

In-N-Out: Towards Good Initialization for Inpainting and Outpainting

Changho Jo
timegate@kaist.ac.kr

Woobin Im
iwb@kaist.ac.kr

Sung-Eui Yoon
sungeui@kaist.edu

School of Computing,
Korea Advanced Institute of Science
and Technology (KAIST),
Daejeon, Korea

1 Experiment Settings

1.1 Network Architecture

In our main paper, except for the task with irregular masks (Section 4.4), we evaluate In-N-Out using a variant of Semantic Regeneration Network (SRN) [8]. We use SRN as our base network since it demonstrates its performance on diverse content extrapolation tasks. Specifically, it shows its performance even to extrapolation of the human body, texture synthesis, and morphing. To bring the context better from the visible region, SRN introduces the context normalization module and it was trained with context adversarial loss and relative spatial variant loss. In our work, we slightly modify SRN to accommodate not only the extrapolation task but various tasks by removing the margin mechanism and let the masked image go directly into the input of the network; it allows us to use various mask shapes as well as rectangular masks. The illustration of SRN [8] is shown in Figure 1 and the illustration of our variant is shown in Figure 2.

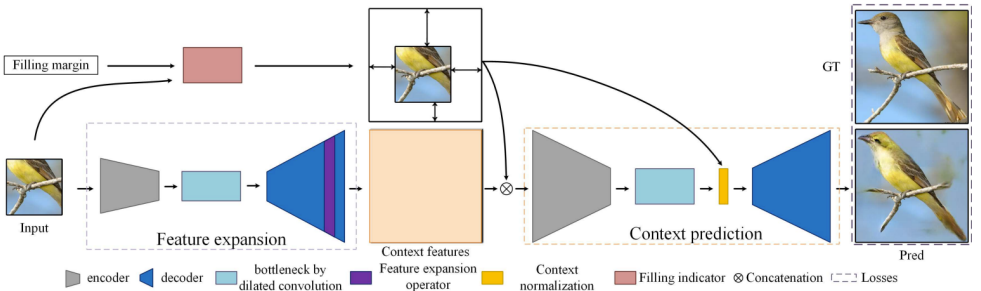


Figure 1: Semantic Regeneration Network (SRN) [8]. This figure is brought from [8].

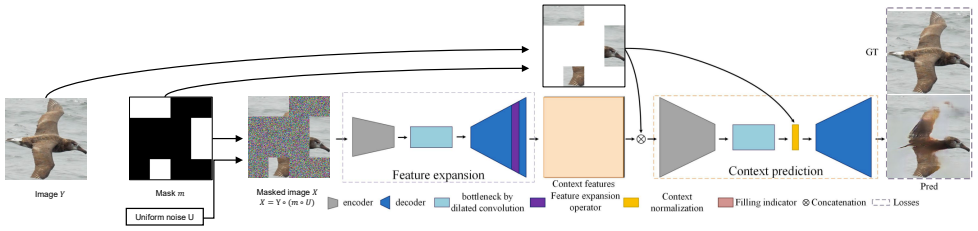


Figure 2: Our variant of SRN. This figure is based on [8].

Method	Easy	Difficult	Extreme
Baseline	33.83	52.46	142.40
In-N-Out	30.48	45.36	128.10

(a) Outpainting task results for each difficulty, on CelebA-HQ dataset [8]

Method	PSNR	SSIM	FID
SRN [8]	18.22	0.513	-
Baseline (Figure 2)	19.46	0.709	33.30
PSL [9]	18.78	0.716	35.86
In-N-Out	19.52	0.711	30.17

(b) Outpainting task compared to baseline(Figure 2), on beach dataset [9].

Table 1: (a) FID results for each difficulty on outpainting task. If the visible area is less than 20% of a masked image, it is classified as an extreme task. Masked images with less than 40% of the visible area are classified as difficult tasks, and others are classified as easy tasks. (b) Quantitative results on outpainting task (beach dataset [9]). Our baseline result is added.

1.2 Masks Used in Experiments

For the outpainting task on CelebA-HQ dataset (Section 4.2), we use rectangular masks with random sizes for the test set. Specifically, we use face locations (used to crop) recursively to mask the cropped images, which enables various mask sizes and ratios. The detailed process can be found in our code. The distribution of the test mask area are shown in Figure 3, and test results according to the difficulty are shown in Table 1a. If the visible area is less than 20% of a masked image, it is classified as an extreme task. Masked images with less than 40% of the visible area are classified as difficult tasks, and others are classified as easy tasks.

For environment map estimation (Section 4.5), we use a b-spline function for the masks of train set and test set; it makes natural-looking partial panoramas. The b-spline function is configured with 4 random points including the center of the image, where each point shows 60-degree FoV. As a result, we could generate images that are similar to input images in lighthouse [4]. Note that In-N-out is trained on inverse masks, where the visible region and masked region are swapped, in training steps. The illustration of partial panorama images are shown in Figure 4, and all illustrations of the train set and test set used in our experiments are shown in Figure 5.

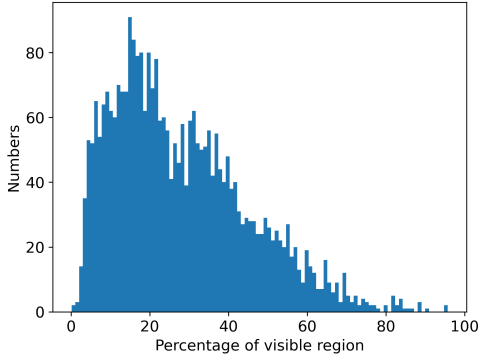


Figure 3: Distribution of test mask area in Section 4.2. This represents the ratio of visible region to the image. We tested our method on masks of various sizes.

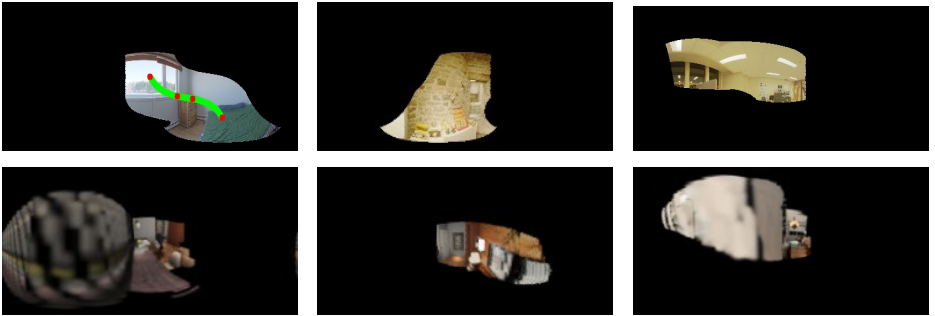


Figure 4: Examples of partial panorama images. Up: Input of our work. Down: Input images used in lighthouse [14]. We use a b-spline function with 4 random points including the center to make natural-looking partial panoramas.

1.3 Setup for Experiments

For Section 4.1 and 4.2, since training iterations are not specified in the existing paper (SRN [8]), we chose (pretrain 40k + train 40k) and (45k + 45k) for each experiment as the loss curves are converged. For Section 4.3, we chose (40k+ 40k), and for Section 4.5, we chose (45k+ 45k). For Section 4.4, we followed the training procedure specified in MEDFE [9] and Shift-Net [10] (30 epochs for each baseline) while In-N-Out does outpainting for half of the iterations (15 epochs: inverse mask, 15 epochs: fine-tune).

For the experiments in Section 4.4, We follow the training and testing split of the dataset, and we use resized images to 256×256 .

For the experiment in Section 4.5, we follow the official split of LAVAL Indoor HDR Dataset [11], and we use resized images to 128×256 (respectively height and width). We use random crop, random flip, and gray-scaled images as data augmentation for our training. We use 45,000 iterations for the training steps and 45,000 iterations for the fine-tuning with batch size 4.


















	Train mask		Train mask (inversed)	
Inpainting task on CUB200 dataset				
Outpainting task on CelebA-HQ dataset				
Outpainting task on beach dataset				
Inpainting task on Paris Street-View dataset				
Environment map estimation task on LAVAL HDR Indoor Dataset				

Figure 5: Illustration of used masks.

1.4 Details of iterations

To choose N and K, we selected appropriate iterations to steps when the training curve is converged. For section 4.1, we found that both baseline and In-N-Out showed similar trends (as shown in Figure 6) and didn’t improve much after additional 40000 iterations. Also, for section 4.2, both baseline and In-N-Out showed similar trends after additional 50000 iterations. Additionally, we did experiments with (Section 4.1, N=20000, K=40000), and (Section Sec. 4.2, N=25000, K=60000) as shown in Figure 7. We have consistently good results with using these different N and K.

2 Additional Results

We provide test results (PSNR and SSIM) of each iteration in Figure 8 for image inpainting and image outpainting, similar to Figure 1 in our main paper. More visual comparisons on image inpainting, image outpainting are given in Figure 9, 10, and 11. Also, we provide our baseline for outpainting task on beach dataset in Table 1b.

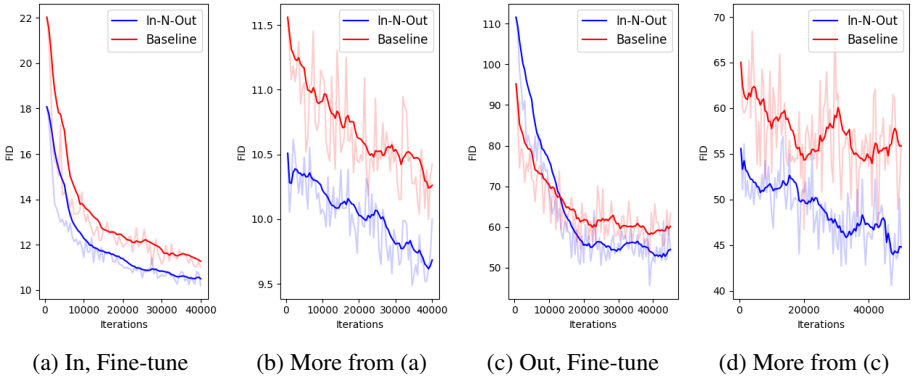


Figure 6: **(b)**: More iterations from Section 4.1. **(d)**: More iterations from section 4.2. Each graph shows FID (Fréchet inception distance) [14] during each stage.

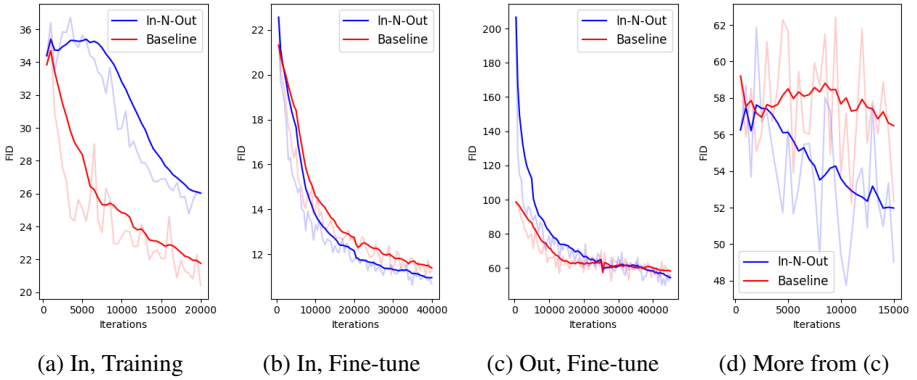


Figure 7: **(a)-(b)**: Section 4.1 with ($N = 20000$, $K = 40000$). **(c)-(d)**: Section 4.2 with ($N = 25000$, $K = 60000$). Each graph shows FID (Fréchet inception distance) [14] during each stage.

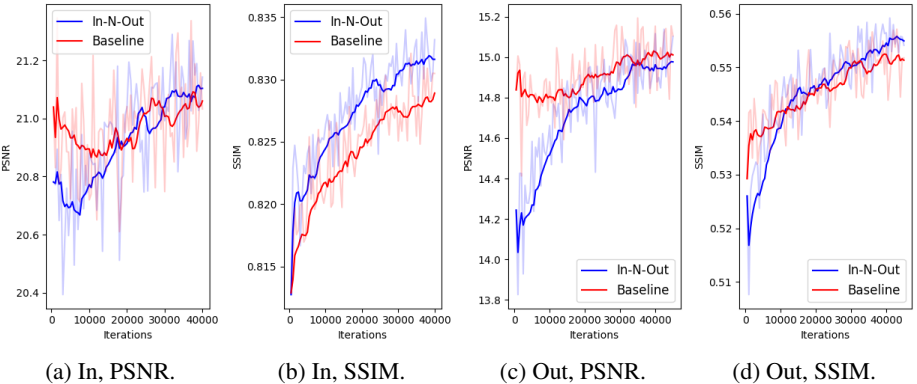


Figure 8: **(a)-(b)**: PSNR and SSIM of each iteration in fine-tuning stage, on inpainting task. **(c)-(d)**: PSNR and SSIM of each iteration in fine-tuning stage, on outpainting task. In-N-Out also shows better performance in terms of SSIM.



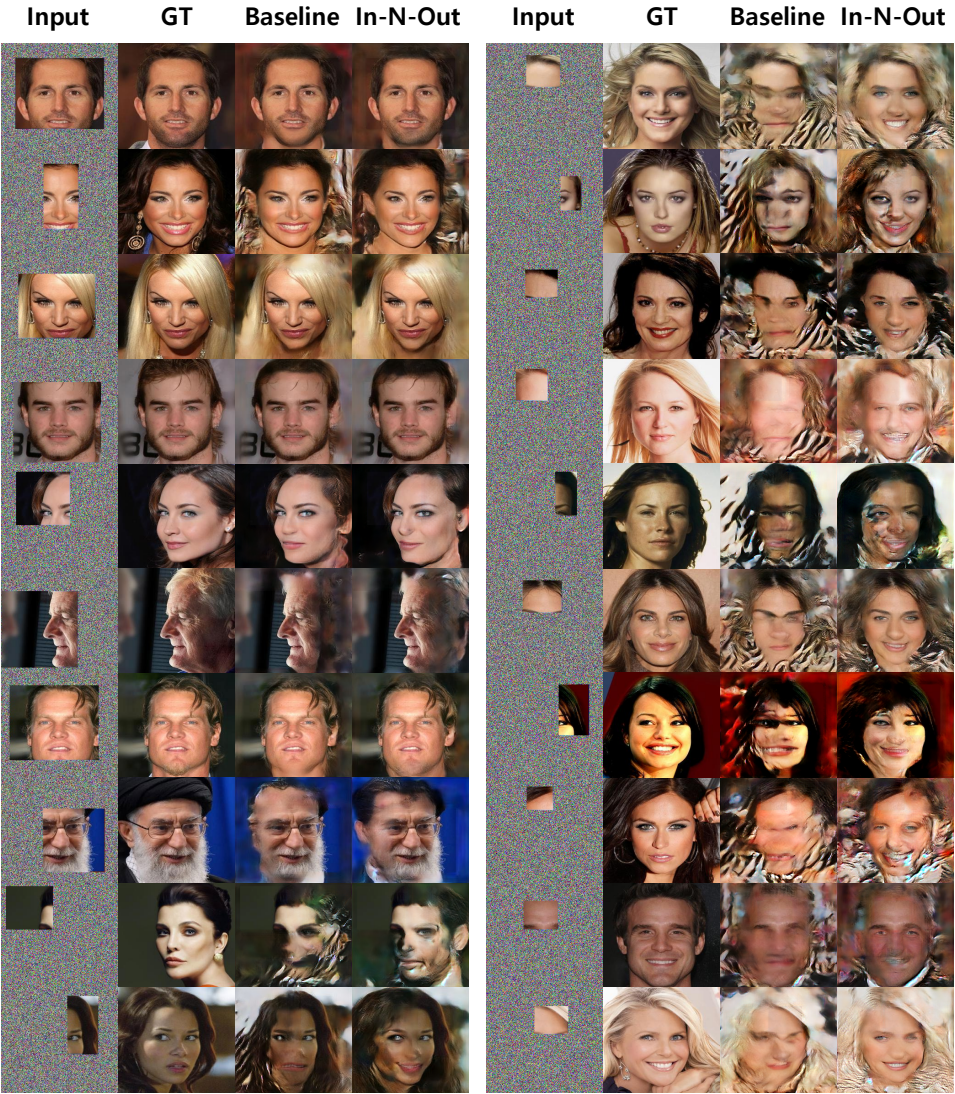


Figure 10: More results on outpainting task (CelebA-HQ dataset [14]).



(a) Outpainting task compared to PSL [9].

Figure 11: More results on outpainting task (beach dataset [9]), compared to PSL [9].

References

- [1] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gamberetto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Trans. Graph.*, 36(6), November 2017. ISSN 0730-0301. doi: 10.1145/3130800.3130891. URL <https://doi.org/10.1145/3130800.3130891>.
- [2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.
- [3] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=Hk99zCeAb>.
- [4] Kyunghun Kim, Yeohun Yun, Keon-Woo Kang, Kyeongbo Kong, Siyeong Lee, and Suk-Ju Kang. Painting outside as inside: Edge guided image outpainting via bidirectional rearrangement with progressive step learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2122–2130, 2021.
- [5] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Proceedings of the European Conference on Computer Vision*, 2020.
- [6] Mark Sabini and Gili Rusak. Painting outside the box: Image outpainting with gans. *arXiv preprint arXiv:1808.08483*, 2018.
- [7] Pratul P Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan T Barron, Richard Tucker, and Noah Snavely. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8080–8089, 2020.
- [8] Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1399–1408, 2019.
- [9] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.
- [10] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In *Proceedings of the European conference on computer vision (ECCV)*, pages 1–17, 2018.