# Distortion-Based Transparency Detection using Deep Learning on a Novel Synthetic Image Dataset

Volker Knauthe[1][0000−0001−6993−5099], Thomas Pöllabauer[2,1][0000−0003−0075−1181], Katharina Faller[1][0000−0003−1502−8463], Maurice Kraus[1][0000−0002−6411−3325], Tristan Wirth[1][0000−0002−2445−9081], Max von Buelow[1][0000−0002−0036−319X], Arjan Kuijper[2][0000−0002−6413−0061], and Dieter W. Fellner[1,2,3][0000−0001−7756−0901]

[1] Technical University of Darmstadt, Darmstadt, Germany
[2] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[3] CGV Institute, Graz University of Technology, Graz, Austria

**Abstract.** Transparency detection is a hard problem, as suggested by animals and humans flying or running into glass. However, humans seem to be able to learn and improve on the task with experience, begging the question, whether computers are able to do so too. Making a computer learn and understand transparency would be beneficial for moving agents, such as robots or autonomous vehicles. Our contributions are threefold: First, we conducted a perception study to obtain insights about human transparency detection methods, when borders of transparent objects are not visible. Second, based on our study insights we created a novel synthetic dataset called *DISTOPIA*, which focuses on the warping properties of transparent objects, placed in a variety of natural scenes and contains over 140.000 high resolution images. Third, we modified and trained a deep neural network classification model with an attention module to detect transparency through warping. Our results show that a neural network trained on synthetic data depicting only distortion effects can solve the transparency detection problem and surpasses human performance.

**Keywords:** Perception · Computer vision · Artificial intelligence · Scene understanding

## 1 Introduction

Transparency is a quite common phenomenon in modern day society. While it can be observed in abundance in urban scenes in the form of glass panes on buildings or cars, it is still not an easy task to recognize transparency for animals and humans. Notable examples are birds flying into windows, insects not being able to find their way back out through window gaps, dogs running into garden doors and even humans bumping into well cleaned shop fronts or

doors. While this phenomenon is most probably due to the combination of different perceptual foci, the lack of transparency recognition seems to play a major role. A transparent material has distinguishable properties due to optical effects like distortion, light absorption or reflectance. While physical transparency can be measured with the appropriate tools and environments (objective measurements), perceived transparency is a vision task that requires information parsing from image input (perception). How this perception task works, is however still an open field in perceptual psychology. Understanding the human capability to detect transparency is not only interesting for psychological purposes, but also for bionic transfer applications. The recognition of transparency by machines and therefore the avoidance of e.g. closed glass doors, or the detection and mitigation of transparency related effects on general vision tasks become more important as moving agents, such robots and autonomous vehicles, become more prevalent. Although, x-junction (contour) detection is a viable cue for transparency detection, we deliberately exclude them from our data, arguing that contours are not a main effect of transparency. Most objects have perceivable contours, but there are use-cases without them, such as looking through a windows or safety glasses. This work provides new insights on the influence of two visual cues derived from transparent objects regarding perception, namely reflection and distortion. Furthermore, the impact of two different distortion objects, one resembling a glass pane, the other a lens, as well as the significance of urban and natural background scenes on human transparency detection were evaluated.
Furthermore, we introduce a large synthetic dataset using ray tracing. Due to our new insights about human perception we chose an image generation process, which enables us to extract varying degrees of the distortion properties of transparency and forgo reflectance. This is of special interest, as reflectance requires complex scene understanding and overlaps strongly with mirroring effects.
Finally we show that an ANN, trained on our dataset, is able to distinguish between transparency and transparency-free scenes. We evaluate our networks for the different scene and transparent object categories, as well as the necessary degree of distortion and elaborate our findings.
In summary our contributions are:

- A perception experiment showing that reflection and refraction play a crucial role in perceiving transparency setting a baseline for human transparency detection capabilities without x-junctions (contours).
- A novel dataset "Distopia" for transparency detection via distortion, depicting a wide variety of scenes and levels of distortion, with more than 140.000 images with a high resolution of more than 5 Megapixels ($2252 \times 2252$ pixels).
- We deploy our dataset to synthetically train an ANN and show that distortion is a sufficient cue to recognize transparancy in real world images. We achieve high classification accuracy of up to 83%, and investigate the performance difference of "Urban" versus "Nature" scenes, as well as difference between our two types of transparency and outperform the human baseline we established in our study.

## 2 Related Work

Physically, transparency is an optical material property that describes the ability to transmit electromagnetic waves depending upon its absorption and refraction [10]. Furthermore, the shape and background of a transparent object can yield additional degrees of visible distortions.
The question of how perceptual transparency works has been a topic of scientific research for over 150 years [1, 2, 19]. The most prominent findings in this area of research are X-junctions, luminance relations, and T-junctions, which occur when a line in the background passes behind a transparent object, undergoing a reduction in contrast [3, 4, 6]. Adelson and Anandan [5] have further developed ordinal relations of transparency in regard to junctions. These junction phenomenons can be easily reproduced even in the absence of physical transparency and have been one of the most common techniques for depicting transparency in art from ancient Egyptian times to the present day [16]. However, those techniques often exploit human perception capabilities. Additional to the research on junctions and luminance relations, there has been research on the constraints of superimposed textures giving rise to perceptual transparency [8] and on color constraints occuring at an objects contours for transparency [9, 11, 12].
Bex [14] showed a high dependency of the sensitivity to distortion on the local image structure and suggested that the detection of distortion is based on a higher-level representation of the image structure. Schlüter and Faul [28] discuss that many shape-related properties of opaque objects cannot be simply transferred to transparent objects, because visual features become less perceivable. Furthermore, they showed that human subjects used distortion and specular reflection cues to compare similar transparent objects, with specular reflections being the dominantly used feature [22].
While stereopsis and motion can play a benefitial role for transparency detection, they are out of scope for our work [24, 7]. This decision is based on the more constrained capturing setup for possible detection applications.
While computer vision-based transparency detection methods already exist, they focus on different aspects or data properties. Most algorithms use additional image information to complement rgb, including known object shapes [15, 20, 25], depth channels [31, 18, 36], structured light [23, 26], light-fields [21, 29] or polarization cues [13, 38]. Furthermore, there are techniques that are exclusively based on rgb input, but regard general image depth estimation with included transparent objects [39] or transparency segmentation [34, 30, 32, 41]. However, the networks of those methods are able to utilize transparency contours and reflection properties, which are excluded from our approach. One big challenge is, that only three public datasets which include transparency exist until now: Trans10k [32], ClearGrasp [31] and TRANSCG [37]. While they are all suitable for their intended purpose, no dataset we know of is applicable for our scenario, because they include objects with contours and do not isolate distortion.

## 3 Transparency Perception User Study

To detect transparency without border cues, we chose to investigate the impact of two transparency cues: specular reflection and distortion. To gather insights for machine vision solutions and to establish a baseline, we conducted a perceptual experiment. For this purpose, stimuli were created, evaluated with a small group of subjects and re-adjusted before the main experiment.

### 3.1 Experiment Design

We used a *yes-no task* setup, a standard method in psychological research to measure a subject's sensitivity to some particular sensory input. This design was selected because it is not prone to subjective scaling biases and provokes spontaneous reactions. The sensory input consisted of visual stimuli, showing a natural or urban scenes with a transparent object in front (signal) or without one. The intensity of each feature varied in steps from barely visible (index of refraction further written IOR of 1.1, reflectiveness 2 %) to clearly visible (IOR 1.8, reflectiveness 16 %). To eliminate the influence of X-junctions (borders), object contours were excluded, by showing the test subjects only the inner part of the image. Subsequently, the stimuli were presented to subjects, who were asked to decide whether they perceive transparency.
The experiment was designed using PsychoPy and conducted online with Pavlovia [42]. All four intensity values of the two cues were shown individually paired with both objects in front of all chosen background scenes. That leads to, $4 \cdot 2 \cdot 2 \cdot 5 = 80$ different stimuli in addition to images without transparency stimuli (*no object*). A proportional usage of images with and without stimuli would have lead to an unacceptable experiment duration, which would exceed the concentration capabilities of the test subjects. Therefore, only 20 % of the stimuli did not contain a transparent object.

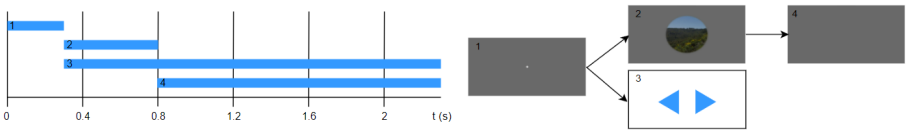


Fig. 1: Timeline of one trial: first, a white cross (1) was shown in the middle of the screen for 0.3 s, to prevent eye fixation and delayed responses. Then the stimulus (2) appeared for 0.5 s in total. After 0.5 s, the stimulus disappeared, and the gray window (4) remained empty for the last 1.5 s. Simultaneously with the stimulus, the subjects could react by pressing either the right arrow key for "Yes, there was an object." or the left arrow key for "No, there was no object." (3). A reaction immediately ended the trial, even if the stimulus has not been shown for the entire 0.5 s yet. If the subjects did not react, the trial was ended after 2.3 s, and the subsequent trial began.

Each of the 80 stimuli showing a transparent object were presented to subjects 21 times, while all five stimuli not containing an object were shown 84 times. This resulted in a total of 2100 trials, each lasting up to 2.3 s (see fig. 1). Those 2100 trials per subject were split into seven blocks of equal size, giving the subjects the opportunity for a short break in between.
Each trial was conducted as described in fig. 1. Stimuli were only shown for a short period of time to prevent subjects from consciously searching for abnormalities. This setup ensures that the reaction was as intuitive as possible.
Before conducting the actual experiment, an iterative preliminary experiment was performed to ensure a suitable process. The main study was conducted on 20 subjects consisted of eleven males and nine females with an average age of 25.6 (with a standard deviation of 7.4). Due to technical issues, the data showing the circular object in the *street* scene was excluded from the analysis.

### 3.2 Stimuli and Data Generation

In order to vary distortion and specular reflections of transparent objects separately, we generated stimuli with synthetic transparent objects. To make the experiment realistic, the objects were presented in different real-world scenes, represented as panorama images. Two of them were located in nature, and three in an urban environment. We used the Poly-Haven dataset, which contains panoramas as high dynamic range images [43].
When selecting the nature panoramas, the focus was to exhibit as few unnatural objects as possible, resulting in two different locations: *Versveldpas* (fig. 2a) and *Desert* (fig. 2c). Two different day times were deliberately selected, especially to get different lighting conditions and thus different specular reflections.
For the urban scenes, it was important that they were as diverse as possible and close to what we encounter in our everyday life. We chose the following three: *Street* (fig. 2f), *Quattro Canti* (fig. 2d) and *Hamburg* (fig. 2e), as they represent different common municipal scenes. We used a ray tracer to generate the synthetic stimuli. The camera faces directly onto the transparent objects, while the panorama is placed as a texture around the scene. We decided to use two sophisticated, but common transparent objects: The first one being a quadratic slice, resembling a window (see fig. 2a) that entails irregularities similar to the deformation of small waves in liquids with low viscosity, such as in water or older window glasses. The circular object (see fig. 2b) is lens-shaped and shows the most deformation near the boundary regions of the object model.
Both objects have the same and constant material properties, except for their IOR and reflection ratio parameters. We decided to use a circular section of the stimuli, as seen in fig. 1. This resulted in all outer points of the stimuli being equidistant from the center, thus providing similar times for saccadic eye movements. The background color on which the stimuli were presented was a neutral gray, to avoid unnecessary eye strain and avoid high contrasts. The outer stimuli boundary was blurred to de-emphasize borderlines.

(a) Versveld. (b) Versveld. (c) Desert (d) 4 Canti (e) Hamburg (f) Street

Fig. 2: Scenes used in our experiments. (a) and (c) shows natural scenes, while (d), (e) and (f) show scenes from urban settings. (a) and (b) additionally show our two transparent object types.

### 3.3 Results

**Data and Analysis** We calculated the mean correct answer ratios of all 20 participants, which is illustrated in table 1. The data shows that images without transparent objects were recognized with a mean correct answer ratio of 92 %. Furthermore, there is a difference in detectability between the stimuli in natural and urban scenes, as well as for our different transparent object types. This effect is especially visible in fig. 3, which combines results from individual transparent object types and scenes. It shows that urban scenes with the squared object distortion cue leads to a significant increase in transparency perception.
For specular reflection, no apparent circumstantial differences can be perceived. There is one clear outlier, which is the quadratic object in *Quattro Canti*. This can be explained by the position of the light, resulting in few specular reflection stimuli. Fig. 3 also reflects this behavior as the perceivability does differ marginally between transparent object types and scenes.

**Transparency Perception through Distortion** We evaluated the effect of different levels of distortion on the perceivability of transparent objects. A Levene tests shows that the variances from our experiment are inhomogeneous. A Shapiro-Wilk test shows that our data is not normally distributed, which, however, can be neglected according to the central limit theorem as the number of samples is greater than 20 in our experiment. A robust Welch test with a resulting p-value below 0.001 shows that there are significant differences between the mean values of the four groups. Given this observation, we further investigated how much these groups differ from each other using a Games-Howell test. This test reveals a significant difference between an IOR of 1.1 and the remaining distortion configurations. Additionally, it showed a significant difference between an IOR of 1.33 and 1.8, but no significant difference between 1.33 and 1.52 and IORs of 1.52 and 1.8, showing that increasing distortions are beneficial for detection, but the distance between distortion values might be to close.
Similar to this analysis, we examined whether our different transparent object types influence the perceivability of transparency. Therefore, we also performed a Welch test ($p < 0.001$), indicating a strong influence on perception. Analogous, the scene type influences the perceivability of transparency ($p < 0.001$).

Table 1: Mean correct answer ratios of all 20 subjects for each stimulus. Values close to 50 % indicate mere guessing. 0 % represents no correct prediction and 100 % means only correct prediction. × denotes images without transparent objects. The circles and Squares denote the type of transparent object.

| | | Distortion (IOR) | | | | Specularity | | | |
|---|---|---|---|---|---|---|---|---|---|
| | × | 1.10 | 1.33 | 1.52 | 1.80 | 2 % | 4 % | 8 % | 16 % |
| Hamb. □ | 92 | 16 | 78 | 85 | 89 | 10 | 43 | 90 | 94 |
| Hamb. ○ | 92 | 10 | 17 | 18 | 23 | 9 | 7 | 28 | 78 |
| 4 Canti □ | 93 | 9 | 53 | 76 | 83 | 7 | 7 | 6 | 9 |
| 4 Canti ○ | 93 | 8 | 11 | 17 | 21 | 7 | 12 | 90 | 96 |
| Street □ | 86 | 18 | 67 | 84 | 90 | 26 | 22 | 79 | 95 |
| Vers. □ | 94 | 8 | 10 | 18 | 34 | 6 | 7 | 51 | 89 |
| Vers. ○ | 94 | 7 | 5 | 6 | 6 | 7 | 8 | 25 | 74 |
| Desert □ | 94 | 6 | 11 | 18 | 36 | 8 | 16 | 77 | 92 |
| Desert ○ | 94 | 10 | 7 | 9 | 10 | 9 | 22 | 77 | 93 |
| $\mu$ | 92 | 10 | 29 | 37 | 44 | 10 | 16 | 58 | 80 |

We can show that the presence of straight line segments in the stimuli, that differs strongly between urban and nature scenes, explains part of the difference between transparency perceivability. Therefor, we calculate the ratio of pixels, that lie on straight line segments according to the deep learning detector *M-LSD* [33, 17], for all distortion levels in each scene. The resulting ratios are depicted in table 2, which correlate with our perception results.
While the perception of transparency is weaker in natural scenes, it might be overcome by machine vision, as humans supposedly do not have an adequate training level for transparent objects in natural environments. Furthermore this cue seems viable, as distortion effects are characteristic and seldom false flags.

Table 2: Normalized differences between the lines in original and the distorted images. The circles and Squares denote the type of transparent object. Natural scenes are at 0, as there are no lines to be lost due to distortion.

| Distortion | Hamburg | | 4 Canti | | Street | Versveldpas | | Desert | |
|---|---|---|---|---|---|---|---|---|---|
| (IOR) | □ | ○ | □ | ○ | □ | □ | ○ | □ | ○ |
| 1.10 | 0.18 | 0.04 | 0.22 | 0.11 | 0.09 | 0 | 0 | 0.01 | 0.01 |
| 1.33 | 0.3 | 0.14 | 0.63 | 0.32 | 0.48 | 0 | 0 | 0.01 | 0.01 |
| 1.52 | 0.35 | 0.07 | 0.85 | 0.25 | 0.51 | 0 | 0 | 0.01 | 0.01 |
| 1.80 | 0.42 | 0.13 | 1 | 0.2 | 0.55 | 0 | 0 | 0.01 | 0.01 |

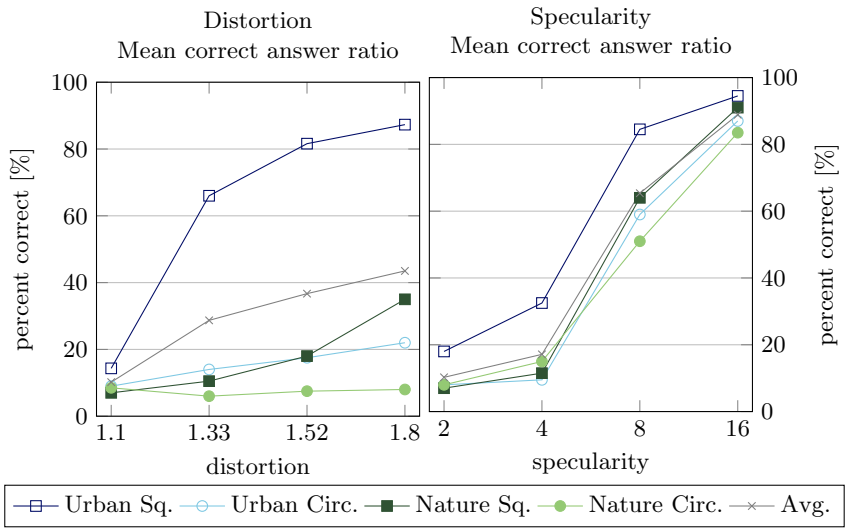


Fig. 3: Mean correct answer ratio of distortion and reflection values grouped by scene and the transparent object type.

**Transparency Perception through Specular Reflections** Similar to the distortion, fig. 3 shows the mean ratios of correct answers for the specular reflection stimuli. A significant Levene test indicates data inhomogeneity. Furthermore, a Welch test with p-value below 0.001 indicates that the results differ significantly between the different groups. In contrast to our previous analysis on different distortion levels, the succeeding Games-Howell test revealed significant differences between all selected reflection levels.
It can be concluded that the levels of specular reflection significantly affect the perceiveability of transparent objects. However, while the results indicated that humans are able work with this cue, specular reflection can also occur due to metallic properties and might not be a strong indicator for transparency. Furthermore, specular reflections require specific light setups to be perceivable.

**Object and scene influence** A comparison of the influence of the chosen transparent object on the participants' ability to perceive the transparent objects showed no significance. Due to an insignificant Levene test, we conducted a one-way analysis of variance, which revealed an insignificant p-value of 0.595. For a comparison of the perceptual influence of the different scenes, we performed a Welch test, which indicated insignificant differences. The marginal effects of specularity changes can be taken from fig. 3.

## 4 Artificial Transparency Detection Method

The insights from our perception study lead us to the conclusion, that distortion is a viable cue for transparency detection, due to its characteristic and rare property. We discard specular reflection and reflection as a whole, due to its dependency on correct lighting and notion to be exhibited by various non transparent objects. To evaluate our idea, we generated a large synthetic dataset, shown in section 4.1 and we evaluate how well a neural network can detect transparency when trained only on distortion in Section 4.2.

### 4.1 *DISTOPIA* Dataset

Following the design choices from our perception study, we synthetically create a large dataset, which only depicts distortion cues. To obtain a broader background base, we decided on a total of 315 panorama scenes (170 urban and 145 nature). To eliminate unintended background distortions we disregarded any scene containing substantial amounts of inherent transparent structures. Furthermore we filtered out natural scenes with large urban influences and vice versa. We kept the window structure and lens object, to have a direct comparison to our study. For rendering we again adopt a ray tracer (Blender's Cycles). We place the camera within a HDRI background sphere and orient it towards our transparent objects. Next we rotate the camera in 5° steps on a horizontal axis and generate one image without transparent object and up to three images per object for each step. Different to our study, the IOR was chosen randomly for each generated image, and sampled from an uniform distribution. This results in a fairer learning set, as humans are also trained on a variety of IOR values. We repeated this procedure for all 315 scenes and two objects, resulting in over 140 000 high resolution images. To test for generalization to real-world photographs, we collect a small set of photos and report results on it alongside our synthetic data.

Table 3: Number of images within our training, validation, and test datasets. Training, Validation, and Testing splits depict disjunct sets of scenes. We collect a small set of photographs to evaluate the generalization to real world images.

| | **Both (A)** | **Circle (B)** | **Rect (C)** | **RectC (D)** | **Real** |
|---|---|---|---|---|---|
| **Training** | 250.426 | 125.192 | 125.218 | 876.514 | - |
| **Validation** | 5.539 | 3.075 | 3.079 | 34.302 | 363 |
| **Testing** | 5.850 | 3.250 | 3.250 | 96.601 | - |

### 4.2 Classification

To show that distortion is a sufficient cue for transparency detection we use the common CNN-based architecture ResNet, adopting the parameterization found

in state-of-the-art GAN literature [35]. We provide each network block with a (bilinearly) down-sampled version of the current tensor, and a residual connection. Our network predicts transparency presence, transparency type (if any), distortion levels. Also, for regularization reasons, we predict whether an image shows a *Nature* or *Urban* scene. We use (binary) cross entropy as loss function for the first three and L1-loss for the distortion strength prediction.
In addition we argue that features relevant for distortion detection without visible corners and edges are located in the earlier layers of the feature extractor. Therefore we add bottleneck attention modules [27] after the first ResNet block. Combining attention with a bottleneck should allow the network to focus on relevant cues and discard the rest. Multiple blocks allow each to attend to different features. Also we use different non-linearities, that is average pooling for half of the blocks and max pooling for the other. We provide a detailed description of the architecture in the appendix for easy reproducibility.
We train ResNet networks for four different splits of our dataset. Next we evaluate our new network architecture (with attention) to the two best performing experiments and compare their performance on real data. The experiments A, B, C, and D are as follows: First, we train a classifier to differentiate between scenes containing rectangular transparent objects, circular transparent objects and scenes containing no transparencies (experiment A, *Both*), second, we train another network to distinguish circular transparencies from images with no transparency (B, *Circle*), a third to classify rectangular with only central crops (C, *Rect*), and finally a fourth for rectangular with additional random crops (D, *RectC*). We central crop the image, to remove the transparency's edges. Not removing the edges would provide exploitable features for the classifier and most likely reduce its generalization capabilities. For experiment D (rectangular with random cropping, before scaling to $512 \times 512$), after applying the central crop, we create an additional 7 random crops (after first removing the edge via central crop) to get different image details. For each crop in all experiments we also include the same image portion without transparencies, giving us a total of 876.513 images. For task A we include circular transparency, for which the circular distortion is an important cue. Therefore we restrict our data processing to only central cropping whenever circles are included. we end up with a total of 250.425 images for experiment A (*Both*). Finally, we want to see the individual performance when training a classifier each to just recognize one kind of transparency (B: circles only, and C: rectangles only). They have 125.192 and 125.218 images for training. The exact number of training, validation, and test images are to be found in table 3. In addition we validate on a set of photographs to evaluate the generalization of our distortion-only training to real world images.
We use conventional image augmentation during training, such as flipping along y, rotating, re-scaling, adjusting brightness, contrast, hue, and saturation. Our experiments showed best performance with small batch sizes and we ended up using 16 images per batch, together with small learning rates between 0.0001 to 0.000 25. As for our attention modules we varied their number and tried as few as 1 and up to 8, and report our results with 8. We adopt an early stopping

strategy, showing our classifiers 50 million images, testing every 0.5 million views against our validation set, and using the best checkpoint on our test data. Our real data is a first portion of our ongoing effort to record a real-world dataset we plan to publish together with our synthetic data.

### 4.3 Evaluation

**Ambiguous Scenes** The quantitative results of our experiments can be seen in table 4. We provide a prediction accuracy for each possible individual subgroup derived from our data splits, concerning the background and object. Furthermore, we chose to show different buckets depending on the IOR each object had and additional buckets for all images without transparency (*None*), as well as all images with transparency (*All*). The reasoning is, that different IORs result in differently hard to detect warping effects. This can be observed as a general trend in our data. While there are maxima in each row (bold), that are not in the $\lfloor 1.7 \rfloor$ category, the prediction differences between those two values are only up to 0.05. Furthermore, the $\lfloor 1.1 \rfloor$ bucket always contains the lowest predictions with a large gap to the $\lfloor 1.2 \rfloor$ and following buckets, suggesting that this IOR is comparatively hard to predict. The emphasized values indicate the highest prediction value over all buckets, including the *None* category. These overlap with the best bucket predictions for the Circle object (B). In all other cases, the None prediction outperform the warping predictions. Especially the Both object (A) category shows large differences of up to 0.54 between the best bucket and None prediction. These findings suggest, that it might be easier to learn the detection of no transparency over learning two different object warping effects for (A). Our findings from the perception experiment, that different background scenery types might influence predictions, did not directly translate to our data.
While the background scene type shows no particular effect on the overall prediction accuracy, the object shape seems to be of high relevance for *All* predictions. The Rect object (C) allows for an accuracy of up to 83 %, with the cropped version dropping to 75 %. The Circle object (B) however, reaches only up to 59 %. This seems to be reflected in the Both (A) set, where only about 62 % of images are classified correctly. Furthermore, all sets containing circles exhibit either strong prediction accuracies in the bucket categories, or the *None* category. This behaviour is not observed in the Rect sets, where evenly high accuracies are reached over all categories. In total, this leads to the conclusion, that object shapes and therefore different warping characteristics play a major role in detecting transparency for our setup.

**Real Data** To check for generalization to real data we validate our checkpoints against our set of photographs during training. We visualize our results in fig. 4. Looking at the results, first, we note strong generalization early on. Second, we see clear outperformance of our model compared to the ResNet version especially for longer training times, as well as a smoothing out of the curve. Finally, when

comparing both of our models between experiments, we see that experiment C (without crops) comes close to experiment D (with crops) after around 17 to 18 million views, stays in a similar range till around 27 to 29 million views and than drops to a similar level as the ResNet version. We argue that this is because of overfitting, since experiment D has around seven times the amount of images and much more variation. All together the drop in performance between our synthetic and our real data is around 15 %.

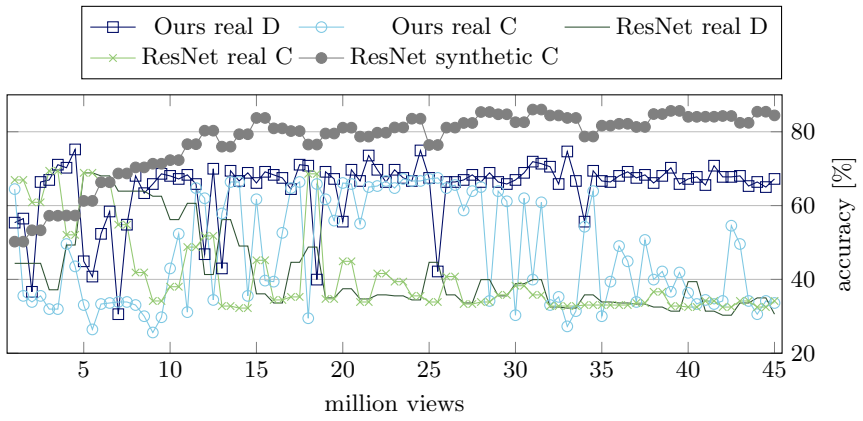


Fig. 4: Accuracy on photographs with ResNet architecture against ours with increasing number of training images (1 to 45 million) shown on experiments C and D. For comparison we include the results of experiment C on synthetic data.

**Failure Cases** We present failure cases of our classifiers, focusing on images with strong distortions, classified incorrectly as depicting no transparency, as well as images with no transparency, classified as containing transparency. Also, we show some interesting effects, for instance, although filtering the scenes beforehand, we missed some nature scenes containing man-made structures, which led to crops containing predominantly urban scenery being classified in the training data as *Nature*. Interestingly, the classifiers predicted some of those images correctly on the test set (that is as *Urban*). In addition, we have some scenes containing transparencies, such as windows. Since we only labeled images as containing transparencies when we ourselves placed some in the rendering, these were also inconsistently labeled. Again, the classifiers predicted some of those confusing samples correctly (containing transparency).
Fig. 5 shows incorrect classifications for circles, rectangles with and rectangles without random cropping. We notice particular problems with skies and landscapes without much detail (such as snowscapes, fields, and wood), as well as with views showing nothing but forest. We also note that strong backlighting

Table 4: For each of our experiments A (rectangular and circular transparency), B (circle only), C (rectangular only), and D (rectangular with random cropping) we report quantified results for both scenes as well as their combination. We subdivide the range of possible refraction parameter choices into buckets $\lfloor\delta\rfloor = [\delta, \delta + 0.1)$. The table also includes performance on images containing no transparency ($\times$) and over all buckets ($\forall$). The emphasized value in each row denotes the highest prediction value in a row and bold values exclude $\times$.

| **Data** | | **Index of Refraction (IOR)** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Object** | **Scene** | $\times$ | $\lfloor$**1.1**$\rfloor$ | $\lfloor$**1.2**$\rfloor$ | $\lfloor$**1.3**$\rfloor$ | $\lfloor$**1.4**$\rfloor$ | $\lfloor$**1.5**$\rfloor$ | $\lfloor$**1.6**$\rfloor$ | $\lfloor$**1.7**$\rfloor$ | $\forall$ |
| **Both (A)** | **Both** | *0.83* | 0.26 | 0.37 | 0.36 | 0.39 | 0.44 | **0.45** | 0.44 | 0.61 |
| | **Urban** | *0.92* | 0.23 | 0.26 | 0.27 | 0.3 | 0.36 | **0.42** | 0.38 | 0.62 |
| | **Nature** | *0.73* | 0.3 | 0.48 | 0.48 | 0.48 | **0.53** | 0.48 | 0.51 | 0.6 |
| **Circle (B)** | **Both** | 0.19 | 0.87 | 0.93 | 0.94 | 0.93 | 0.93 | 0.96 | ***0.96*** | 0.56 |
| | **Urban** | 0.15 | 0.86 | 0.91 | 0.92 | 0.91 | 0.9 | 0.93 | ***0.94*** | 0.53 |
| | **Nature** | 0.23 | 0.88 | 0.96 | 0.96 | 0.96 | 0.96 | 0.98 | ***0.98*** | 0.59 |
| **Rect (C)** | **Both** | *0.97* | 0.45 | 0.61 | 0.65 | 0.68 | 0.76 | **0.79** | 0.76 | 0.82 |
| | **Urban** | *0.97* | 0.47 | 0.64 | 0.66 | 0.65 | 0.76 | **0.85** | 0.8 | 0.83 |
| | **Nature** | *0.97* | 0.43 | 0.59 | 0.64 | 0.72 | **0.76** | 0.72 | 0.72 | 0.81 |
| **RectC (D)** | **Both** | *0.98* | 0.43 | 0.49 | 0.53 | 0.57 | 0.62 | 0.63 | **0.63** | 0.75 |
| | **Urban** | *0.99* | 0.41 | 0.49 | 0.54 | 0.57 | 0.62 | **0.64** | 0.62 | 0.73 |
| | **Nature** | *0.97* | 0.44 | 0.49 | 0.52 | 0.57 | 0.62 | 0.62 | **0.65** | 0.76 |

(sun, human-made lights) leads to errors. Finally, we include a sample of circle-classifier failures on an image depicting vignetting-like characteristics (top right), incorrectly classifying the image as containing a circular transparency. Although we handpicked our scenes to have a clean split between scenes depicting *Urban* structures and those, showing *Nature* such as forests and plains, some samples slipped through our filtering. This becomes especially apparent when looking at our use case D (rectangles with cropping) in which a small man-made structure (such as a house) in the background, can become the central element of the image. An interesting effect we witnessed concerns transparencies within the background image, as can be seen in fig. 5: Our classifier (correctly) predicts the presence of the transparancy, even though our ground truth label says otherwise. To combat this problem, one needs to filter even more rigorously for use case D.

## 5 Conclusion and Future Work

We motivated the problem of transparency detection for e.g. intelligent moving agents. In order to investigate whether specular reflections and distortions are possible cues for transparency detection and to establish a baseline, we designed and carried out a perception experiment. Our subsequent analysis shows that the two cues have a significant positive effect on the perceivability of transparency. While specular reflection is the more robust cue for humans, it is also strongly dependent on scene understanding, lighting conditions and overlaps with reflection properties. Therefore, in our second experiment we investigate how well

(a) Building (*Nature* scene) (b) Tree (*Nature* scene) (c) Brick wall (*Urban* scene)

Fig. 5: Although we checked by hand, some scenes of our dataset contain some ambiguity, which makes it interesting to see, what our classifiers do with such samples. For instance, the first image (a) shows a building within a *Nature* scene, blurring the line between our two splits of *Nature* versus *Urban* scenes. This becomes a problem when, during our random cropping, the house becomes the main subject of the resulting image. Similarly, the second image (b) shows a tree (*Nature*) in front of a brick wall (*Urban*). The last image (c) shows strong transparencies in the background (glass).

computers can detect transparency when trained on distortion only. We created *DISTOPIA*, our novel dataset for distortion-only-based transparency detection containing more than 140 000 high-resolution images. This dataset will be released, to allow the reproducibility of our findings and further research by the community. We then evaluated the performance of a convolutional neural network on our dataset and introduce a modification to better suit our applications and improve results on real data. We achieve high performance, both on our synthetic test set, as well as on our real photographs. This leads us to conclude, that distortion is also a viable cue to achieve high classification accuracy for computer-based vision models, that surpasses human transparency perception.
In the future, we want to expand our dataset by including more kinds of transparent objects, especially regarding their overall shape and material properties (e.g. acrylic glass and toned glass) and create a bigger and more diverse real world dataset for testing. Also, after demonstrating the validity of using distortion-only for transparency detection via a CNN, we want to adopt additional network architecture such as the vision transformer and new developments such as FocalNets [40] to improve upon our results.

## Acknowledgements

Part of the research in this paper was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 407 714 161. We thank Frank Jäkel for his generous support and the anonymous reviewers whose comments helped improve this manuscript.

## References


[1] von Helmholtz Hermann. *Allgemeine Encyklopädie der Physik / 9 Handbuch der physiologischen Optik*. Vol. 9. Leopold Voss, 1867. DOI: 10.3931/E-RARA-21259.

[2] Beatrix Tudor-Hart. "Beiträge zur Psychologie der Gestalt. XVI. Studies in transparency, form and colour". In: *Psychologische Forschung* 10.1 (Jan. 1928), pp. 255–298. DOI: 10.1007/bf00492012.

[3] Fabio Metelli. "An Algebraic Development of the Theory of Perceptual Transparency". In: *Ergonomics* 13.1 (Jan. 1970), pp. 59–66. DOI: 10.1080/00140137008931118.

[4] Fabio Metelli. "The Perception of Transparency". In: *Scientific American* 230.4 (Apr. 1974), pp. 90–98. DOI: 10.1038/scientificamerican0474-90.

[5] Edward H Adelson and Padmanabhan Anandan. *Ordinal characteristics of transparency*. 1990.

[6] Takeo Watanabe and Patrick Cavanagh. "Transparent surfaces defined by implicit X junctions". In: *Vision Research* 33.16 (Nov. 1993), pp. 2339–2346. DOI: 10.1016/0042-6989(93)90111-9.

[7] Barton L. Anderson and Bela Julesz. "A theoretical analysis of illusory contour formation in stereopsis." In: *Psychological Review* 102.4 (Oct. 1995), pp. 705–743. DOI: 10.1037/0033-295x.102.4.705.

[8] Takeo Watanabe and Patrick Cavanagh. "Texture Laciness: The Texture Equivalent of Transparency?" In: *Perception* 25.3 (Mar. 1996), pp. 293–303. DOI: 10.1068/p250293.

[9] Michael D'Zmura et al. "Color Transparency". In: *Perception* 26.4 (Apr. 1997), pp. 471–492. DOI: 10.1068/p260471.

[10] M Tavel. "What determines whether a substance is transparent? For instance, why is silicon transparent when it is glass but not when it is sand or a computer chip". In: *Scientific American* (1999). URL: https://www.scientificamerican.com/article/what-determines-whether-a/.

[11] Franz Faul and Vebjørn Ekroll. "Psychophysical model of chromatic perceptual transparency based on substractive color mixture". In: *Journal of the Optical Society of America A* 19.6 (June 1, 2002), p. 1084. DOI: 10.1364/josaa.19.001084.

[12] Jacqueline M. Fulvio, Manish Singh, and Laurence T. Maloney. "Combining achromatic and chromatic cues to transparency". In: *Journal of Vision* 6.8 (July 7, 2006), p. 1. DOI: 10.1167/6.8.1.

[13] Vimal Thilak, David G. Voelz, and Charles D. Creusere. "Polarization-based index of refraction and reflection angle estimation for remote sensing applications". In: *Applied Optics* 46.30 (Oct. 18, 2007), p. 7527. DOI: 10.1364/ao.46.007527.

[14] Peter J. Bex. "(In) Sensitivity to spatial distortion in natural scenes". In: *Journal of Vision* 10.2 (2010), pp. 1–15. DOI: 10.1167/10.2.23.

[15] Ulrich Klank, Daniel Carton, and Michael Beetz. "Transparent object detection and reconstruction on a mobile platform". In: *2011 IEEE Interna-*

*tional Conference on Robotics and Automation* (Shanghai, China, May 9–13, 2011). IEEE. IEEE, May 2011, pp. 5971–5978. DOI: 10.1109/icra.2011.5979793.

[16] Bilge Sayim and Patrick Cavanagh. "The Art of Transparency". In: *i-Perception* 2.7 (Jan. 1, 2011), pp. 679–696. DOI: 10.1068/i0459aap.

[17] Rafael Grompone von Gioi et al. "LSD: a Line Segment Detector". In: *Image Processing On Line* 2 (Mar. 24, 2012), pp. 35–55. DOI: 10.5201/ipol.2012.gjmr-lsd.

[18] Nicolas Alt, Patrick Rives, and Eckehard Steinbach. "Reconstruction of transparent objects in unstructured scenes with a depth camera". In: *2013 IEEE International Conference on Image Processing* (Melbourne, Australia, Sept. 15–18, 2013). IEEE. IEEE, Sept. 2013, pp. 4131–4135. DOI: 10.1109/icip.2013.6738851.

[19] K Koffka. *Principles Of Gestalt Psychology*. Routledge, Oct. 8, 2013. DOI: 10.4324/9781315009292.

[20] Ilya Lysenkov and Vincent Rabaud. "Pose estimation of rigid transparent objects in transparent clutter". In: *2013 IEEE International Conference on Robotics and Automation* (Karlsruhe, Germany, May 6–10, 2013). IEEE. IEEE, May 2013, pp. 162–169. DOI: 10.1109/icra.2013.6630571.

[21] Kazuki Maeno et al. "Light Field Distortion Feature for Transparent Object Recognition". In: *2013 IEEE Conference on Computer Vision and Pattern Recognition* (Portland, OR, USA, June 23–28, 2013). IEEE, June 2013, pp. 2786–2793. DOI: 10.1109/cvpr.2013.359.

[22] Nick Schlüter and Franz Faul. "Are optical distortions used as a cue for material properties of thick transparent objects?" In: *Journal of Vision* 14.14 (2014), pp. 2–2. DOI: doi.org/10.1167/14.14.2.

[23] Kai Han, Kwan-Yee K. Wong, and Miaomiao Liu. "A fixed viewpoint approach for dense reconstruction of transparent objects". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Boston, MA, USA, June 7–12, 2015). IEEE, June 2015, pp. 4001–4008. DOI: 10.1109/cvpr.2015.7299026.

[24] Takahiro Kawabe, Kazushi Maruya, and Shin'ya Nishida. "Perceptual transparency from image deformation". In: *Proceedings of the National Academy of Sciences* 112.33 (Aug. 3, 2015). DOI: 10.1073/pnas.1500913112.

[25] Cody J. Phillips, Matthieu Lecce, and Kostas Daniilidis. "Seeing Glassware: from Edge Detection to Pose Estimation and Shape Recovery". In: *Robotics: Science and Systems XII*. Vol. 3. Michigan, USA. Robotics: Science and Systems Foundation, 2016, p. 3. DOI: 10.15607/rss.2016.xii.021.

[26] Yiming Qian, Minglun Gong, and Yee-Hong Yang. "3D Reconstruction of Transparent Objects with Position-Normal Consistency". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV, USA, June 27–30, 2016). IEEE, June 2016, pp. 4369–4377. DOI: 10.1109/cvpr.2016.473.

[27] Jongchan Park et al. "BAM: Bottleneck Attention Module". In: *British Machine Vision Conference*. arXiv, 2018. DOI: 10.48550/ARXIV.1807.06514. Pre-published.

[28] Nick Schlüter and Franz Faul. "Visual shape perception in the case of transparent objects". In: *Journal of Vision* 19.4 (Apr. 22, 2019), p. 24. DOI: 10.1167/19.4.24.

[29] Dorian Tsai et al. "Distinguishing Refracted Features Using Light Field Cameras With Application to Structure From Motion". In: *IEEE Robotics and Automation Letters* 4.2 (Apr. 2019), pp. 177–184. DOI: 10.1109/lra.2018.2884765.

[30] Haiyang Mei et al. "Don't Hit Me! Glass Detection in Real-World Scenes". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, June 13–19, 2020). IEEE, June 2020, pp. 3687–3696. DOI: 10.1109/cvpr42600.2020.00374.

[31] Shreeyak Sajjan et al. "Clear Grasp: 3D Shape Estimation of Transparent Objects for Manipulation". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)* (Paris, France, May 31–Aug. 31, 2020). IEEE. IEEE, May 2020, pp. 3634–3642. DOI: 10.1109/icra40945.2020.9197518.

[32] Enze Xie et al. "Segmenting Transparent Objects in the Wild". In: *Computer Vision – ECCV 2020*. Springer. Cham: Springer International Publishing, 2020, pp. 696–711. DOI: 10.1007/978-3-030-58601-0_41.

[33] Geonmo Gu et al. "Towards Light-weight and Real-time Line Segment Detection". In: (2021). DOI: 10.48550/ARXIV.2106.00186. Pre-published.

[34] Hao He et al. "Enhanced Boundary Learning for Glass-like Object Segmentation". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC, Canada, Oct. 10–17, 2021). IEEE, Oct. 2021, pp. 15859–15868. DOI: 10.1109/iccv48922.2021.01556.

[35] Tero Karras et al. "Alias-Free Generative Adversarial Networks". In: *Proc. NeurIPS*. arXiv, 2021. DOI: 10.48550/ARXIV.2106.12423. Pre-published.

[36] Luyang Zhu et al. "RGB-D Local Implicit Function for Depth Completion of Transparent Objects". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Nashville, TN, USA, June 20–25, 2021). IEEE, June 2021, pp. 4649–4658. DOI: 10.1109/cvpr46437.2021.00462.

[37] Hongjie Fang et al. "TransCG: A Large-Scale Real-World Dataset for Transparent Object Depth Completion and a Grasping Baseline". In: *IEEE Robotics and Automation Letters* 7.3 (July 2022), pp. 7383–7390. DOI: 10.1109/lra.2022.3183256.

[38] Haiyang Mei et al. "Glass Segmentation using Intensity and Spectral Polarization Cues". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA, USA, June 18–24, 2022). IEEE, June 2022, pp. 12622–12631. DOI: 10.1109/cvpr52688.2022.01229.

[39] Tristan Wirth et al. “Fitness of General-Purpose Monocular Depth Estimation Architectures for Transparent Structures”. In: *Eurographics 2022 - Short Papers* (Reims, France). Ed. by Nuria Pelechano and David Vanderhaeghe. The Eurographics Association, 2022, pp. 9–12. DOI: 10.2312/EGS.20221020.

[40] Jianwei Yang et al. “Focal Modulation Networks”. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2022). DOI: 10.48550/ARXIV.2203.11926. Pre-published.

[41] Letian Yu et al. “Progressive Glass Segmentation”. In: *IEEE Transactions on Image Processing* 31 (2022), pp. 2920–2933. DOI: 10.1109/tip.2022.3162709.

[42] *Pavlovia*. URL: https://pavlovia.org/.

[43] *Poly Haven*. URL: https://polyhaven.com/.

## Supplementary Material


Fig. 6: The images show an urban scene from our study. The first row visualizes four degrees of reflection. The percentual number denotes the reflectiveness of a material between the scene and the camera. The green colored mask emphasizes the influence of the reflectiveness on the base image. The second row visualizes four degrees of refraction. The number denotes the IOR and the green lines highlight straight lines. The stronger the refraction, the more lines disappear.

Table 5: Network architecture in detail.

| **Function** | **Parameters** | **Shape** | **Precision** |
|---|---|---|---|
| b512.fromrgb | 256 | 64, 512, 512 | float16 |
| b512.skip | 8192 | 128, 256, 256 | float16 |
| b512.conv0 | 36928 | 64, 512, 512 | float16 |
| b512.conv1 | 73856 | 128, 256, 256 | float16 |
| b512 | - | 128, 256, 256 | float16 |
| (bottleneck attention) | - | 8, 128, 256, 256 | float32 |
| b256.skip | 32768 | 256, 128, 128 | float16 |
| b256.conv0 | 147584 | 128, 256, 256 | float16 |
| b256.conv1 | 295168 | 256, 128, 128 | float16 |
| b256 | - | 256, 128, 128 | float16 |
| b128.skip | 131072 | 512, 64, 64 | float16 |
| b128.conv0 | 590080 | 256, 128, 128 | float16 |
| b128.conv1 | 1180160 | 512, 64, 64 | float16 |
| b128 | - | 512, 64, 64 | float16 |
| b64.skip | 262144 | 512, 32, 32 | float16 |
| b64.conv0 | 2359808 | 512, 64, 64 | float16 |
| b64.conv1 | 2359808 | 512, 32, 32 | float16 |
| b64 | - | 512, 32, 32 | float16 |
| b32.skip | 262144 | 512, 16, 16 | float32 |
| b32.conv0 | 2359808 | 512, 32, 32 | float32 |
| b32.conv1 | 2359808 | 512, 16, 16 | float32 |
| b32 | - | 512, 16, 16 | float32 |
| b16.skip | 262144 | 512, 8, 8 | float32 |
| b16.conv0 | 2359808 | 512, 16, 16 | float32 |
| b16.conv1 | 2359808 | 512, 8, 8 | float32 |
| b16 | - | 512, 8, 8 | float32 |
| b8.skip | 262144 | 512, 4, 4 | float32 |
| b8.conv0 | 2359808 | 512, 8, 8 | float32 |
| b8.conv1 | 2359808 | 512, 4, 4 | float32 |
| b8 | - | 512, 4, 4 | float32 |
| b4.mbstd | - | 513, 4, 4 | float32 |
| b4.conv | 2364416 | 512, 4, 4 | float32 |
| b4.fc | 4194816 | 512 | float32 |
| b4.out | 262656 | 512 | float32 |
| b4.scene_classr | 1026 | 2 | float32 |
| b4.trans_class | 1539 | 3 | float32 |
| b4.trans_presence_class | 1026 | 2 | float32 |
| b4.strength_class | 513 | 1 | float32 |

Table 6: Bottleneck attention architecture in detail.

| Function | Parameters | Shape | Precision |
|---|---|---|---|
| b512.conv1 | 73856 | 128, 256, 256 | float16 |
| b512 | - | 128, 256, 256 | float16 |
| cbams.0.ChannelGate.mlp | 2184 | 128 | float32 |
| cbams.0.ChannelGate | - | 128, 256, 256 | float32 |
| cbams.0.SpatialGate.compress | - | 2, 256, 256 | float32 |
| cbams.0.SpatialGate.spatial | 100 | 1, 256, 256 | float32 |
| cbams.0.SpatialGate | - | 128, 256, 256 | float32 |
| cbams.1.ChannelGate.mlp | 2184 | 128 | float32 |
| cbams.1.ChannelGate | - | 128, 256, 256 | float32 |
| cbams.1.SpatialGate.compress | - | 2, 256, 256 | float32 |
| cbams.1.SpatialGate.spatial | 100 | 1, 256, 256 | float32 |
| cbams.1.SpatialGate | - | 128, 256, 256 | float32 |
| cbams.2.ChannelGate.mlp | 2184 | 128 | float32 |
| cbams.2.ChannelGate | - | 128, 256, 256 | float32 |
| cbams.2.SpatialGate.compress | - | 2, 256, 256 | float32 |
| cbams.2.SpatialGate.spatial | 100 | 1, 256, 256 | float32 |
| cbams.2.SpatialGate | - | 128, 256, 256 | float32 |
| cbams.3.ChannelGate.mlp | 2184 | 128 | float32 |
| cbams.3.ChannelGate | - | 128, 256, 256 | float32 |
| cbams.3.SpatialGate.compress | - | 2, 256, 256 | float32 |
| cbams.3.SpatialGate.spatial | 100 | 1, 256, 256 | float32 |
| cbams.3.SpatialGate | - | 128, 256, 256 | float32 |
| cbams.4.ChannelGate.mlp | 2184 | 128 | float32 |
| cbams.4.ChannelGate | - | 128, 256, 256 | float32 |
| cbams.4.SpatialGate.compress | - | 2, 256, 256 | float32 |
| cbams.4.SpatialGate.spatial | 100 | 1, 256, 256 | float32 |
| cbams.4.SpatialGate | - | 128, 256, 256 | float32 |
| cbams.5.ChannelGate.mlp | 2184 | 128 | float32 |
| cbams.5.ChannelGate | - | 128, 256, 256 | float32 |
| cbams.5.SpatialGate.compress | - | 2, 256, 256 | float32 |
| cbams.5.SpatialGate.spatial | 100 | 1, 256, 256 | float32 |
| cbams.5.SpatialGate | - | 128, 256, 256 | float32 |
| cbams.6.ChannelGate.mlp | 2184 | 128 | float32 |
| cbams.6.ChannelGate | - | 128, 256, 256 | float32 |
| cbams.6.SpatialGate.compress | - | 2, 256, 256 | float32 |
| cbams.6.SpatialGate.spatial | 100 | 1, 256, 256 | float32 |
| cbams.6.SpatialGate | - | 128, 256, 256 | float32 |
| cbams.7.ChannelGate.mlp | 2184 | 128 | float32 |
| cbams.7.ChannelGate | - | 128, 256, 256 | float32 |
| cbams.7.SpatialGate.compress | - | 2, 256, 256 | float32 |
| cbams.7.SpatialGate.spatial | 100 | 1, 256, 256 | float32 |
| cbams.7.SpatialGate | - | 128, 256, 256 | float32 |
| b256.skip | 32768 | 256, 128, 128 | float16 |
| b256.conv0 | 147584 | 128, 256, 256 | float16 |