

Figure 9: A full explanatory diagram of the Rate-Distortion Optimization Procedure inspired from both Ballé et al. (2016) and Deza et al. (2019). The goal is to find the equivalent ‘perceptual transmission rate’ for a given distortion  $\sigma$  to find a matched-resource perceptual input for Foveation-Texture that is non-foveated. This optimization pipeline produces Uniform-Blur, a perceptual system that receives as input uniformly blurred images as a way to loosely mimic uniform retinal ganglion cell re-distribution in as if it were to occur in humans. We now have a proper control to evaluate how a foveated texture based model (Foveation-Texture) compares to a non-foveated model (Uniform-Blur) when restricted with the *same* amount of perceptual resources under the aggregate SSIM matching constraint.

## 557 A Description of All Perceptual Systems

558 **Foveation-Texture:** We adjusted the parameters of the foveation texture transform to have stronger  
 559 distortions in the periphery that can consequently amplify the differences between a foveated and  
 560 non-foveated system. This was done setting the rate of growth of the receptive field size (scaling  
 561 factor)  $s = 0.4$ .

562 This value ( $s = 0.4$ ) was used instead of  $s = 0.5$ , given that experiments of Freeman & Simoncelli  
 563 (2011); Deza et al. (2019) have shown that this scaling factor yields a match with physiology but  
 564 only when human observers are psychophysically tested *between* pairs of synthesized/rendered  
 565 image metamers. Works of Wallis et al. (2017, 2019); Deza et al. (2019); Shumikhin (2020) have  
 566 suggested that the when comparing a non-foveated *reference image* to its foveated texturized version,  
 567 the scaling factor is actually much smaller than 0.5 (0.24, or in some cases as small as 0.20; See  
 568 Table 3). We thus selected a smaller factor of  $s = 0.4$  (that is still metameric to a human observer  
 569 between synthesized pairs), as smaller scaling factors significantly reduced the crowding effects.  
 570 Ultimately, the selection of this value is not critical in our studies as: 1) we are interested in grossly  
 571 exaggerating the distortions beyond the human metamer boundary to test if the perceptual system  
 572 will learn something new or different from the highly manipulated images that use a new family of  
 573 transformations; 2) we are not making any comparative measurements to human psychophysical  
 574 experiments where matching such scaling factors would be critical *e.g.* Deza & Eckstein (2016);  
 575 Eckstein et al. (2017); Geirhos et al. (2018).

576 **Reference:** We use the same image transform at the foveation stage for Reference but set the scaling  
 577 factor set to  $s = 0$ . In this way, any potential effects of the compression/expansion operations of the  
 578 image transform stage in the perceptual system is tightly upper-bounded by Reference over Foveation-  
 579 Texture. Thus, the only difference after stage 1 is whether the image statistics were texturized in  
 580 increasingly large pooling windows (Foveation-Texture), or not (Reference) – however note that the  
 581 texturization procedure comes at a computational cost and modifies the amount of resources allocated  
 582 in the image.

583 Indeed, the Reference system does not provide a matched-resource non-foveated control – the  
 584 Reference model only provides a non-foveated *upper bound* that removes the effects of crowding that  
 585 Foveation-Texture does have (See Theorem 1). In fact, the matched-resource control – under certain  
 586 constraints (See Table 2) – that is also non-foveated is the Uniform-Blur system as described earlier  
 587 in the paper, and in more detail as follows.

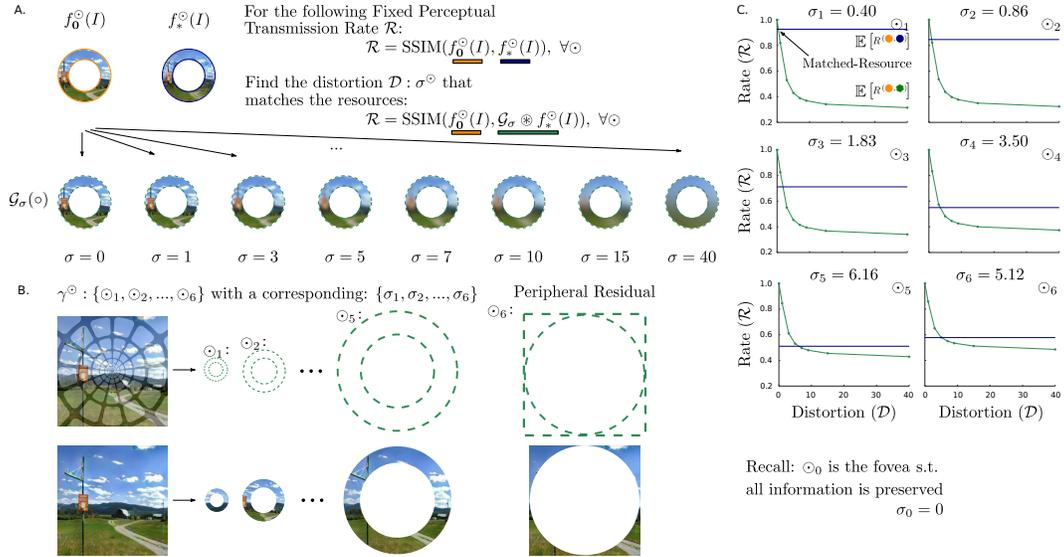


Figure 10: A. The full explanatory diagram of the Rate-Distortion Optimization Procedure adapted for Foveation-Blur. B. The goal is to find the equivalent ‘perceptual transmission rate’ for a given distortion  $\sigma$  to find a matched-resource perceptual input for Foveation-Texture that is foveated but with adaptive Gaussian blurring, *i.e.* we must find the standard deviation of the Gaussian blurring kernel which is computed over a set of eccentricity rings that have been windowed with cosine functions. C. The full Rate-Distortion curves as a function of retinal eccentricity rings.

588 **Uniform-Blur:** Uniform-Blur provides a non-foveated resource matched control with respect to  
589 Foveation-Texture. This perceptual system is essentially computed via finding the optimal standard  
590 deviation  $\sigma$  of the Gaussian filtering kernel  $\mathcal{G}_\sigma$  as shown in Figure 4. This distortion image is  
591 computed via the convolution ( $\otimes$ ) of the Gaussian filter  $\mathcal{G}_\sigma$  with the image  $f_0(I)$ . Here, Wang et al.  
592 (2004)’s SSIM is our candidate perceptual metric as it will take into consideration the luminance,  
593 contrast and structural changes locally for the entire image and pool them together for an aggregate  
594 perceptual score (and also the rate  $\mathcal{R}$ ) that is upper bounded by 1 and correlated with human perceptual  
595 judgments. As SSIM operates on the luminance of the image, all validation images over which the  
596 RD curve (right) was computed were transformed to grayscale to find the optimal standard deviation  
597 ( $\sigma = 3.4737$ ).

598 It is also worth emphasizing that the previous matching procedure is done over an aggregate family of  
599 images in the validation set (hence the use of the expected value ( $\mathbb{E}[\cdot]$ ) in Figure 4). This gives us a  
600 single standard deviation that will be used to filter *all* the images corresponding to the Uniform-Blur  
601 transform the same way.

602 **Foveation-Blur:** Is a foveated perceptual system that receives Rate-Distortion optimized images  
603 that have been blurred with different standard deviations of the gaussian kernel  $\mathcal{G}_\sigma$  as a function of  
604 retinal eccentricity. We picked the same eccentricity rings (collection of pooling regions that lie  
605 along the same retinal eccentricity) as Foveation-Texture given that we did not want to include a  
606 potential effect that is driven by differences in receptive field sizes rather than differences in type of  
607 computation. Figure 10 shows the full set of distortion strengths ( $\sigma$ ) of each receptive field ring to  
608 match the perceptual transmission rate of the Foveation Texture Transform ( $f_*(\circ)$ ).

609 There are other alternatives to potentially find the set of standard deviation coefficients that are not  
610 driven by a rate-distortion optimization procedure. One possibility could have been to find a mapping  
611 between pixels and degrees of visual angle as done in Pramod et al. (2018) and derive the coefficients  
612 by fitting a contrast sensitivity function given the visual field. While this approach is appealing, the  
613 coefficients for object recognition such as in ImageNet Russakovsky et al. (2015) can not be extended  
614 to scenes such as Places Zhou et al. (2017). In addition, the coupling of the RD-optimization with  
615 SSIM provides a perceptual guarantee to compare Foveation-Blur-Net to either Foveation-Texture or  
616 Uniform-Blur.

## 617 B Reference as a perceptual Upper Bound

618 **Theorem 1.** *Reference is a perceptual upper bound, and its generalization performance can be*  
 619 *matched, but can not be exceeded (due to possession of maximum image information).*

620 *Proof.* Let  $I' = \mathcal{D}(M)$  be the decoded image to be received by the second stage  $g(\circ)$  of any  
 621 perceptual system, where  $M_{\theta_i, \psi_i} = \alpha_i Q_{\theta_i, \psi_i} + (1 - \alpha_i) T_{\theta_i, \psi_i}$  is the convex combination between  
 622 structure and texture for the collection of pooling regions  $i$  (Figure 2 B.). It can be observed that  
 623 for Reference the values of  $\alpha$  yield  $\alpha_i = 0, \forall i$ , thus any other system that has at least 1 value of  
 624  $\alpha_i \neq 0$  will render a decoded image with a non-zero distortion in pixel space, thus making the  
 625 resources (amount of information) of Reference greater or equal than any other system with non-zero  
 626 coefficients (e.g. Foveation-Texture).  $\square$

627 **Remark 1.** *An example of a theoretically matched generalization performance system to Reference*  
 628 *from another non-zero distortion network is possible if the family of pre-distorted images were based*  
 629 *on textures (also see Gatys et al. (2015) Figure 5).*

630 **Remark 2.** *The resulting transformed images from  $f_0(\circ)$  and  $f_*(\circ)$  are not diffeomorphic to each*  
 631 *other.*

## 632 C Full set of IQA Metrics

(mean $\pm$ std)	SSIM ( <i>Matched</i> )	MSE ( $\uparrow$ )	Mutual Information ( $\downarrow$ )
Reference	1.0	0.0	7.39 $\pm$ 0.52
Foveation-Texture	<b>0.58 <math>\pm</math> 0.11</b>	<b>976.78 <math>\pm</math> 522.22</b>	<b>1.40 <math>\pm</math> 0.42</b>
Uniform-Blur	<b>0.57 <math>\pm</math> 0.15</b>	458.67 $\pm$ 277.13	1.86 $\pm$ 0.58
Foveation-Blur	<b>0.58 <math>\pm</math> 0.15</b>	507.35 $\pm$ 302.71	1.84 $\pm$ 0.56
(mean $\pm$ std)	MS-SSIM (Wang et al., 2003)( $\downarrow$ )	CW-SSIM (Wang & Simoncelli, 2005) ( $\downarrow$ )	FSIM (Zhang et al., 2011)( $\downarrow$ )
Reference	1.0	1.0	1.0
Foveation-Texture	<b>0.20 <math>\pm</math> 0.03</b>	<b>0.74 <math>\pm</math> 0.05</b>	<b>0.76 <math>\pm</math> 0.05</b>
Uniform-Blur	0.36 $\pm$ 0.03	0.98 $\pm$ 0.01	<b>0.69 <math>\pm</math> 0.09</b>
Foveation-Blur	0.36 $\pm$ 0.03	0.98 $\pm$ 0.01	<b>0.67 <math>\pm</math> 0.10</b>
(mean $\pm$ std)	VSI (Zhang et al., 2014) ( $\downarrow$ )	GMSD (Xue et al., 2013) ( $\uparrow$ )	NLPD (Laparra et al., 2016) ( $\uparrow$ )
Reference	1.0	0.0	0.0
Foveation-Texture	<b>0.93 <math>\pm</math> 0.02</b>	<b>0.19 <math>\pm</math> 0.03</b>	<b>0.75 <math>\pm</math> 0.16</b>
Uniform-Blur	<b>0.91 <math>\pm</math> 0.04</b>	<b>0.19 <math>\pm</math> 0.03</b>	0.40 $\pm$ 0.09
Foveation-Blur	<b>0.91 <math>\pm</math> 0.04</b>	<b>0.22 <math>\pm</math> 0.04</b>	0.45 $\pm$ 0.11
(mean $\pm$ std)	MAD (Larson & Chandler, 2010) * ( $\uparrow$ )	VIF (Sheikh & Bovik, 2006) ( $\downarrow$ )	LPIPSvgg (Zhang et al., 2018) * ( $\uparrow$ )
Reference	0.0	1.0	0.0
Foveation-Texture	166.77 $\pm$ 19.46	<b>0.12 <math>\pm</math> 0.03</b>	0.35 $\pm$ 0.05
Uniform-Blur	<b>182.19 <math>\pm</math> 16.50</b>	<b>0.12 <math>\pm</math> 0.03</b>	<b>0.52 <math>\pm</math> 0.07</b>
Foveation-Blur	<b>185.90 <math>\pm</math> 18.60</b>	<b>0.16 <math>\pm</math> 0.03</b>	<b>0.54 <math>\pm</math> 0.08</b>
(mean $\pm$ std)	DISTS (Ding et al., 2020) * ( $\uparrow$ )		
Reference	0.0		
Foveation-Texture	0.20 $\pm$ 0.03		
Uniform-Blur	<b>0.36 <math>\pm</math> 0.03</b>		
Foveation-Blur	<b>0.35 <math>\pm</math> 0.03</b>		

Table 2: List of Full IQA Metrics from Ding et al. (2020) where we compare Image Transforms  $f(\circ)$  w.r.t. Reference for the *testing* set. Arrows ( $\uparrow$  /  $\downarrow$ ) indicate the direction of the *greatest* distortion according to the metric thus values further away from the Reference place a specific transform at a resource disadvantage. We observe matched distortion via virtual ties for SSIM (matched and optimized in the *validation* set), VSI, GMSD FSIM, and VIF; greater distortion (Foveation-Texture at a disadvantage) for MSE, Mutual Information, MS-SSIM, CW-SSIM, NLPD; and lower distortion (Foveation-Texture at an advantage) for MAD, and texture-based tolerance methods such as DISTS and LPIPSvgg – hence implicitly proving that our transform does indeed preserve local texture. Scores were computed over 5000 images. Numbers in bold represent highest/lowest IQA scores; virtual ties were declared if highly overlapping standard deviations are noticeable e.g.: FSIM, VIF.

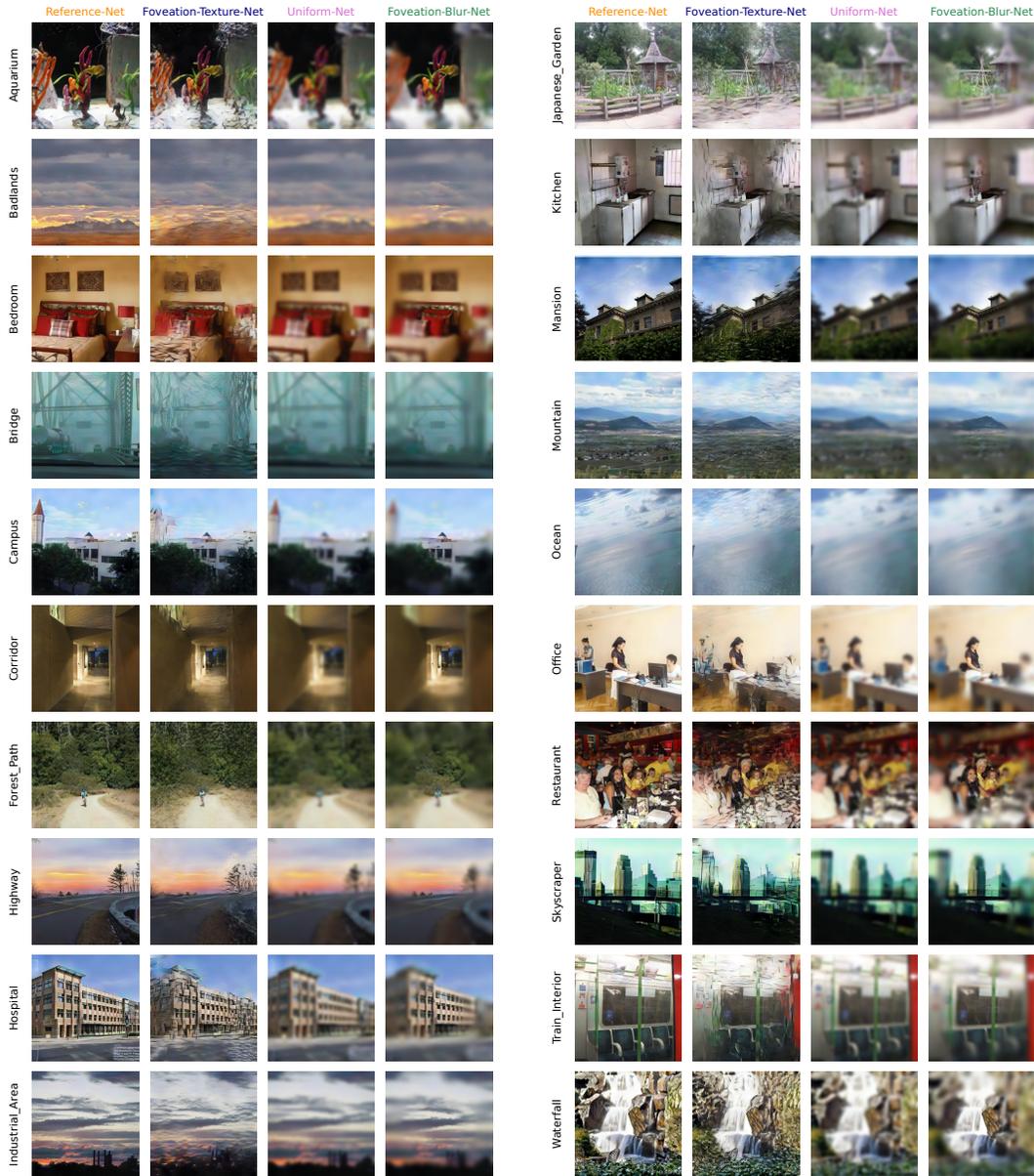


Figure 11: Sample Testing Image Mosaics.

## 633 D Image Transform Samples

634 Figure 11 is an extension of Figure 4 which shows a collection of randomly sampled images from  
 635 each one of the 20 scene classes and how they look under each image transform before being fed  
 636 to each network. Details worth noticing include: 1) Reference images are not full high resolution,  
 637 and are slightly compressed given the encoder/decoder pipeline of the transform to operate as a  
 638 tighter upper bound (observable when zooming in); 2) The foveal area is preserved and *identical*  
 639 for Reference, Foveation-Texture and Foveation-Blur; 3) The peripheral distortions are more or less  
 640 apparent contingent on the image structure; 4) All images used in our experiments were rendered at  
 641  $256 \times 256$  px.

Model	Freeman & Simoncelli (2011)	Wallis et al. (2019)	Fridman et al. (2017)	Deza et al. (2019)
Feed-Forward	-	-	✓	✓
Input	Noise	Noise	Image	Image
Multi-Resolution	✓	✓	-	-
Texture Statistics	Steerable Pyramid	VGG19 conv-1 <sub>1</sub> , 2 <sub>1</sub> , 3 <sub>1</sub> , 4 <sub>1</sub> , 5 <sub>1</sub>	Steerable Pyramid	VGG19 relu4 <sub>1</sub>
Style Transfer	Portilla & Simoncelli (2000)	Gatys et al. (2016)	Rosenholtz et al. (2012)	Huang & Belongie (2017)
Foveated Pooling	✓	✓	(Implicit via FCN)	✓
Decoder (trained on)	-	-	metamers/mongrels	images
Moveable Fovea	✓	✓	✓	✓
Use of Noise	Initialization	Initialization	-	Perturbation
Non-Deterministic	✓	✓	-	✓
Direct Computable Inverse	-	-	(Implicit via FCN)	✓
Rendering Time	hours	minutes	milliseconds	seconds
Image type	scenes	scenes/texture	scenes	scenes
Critical Scaling (vs Synth)	0.46	~ {0.39/0.41}	Not Required	0.5
Critical Scaling (vs Reference)	Not Available	~ {0.2/0.35}	Not Required	0.24
Experimental design	ABX	Oddball	-	ABX
Reference Image in Exp.	Metamer	Original	-	Compressed via Decoder
Number of Images tested	4	400	-	10
Trials per observers	~ 1000	~ 1000	-	~ 3000

Table 3: Foveated Texture-based transform comparison. Redrawn from Deza et al. (2019).

## 642 E Differences across other Foveation models

643 There are currently 4 foveation models that implement texture-like computation in the peripheral  
644 field of view as shown in Table 3. We selected the Foveation Texture Transform model of Deza  
645 et al. (2019) given that it is computationally tractable to render a foveated image dataset (100'000) at  
646 a rate of 1 image/second (rather than hours Freeman & Simoncelli (2011) or minutes Wallis et al.  
647 (2017)). We did not use the highly accelerated model of Fridman et al. (2017) (order of milliseconds,  
648 that was based on the Texture-Tiling Model of Rosenholtz et al. (2012)) because it was: 1) Not  
649 psychophysically tested with human observers thus there is no guarantee of visual metamerism via  
650 the choice of texture statistics (although see the recent work of Shumikhin (2020)); 2) But most  
651 importantly, it does not provide an upper-bound computational baseline (similar to Reference).

652 Altogether, we think that re-running our experiments and testing them with all other foveated models  
653 such as the before-mentioned is a direction of future work as we would be curious to see the  
654 replicability of our pattern of results across other texture-based peripheral models. Naturally, the type  
655 of texture-based foveation used will also yield different matched resource systems (Uniform-Blur  
656 and Foveation-Blur), as different models rely on texture computation in different ways – and thus  
657 will affect the IQA metric scores when performing the perceptual optimization.

Model	Wang & Cottrell (2017))	Wu et al. (2018)	Pramod et al. (2018)	(Ours)
Image input type	scenes	objects	objects	scenes
Single/Dual Stream	Dual + Gating	Dual + Concatenation	Single	Single
Role of Single/Dual Stream	Coupling the fovea + periphery	Contextual modulation (scene gist)	Serializing the (single)	two-stage model
Foveated Transform (F.T.)	log-polar + adaptive gaussian blurring	Region Selection	adaptive gaussian blurring	Visual Metamer w/ texture-distortion
Stochastic F.T.	-	-	-	✓ Deza et al. (2019)
Representational Stage of F.T.	retinal (Geisler & Perry, 1998)	"Overt Attention"	retinal (Geisler & Perry, 1998)	V2 (Freeman & Simoncelli, 2011)
Moveable Fovea	✓	✓	✓	✓
Accounts for pooling regions	Implicit via adaptive gaussian blurring	-	Implicit via adaptive gaussian blurring	✓
Accounts for visual crowding	-	-	-	✓
Accounts for retinal eccentricity	✓	Implicit via cropping	✓	✓
Accounts for loss of visual acuity	✓	-	✓	Implicit via visual crowding
Critical Radius (Larson & Loschky, 2009)	8 deg	Not Applicable (Objects)		~ 8.67 deg (Estimated from Fig. 8)
Out of Distribution Generalization	-	-	-	✓
Robustness to Distortion Type	-	Blurring	Blurring	Occlusion
Spatial Frequency Preference	High (Fovea), Low (Periphery)	Low (Global)	High (Fovea), Low (Periphery)	High (Global)
Weighted Bias Emerges	Center/Fovea	Center/Fovea	Center/Fovea	Center/Fovea
Goal of Foveal-Peripheral Architecture	Fit Behavioural Results	Increase Recognition Accuracy		Explore Perceptual Properties
Model System Focus	Human	Machine	Human	Hybrid

Table 4: A summary set of Foveal-Peripheral CNN model characteristics.

## 658 F Differences to other Relevant Work

659 There are several works that have used foveation to show a type of representational advantage over  
660 non-foveated systems. Mainly Pramod et al. (2018) with adaptive gaussian blur, and Wu et al. (2018)  
661 with scene gist, that have been targeted towards a computational goal in increasing object recognition  
662 performance. For scene recognition, only Wang & Cottrell (2017) has successfully modelled known  
663 behavioural results of Larson & Loschky (2009) via a dual-stream neural network that uses adaptive  
664 gaussian blurring and a log-polar transform. One key difference however is that we are interested  
665 in exploring the effects of peripheral texture-base computation that give rise to *visual crowding* and  
666 that is also linked to area V2 in the primate ventral stream – rather than retinal as in Wang & Cottrell  
667 (2017) which resembles our control condition: Foveation-Blur.

668 In general, we are taking a complimentary approach to Wang & Cottrell (2017) & Wu et al. (2018),  
669 and a similar one to Pramod et al. (2018) where we *a priori do not know of a functional role of*  
670 *texture-based computation or prime ourselves to fit our model to a reference behavioural result.*  
671 Thus we explore what perceptual properties it may have in comparison to a non-foveated system  
672 (Uniform-Blur, Reference) or a foveated system that only implements adaptive gaussian blurring  
673 (Foveation-Blur). Table 4 highlights key similarities & differences between these papers and ours.

674 **G Training, Testing and Learning Dynamics**

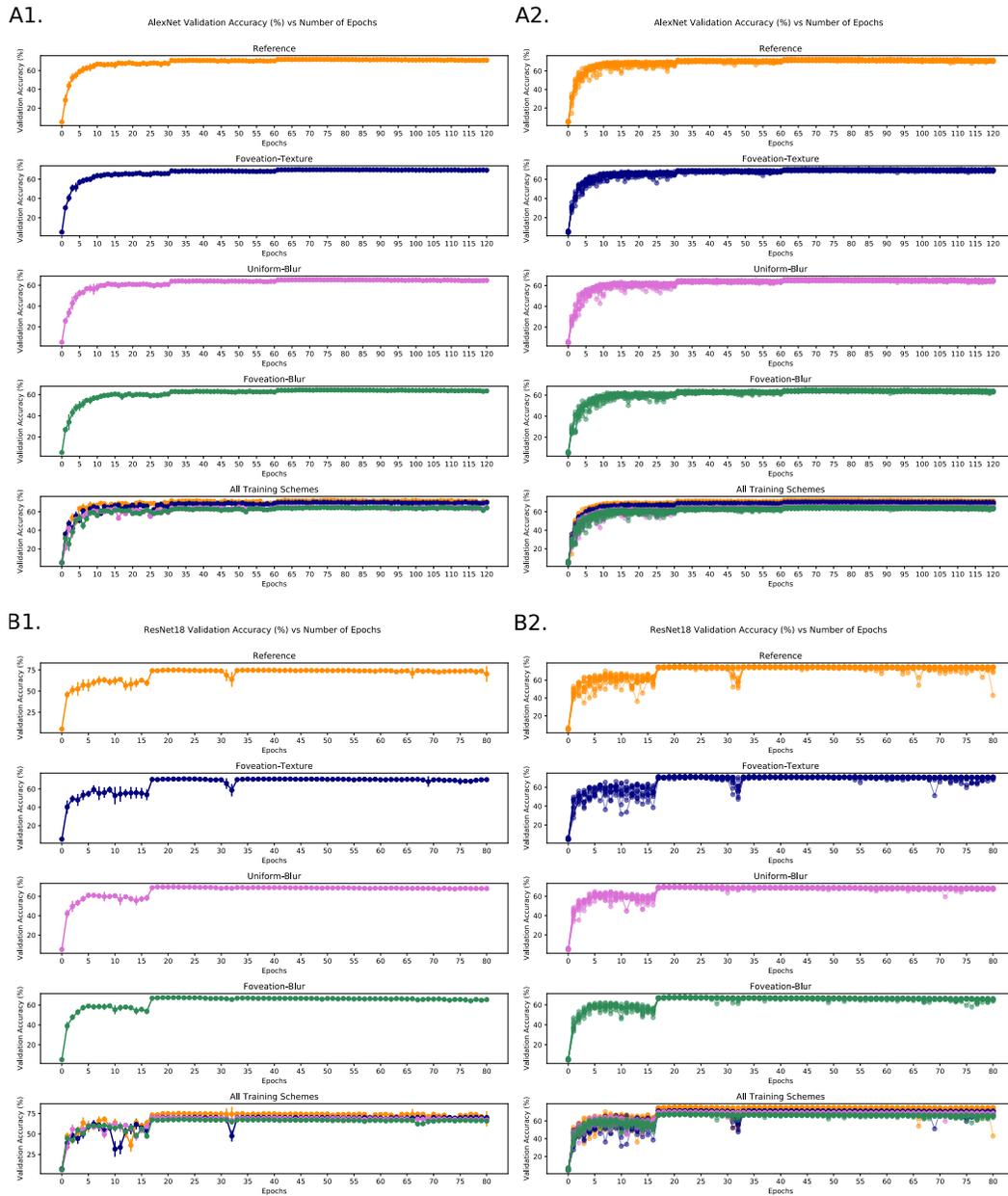


Figure 12: Learning Dynamics visualized via the Validation Accuracy over all epochs for AlexNet and ResNet18 as  $g(\circ)$ . Left: A1/B1 we see the aggregate Validation Accuracy. Right: A2/B2 the individual Validation Accuracies are shown for each network.

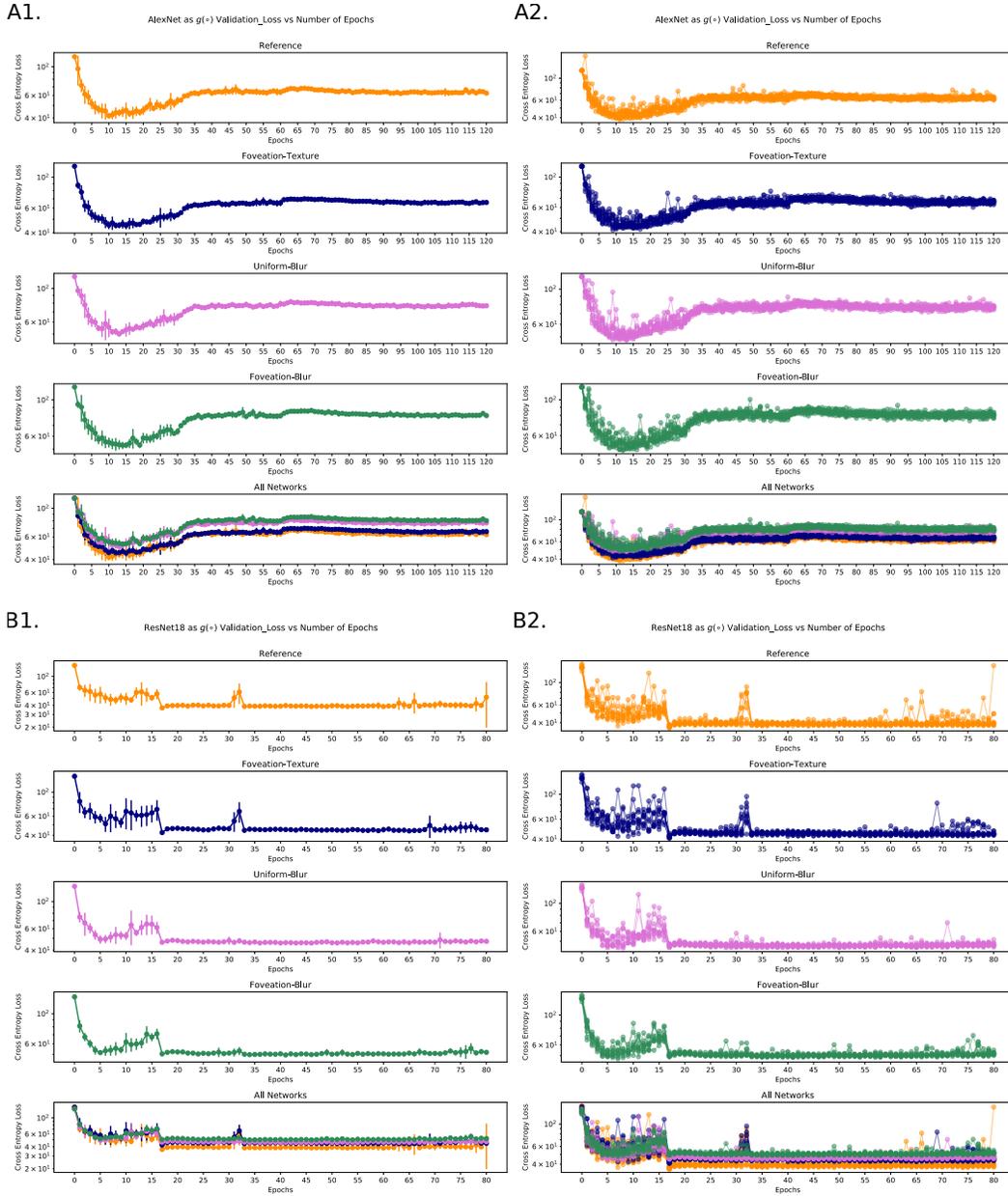


Figure 13: Validation Loss (Cross Entropy) over all epochs for AlexNet and ResNet18 as  $g(o)$ . Left: A1/B1 we see the aggregate Validation Loss. Right: A2/B2 the individual Validation Losses for each network. It is interesting to see that despite re-bound effects in the validation loss, that the validation accuracy continues to increase (See Figure 12).

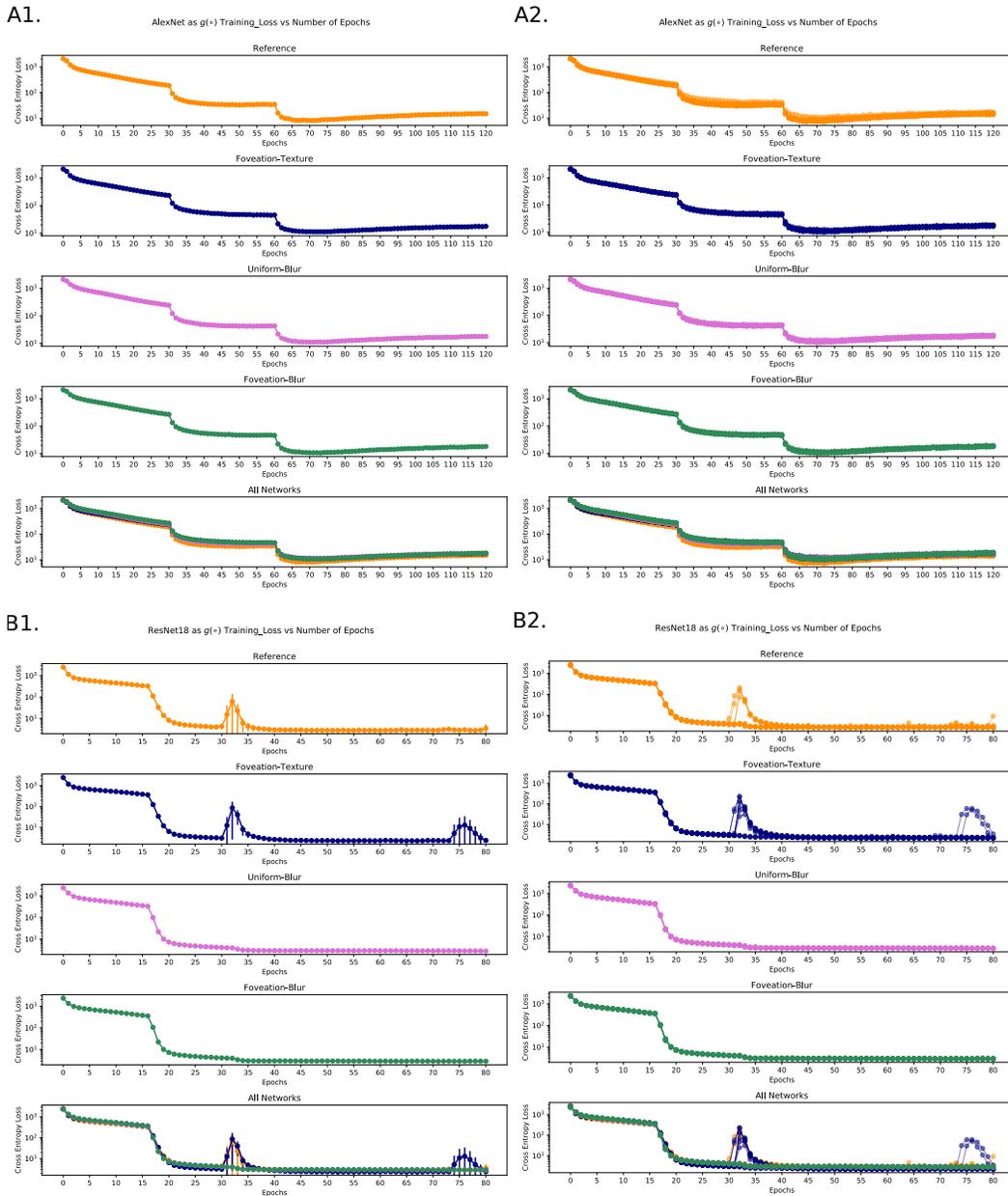


Figure 14: Training Loss (Cross Entropy) over all epochs for AlexNet and ResNet18 as  $g(\circ)$ . Left: A1/B1 we see the aggregate training loss. Right: A2/B2 the individual training losses for each network.

675 Perceptual Systems were trained with SGD, nestorov momentum, no dampening, weight decay =  
 676 0.0005, momentum = 0.9, a batch size of 128, Color Normalization of mean = (0.485, 0.456, 0.406),  
 677 and std = (0.229, 0.224, 0.225). Systems that used AlexNet as  $g(\circ)$  were trained for 120 epochs with  
 678 a scheduled learning rate, where the initial learning rate of 0.01 was halved after the 30th epoch, and  
 679 halved again after 60th epoch. Systems that used ResNet18 as  $g(\circ)$  were trained for 80 epochs and  
 680 with an initial learning rate of 0.05, which was multiplied by 0.25 after the first 16 epochs, and then  
 681 multiplied again by 0.25 after the 32nd epoch. All systems were trained with a cross-entropy loss  
 682 and received images size of  $256 \times 256 \times 3$ . No data-augmentation or cropping was used at training  
 683 or testing.

684 **H Generalization**

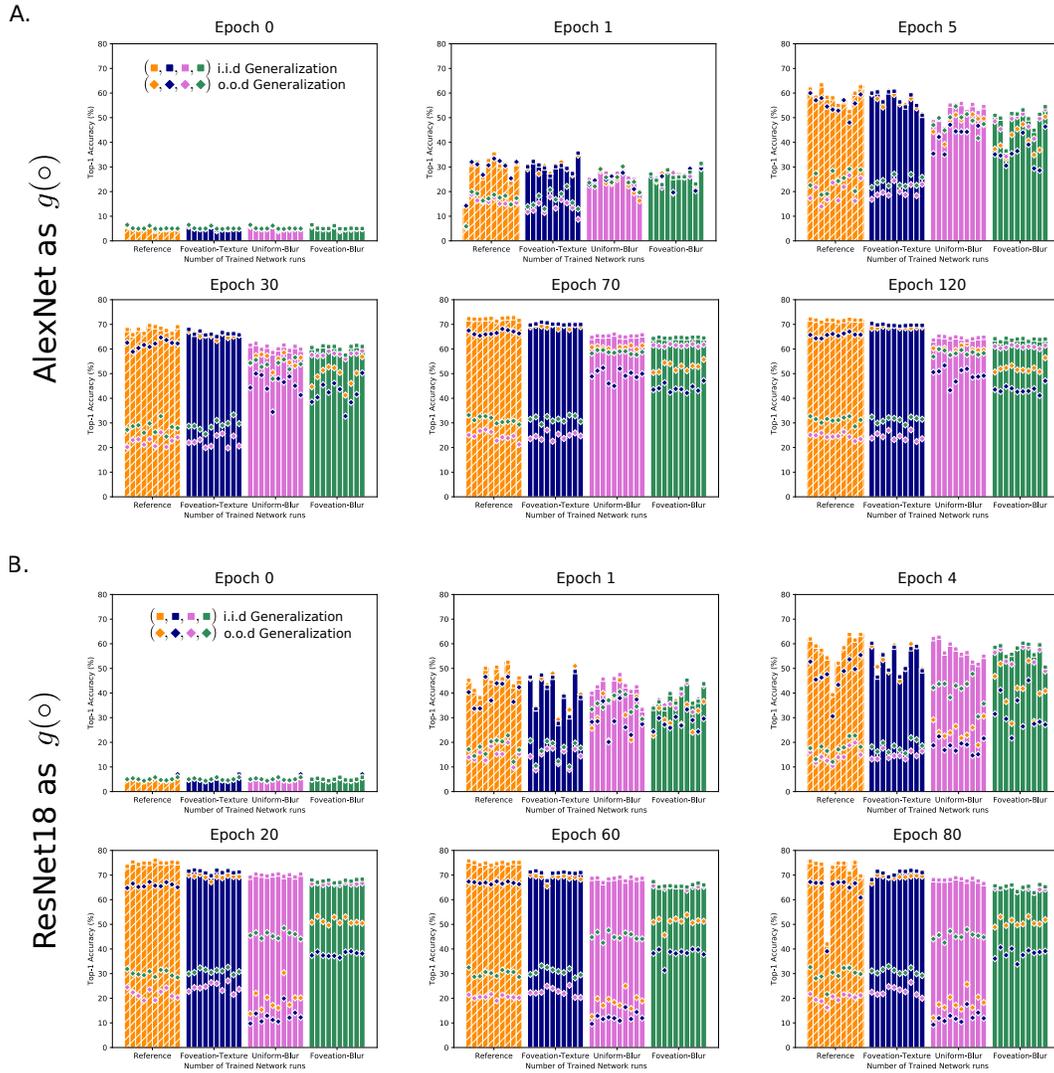


Figure 15: Generalization Dynamics over a set of multiple epochs for AlexNet and ResNet18 as  $g(o)$ .

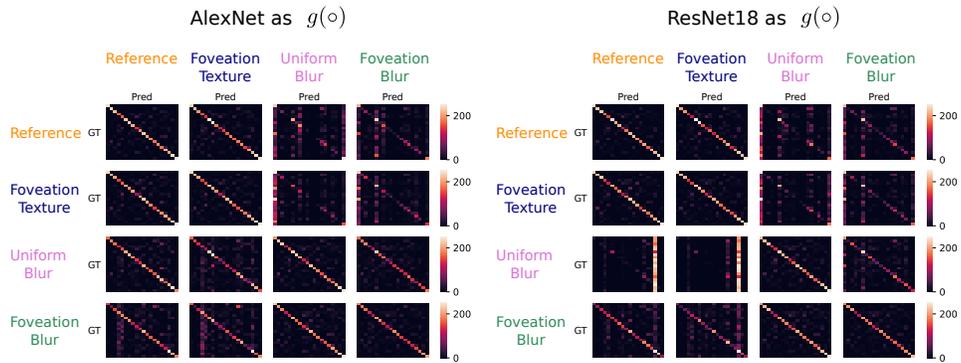


Figure 16: A sample collection of Confusion Matrices for the first of the 10 randomly initialized networks for each of the 4 perceptual systems with their transforms for both AlexNet and ResNet18 as  $g(\circ)$ . We see similar classification patterns between Foveation-Texture and the Reference, and also similar classification strategies between the Foveation-Blur and Uniform-Blur system. The asymmetry in the upper and lower off-diagonal quadrants highlight the differences between Foveation-Texture & Reference vs Foveation-Blur & Uniform-Blur. Each row/column per confusion matrix represents each of the scene classes in alphabetical order. These classes are: aquarium, badlands, bedroom, bridge, campus, corridor, forest path, highway, hospital, industrial area, japanese garden, kitchen, mansion, mountain, ocean, office, restaurant, skyscraper, train interior, waterfall.

# Generalization

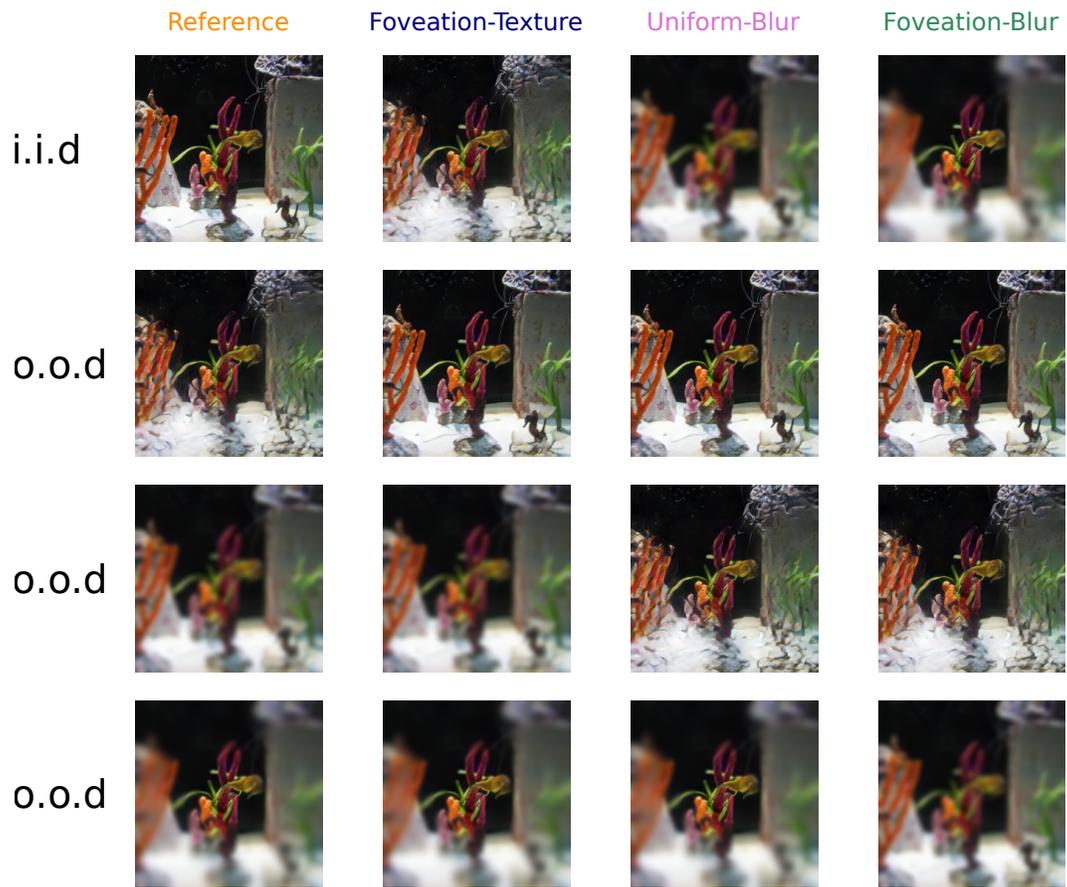


Figure 17: Sample image used in a full i.i.d and o.o.d evaluation.

# I Filter Visualization & Spatial Frequency Sensitivity

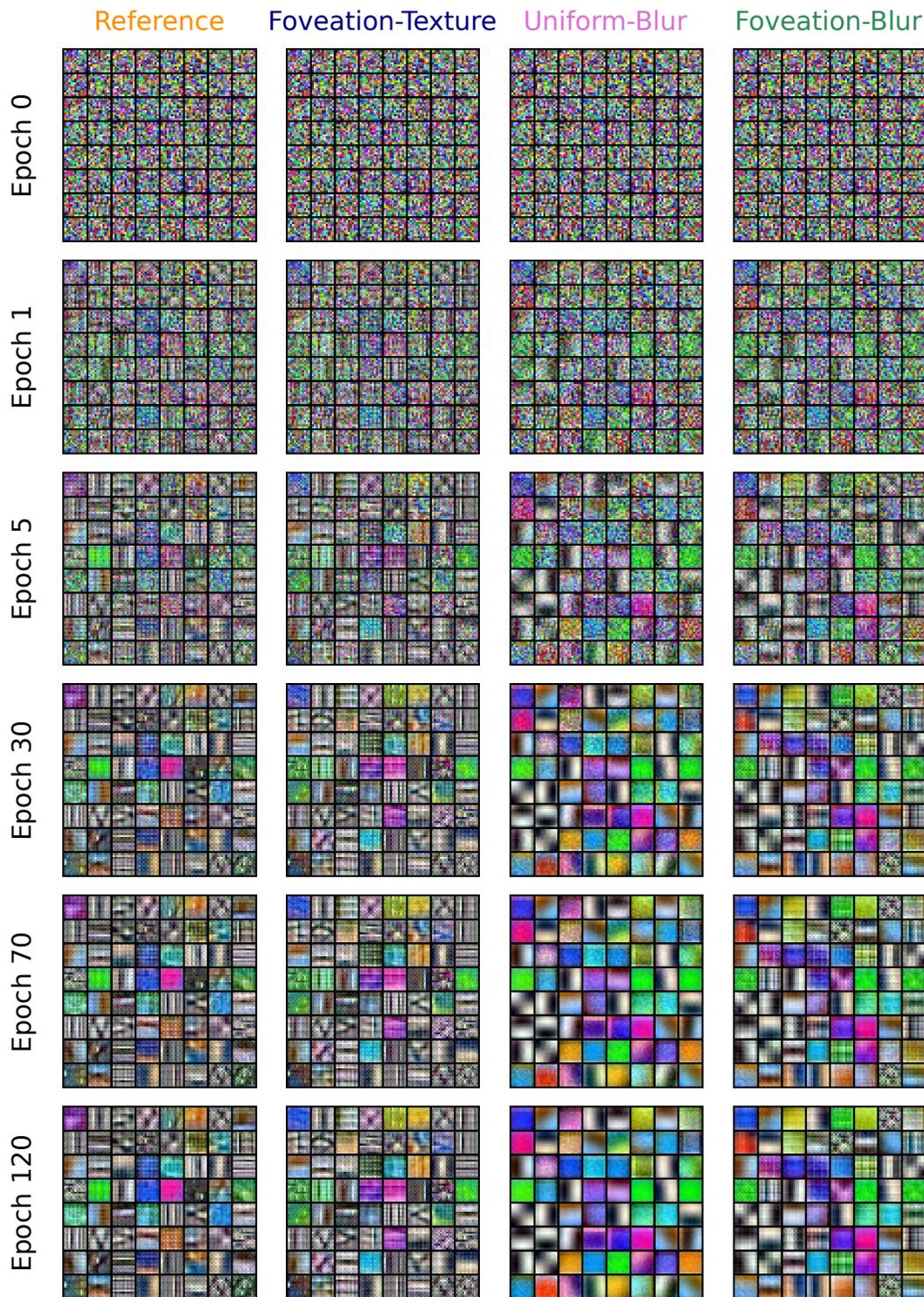


Figure 18: Evolution of AlexNet as  $g(\circ)$  Conv-1 Filters from 1st Random Weight Initialization.

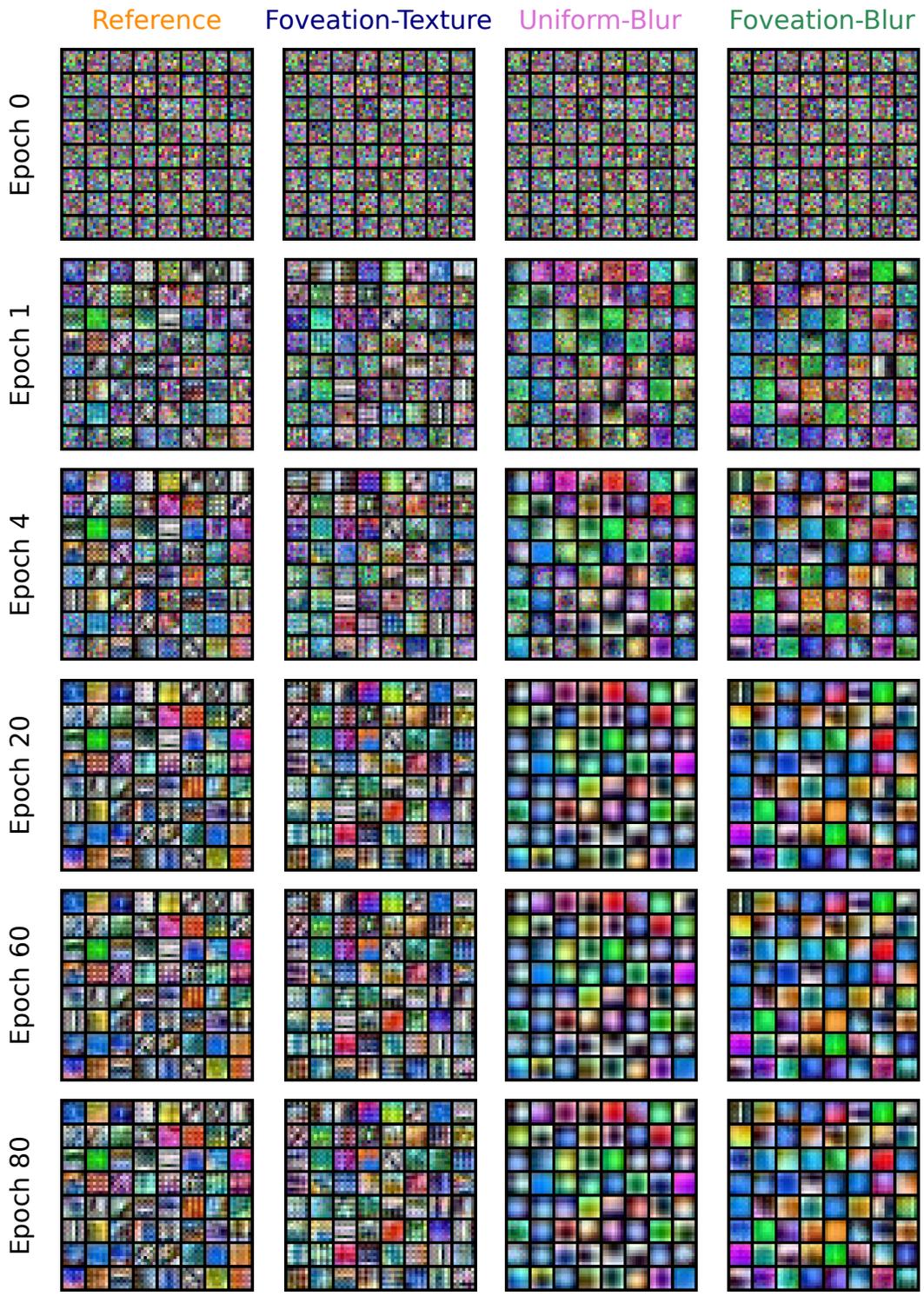


Figure 19: Evolution of ResNet18 as  $g(\circ)$  Conv-1 Filters from 1st Random Weight Initialization.

686 The size of all shown images was  $256 \times 256 \times 3$ , thus the units of the gaussian filters specified from  
687 Section 3.2 are in pixels. For a given Gaussian filtering operation  $\mathcal{G}_\sigma$  for a given standard deviation  $\sigma$ ,  
688 low pass spatial frequency (LF) images were computed via:

$$LF(I^C) = \mathcal{G}_\sigma \star I^C \quad (2)$$

689 for each channel  $C$ . Similarly, High Pass Spatial Frequency (HF) image stimuli were computed via:

$$HF(I^C) = I^C - \mathcal{G}_\sigma \star I^C + \text{mean}_{\text{val}}^C \quad (3)$$

690 where  $\text{mean}_{\text{val}}^C$  (which we call the residual in the main body of the paper) is the average of image  
691 intensity over the held-out validation set for each channel  $C$ , a small extension from Geirhos et al.  
692 (2019) as our image stimuli is in both color and grayscale.

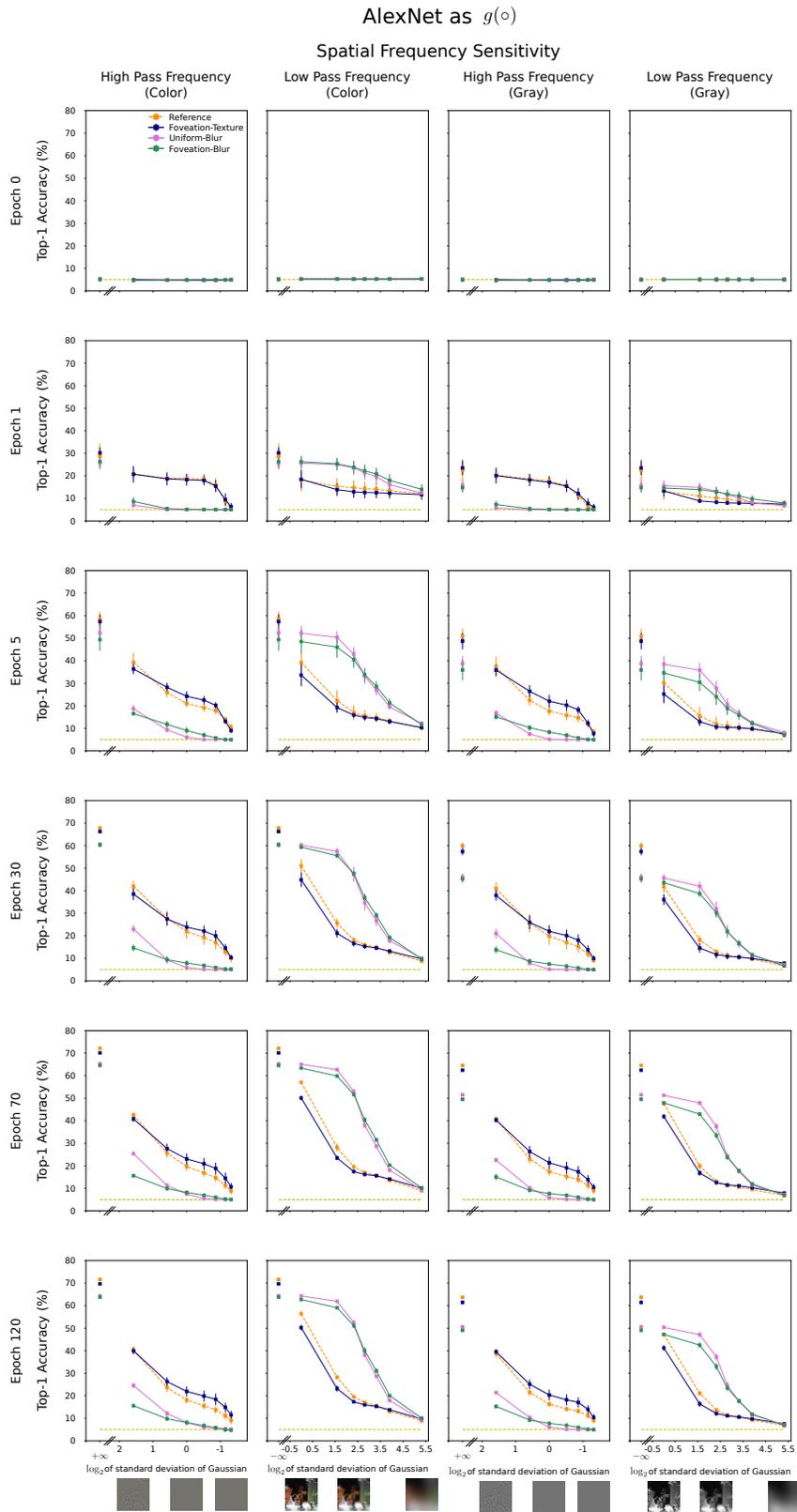


Figure 20: Aggregate Spatial Frequency Sensitivity for AlexNet as  $g(\circ)$  after epochs 0, 1, 5, 30, 70, 120.

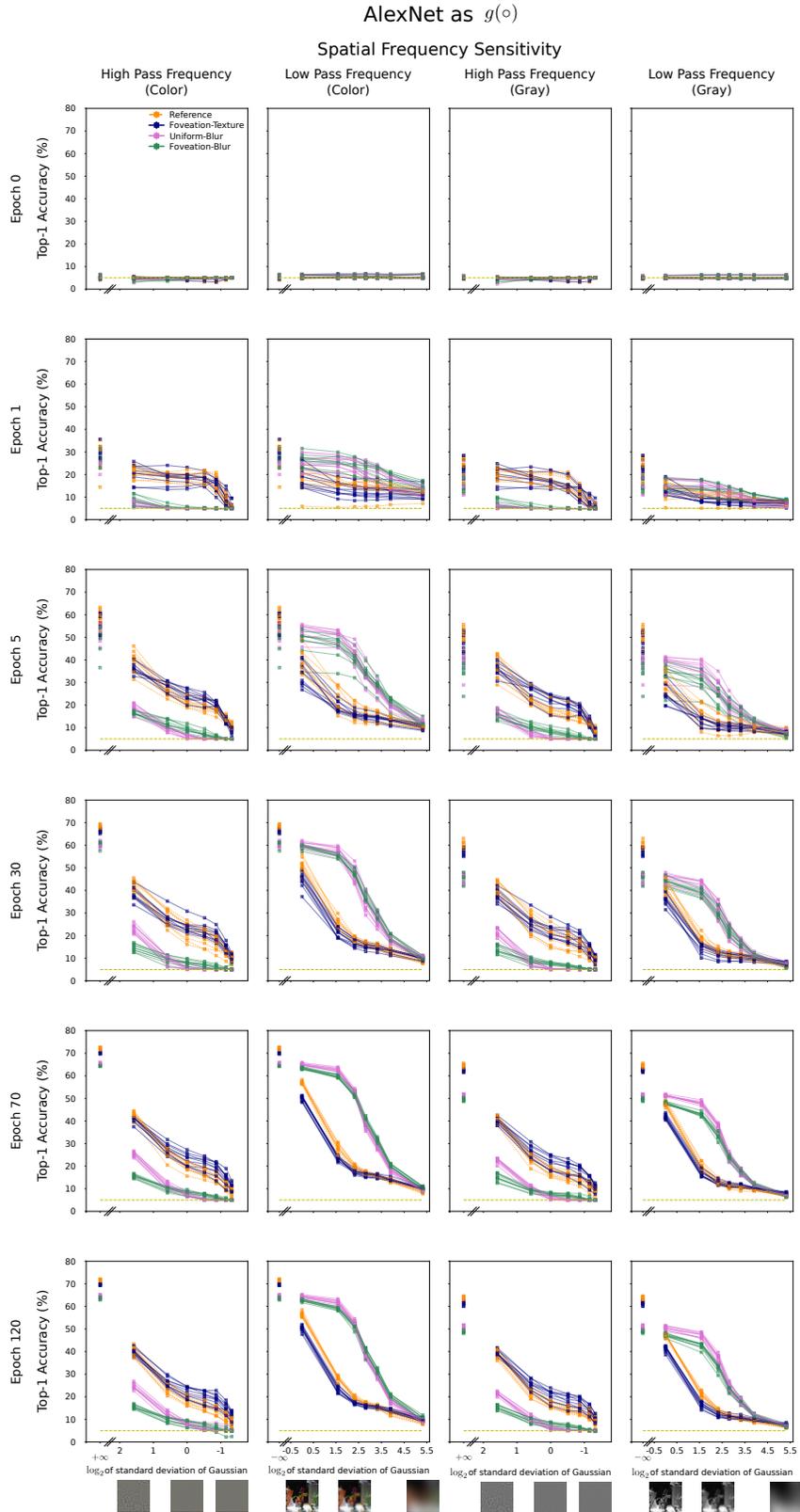


Figure 21: Individual Spatial Frequency Sensitivity for AlexNet as  $g(\circ)$  after epochs 0, 1, 5, 30, 70, 120.



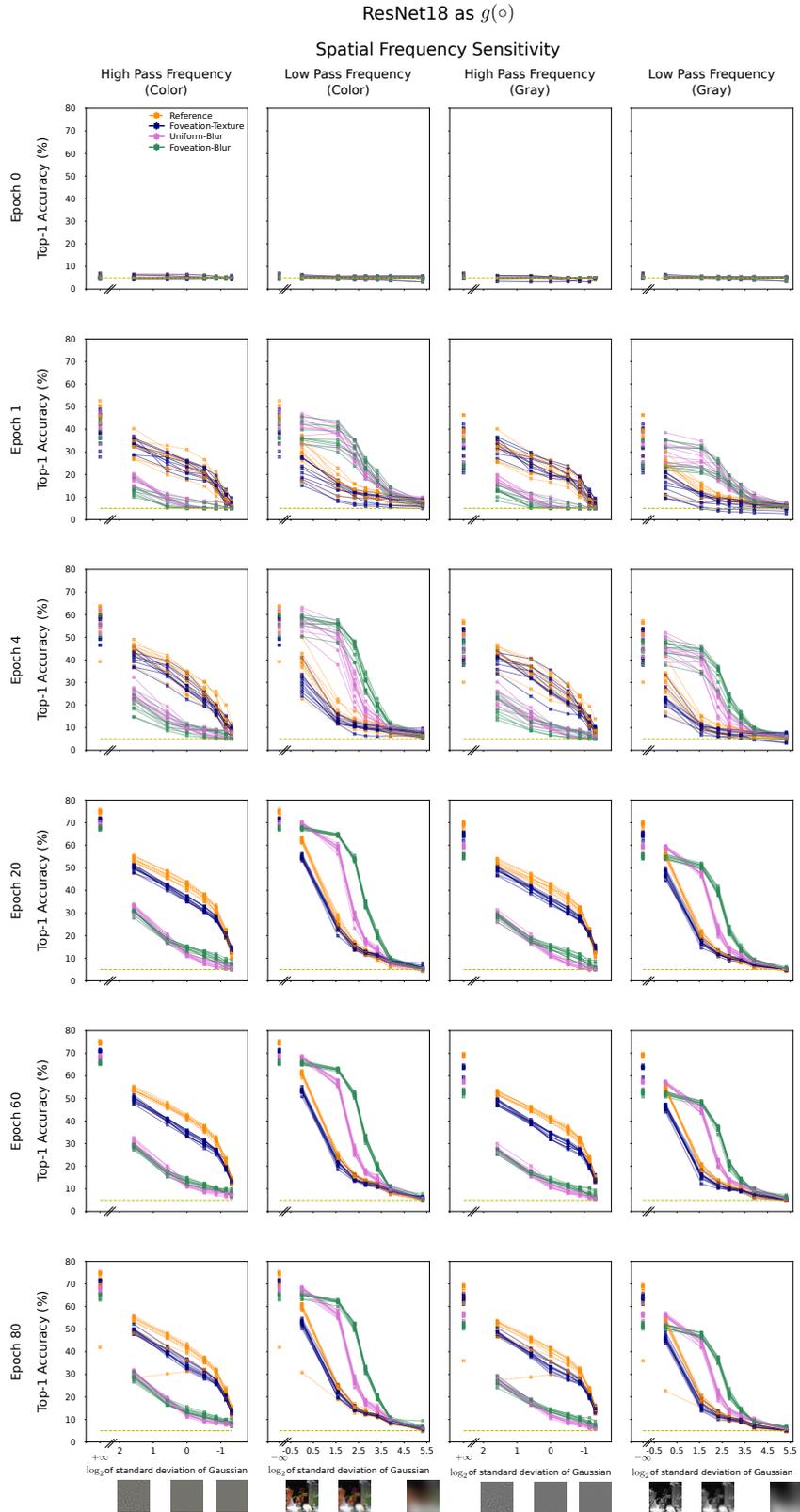


Figure 23: Individual Spatial Frequency Sensitivity for ResNet18 as  $g(\circ)$  after epochs 0, 1, 4, 20, 60, 80.

# High Pass Spatial Frequency (Color)

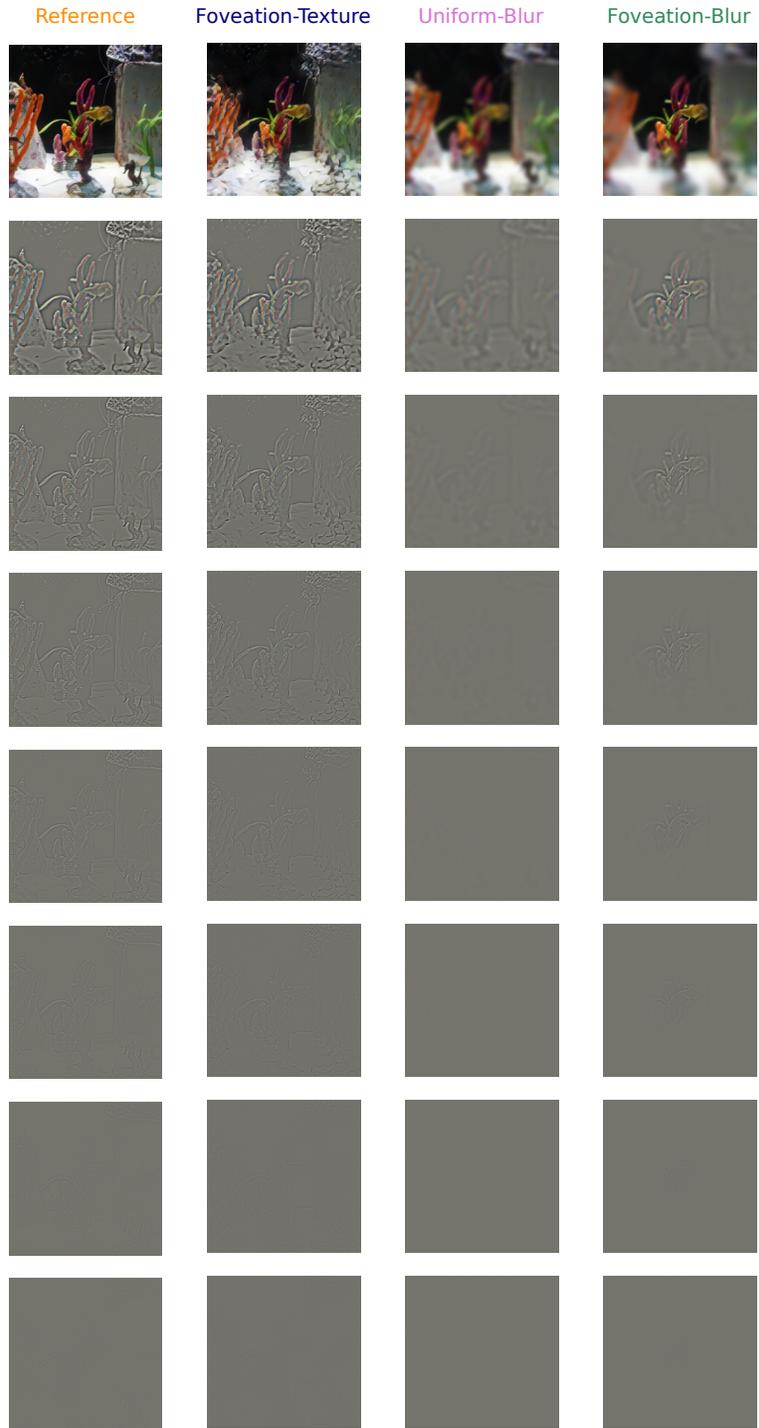


Figure 24: Sample High Pass Spatial Frequency Color Stimuli.

# High Pass Spatial Frequency (Gray)

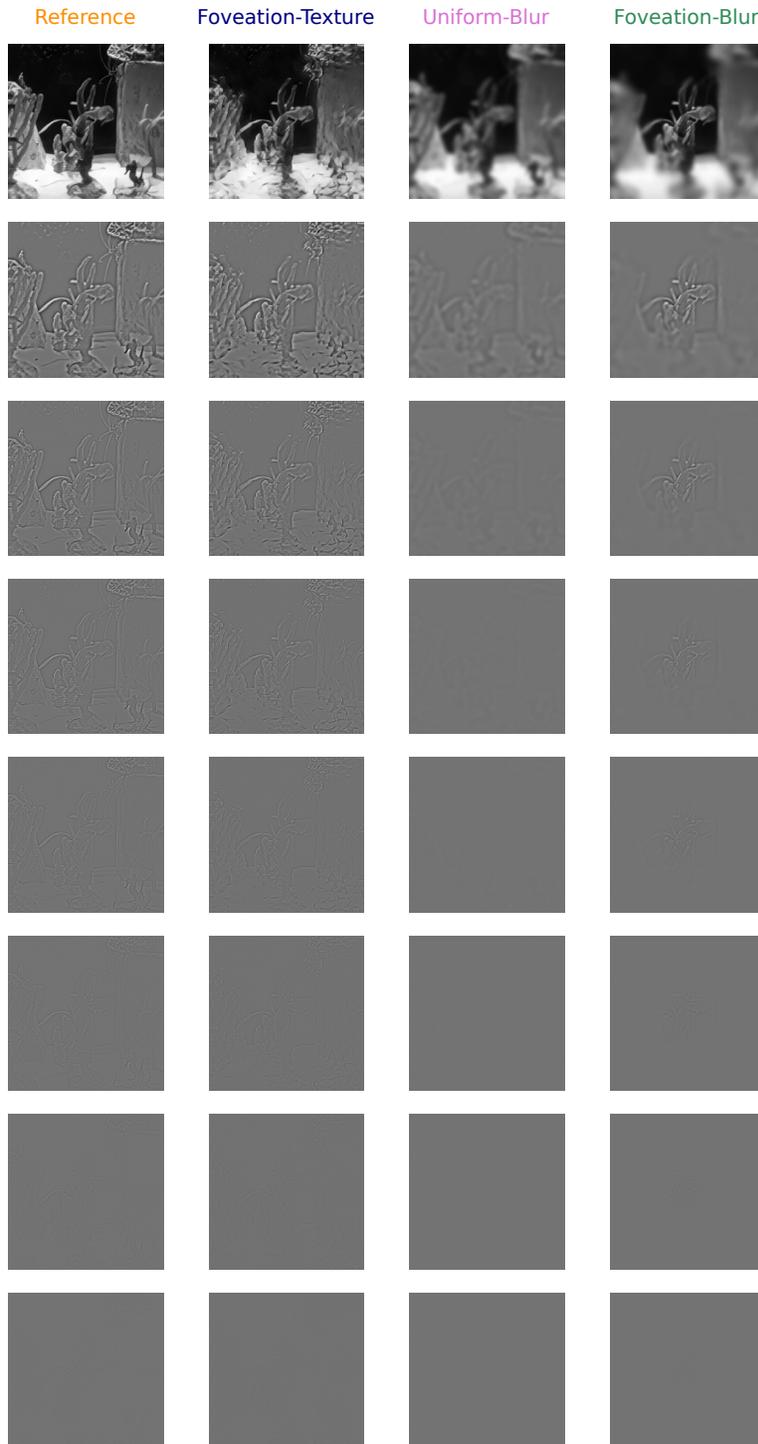


Figure 25: Sample High Pass Spatial Frequency Gray Stimuli.

# Low Pass Spatial Frequency (Color)

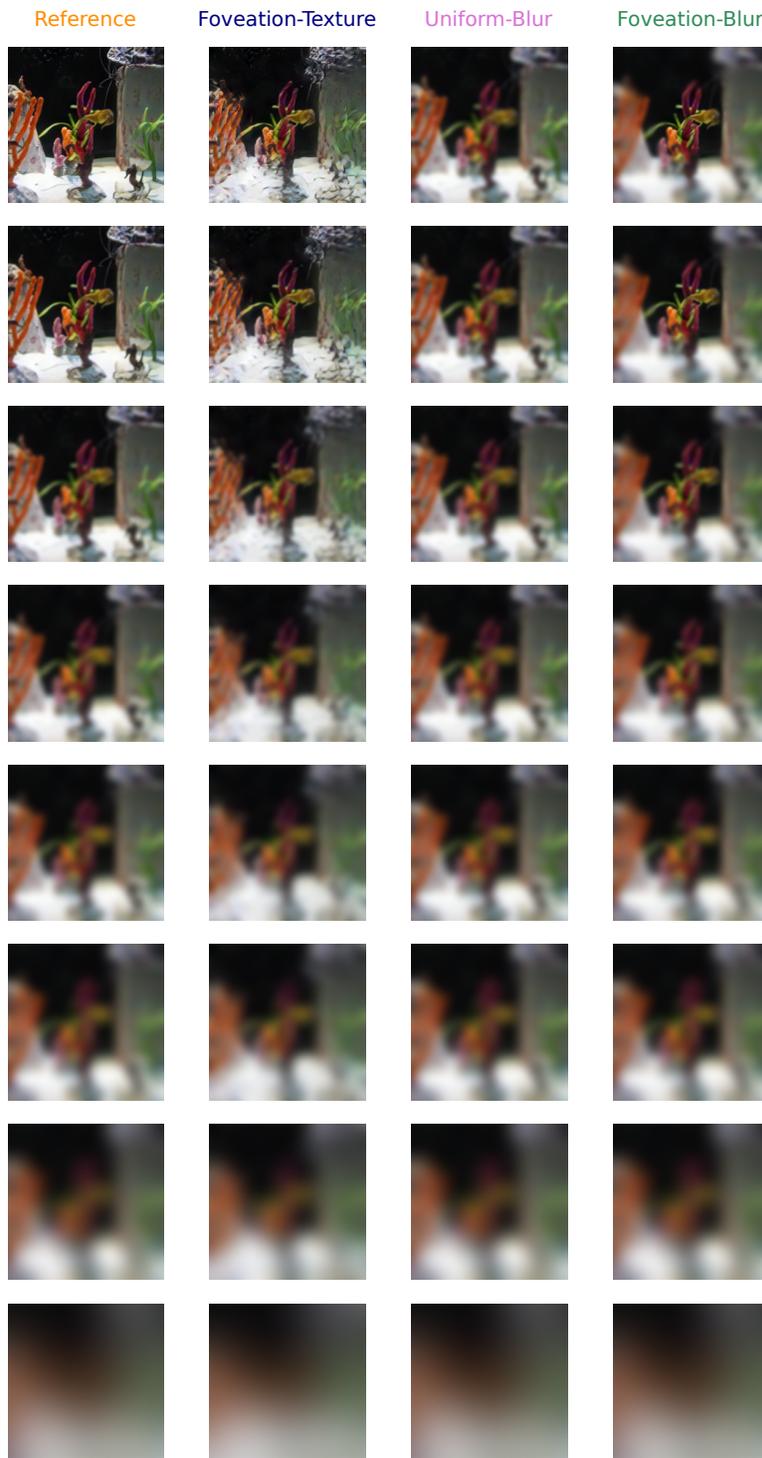


Figure 26: Sample Low Pass Spatial Frequency Color Stimuli.

# Low Pass Spatial Frequency (Gray)



Figure 27: Sample Low Pass Spatial Frequency Gray Stimuli.

693 **J Robustness to Occlusion**

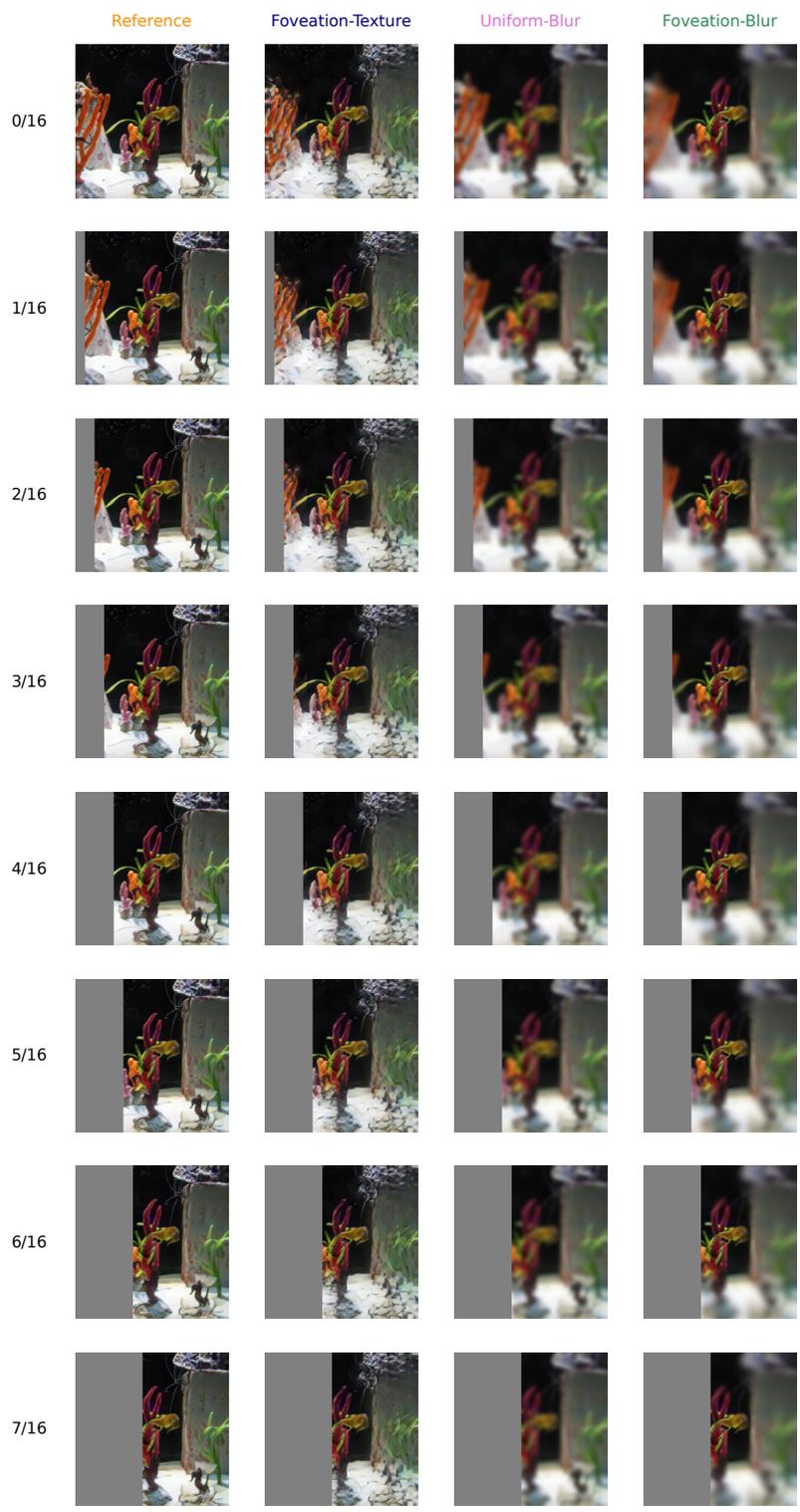


Figure 28: Left2Right Occlusion Sample Stimuli.

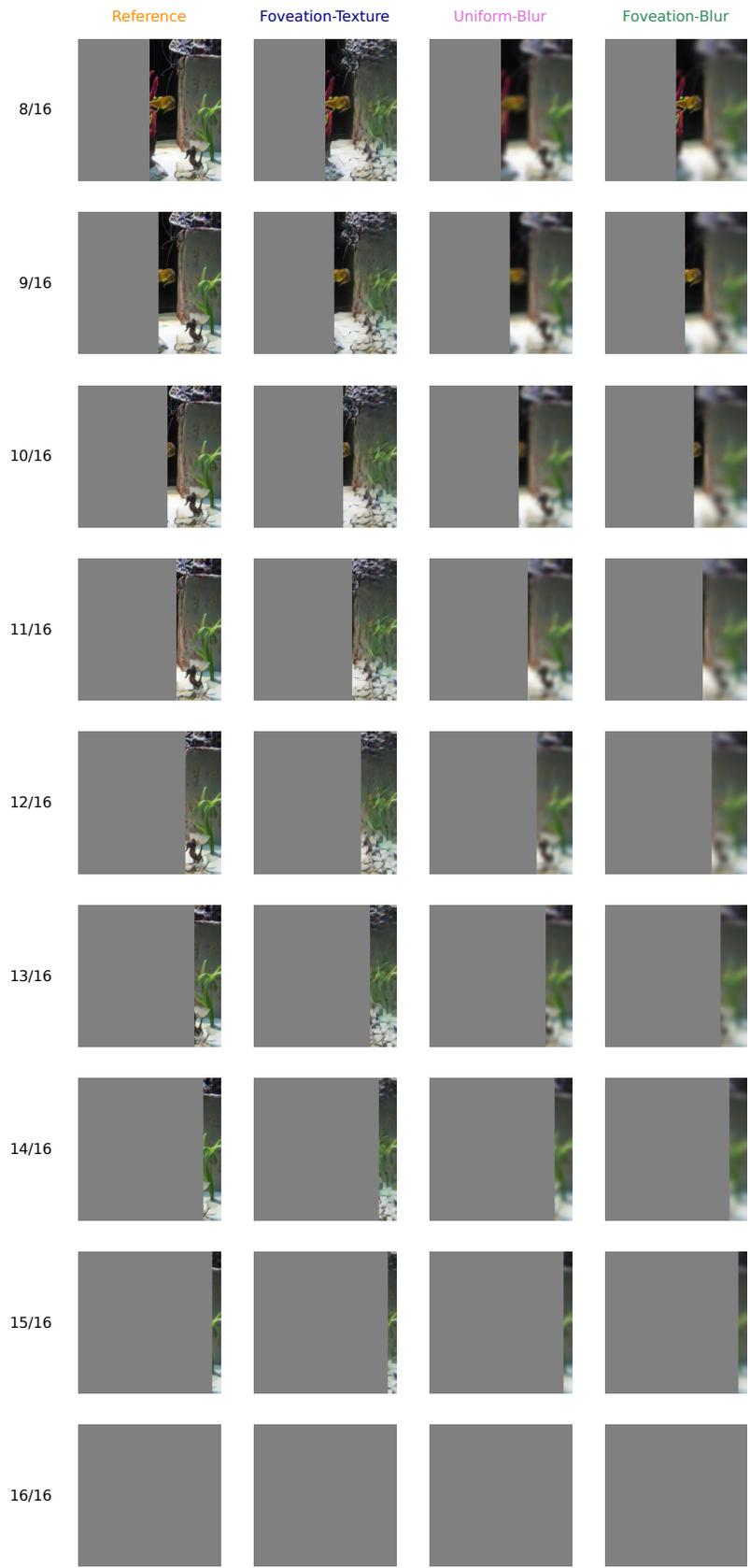


Figure 29: Left2Right Occlusion Sample Stimuli.

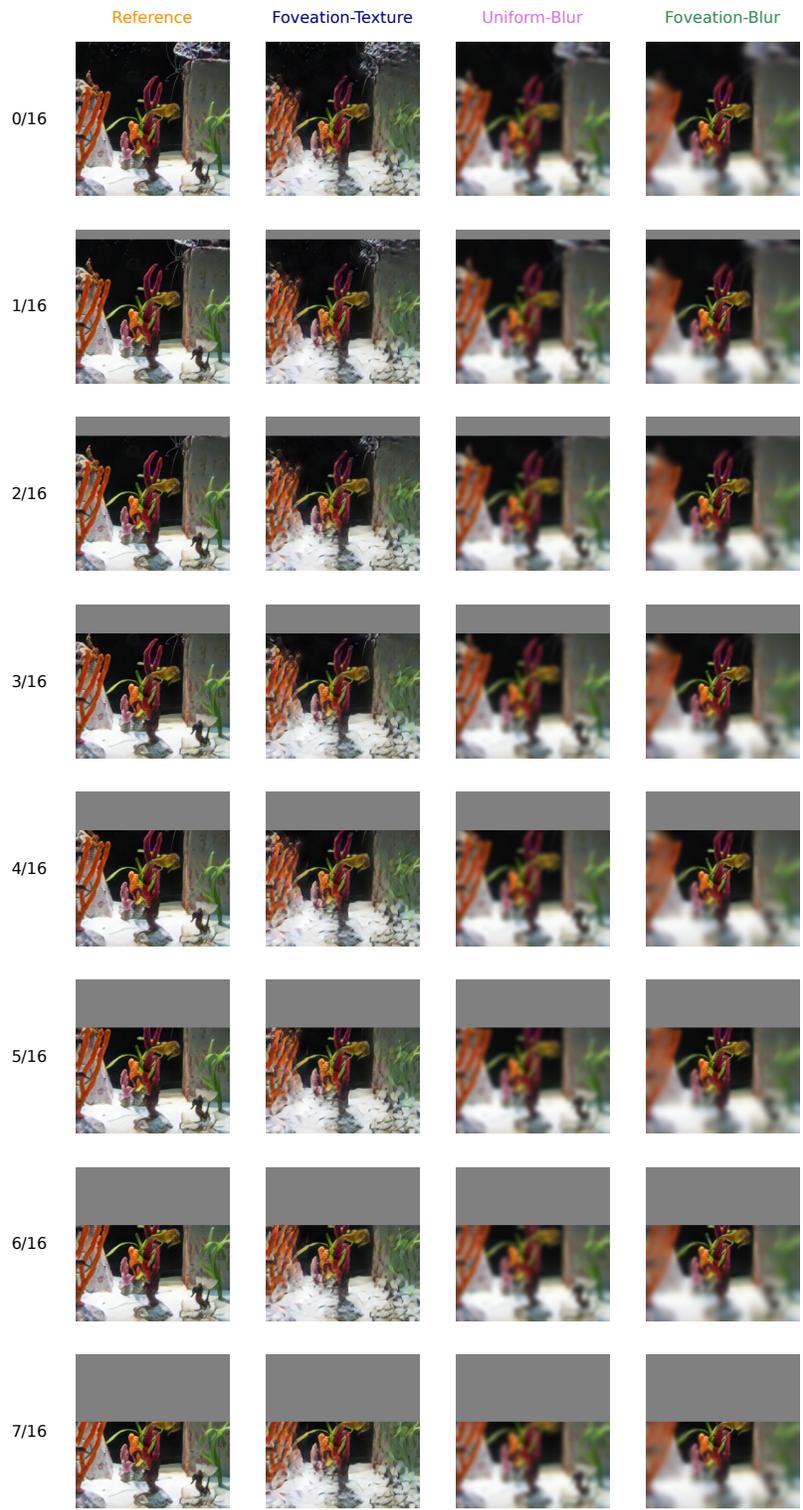


Figure 30: Top2Bottom Occlusion Sample Stimuli.

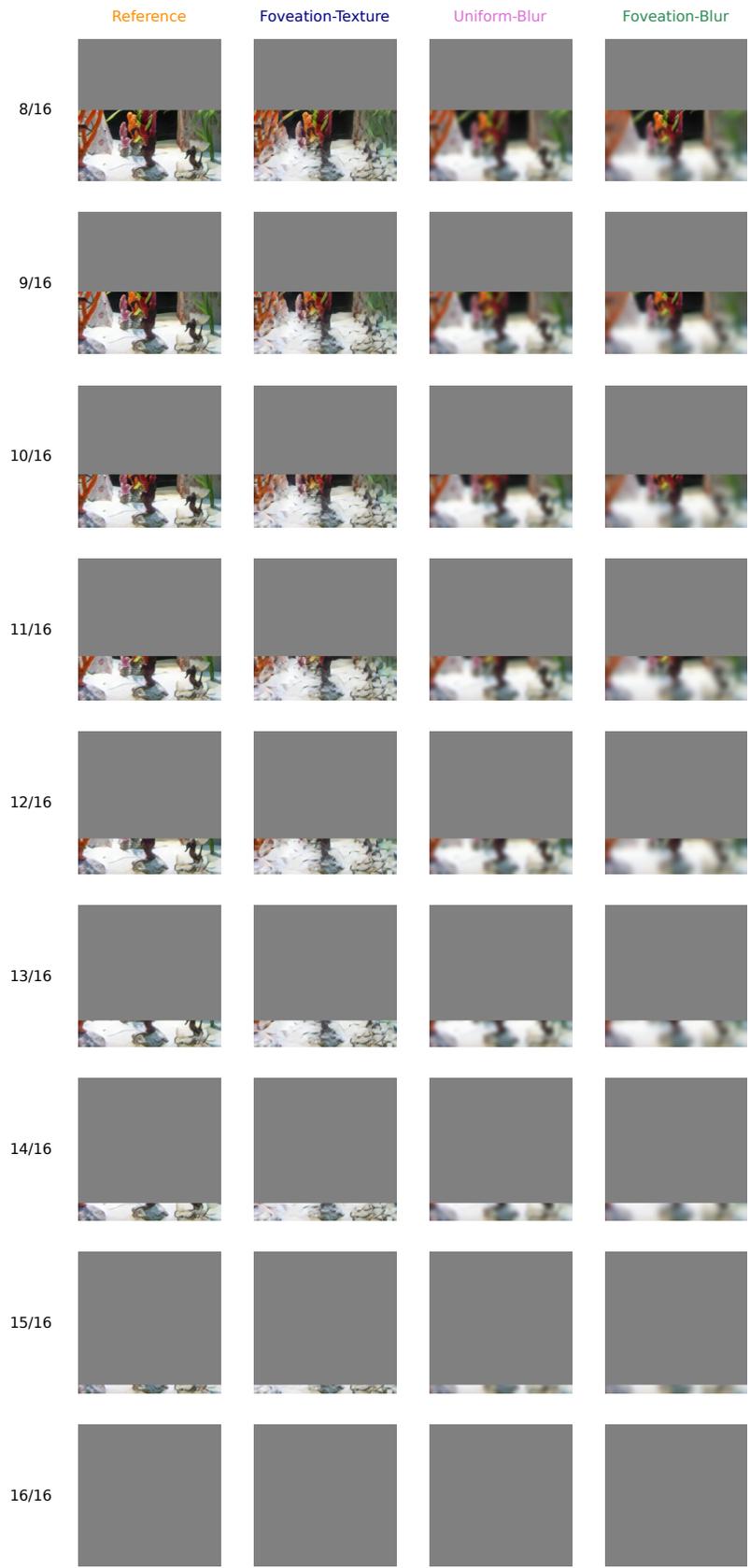


Figure 31: Top2Bottom Occlusion Sample Stimuli.

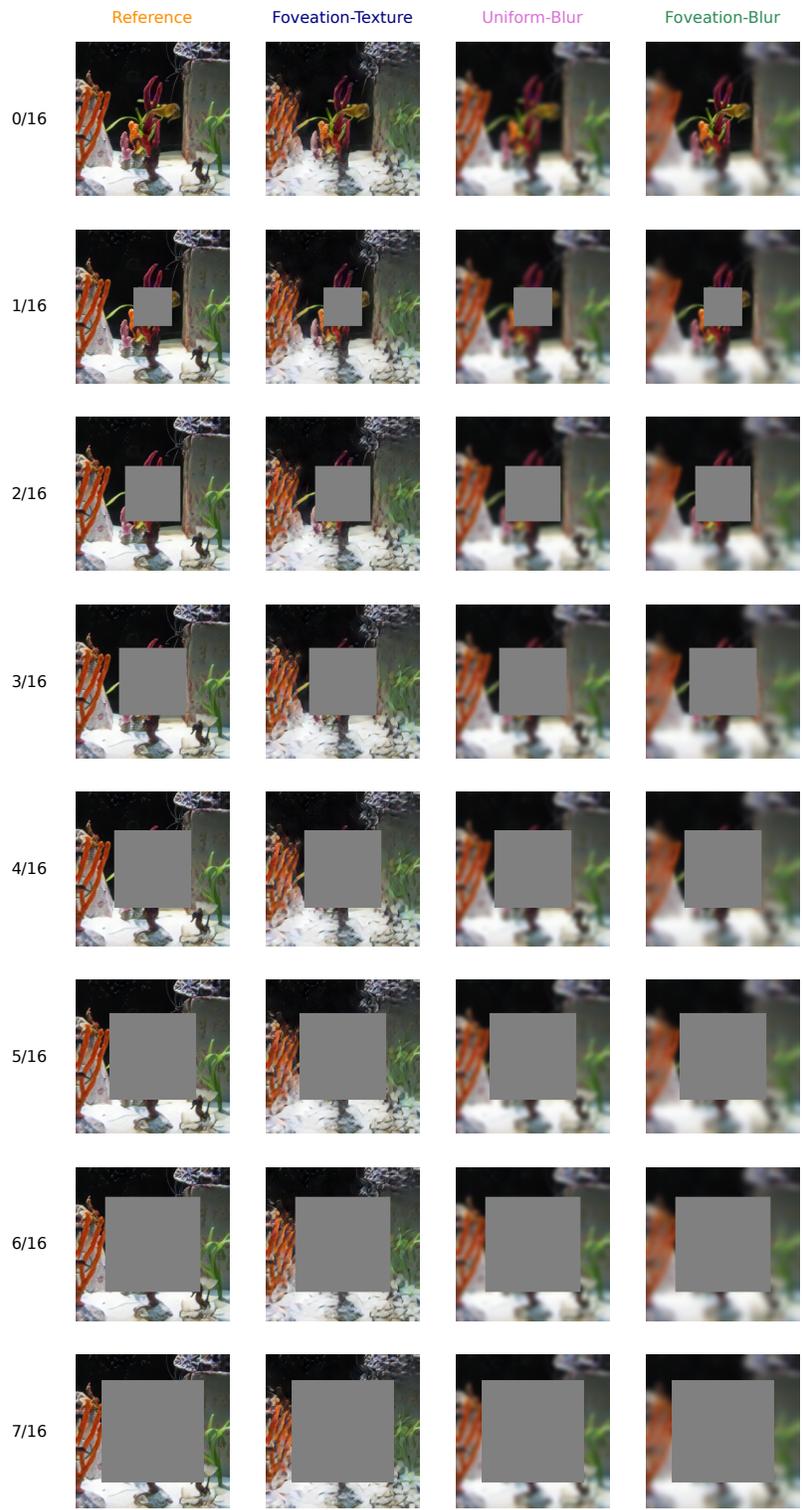


Figure 32: Scotoma Occlusion Sample Stimuli.

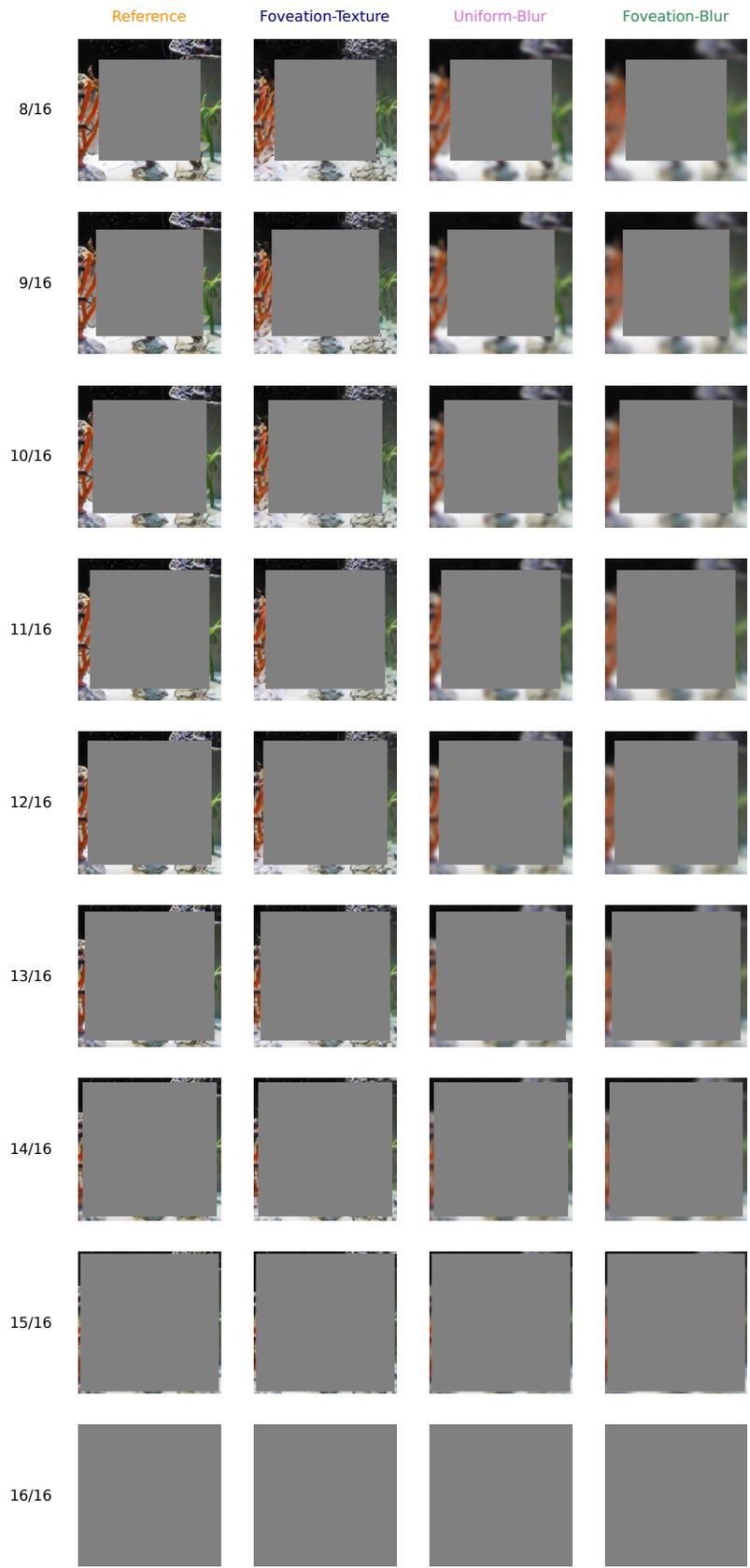


Figure 33: Scotoma Occlusion Sample Stimuli.

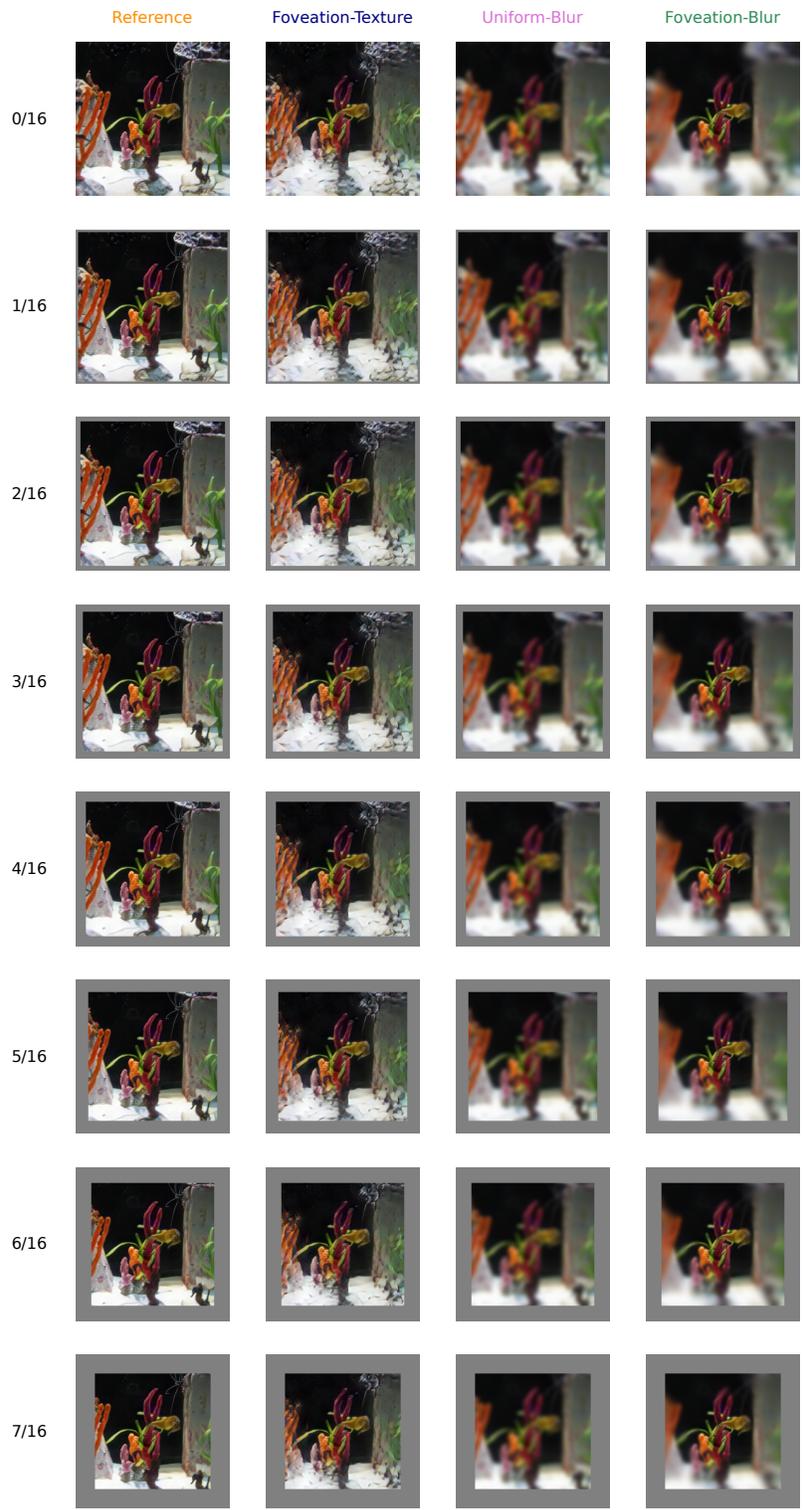


Figure 34: Glaucoma Occlusion Sample Stimuli.

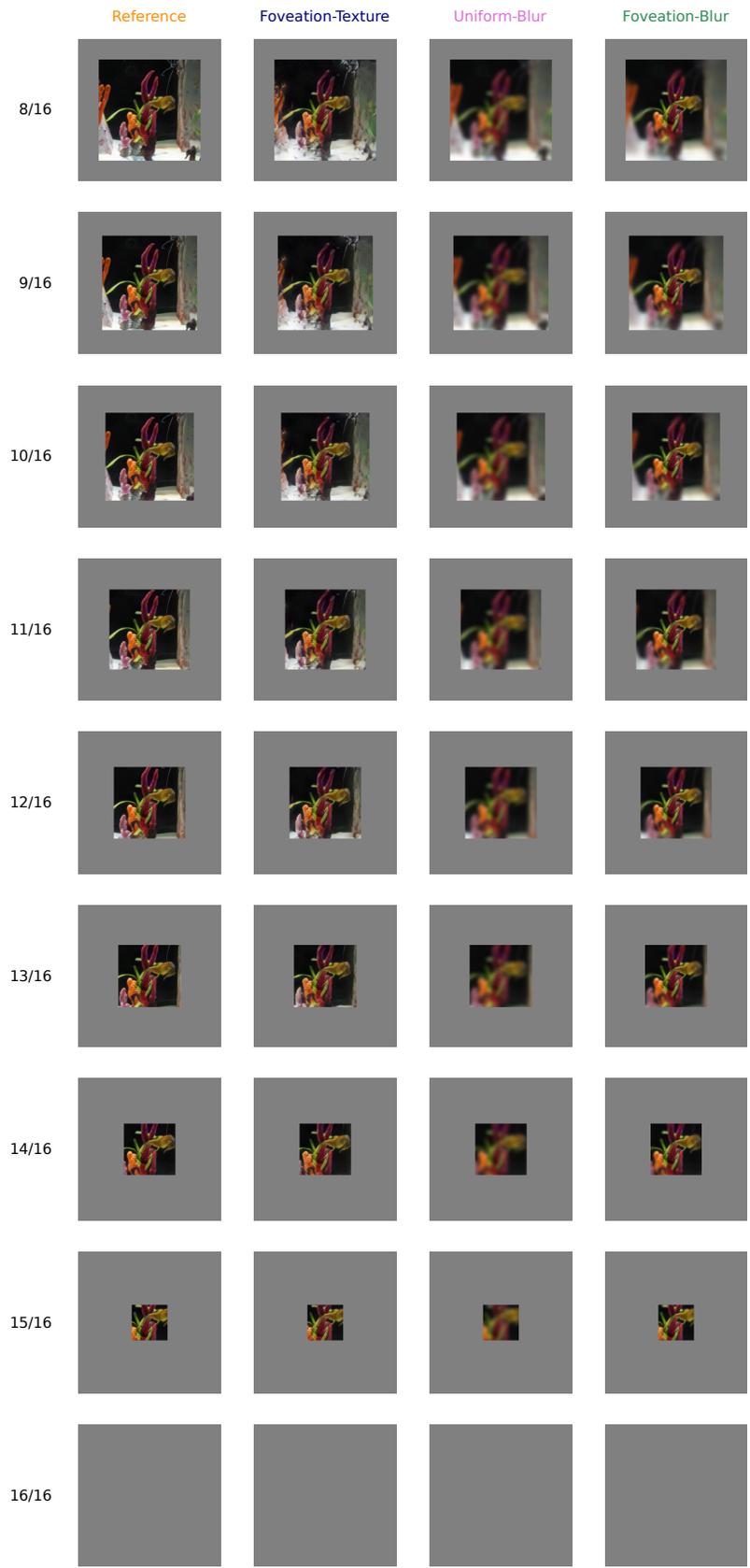


Figure 35: Glaucoma Occlusion Sample Stimuli.

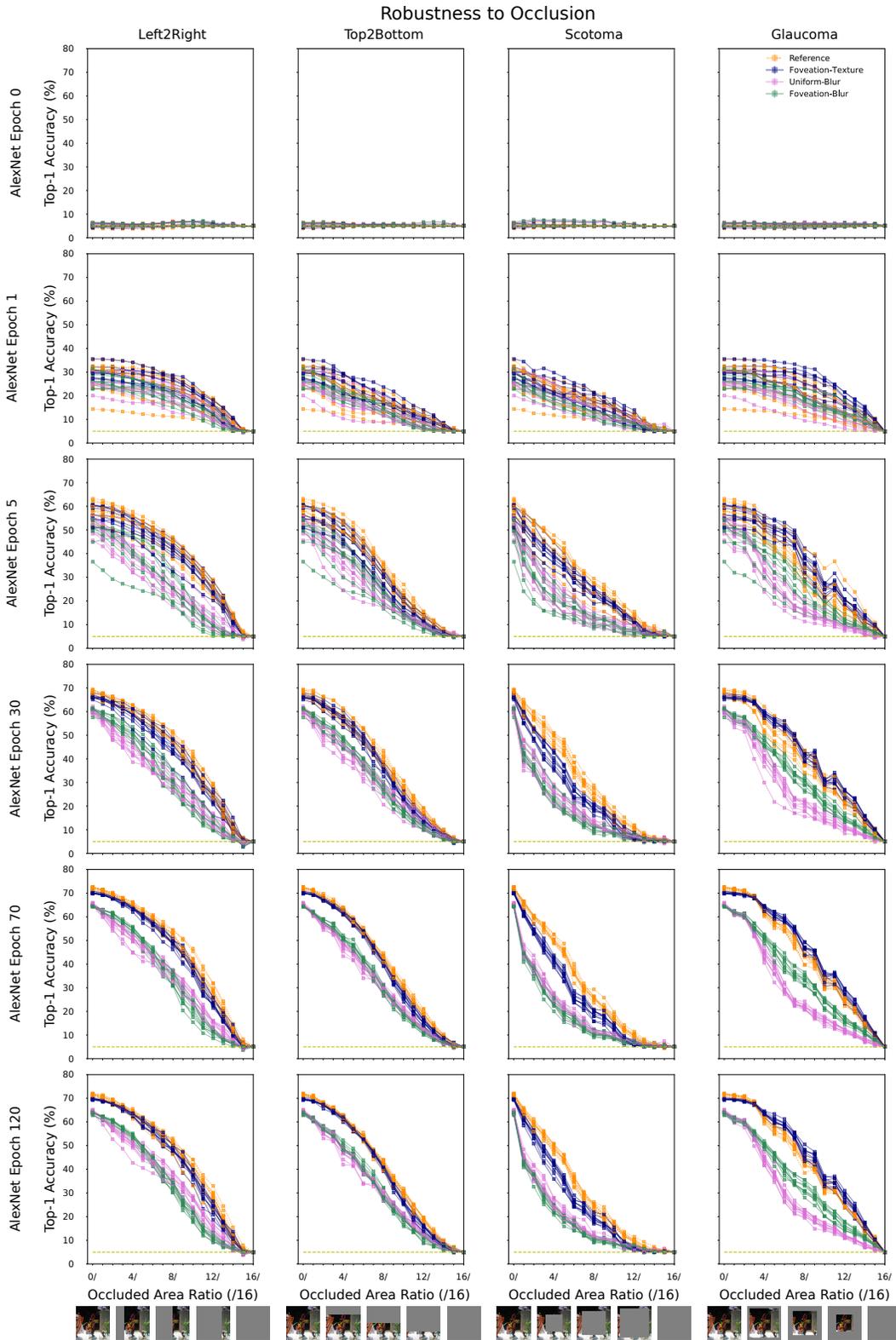


Figure 36: Aggregate Robustness to Occlusion plots for AlexNet as  $g(o)$  after epochs 0, 1, 5, 30, 70, 120.

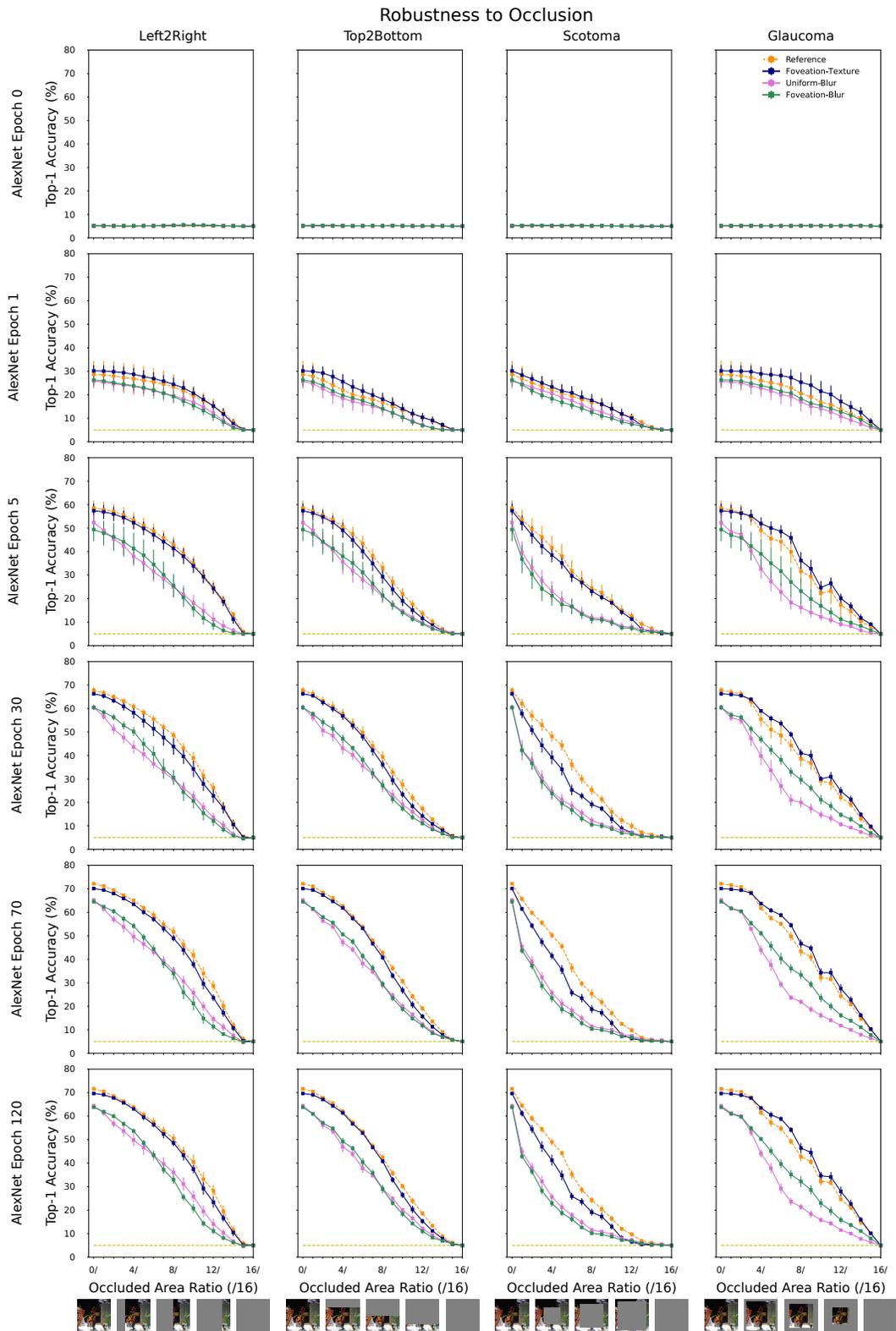


Figure 37: Aggregate Robustness to Occlusion plot for AlexNet as  $g(\circ)$  after epochs 0, 1, 5, 30, 70, 120.

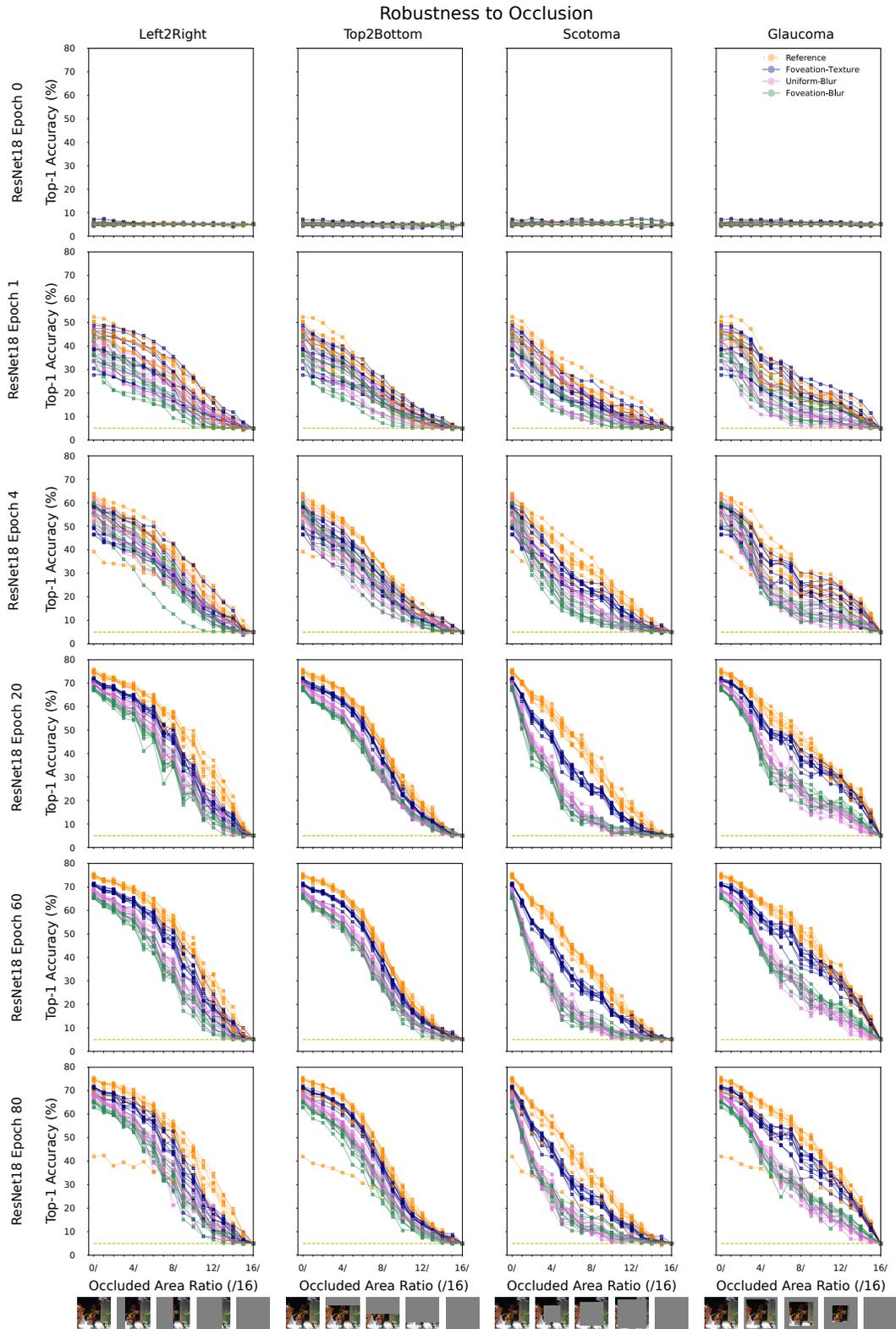


Figure 38: Individual Robustness to Occlusion plots for ResNet18 as  $g(\circ)$  after epochs 0, 1, 4, 20, 60, 80.

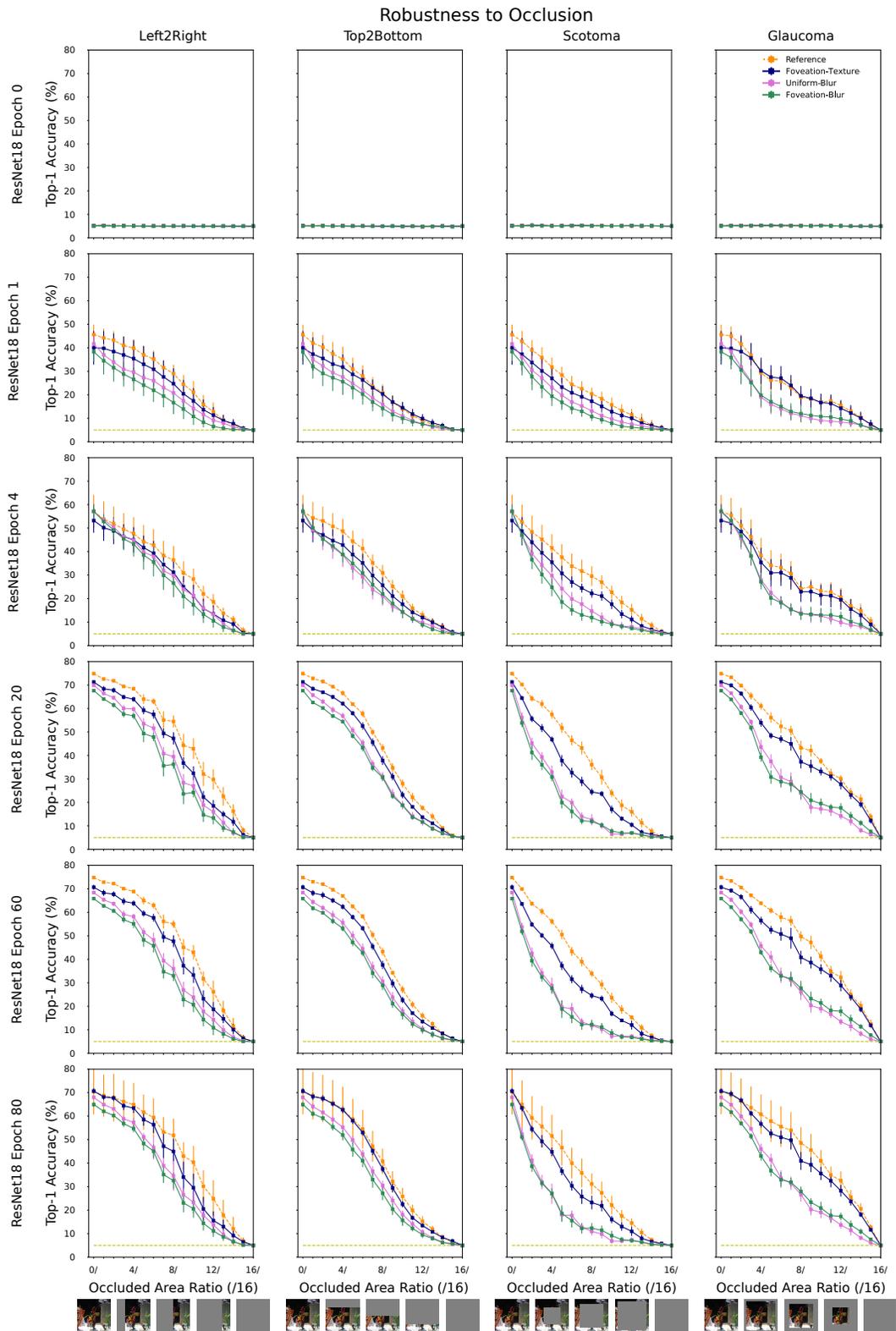


Figure 39: Average Robustness to Occlusion plot for ResNet18 as  $g(\circ)$  after epochs 0, 1, 4, 20, 60, 80.

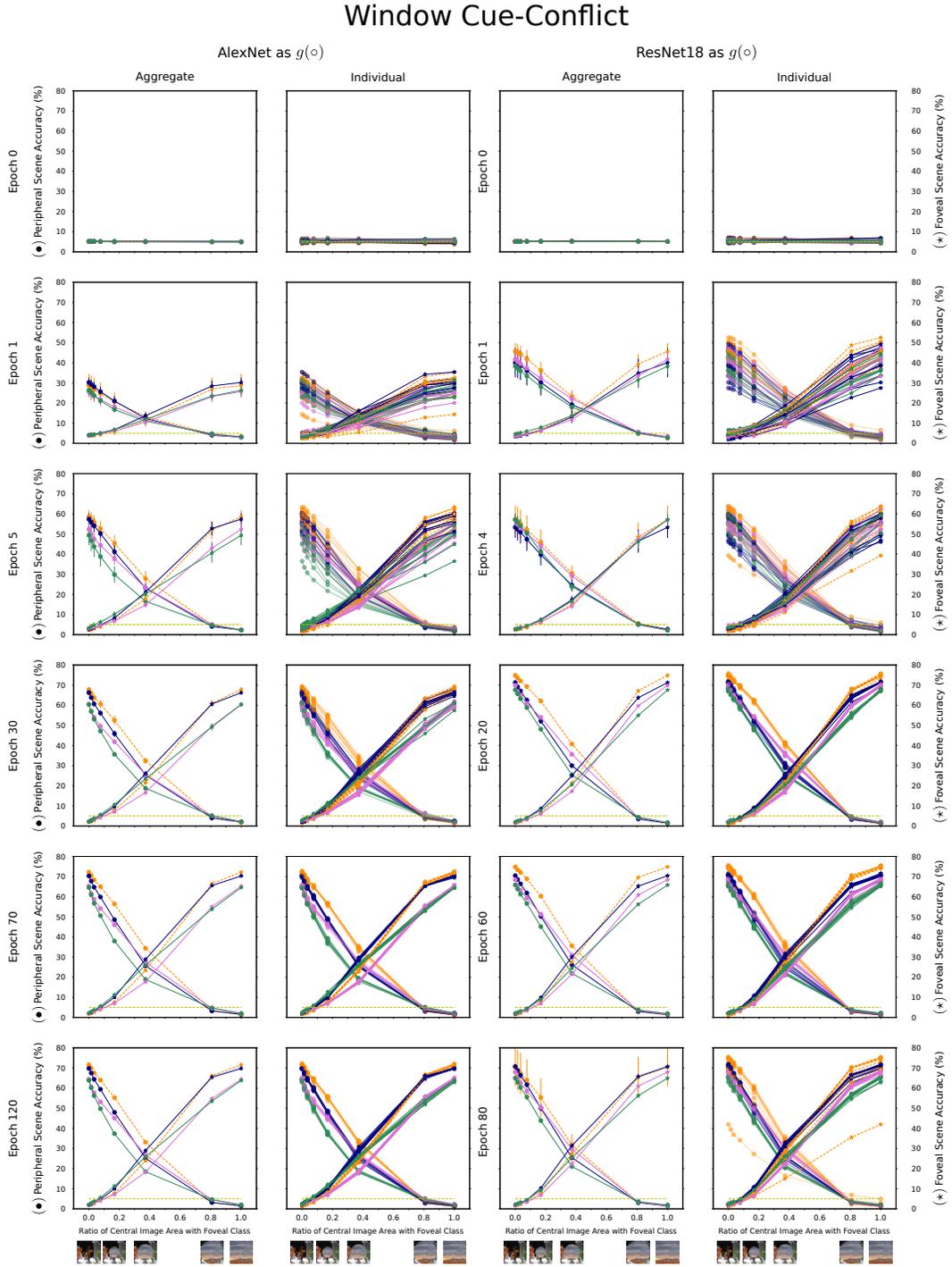


Figure 40: Aggregate and Individual Window Cue-Conflict plots for AlexNet and ResNet18 as  $g(\circ)$  after epochs 0, 1, 5, 30, 70, 120 and 0, 1, 4, 20, 60, 80. respectively

# Window Cue Conflict

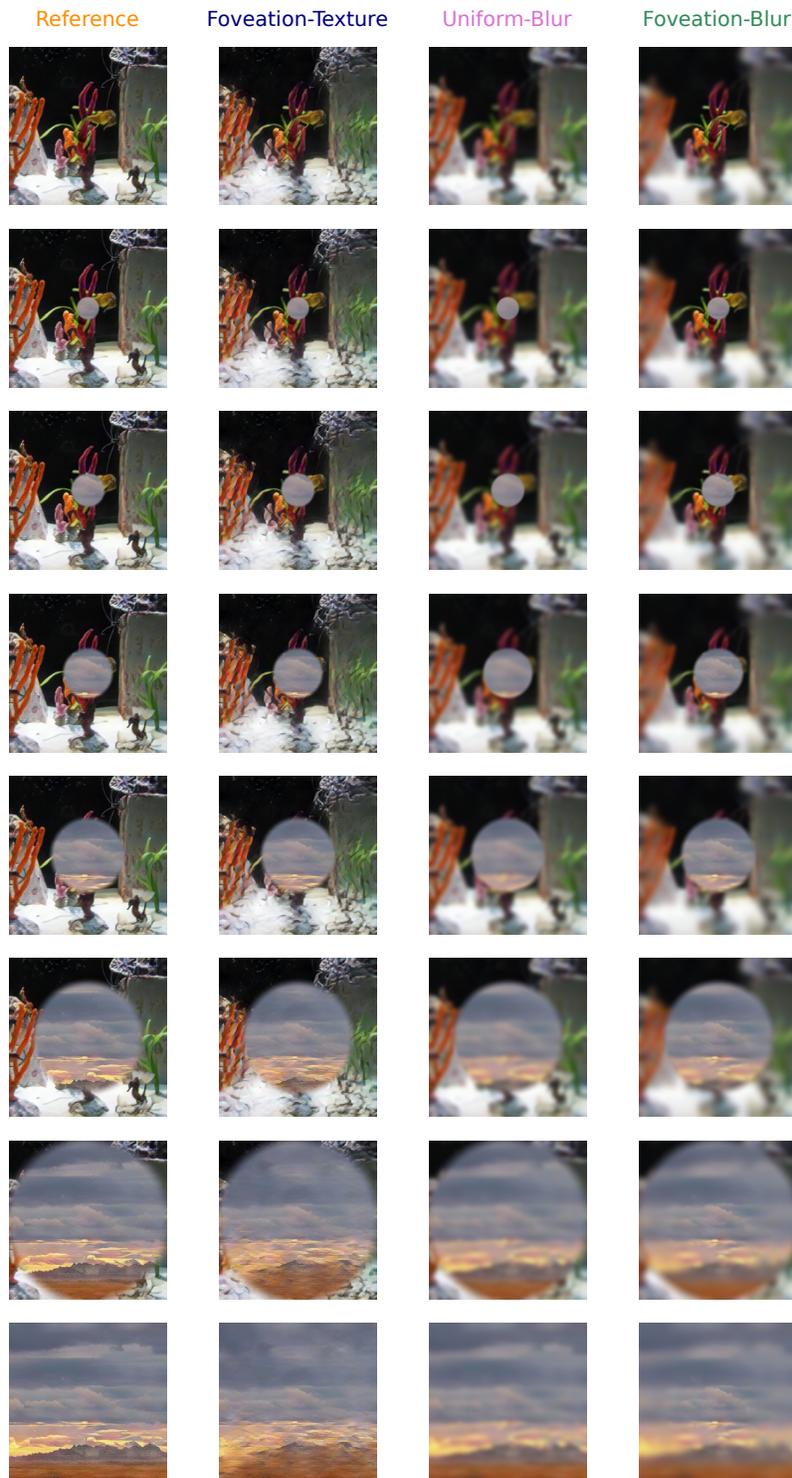


Figure 41: Sample Window Cue Conflict Stimuli.

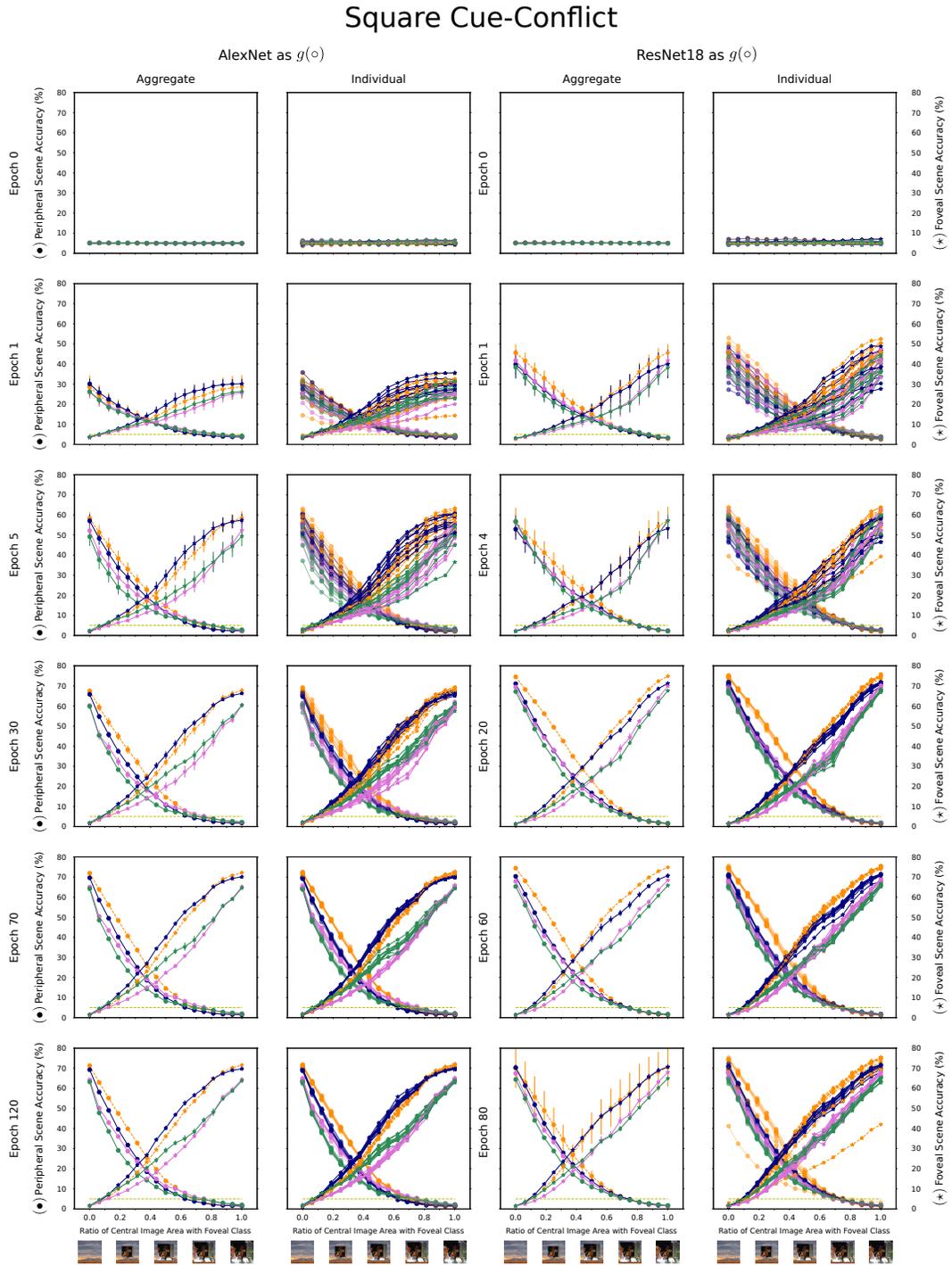


Figure 42: Aggregate and Individual Square Cue-Conflict plots for AlexNet and ResNet18 as  $g(o)$  after epochs 0, 1, 5, 30, 70, 120 and 0, 1, 4, 20, 60, 80 respectively

# Square Cue Conflict

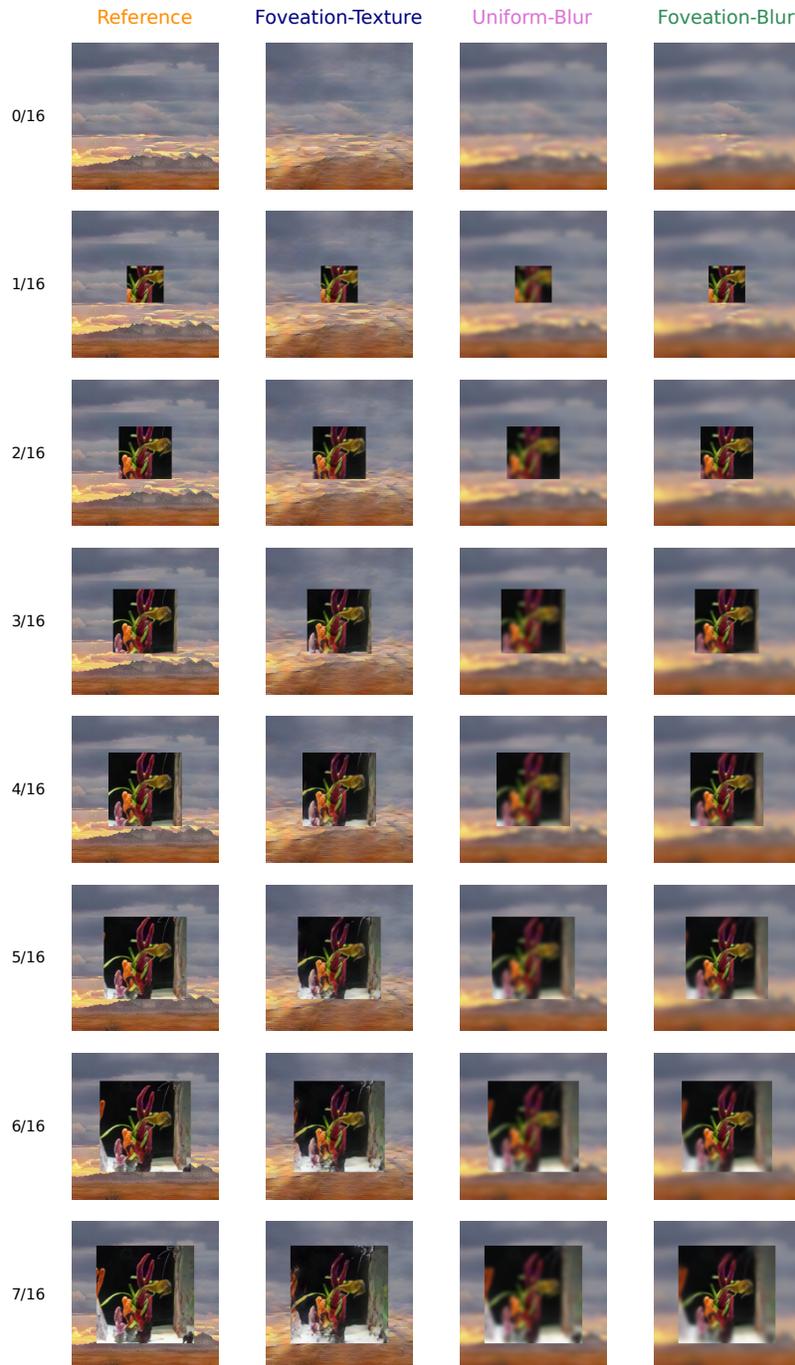


Figure 43: Sample Square Cue Conflict Stimuli.

# Square Cue Conflict

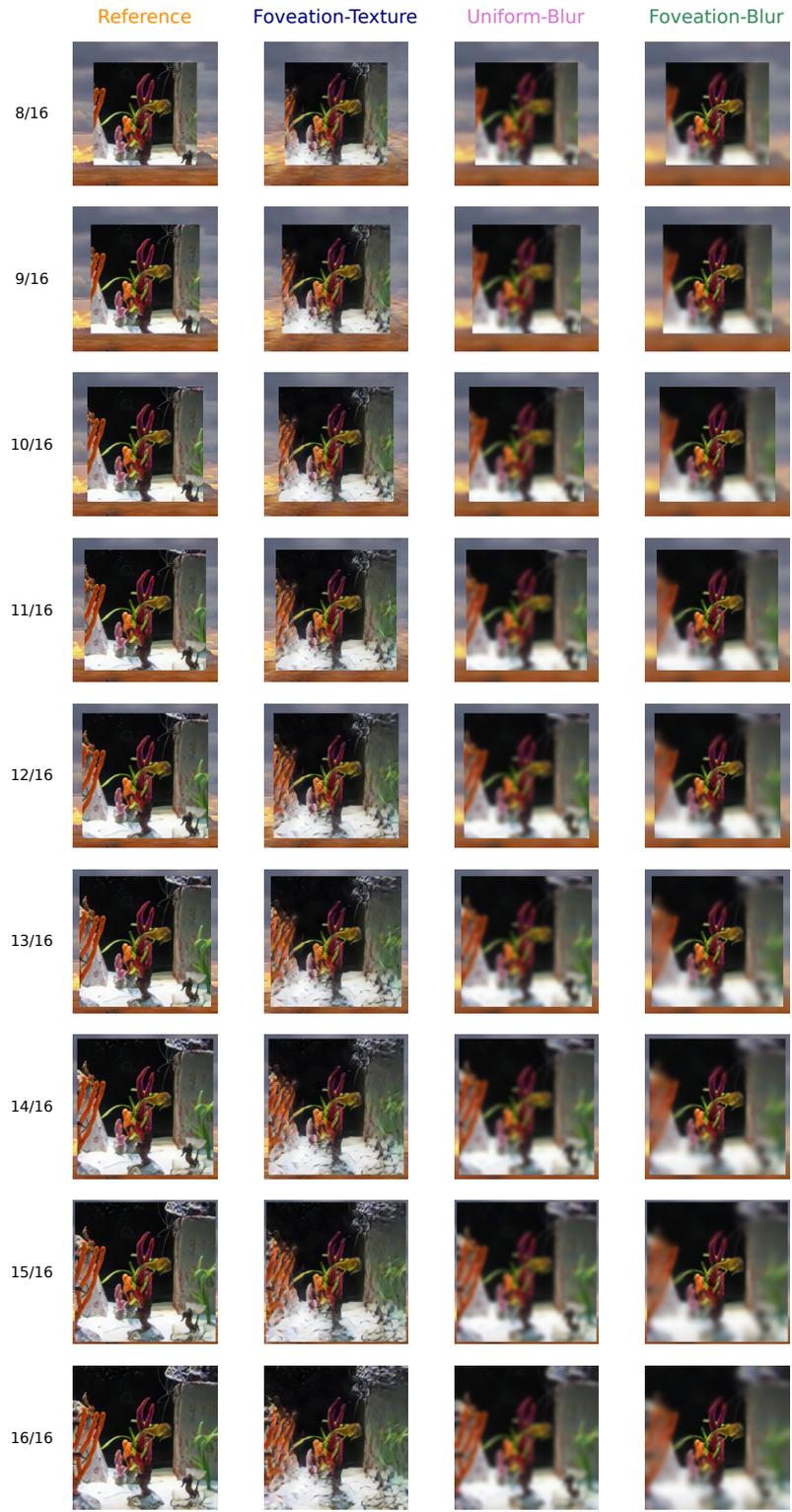


Figure 44: Sample Square Cue Conflict Stimuli.

696 **M Code and Data**

697 All Code and Image Databases for replicability is available for download here: Code and Data

698 Code and Data will also be released on GitHub.

## 699 **N Differences from Previous Manuscript Versions**

700 [Added; this submission] Improved training and convergence of stage 2 neural networks. AlexNet +  
701 ResNet18 now have scheduled learning rates, weight decay and Nesterov momentum when trained  
702 with SGD for each image distribution.

703 [Added; this submission] High Pass and Low Pass Spatial Frequency experiments for grayscale  
704 stimuli as suggested in round of review from ICML 2021.

705 [Added; this submission] Square Uniform cue-conflict experiment to re-verify center image bias as  
706 suggested in round of review from ICML 2021.

707 [Added; this submission] Left2Right & Top2Bottom experiments moved to main body.

708 [Added; this submission] both Aggregate and Individual plots for each system to qualitatively check  
709 for variance in individual network differences.

710 [Added; this submission] Visualization of filters from the first convolutional layer for each system.

711 [Added; this submission] Additional use of Mean Square Error, Mutual Information and 10 more  
712 IQA metrics from Ding et al. (2020) as supporting Image Quality Assessment metrics to compare to  
713 SSIM for Rate-Distortion Optimization as suggested through reviews in ICML 2021.

714 [Added; for ICML 2021] Sketched proof of Reference being a Perceptual Upper Bound.

715 [Added; for ICLR 2021] Rate-Distortion Optimization procedure to compute Uniform-Blur and  
716 Foveation-Blur.

717 [Added; for ICLR 2021] Improved written clarity, and re-emphasized focus of paper on Foveation  
718 w.r.t Machines (not humans – which caused misinterpretation and rejection from NeurIPS 2020).

719 [Removed; for ICML 2021] Claims about Foveation-Texture inducing a shape bias (currently parallel  
720 work) from Submission to ICLR 2021.

721 [Removed; for ICLR 2021] Experiments about data-augmentation via eye-movements + classical  
722 augmentation schemes such as random cropping + rescaling (parallel work) from Submission to  
723 NeurIPS 2020.

724 [Bug fix; this submission] Even runs were continuations of odd runs in 10 run randomization across  
725 networks due to bug w.r.t distributed parallelization, from submission to ICML 2021. Note: General  
726 pattern of results did not change, and all curves have been re-plotted.

727 [Previous paper scores, decisions, meta-reviews and author opinions:]

- 728 1. NeurIPS 2020: 5,4,3,4 (reject: Unanimous bad reviews, focus of all reviewers was a need for  
729 human psychophysical studies even though the paper was not about human vision – which  
730 prompted us to re-write the paper to make our goals more clear: “What is the impact of  
731 texture-based foveation on machines?; and what can these results tell us about the human  
732 visual system – mainly the visual periphery that has texture-like computation – from a  
733 computational perspective?”. [fixed])
- 734 2. ICLR 2021: 7,7,7,3,5 (reject: Mixed reviews & needed to tone down claims and re-  
735 emphasize why texture was used in the periphery [fixed])
- 736 3. ICML 2021: 3 Weak Rejects (1 Accept + 1 Weak Accept downgraded their scores post-  
737 rebuttal suggesting the work was not a good fit for ICML), 1 Strong Reject (withdrawn: we  
738 caught a bug post-rebuttal phase in the process of code/data release that did not affect the  
739 main pattern or results, but required re-running all the experiments and overall improved  
740 the current version of the paper. Reviewers suggested different IQA metrics beyond SSIM  
741 to make comparisons for matched perceptual compression (we added MSE, Mutual In-  
742 formation, and 10 more IQA metrics). This has been added and addressed in our current  
743 version.).

744 A recurrent theme in negative reviews has been that the model does not (in its current state) advance  
745 the state of the art by beating a baseline. While these hallmarks are pivotal for computer vision, our  
746 goal is complimentary, as we would like to model, and understand the representational consequences  
747 – beyond accuracy – of spatially-adaptive computation in machines inspired by the foveated visual  
748 system of humans.