, we complement these standard measurements with image quality assessment performed on a lab scene, shown in figure 6 (a). The scene unites several challenging conditions (homogeneous areas, gradients, textures, color details, portraits, etc.) and is—at least—as complex as the real world. The advantage compared to real-world snapshots is that both the scene and the lighting conditions in our lab are perfectly repeatable. When a new smartphone comes out, all we have to do is shooting our lab scene with than single device. This allows instant comparison

Figure 6. *Lab scenes uniting challenging elements for detail and texture rendering. (a) ours, (b) [12]'s.*

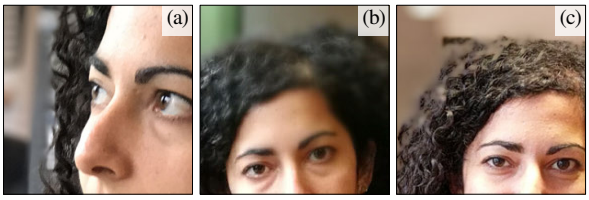


Figure 7. *Issues in subject/background segmentation: (a) approximate edge; (b–c) blurred subject or sharp background.*

with all smartphones we have ever evaluated. The scene is flat, i.e. it has no perspective, which is convenient for comparing devices with different equivalent focal lengths. The same method is applied by [12] in their studio scene, shown in figure 6 (b). They shoot it with all camera bodies, at all ISO speeds, always the same framing. Their web interface allows users to quickly and comfortably compare details of their studio scene in different camera body and ISO speed configurations.

The inconvenient of such complex scenes is that their quality is complex to assess. For instance, when image noise is measured on a homogeneous area, one can easily compute its standard deviation or even its frequency power spectrum. But how to quantify noise on the picture of a human face? We found that the most reliable and robust "tool" for evaluating such complex concepts as grain on human skin or texture on human hair is the human visual system. The challenge with humans, however, is repeatability. We observed that when they are presented with two images from different devices and asked which they prefer, the results are not very repeatable. This is because individuals do not look at the same parts of the image and because they do not balance different criteria (color, sharpness, grain, etc.) in the same way. But when we break down the complexity and ask well-defined questions, perceptual evaluation yields highly repeatable results. For instance, to assess noise, we present a particular detail in our lab scene and ask people to evaluate the noise (and only the noise) compared to a stack of reference images. The repeatability can be further improved when this test is repeated for several details in the scene.

This "lab scene method" combines the repeatability of traditional test charts with the robustness of human perception against all kinds of image processing tricks. It better correlates with the real-world quality experience than evaluations purely based on test charts.

Proposed method

The method and laboratory setup we propose is basically an application of the "lab scene method" to the problem of computational bokeh. The similarity between our own lab scene and [12] suggests that there exists some consensus about what objects are pertinent for evaluating noise and detail. To the best of our knowledge, we are the first to propose a lab scene for evaluating computational bokeh. Unlike a scene for noise and detail, a bokeh scene must obviously be three-dimensional.

Evaluation criteria

The criteria we want to evaluate are based on observations by reviewers and on our own field testing.

Subject/background segmentation. Precise segmentation between subject and background is key to obtaining DSLR-like results. Current implementations suffer from depth maps at too low resolutions, leading to approximate edges (figure 7 (a)), and depth estimation errors, leading to sharp zones on the background or blurred zones on the subject (figure 7 (b) and (c)). These artifacts are today the most important difference between smartphones and DSLRs—avoiding them should be the highest priority for the smartphone industry.

One particular case where subject/background segmentation can fail is scene motion during capture. In particular, structure-from-motion based approaches are more vulnerable to scene motion than dual camera approaches. Since we consider subjects in motion a common real world use case, we want to evaluate the impact.

Equivalent aperture. In the paragraph on depth of field we showed that an equivalent aperture in the 35 mm format could be computed by taking the sensor size into account. We propose to extend this concept to the domain of computational bokeh. In this case, the equivalent aperture no longer depends on the physical characteristics of the camera, but on the image processing applied. We estimate it's value by comparing the blur of certain image details to that of a full-frame DSLR at different apertures.

Blur gradient. When depth changes continuously, blur intensity should also change continuously. Basic portrait scenes may consist of only two planes more or less parallel to the sensor plane, but most scenes contain a more complex three-dimensional composition. Smooth blur gradients are therefore important for DSLR-like bokeh.

Noise consistency. When looking at the images in more detail, we observe an interesting side-effect of the bokeh simulation: the computationally blurred areas are totally free of grain. This comes at little surprise—blurring is a well-known denoising technique. Nevertheless it leads to an unnatural appearance. In a DSLR, the blur is created optically before the light rays even hit the sensor. The noise is therefore strictly the same in both in-focus and out-of-focus areas and does not guide attention to either. But a noise-free background draws attention away from a noisy subject and in this sense counteracts the intention behind bokeh simulation, which is to draw attention to the subject.

Character of the bokeh. This is the criteria that photographers first think about when discussing bokeh. Unfortunately, there is no general agreement among them on how perfect bokeh should be like. But computational bokeh seems to be simpler in this respect than optical bokeh. From what we observe, none of the smartphones simulates optical vignetting (causing the bokeh shape to vary in the field) or non-circular irises (causing bokeh of non-circular shape). Neither did we ob-

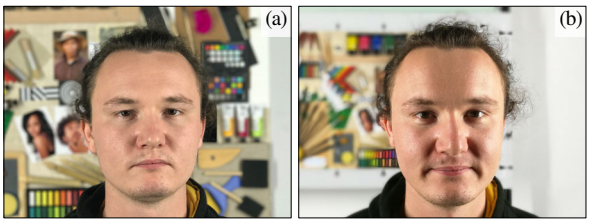


Figure 8. *Attention to perspective: the hair above the model's left ear is blurred away in (a) and well preserved in (b). But that comparison is unfair because the background in (a) is complex and that in (b) is simply a homogeneous white surface.*

serve purple or green fringes (caused by chromatic aberrations), "soap-bubbles" (caused by over-corrected spherical aberrations) or "donuts" (caused by catadioptric lenses). While certain photographers rely on these effects to obtain a particular look, the smartphone manufacturers seem to suppose that they do not appeal to their average user. What we observe are circular shapes, the sharpness of which varies between smartphones. We want to report this sharpness in our evaluation. And it seems that it can be observed at any point of the image, because unlike for optical bokeh, the character of computational bokeh does not vary in the field.

Repeatability. Unlike optical bokeh, which is independent of exposure, computational bokeh tends to work better when there is more light. More light increases the signal-to-noise ratio in the images and facilitates the computation of the depth map. But even when the lighting does not change, we observe that many devices have rather unstable behavior. Some artifacts appear on one image and disappear on another, without any obvious reason. Sometimes the shallow depth of field mode fails altogether and the image is captured without any blur applied. In this respect, computational bokeh is much more challenging to evaluate than noise and detail, which are perfectly repeatable from one image to the other. In any case it will be necessary to base the evaluation on several images.

Evaluating subject/background segmentation

Laboratory setup. Segmenting the image into subject and background can be more or less complicated, depending on the textures of both the subject and the background. For devices with different equivalent focal lengths, the relation between subject and background of a given scene typically changes, thereby changing the difficulty of the exercise, as illustrated in figure 8.

To achieve a fair comparison between all devices, we must ensure that the difficulty does not depend on the focal length. We could have achieved this utilizing a fractal pattern as background—for instance a dead leaves pattern [13]. But while their power spectrum corresponds to that of natural images, spatially, they are quite different from most real-world backgrounds. Their self-similarity makes depth estimation more challenging than it is in most real-world scenarios, which is a bias that we prefer to avoid.

Our proposed setup, shown in figure 9 (a) and (b), consists of two planes—a subject and a background—the distance between which can be adjusted. First we measure the equivalent focal length at portrait distance of the smartphone under test. Then we adjust the distances between device, subject and background accordingly (figure 10). Computer controlled lighting ensures uniform illumination on both planes whatever the focal length. As subject we use a mannequin head surrounded by a complex shape to simulate a person's face, headdress and a waving hand. As background we use a large format print of our noise and detail lab scene, to which we have added several white straight lines. Figure 10 shows photos taken of this scene with two smartphones having different equivalent focal lengths. Note how the scene appears almost in the same way in both cases.

Measure. The numerous features on both subject plane and background plane allow us to quantitatively determine the precision and reliability of the subject/background segmentation.

As explained in the section *Related work*, we employ perceptual evaluation since it is very robust against processing artifacts. To obtain repeatable results, we have defined a detailed protocol describing exactly what features should be observed and how they should be judged. By observing various kinds of features, we attempt to obtain an unbiased measure of the real-world capacity of the device to distinguish between subject and background.

Examples:

- For each of the white straight lines on the background, we give one point if it is melted away on its entire length.
- For each of the holes in the "hand" on the top right, we give one point if its content is as blurry as the rest of the background.
- On each of the spikes surrounding the face are printed small numbers from 1 to 5. The device gets one point for each number that is readable and an extra point when the number is as sharp as the rest of the subject.

Having features of different scale allows to measure the resolution and precision of the depth map. In the case of the spikes, most devices get at least one point for each of the numbers 1 to 3. But only very good devices manage to get the 4s as sharp as the subject—and currently no smartphone masters the 5s. On the other hand, any DSLR we tested obtained the maximum score. Examples are shown in figure 13.

Evaluating blur quality

Laboratory setup. While a scene consisting of two planes is perfect for comparing subject/background segmentation between cameras with different focal lengths, it is not suitable for determining the depth of field and only to a limited extent for evaluating other characteristics of the bokeh. This is why our proposal includes a second scene, shown in figure 9 (c). To simplify the installation of the two scenes in our lab, they are actually both included in one setup: you can see in figure 9 (a) that the second scene is situated in the lower part of the background canvas—a part that is hidden behind the subject in the first scene.

This scene consists of a subject (mannequin head) or macro object (plastic flower) in the foreground and two planes (on the right and at the bottom) that are almost parallel to the optical axis. This time, these planes are covered with regular patterns, and they also extend in front of the subject. In the back, far from the focus plane, we place some tiny LEDs serving as points of light. This scene allows us to evaluate all the remaining criteria expressed above.

Measure. For determining the *equivalent aperture*, we have shot the same scene using a full-frame DSLR, at various focal lengths, for various apertures each. Given the equivalent fo-

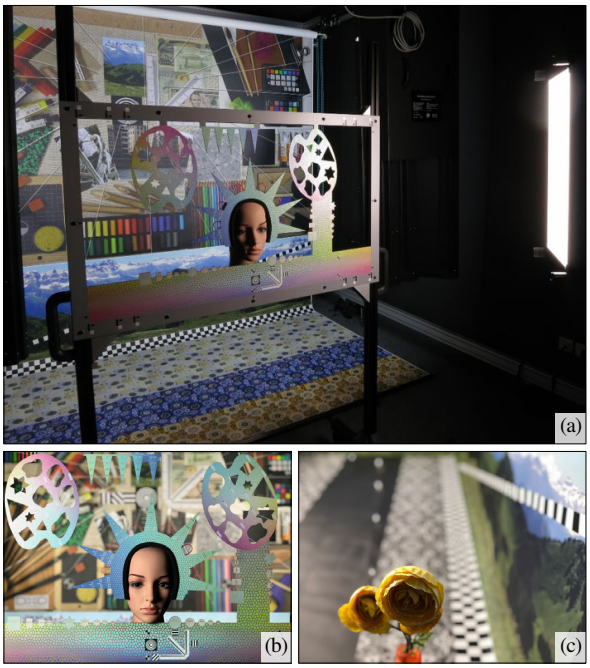


Figure 9. *The proposed lab setup (a) containing a first scene (b) consisting of two planes the distance between which can be adjusted, as well as a second scene (c) with continuously changing depth.*

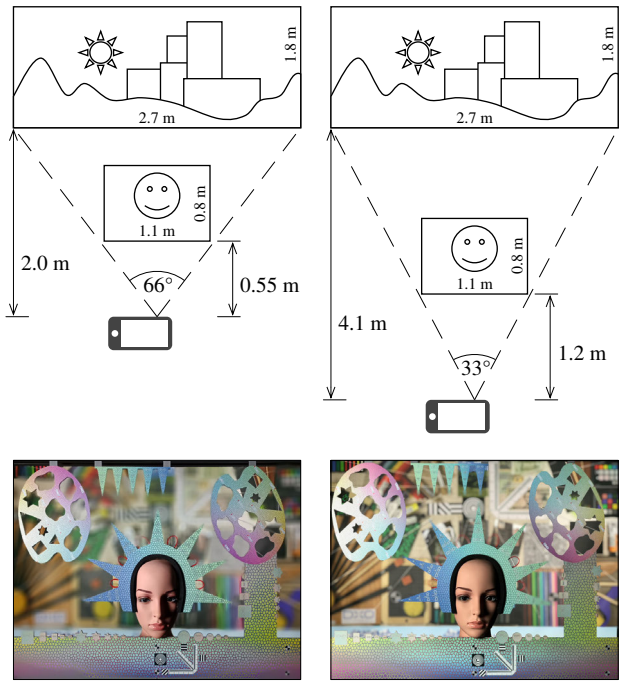


Figure 10. *By adjusting the distances between device, subject and background in function of the equivalent focal length (i.e. the field of view), our test scene appears the same for all devices. This allows to compare subject/background segmentation independently of the equivalent focal length. The above example shows the distances and resulting images for smartphones with equivalent focal lengths of 24 and 54 mm.*

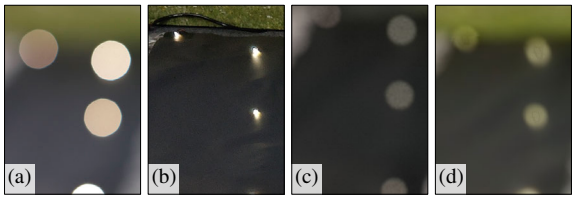


Figure 11. *Out-of-focus points of light. (a) DSLR (full frame, 50 mm, f/2.8); (b) smartphone without computational bokeh; (c) good computational bokeh; (d) computational bokeh affected by processing artifacts.*

cal length of the smartphone, we pick the appropriate stack of DSLR images and compare the blurriness produced by the smartphone to these references. The aperture of the DSLR image that comes closest defines the equivalent aperture of the computational bokeh.

The *blur gradient smoothness* is observed on the regular patterns at the bottom and on the line of black and white squares on the right. Having the same pattern at all distances reveals even small discontinuities that would go unnoticed in many real-world scenes.

Noise consistency is observed between the in-focus subject and the background. Although the depth differs greatly between the two image patches we compare, they appear close to each other in the image, and have similar gray levels, so that there is no reason besides bokeh simulation for them to show different grain.

The *bokeh shape* can be observed on the LED light spots in the background. Assigning a score is difficult because of the missing general agreement on what perfect bokeh should look like. Our current approach is to demand a circular shape with rather sharp borders. Artifacts like in figure 11 (d) should be avoided because they do not resemble optical bokeh of any lens. And DSLRs do not necessarily obtain the highest score in this category because their optical bokeh is not always circular.

Evaluating repeatability

Laboratory setup. We apply some variations to both scenes, to observe the impact of the presence or absence of faces and that of moving parts in the scene. We test at two different levels of illumination, at 1000 and 50 lux. For every variation, we take five identical pictures, to observe repeatability issues that are independent of the scene content.

Measure. Repeatability is assessed perceptually by comparing the images showing the same scene.

Results

The bokeh evaluation protocol we apply in our DxOMark Mobile reviews is heavily inspired by the described method. The multitude of individual measurement results it provides for each criteria are aggregated into subscores via empiric formulas, which are then aggregated into a single-value bokeh score via another empiric formula. While the individual measurements are inherently objective, the aggregation formulas may seem a bit subjective at first.

For instance, how should repeatability and equivalent aperture be balanced? Obviously we want both, so the maximum score must be given to a device that produces strong bokeh *and* has high repeatability. But do you prefer a device that repeatably produces

weak bokeh—or a device that produces artifacts most of the time but sometimes manages to obtain really stunning results? This is definitely a question of preference and there is no universal answer. When designing our aggregation formulas, our goal is to guide smartphone manufacturers to make trade-offs that please *the majority* of their users. We therefore base our formulas on assumptions on what the majority of people want.

For the DxOMark Mobile protocol, we complete the lab measurements with a set of natural test scenes. They are less repeatable, but having a large diversity of image content allows to confirm the lab results.

Figure 12 shows the scores for three smartphones released in 2017. The theoretical best score, attained by a DSLR with a very good lens, is 100. The smartphones differ by the hardware they use and by the processing and tuning applied. Overall they all obtain impressive results, even though they cannot (yet) match the DSLR. In figure 13 we reproduce some details from our lab scenes to illustrate how the quality difference measured by our method manifests itself in the lab images. In figure 14, we show a real-world use case, an indoor portrait, shot with the three devices.

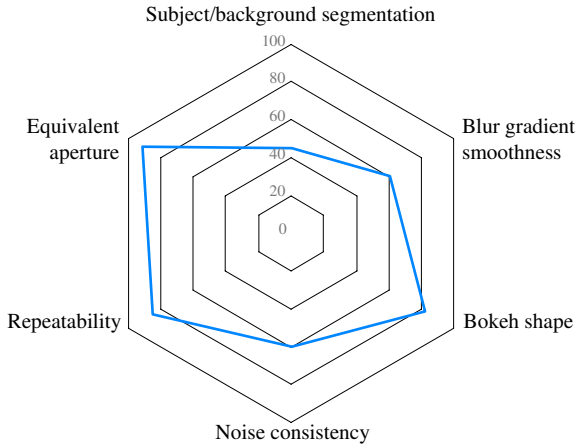
Conclusion

We have presented a laboratory setup and assessment method to evaluate the quality of computational bokeh. Our method does not require shooting all test images with all smartphones on the same day—the score can be determined independently for each smartphone. This flexibility allows to rank all smartphones on a common scale. We have tested all major smartphones proposing shallow depth of field modes, from 2015 to today. The scores are shown in figure 15. The progress between 2015 and 2017 is remarkable and makes us optimistic that most of the teething troubles currently connected with computational bokeh will be solved during the next years.

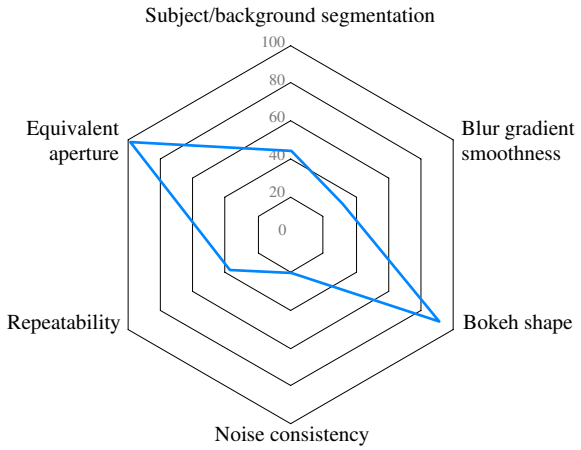
Our evaluation shows that precise segmentation of subject and background is key for obtaining convincing results. Interestingly, we observe no significant correlation between bokeh quality and the hardware technology employed. On the other hand, we observe obvious differences in tuning and in the manufacturers readiness to assume risk: some apply strong bokeh aiming to reproduce the very shallow depth of field of DSLR portrait lenses. Others prefer blurring a little less, thereby hiding artifacts caused by perfectible subject/background segmentation and obtaining more natural look.

The proposed method is designed so that a DSLR obtains the highest score, provided it has a circular aperture of $f/2.8$ or wider. One might however imagine situations where smartphones and their computational bokeh outperform DSLRs. Suppose a group portrait where different faces lie in slightly different planes. Using a DSLR, for obtaining a photograph where all faces appear perfectly sharp, you would have to stop down, which would significantly reduce background blur. This is because depth of field and background blur are related by physical laws. Computational bokeh is not subject to these laws. A smartphone could combine a depth of field large enough to have all faces sharp *and* strong background blur. An evaluation method that takes such use cases into account is subject to future research.

Device A (DxOMark Mobile bokeh score 55)



Device B (DxOMark Mobile bokeh score 30)



Device C (DxOMark Mobile bokeh score 45)

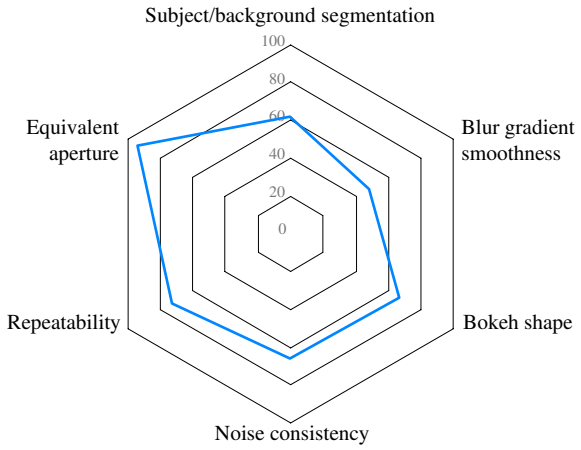
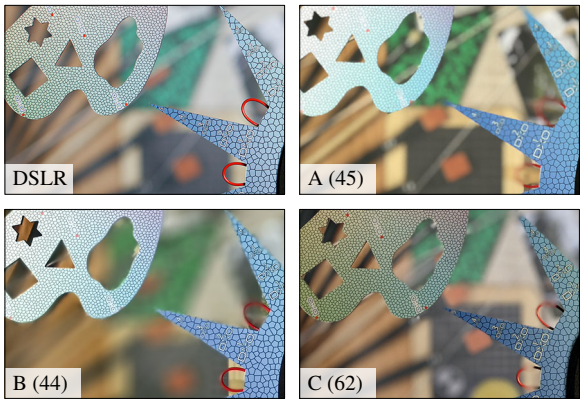
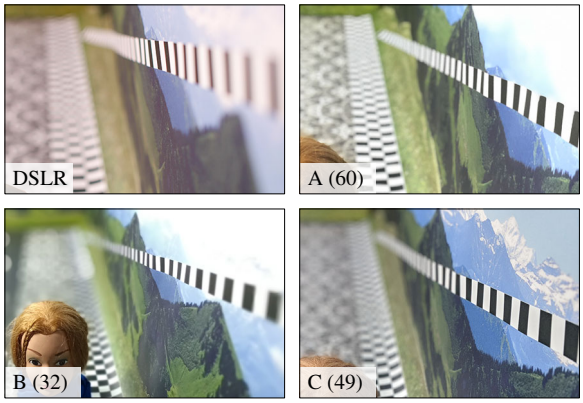


Figure 12. *Sub- and total scores for three smartphones released in 2017. Device A: wide-angle + telephoto dual camera. Device B: color + monochrome dual camera. Device C: single camera with dual pixels.*

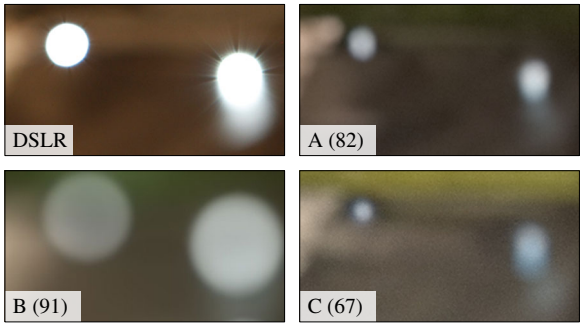
Subject/background segmentation



Blur gradient smoothness



Bokeh shape



Noise consistency

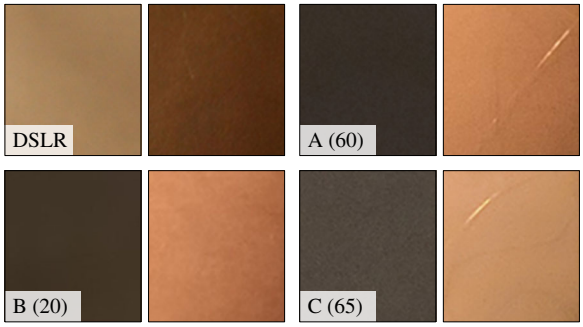


Figure 13. *Examples that illustrate the differences between a DSLR and the smartphone devices A, B and C. The scores in figure 12 are computed from dozens of unitary evaluations and give a more nuanced picture than any of these examples.*

Device A

Device B

Device C

Figure 14. *Real-world use case: an indoor portrait shot in shallow depth of field mode. Note that the comparison is tricky due to the different fields of view. Photographers usually prefer the narrower field of view of the devices A and C. While A is equipped with a dedicated telephoto lens, C applies digital zoom. Device A achieves good segmentation between the subject and the background and it applies rather weak bokeh. While this limits the wow-factor, it also limits the visibility of segmentation artifacts. This device does not attempt to blur the table in front of the subject. The only annoying artifact are the sharp points of light behind the subject's waving hand. Device B simulates strong bokeh, but the result suffers from segmentation errors, which are amplified by the strength of the bokeh. The upper part of the refrigerator is blurry, the lower part is perfectly sharp. The microwave oven becomes sharper as it approaches the subjects head, and parts of the hair are blurred. Device C, despite having only one camera, comes very close to device A. But the control on the refrigerator is sharp and the transition between the table and the subject's arm is too abrupt.*

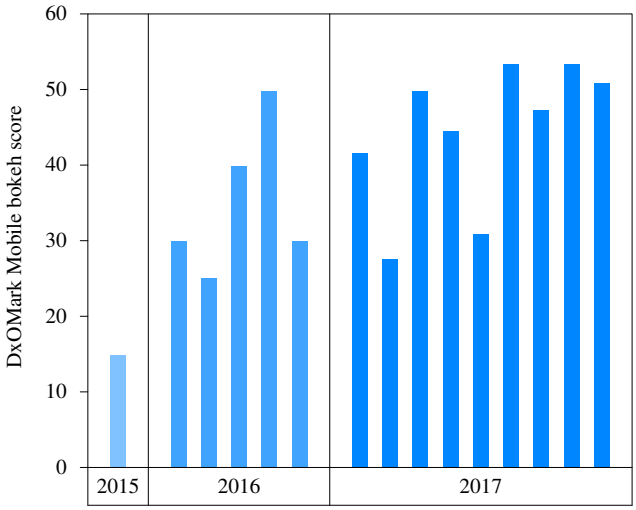


Figure 15. *DxOMark Mobile bokeh scores for major smartphones offering shallow depth of field modes, in chronological order. The progress made in only two years is remarkable and suggests that computational bokeh will further improve during the coming years.*

References

[1] Ralph E. Jacobson, Sidney F. Ray, Geoffrey G. Attridge and Norman R. Axford, *The Manual of Photography*, 9th edition, Focal Press (Elsevier), 2000.

[2] James L. Mannos, David J. Sakrison, "The Effects of a Visual Fidelity Criterion on the Encoding of Images", *IEEE Transactions on Information Theory*, vol. 20(4), 1974.

[3] "Depth of Field—An Insider's Look Behind the Scenes", *Camera Lens News*, Carl Zeiss AG, 1997.

[4] Hubert H. Nasse, *Schärfentiefe und Bokeh*, Carl Zeiss AG, 2010.

[5] FE 100mm F2.8 STF GM OSS, Sony, product page, https://www.sony.com/electronics/camera-lenses/sel100f28gm, consulted 2017-11-24.

[6] Adam P. Murray, *Tested: Galaxy Note 8 Live Focus vs iPhone 7 Plus Portrait Mode*, Macworld, https://www.macworld.com/article/3221403/android/galaxy-note-8-live-focus-vs-iphone-7-plus-portrait-mode.html, 2017, consulted 2017-11-24.

[7] Raymond Wong, *Which phone takes the best portrait photos: iPhone X, Pixel 2, or Note 8?*, Mashable, http://mashable.com/2017/11/09/apple-iphone-x-pixel-2-xl-note-8-portrait-mode-comparison, 2017, consulted 2017-11-24.

[8] *Full camera shootout: Galaxy Note8 vs iPhone 8 Plus*, GSM Arena, https://www.gsmarena.com/iphone_8_plus_vs_galaxy_note8-review-1673.php, 2017, consulted 2017-11-24.

[9] *ColorChecker Classic*, X-Rite, product page, https://www.xrite.com/categories/calibration-profiling/colorchecker-classic, consulted 2017-11-24.

[10] *USAF 1951 Resolution Target*, SilverFast, product page, http://www.silverfast.com/show/resolution-target/en.html, consulted 2017-11-24.

[11] ISO 12233, *Photography – Electronic still picture imaging – Resolution and spatial frequency responses*, 2017.

[12] Kelcey Smith, *Studio Test Scene*, DPReview, https://www.dpreview.com/articles/2601653565/studio-test-scene, 2013, consulted 2017-11-24.

[13] Ann B. Lee, David Mumford, Jinggang Huang, "Occlusion Models for Natural Images: A Statistical Study of a Scale-Invariant Dead Leaves Model", *International Journal of Computer Vision*, vol. (1/2), 35–59, 2001.