3D Unsharp Masking for Scene Coherent Enhancement
Supplemental Material 1: Experimental Validation of the Algorithm

Tobias Ritschel Kaleigh Smith Matthias Ihrke Thorsten Grosch Karol Myszkowski Hans-Peter Seidel
Computer Graphics Group, MPI Informatik Saarbrücken

The algorithm described in the main paper provides a general technique of enhancing the perceived contrast in synthesized scenes. Because the procedure operates on the lighting function, the enhancement is done in a scene coherent way. To achieve best possible results, it is crucial that the core parameters of the algorithm are set optimally.

The aims of the study described here are therefore:

1. to provide empirical evidence for the assumption that the enhanced scene is superior to the original in terms of perceived contrast and preference
2. to provide guidelines for the setting of the algorithm's parameters λ and σ in a variety of different scenes such that they do not produce objectionable artifacts.

1 Method

To study the meaningful range of settings for the parameters of interest, a standard psychophysical procedure, the "method of adjustment" was used [Gescheider 1997]. This was deemed to be the most appropriate choice for the current study since it allows both for an efficient implementation of the experiment and provides a single means of estimating all thresholds and preference values of interest.

In this framework, the kernel width σ and the gain value λ were considered by allowing the subjects to adjust λ for a given scene under different tasks and varying σ. Of special interest for determining the reasonable parameter range are the values for λ which result in a barely visible enhancement (λ_{low}) and objectionable artifacts (λ_{obj}), respectively. To find out about the preference of the users, the subjectively "best contrast" setting of the gain value (λ_{best}) for a given scene and a given σ was also obtained.

In the experimental setup, two stimuli were presented next to each other, where both images depicted the same scene seen from the same viewpoint. One of the two images showed an enhanced version of the scene while the other showed the original, non-enhanced scene for comparison. The space of the kernel-width σ was discretized in three steps $\sigma \in \sigma_{\text{low}}, \sigma_{\text{medium}}, \sigma_{\text{high}}$ which were chosen in a pilot study to produce a perceptually low, medium and strong effect, respectively. For the 3 different settings of σ, the subjects than adjusted λ according to their preference.

1.1 Participants

15 participants (9 male, 6 female) with normal or corrected-to-normal vision took part in the experiments. Subjects were compensated for their efforts with a small fee (15 USD). Participants were recruited from the university campus and were mostly students of computer science and their mean age was 24 years (range 19 to 30 years). Subjects were naïve regarding the goal of the experiment and inexperienced in the field of computer graphics and photography.

1.2 Design

The study implemented a 4×3 within-subject design by varying 4 scenes and the kernel size σ (low, medium, high) as independent variables. The gain values $\lambda_{\text{low}}, \lambda_{\text{best}}, \lambda_{\text{obj}}$ adjusted by the user under the 3 different instructions functioned as dependent variables.

The experiment was divided into 32 trials. The three steps of the kernel width σ were presented 2 times in randomized order for each scene in order to provide the possibility to assess the reliability of the method. The sequence of the scenes was also randomized across subjects. In addition, the first two trials for each scene were repeated as they functioned as practicing phase for each scene. The data from these first two trials did not enter in the final analysis.

1.3 Materials and Apparatus

All stimuli were presented at a resolution of 2048×1536 on a 20.8 inch (diagonal) Barco Coronis Color 3MP (MDCC 3120-DL) display that was connected to a personal computer running the Psychtoolbox software [Brainard 1997]. The monitor was viewed by the subjects orthogonally at a distance of 80 cm, the whole trial-display occupying 29.68 visual degrees. The stimuli were generated by rendering 4 different scenes and then enhancing them using the algorithm described in the main paper. The images were displayed on the screen at a resolution of 960×720 each. The four tested scenes are depicted in Figure 10 – 13.

Between two consecutive steps in the adjustment of λ, a mask was displayed for 300 ms in order to prevent subjects from judging the adjustment temporally (between successive steps of the enhancement) instead of comparing to the original image. A lowpass version of the image blurred with a gaussian filter of size 50×50 pixels was used as a mask.

1.4 Procedure

On entering the laboratory, subjects were asked for their experience in the areas of computer graphics and/or photography. Subjects that judged their experience within one of these fields as "above average" were excluded from the study. The participants were then seated in front of a monitor running the experimental software in an otherwise darkened room. They received standardized on-screen instructions regarding the procedure of the experiment. Before the actual experiment started, the participants were shown all scenes that were to appear in the course of the session.

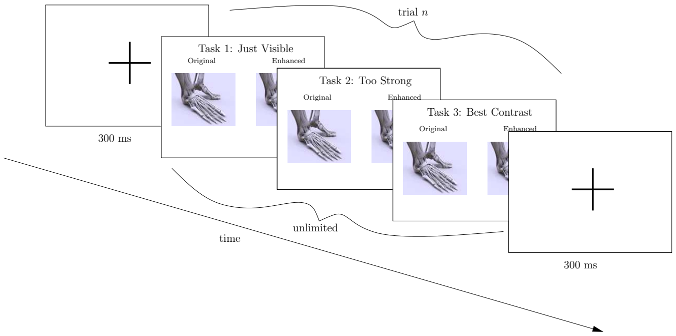

Figure 1: *Experimental procedure for a single trial.*

Each trial started with a fixation cross in the middle of the screen (1500 ms) that was followed by the trial-display, consisting of images of the original and the enhanced scene (see fig. 1). The location of the enhanced image (left or right) was balanced across subjects. At the beginning of each trial, these images were identical as the gain value λ was initially set to 0. Subjects were then asked to adjust the enhancement by pressing the up- and down-keys. A discrete beep indicated that the subjects reached either $\lambda = 0$ or the upper limit $\lambda = 3$. The participants confirmed their choice of the image by pressing the Return-key.

For each display, the subject had to complete 3 tasks:

1. adjust the image until there was a just noticable difference in contrast between the original and the enhanced image,
2. adjust the image until it looked too strong and artifacts appeared,
3. find the subjectively best-contrast setting (this task was made more comprehensible by giving the analogy of adjusting the contrast of a TV-set).

The current task was indicated above the two images.

After the participants had completed the final 8th trial for each scene, they were presented with a hardcopy showing the same image and asked to mark the parts that they used most frequently to evaluate the strength of the enhancement.

The complete experimental session took approximately one hour. In spite of the rather long time required to complete the experiment, subjects did not complain about a loss of concentration, an effect that might be due to the good sense of control induced by the adjustment procedure.

2 Results

For a descriptive summary of the data, please refer to tables 1 and 2 and to Figures 4 and 5. Thumbnails of the images chosen by the subjects for all tasks are depicted in Figures 6 to 9 along with the corresponding values for the parameters.

2.1 Statistical Analysis

Analysing the two repetitions of the settings for λ chosen by the user for the same parameter set by statistically testing for the significance of their correlation, the experimental procedure proves to be quite reliable (all $r > .75$, all $p < .001$). In order to decrease the impact of random fluctuations, the mean of the two repetitions is used in the further analysis as dependent variable. Also, to correct for outliers, all values whose difference to the mean value (for each scene, σ and task) exceeded that of two times the standard deviation were replaced by the mean of the remaining values (less than 5% of the data). Because the gain value λ scales differently for all scenes, independent 3×3 ($\sigma \times$ task) ANOVAs were computed for each scene, treating both factors as repeated measures because they were obtained from the same subject.

Chamfer Plane. The ANOVA reveals only a main effect of task, $F(2, 28) = 58.99, p < .001$, all other main and interaction effects are non-significant. Using adjusted pairwise contrasts (following Holm's [Holm 1979] proposal), the found effect can readily be shown to stem from the best contrast setting being being larger than the lower threshold ($\lambda_{\text{low}} < \lambda_{\text{best}}$: $t(14) = 2.14, p < .05$) and the higher threshold being larger than the preferred setting ($\lambda_{\text{best}} < \lambda_{\text{obj}}$: $t(14) = 4.03, p < .01$). As an indicator for the size of the effect, Cohen's d [Cohen 1988] is computed for all of the pairwise contrasts, yielding medium to strong effects according to Cohen's conventions ($d(\lambda_{\text{low}}, \lambda_{\text{best}}) = .78$, $d(\lambda_{\text{best}}, \lambda_{\text{obj}}) = 1.47$).

Dice. The corresponding ANOVA for the *Dice* scene also reveals a single main effect of task, $F(2, 28) = 32.06, p < .001$. Again, pairwise comparisons indicate the pattern of results outlined above ($\lambda_{\text{low}} < \lambda_{\text{best}}$: $t(14) = 1.83, p = .08$; $\lambda_{\text{best}} < \lambda_{\text{obj}}$: $t(14) = 2.75, p < .03$). Again, the effect sizes are moderate to strong ($d(\lambda_{\text{low}}, \lambda_{\text{best}}) = .67$, $d(\lambda_{\text{best}}, \lambda_{\text{obj}}) = 1.01$).

Feet. The ANOVA for the *Feet* scene reveals again the same main effect of task, $F(2, 28) = 49.01, p < .001$ but also a main effect of the kernel width σ, $F(2, 28) = 16.46, p < .001$ as well as an interaction $\sigma \times$ task, $F(4, 56) = 7.53, p < .001$. The main effect of task can be shown to stem from the same pattern discussed above, $\lambda_{\text{low}} < \lambda_{\text{best}}$: $t(14) = 1.16, p < .03$ ($d = 0.88$); $\lambda_{\text{best}} < \lambda_{\text{obj}}$: $t(14) = 4.51, p < .01$, ($d = 1.65$). From Figure 4c) the expectation that the main effect for σ is caused by a preference for higher λ-values for low σ and only for the preferred and upper threshold, can be derived. However, the planned contrasts do not reach significance for the pairwise comparisons (σ_{low}, σ_{medium}: $t(14) = 1.77, p = .09$, σ_{low}, σ_{high}: $t(14) = 1.38, p = .18$). The interaction effect is due to the visibility threshold's independence on σ (see fig. 4), because the λ_{low}-values did not show any effect of σ ($F(2, 28) = 0.05, p > .9$).

Keys. Again, the ANOVA yields a main effect of task ($F(2, 28) = 32.15, p < .001$) and σ ($F(2, 28) = 7.85, p < .002$) as well as the interaction $\sigma \times$ task, $F(4, 56) = 4.54, p < .05$. The effect of task however does not manifest itself in a significant difference between lower threshold and preferred value ($t(14) = 1.28, p = .21$), so users preferred a rather subtle enhancement for that scene that was barely above visibility threshold. There is however still a significant range of parameter values before the occurence of artifacts ($\lambda_{\text{low}} < \lambda_{\text{obj}}$: $t(14) = 4.86, p < .001$, $d(\lambda_{\text{low}}, \lambda_{\text{obj}}) = 1.70$). The overall σ-effect as well as the interaction effect from the ANOVA cannot be tracked down to single differences, due to missing statistical power.

The reported analyses revealed an effect of the kernel size σ only for two of the scenes, *Feet* and *Keys* while for scenes *Dice* and *Chamfer Plane* it had a zero-effect. A possible explanation for this result could be that the scenes differed according to the mean distances between objects, thereby producing a higher amount of "cluttering" for the *Feet* and *Keys* scenes (in the other two scenes, larger shapes that are relatively distant from each other dominate). For higher values of σ, parts of objects close to each other in the scene could have contributed to the local enhancement, producing objectionable artifacts already for relatively low values of λ.

2.1.1 Conversion to JND-Units

In order to allow for an inter-scene comparison, the λ-values are rescaled to just-noticable-difference (JND) units, based on the value of the lower threshold

$$\hat{\lambda} = \frac{\lambda}{\lambda_{\text{low}}}.$$

This implies the assumption of a linear relationship between perceived contrast and λ-value. Because the range of interesting values is relatively small, this approximation can be assumed to be valid.

To determine whether the adjustment of the parameter shows a dependency on the scenes, an overall $4 \times 3 \times 3$ (scenes $\times$ σ $\times$ task) ANOVA is conducted with the rescaled $\hat{\lambda}$-values. A main effect of scene reaches significance ($F(3, 42) = 5.09, p < .005$). Pairwise comparisons reveal however, that only scene *Dice* and *Keys* differ in a statistically significant way ($t(14) = 2.89, p < .05$) while all other comparisons did not approach significance. Because these two scenes were very different from each other in terms of

Table 1: *Summary of the results from the study. User-chosen λ values are given for each setting of σ, scene and task.*

	User-adjusted λ								
	λ_{low}			λ_{best}			λ_{obj}		
	σ_{low}	σ_{medium}	σ_{high}	σ_{low}	σ_{medium}	σ_{high}	σ_{low}	σ_{medium}	σ_{high}
Dice	0.05	0.06	0.05	0.11	0.11	0.10	0.22	0.22	0.23
Keys	0.25	0.22	0.21	0.34	0.31	0.26	0.75	0.63	0.56
Chamfer Plane	0.34	0.36	0.32	0.46	0.55	0.55	0.96	1.04	1.05
Feet	0.23	0.22	0.20	0.43	0.27	0.33	0.85	0.69	0.71

Table 2: *Summary of the results from the study. User-chosen and rescaled $\hat{\lambda}$ values are given for each setting of σ, scene and task.*

	User-adjusted λ								
	$\hat{\lambda}_{\text{low}}$			$\hat{\lambda}_{\text{best}}$			$\hat{\lambda}_{\text{obj}}$		
	σ_{low}	σ_{medium}	σ_{high}	σ_{low}	σ_{medium}	σ_{high}	σ_{low}	σ_{medium}	σ_{high}
Dice	1.00	1.00	1.00	1.81	1.82	2.28	4.83	4.20	5.55
Keys	1.00	1.00	1.00	1.48	1.46	1.09	3.39	2.89	2.73
Chamfer Plane	1.00	1.00	1.00	1.78	1.83	1.91	3.52	3.22	3.63
Feet	1.00	1.00	1.00	2.06	1.39	1.76	3.99	3.38	3.73

lighting, assembly and spatial frequency (on the mesh), this result indicates that the enhancement parameter's space is perceptually rather uniform and only diverges when scenes show very different characteristics.

The other statistical comparisons do not produce any new conclusions, the pattern of results already discussed remains stable also for the transformed values. The transformation of the λ-values to JND-units emphasizes the finding, that the preferred setting of λ is relatively stable across scenes and settings of σ. The emergence of artifacts however, is apparently more scene-dependent. As a rule of thumb it can be formulated, that coming from a just visible enhancement, the strength can be doubled to yield close-to-optimal results and multiplied by four to yield an approximation of the upper threshold.

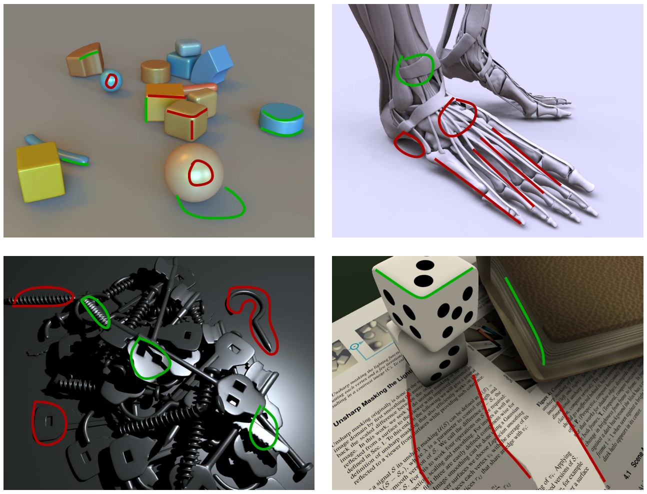

Figure 2: *"Regions of Interest" marked by the users as guiding their adjustment process. Red color indicates a very common markup, green a less frequent occurence.*

2.2 Regions of Interest

The regions indicated by the subjects as being the most important for their decision of the enhancement strength are depicted in Figure 2. All subjects focussed on regions that were expected to be sensitive to the algorithm (e.g. the shadows in the *Dice* scene or the edges of the objects in the *Chamfer Plane*).

It is therefore concluded, that subjects were able to grasp the meaning of the enhancement and to relate it appropriately to their response.

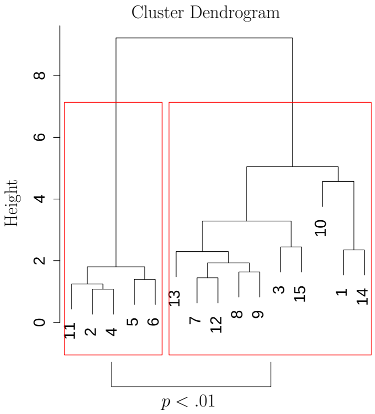

Figure 3: *Cluster Analysis for the subjects using Ward's method and euclidean distances. Numbers indicate different subjects, clusters are marked in red.*

2.3 Individual Differences

To investigate individual differences in the preference for various strengths of the enhancement, a hierarchical cluster analysis applying Ward's method with euclidean distances is computed [Hastie et al. 2001]. The cluster-dendrogram (see Fig. 3) reveals two groups of subjects, distinguished by their preference of the adjustment. It is assumed, that the two subgroups are separating persons with strong and weak preference for contrast, respectively. A statistical test comparing these two groups of subjects according to their mean adjustment of λ_{low}, λ_{best} and λ_{obj} shows that the subjects assigned to cluster 1 prefer lower values of λ for all tasks compared to subjects from cluster 2 (λ_{low}: $t(14) = -3.23, p < .01$, λ_{best}: $t(14) = -7.12, p < .01$, λ_{obj}: $t(14) = -4.71, p < .01$).

References

BRAINARD, D. 1997. The psychophysics toolbox. *Spatial Vision 10*, 437–442.

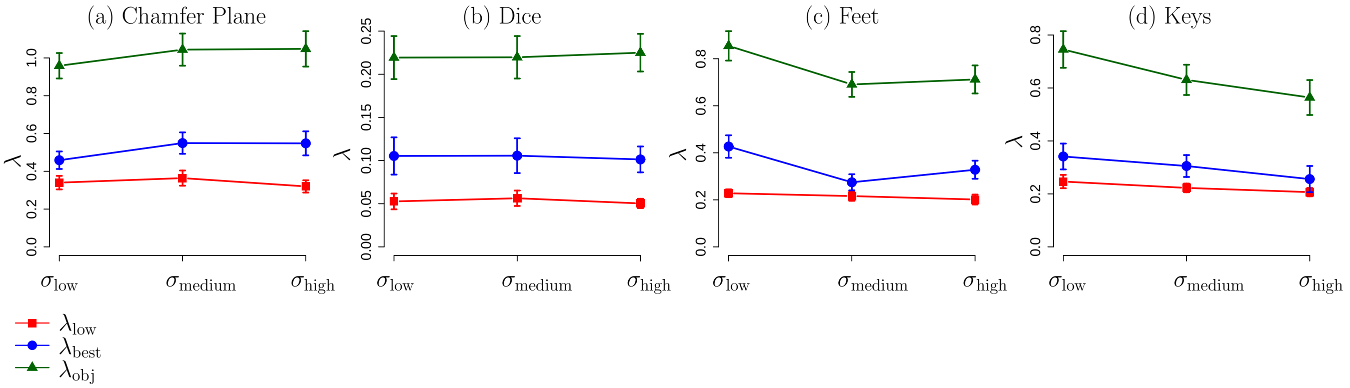

Figure 4: *User adjustments of the gain value λ varied as a function of kernel width σ for the 4 scenes. Error-bars indicate the standard error of mean (SEM).*

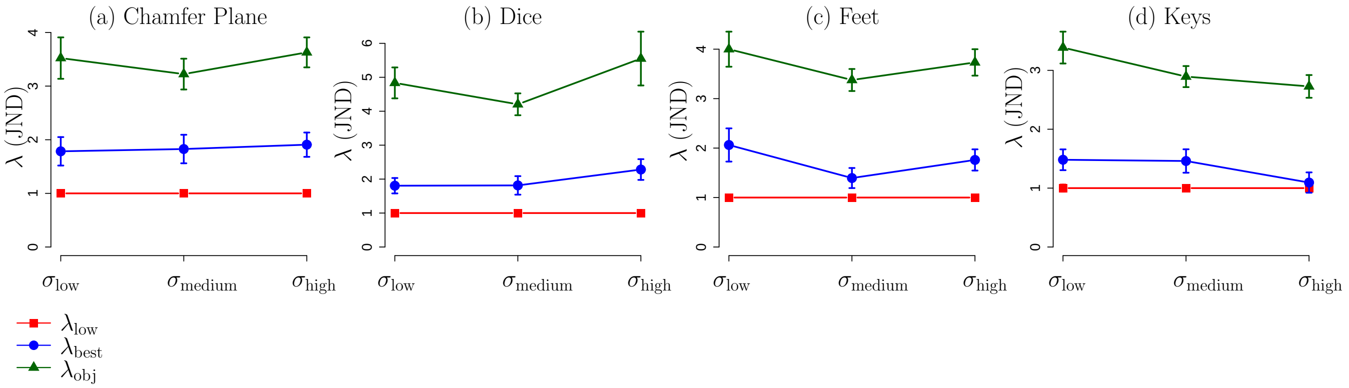

Figure 5: *User adjustments of the gain value λ varied as a function of kernel width σ for the 4 scenes in JND units. Error-bars indicate the standard error of mean (SEM).*

COHEN, J. 1988. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates.

GESCHEIDER, G. A. 1997. *Psychophysics: The Fundamentals*. Lawrence ErlbaumAssociates.

HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

HOLM, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

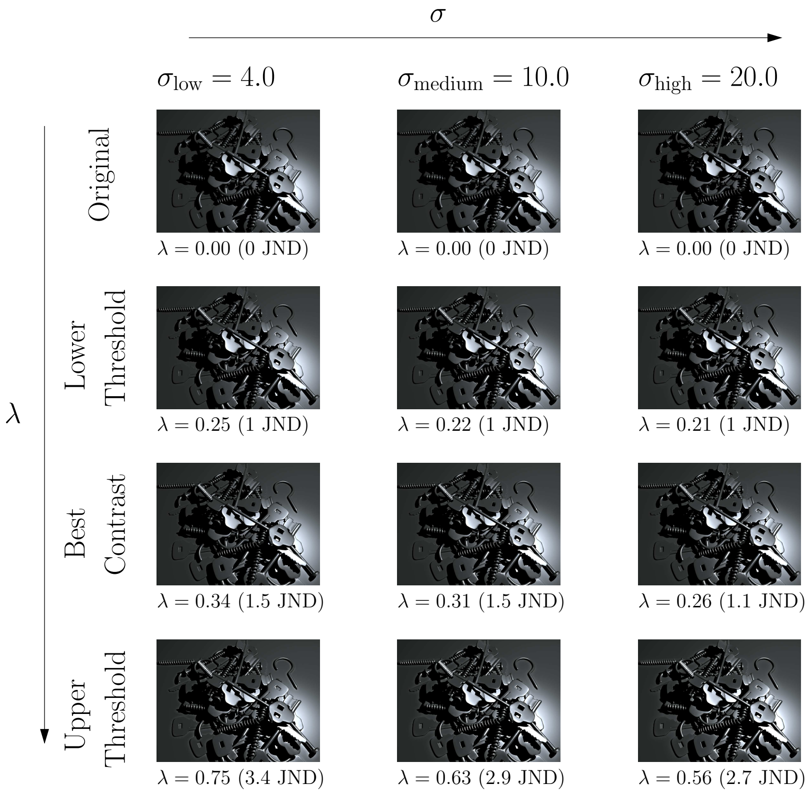

Figure 6: *User chosen λ for scene* Keys*. The images chosen by the users are shown for all tasks and values of σ.*

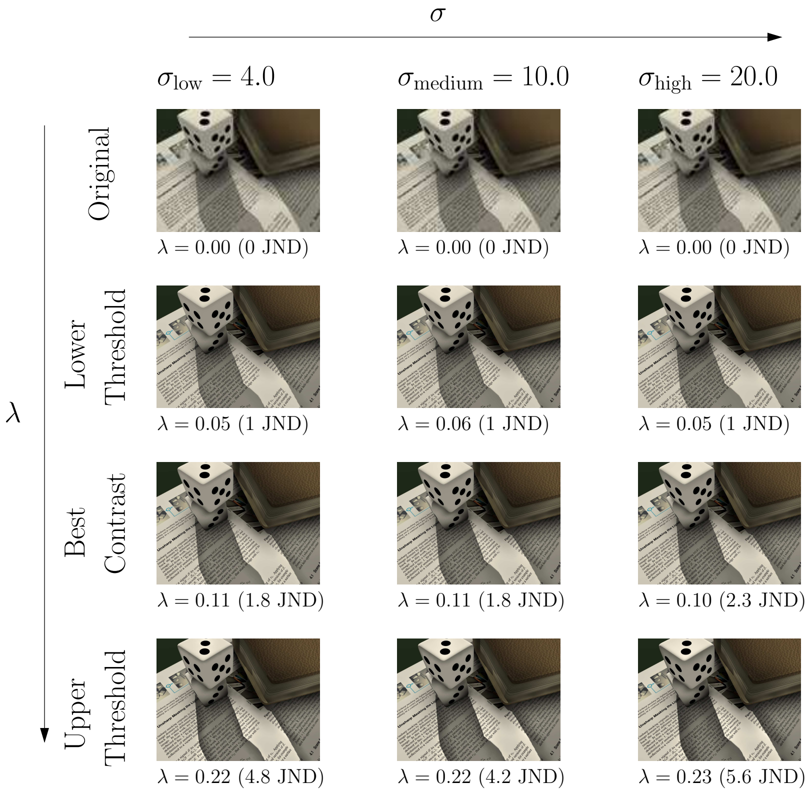

Figure 7: *User chosen λ for scene* Dice*. The images chosen by the users are shown for all tasks and values of σ.*

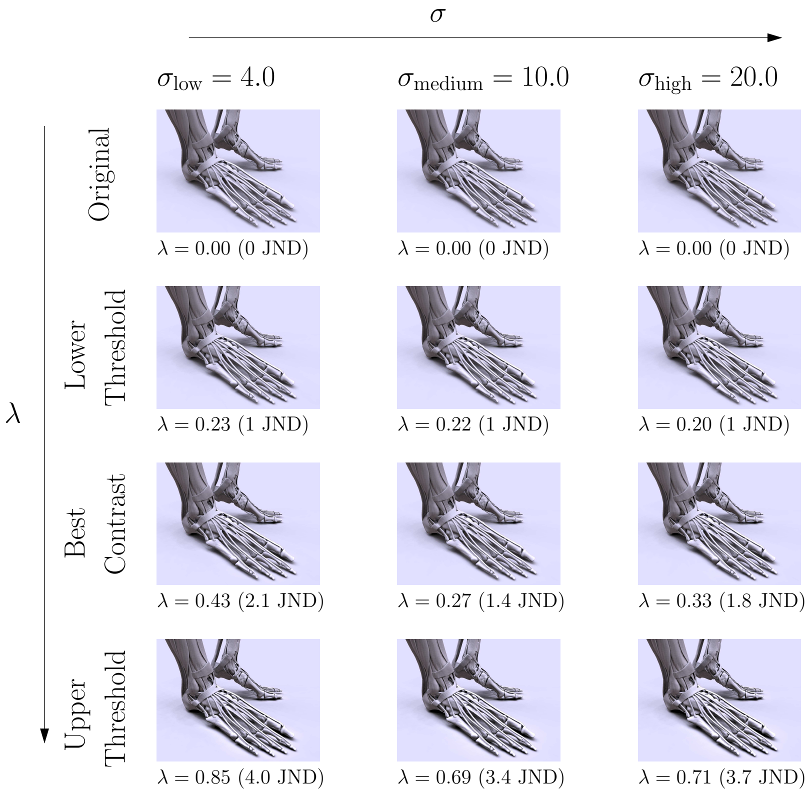

Figure 8: *User chosen λ for scene* Feet. *The images chosen by the users are shown for all tasks and values of σ.*

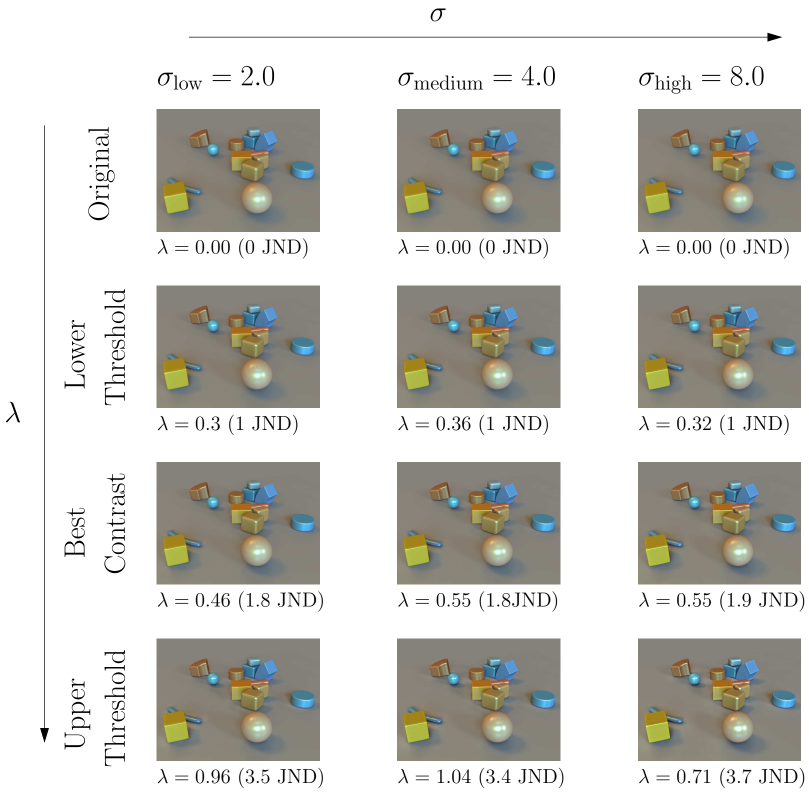

Figure 9: *User chosen λ for scene* Chamfer Plane. *The images chosen by the users are shown for all tasks and values of σ.*

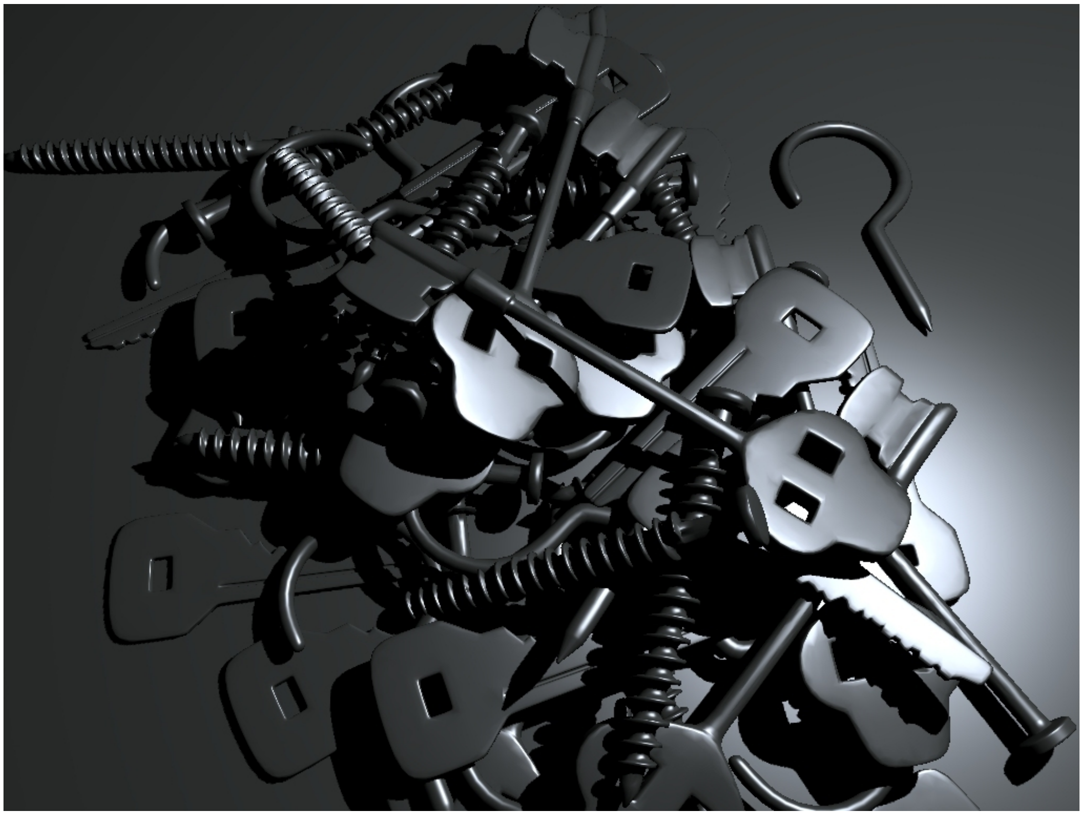

Figure 10: *Scene* Keys.

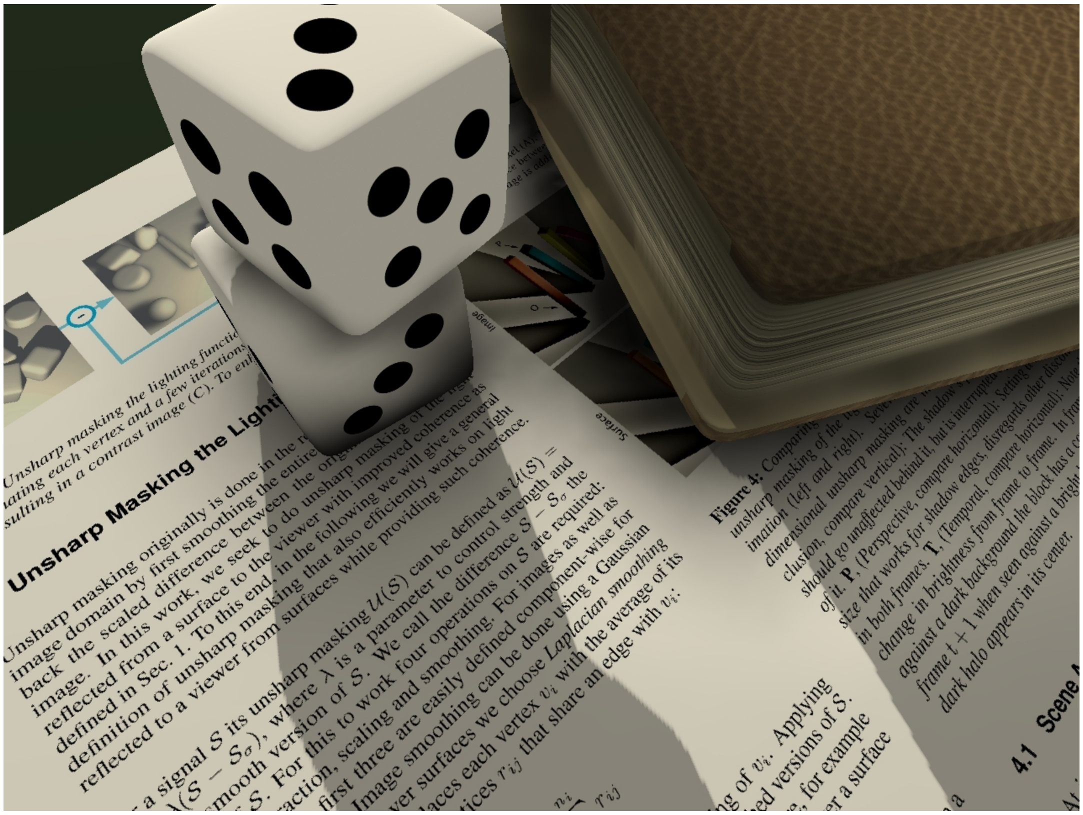
Unsharp Masking the Lighti
Unsharp masking the lighting functio
Unsharp masking originally is done in the r
image domain by first smoothing the entire
back the scaled difference between the orig
image. In this work, we seek to do unsharp masking of the light
reflected from a surface to the viewer with improved coherence as
defined in Sec. 1. To this end, in the following we will give a general
definition of unsharp masking that also efficiently works on light
reflected to a viewer from surfaces while providing such coherence.
Image
Surface
Figure 4: Comparing
unsharp masking of the
imation (left and right).
dimensional unsharp masking
clusion, compare vertical)
should go unaffected behind it
of it. P, (Perspective, compare horizontal)
size that works for shadow edges
in both frames. T, (Temporal, compare horizontal)
change in brightness from frame to frame
against a dark background the block has
frame t + 1 when seen against a bright
dark halo appears in its center.
4.1 Scene

Figure 11: *Scene* Dice.

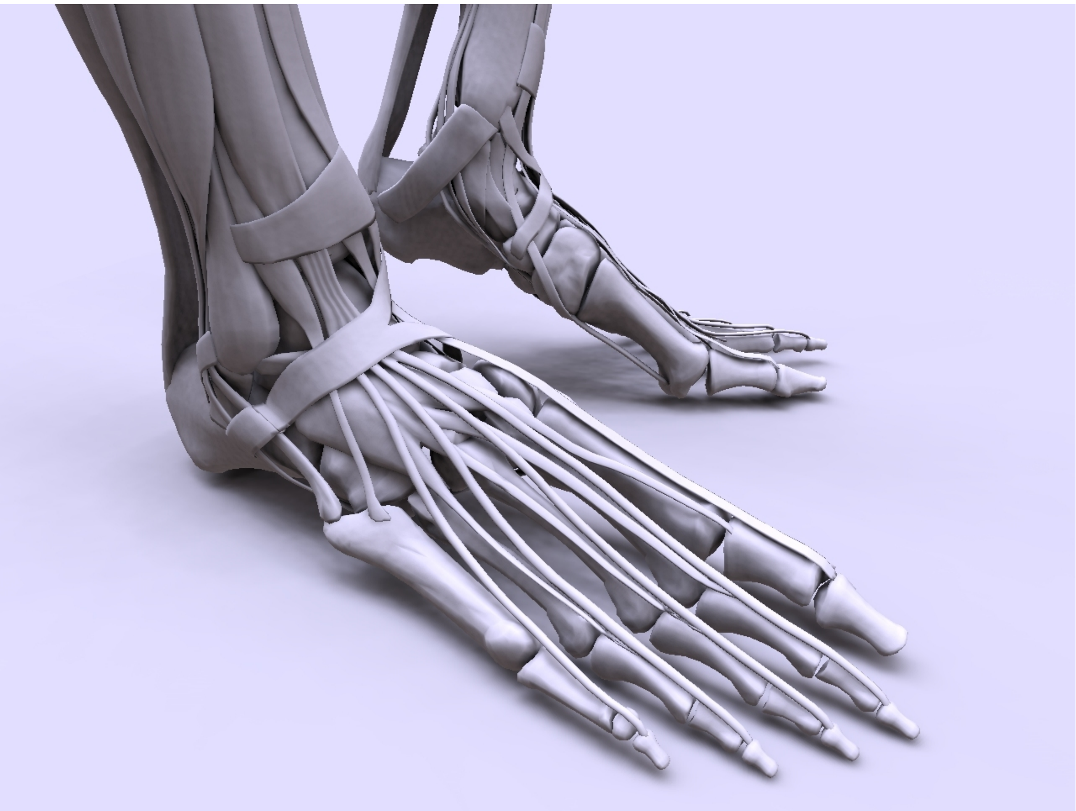

Figure 12: *Scene* Feet.

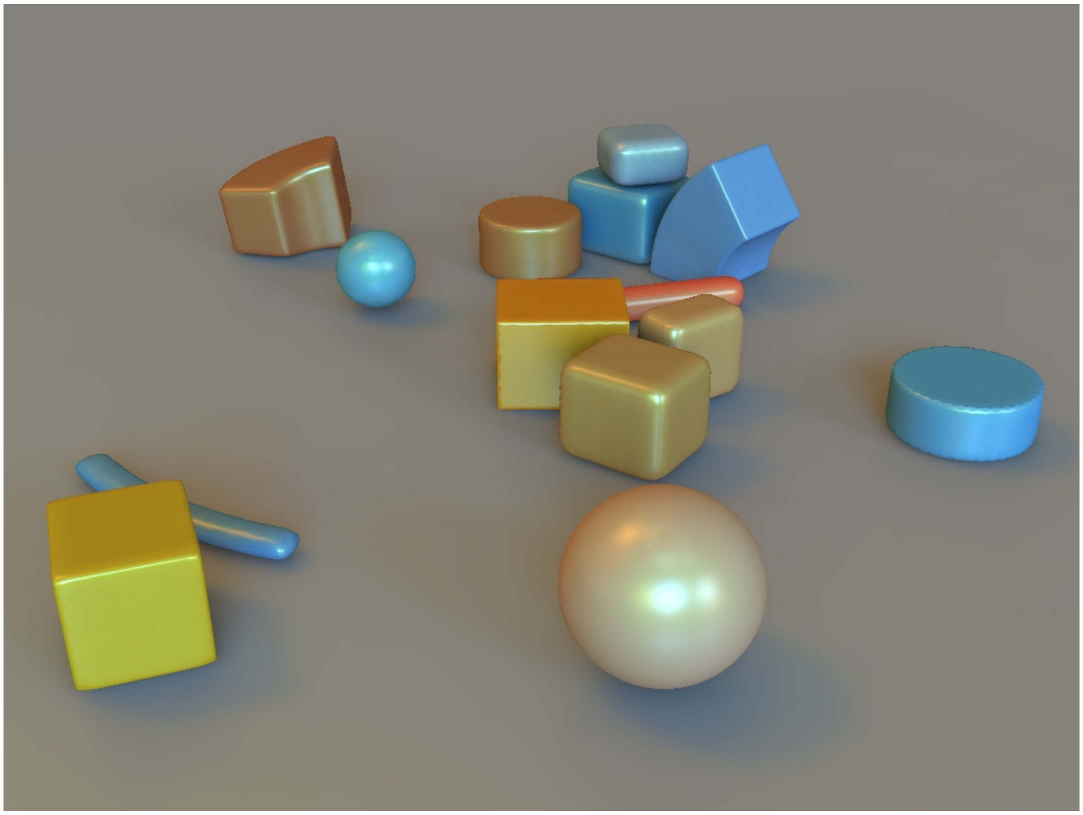

Figure 13: *Scene* Chamfer Plane.