Evaluating High-Resolution Vessel Mask-to-Fundus Translation under Non-Monotonic GAN Dynamics

Jiaqiang Yang

College of Liberal Arts and Sciences, University of Connecticut, Storrs, CT, 06269, the United States
tbg24003@uconn.edu

Abstract. Clinically viable retinal fundus synthesis from vessel masks requires photorealistic appearance while maintaining anatomical agreement with the input structure. The task is treated as a structure-swe nsitive, high-resolution (512×512) paired translation problem, with a Pix2Pix-style model as a baseline. Evaluation uses distribution metrics (FID, KID) alongside paired measures (LPIPS, MS-SSIM) to separate set-level realism from target-aligned fidelity. Because adversarial training can vary substantially across epochs, several late-stage checkpoints are compared. The checkpoint with the best FID/KID often differs from the one with the best LPIPS/MS-SSIM, so checkpoint choice depends on whether the goal is realism for augmentation or stricter per-sample correspondence. Qualitative inspection is supported by an auxiliary vessel segmenter that visualizes the input mask, generated fundus image, and re-segmented vessels in a single layout. All experiments follow a fixed protocol on the combined DRIVE+CHASE DB1 training set (48 image–mask pairs) as an in-sample reference. Later checkpoints (e.g., epoch 190) show fewer texture/color artifacts and more consistent vessel structure than earlier ones (e.g., epoch 70), indicating that training stage can strongly affect the realism–fidelity balance under limited paired supervision.

Keywords: Retinal fundus synthesis; Vessel mask; Conditional GAN; Pix2Pix; Evaluation metrics.

1 Introduction

Generative adversarial networks (GANs) can model complex image distributions and generate realistic samples, which have motivated their use in medical image synthesis in recent years [1]. In retinal fundus imaging especially, GANs have been used for applications like limited data augmentation and robustness analysis [2, 3]. However, when synthesis is conditioned on an anatomical mask (e.g., a vessel network), focusing only on visual plausibility will be a huge disaster. The generated fundus image must also preserve the vascular geometry specified by the input, while one reason for clinical misinterpretation is that an incorrect quantitative result is from a structural deviation.

Mask-conditioned generation can be performed using conditional GANs (cGANs), employing the mask as a direct conditioning input during generation [4]. But in vessel-

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_98

guided fundus synthesis, existing studies typically emphasize the image quality due to the direct requirements from surgical settings, whereas precise assessment of mask-conditioned structural fidelity at high resolution—and the robustness of adversarial training, in this context—is less frequently reported [3]. For image-to-image transformations, paired methods such as Pix2Pix use pixel alignment supervision to facilitate correspondence between samples [5], while unpaired methods such as CycleGAN remove the alignment requirement, which may reduce target-specific fidelity [6]. The problem of training stability has also been addressed by other adversarial objectives, including WGAN and WGAN-GP, where gradient penalty is used to regularize the discriminator [7, 8].

This paper designs a paired cGAN framework based on these considerations. The framework uses 512×512 vessel-mask-to-fundus translation under a fixed training and reporting protocol, evaluates both distribution-level realism and mask-aligned fidelity, and reports training behavior across checkpoints. All evaluations are conducted in-sample on the training set to provide a controlled internal reference.

This study proposes a controlled mask-to-fundus synthesis formulation with a fixed evaluation protocol. It specifies a paired training pipeline from mask to fundus, and an evaluation suite that includes distribution-level metrics (FID, KID). All the metrics are computed on the same paired set used for training and are intended for within-protocol comparisons rather than out-of-sample generalization [9, 10]

In order to reflect non-monotonic GAN dynamics, multiple checkpoints are reported in the results. Because adversarial optimization is usually non-monotonic, this paper reports multiple late-stage checkpoints under the same training and evaluation protocol and discusses cross-metric trade-offs, instead of relying solely on the final epoch [8, 9].

Structure-aware qualitative protocol using an auxiliary segmenter. A separately trained vessel segmenter supports a four-column layout $(x, m, \hat{x}, \tilde{m})$ for inspection, where x is the real fundus image, m is the ground-truth vessel mask, $\hat{x}$ is the synthesized fundus image, and $\tilde{m}$ is the predicted vessel mask [10].

2 Model and Evaluation Metrics

2.1 Pix2Pix-style Conditional GAN Baseline (Mask→Fundus)

The baseline follows the paired image-to-image translation setting, where a vessel mask $m \in R^{1\times H\times W}$ (single-channel) is used to synthesize a corresponding color fundus image $x \in R^{3\times H\times W}$ (RGB). This study supposes G denote the generator and D denote the discriminator. The generator predicts a synthesized fundus image $\hat{x} = G(m)$.

Input/Output and normalization. Both the RGB fundus image x and the mask m are converted to tensors in [0,1] and linearly scaled to [-1,1] via. The generator outputs using a final tanh activation [5].

Architecture. The generator adopts a U-Net encoder–decoder with skip connections mapping a 1-channel mask to a 3-channel RGB output [10]. Skip connections preserve

fine vessel structure by propagating high-resolution features [10]. The discriminator is a conditional PatchGAN-style convolutional network operating on the channel-wise concatenation [m, x] (4-channel input), producing a patch-wise logit map (no sigmoid inside D) [5].

Objective. The baseline uses a logistic adversarial loss implemented with binary cross-entropy (BCE) with logits (BCEWithLogits), combined with an L_1 reconstruction term [5]. In addition, the implementation optionally includes an edge-consistency penalty computed from grayscale Sobel gradient magnitudes, and a total-variation (TV) regularizer to suppress high-frequency artifacts [5].

Discriminator loss:

$$L_D = BCEWithLogits(D([m,x]),1) + BCEWithLogits(D([m,\hat{x}]),0) \quad (1)$$

Generator loss:

$$L_G = BCEWithLogits(D([m,\hat{x}]),1) + \lambda_1 \| \hat{x} - x \|_1 + \lambda_{tv} L_{tv}(\hat{x}),\ \hat{x} = G(m) \quad (2)$$

In all experiments, $\lambda_1 = 120,\ \lambda_{tv} = 0.05, \lambda_{edge} = 0$

2.2 Evaluation Metrics and Auxiliary Vessel Segmentation

Evaluation Metrics and Auxiliary Vessel Segmentation. Distribution-level similarity between real and synthesized fundus sets is measured by Fréchet Inception Distance (FID) [9, 11] and Kernel Inception Distance (KID). Paired perceptual and structural similarity between is measured by Learned Perceptual Image Patch Similarity (LPIPS) and multi-scale structural similarity (MS-SSIM) [12, 13]. Lower values are better for FID/KID/LPIPS, while a higher MS-SSIM value is better.

Auxiliary segmenter. To conduct structure-aware inspection of synthesized images, an auxiliary vessel segmenter is used in the study. A U-Net-style model takes an RGB fundus image as input and predicts a single-channel vessel probability map, which is thresholder (0.5) to obtain the binary mask m̃ [10]. The segmenter is used only to generate m̃ for the four-column visualization and does not participate in GAN training or the computation of the reported quantitative metrics. To quantify mask-conditioned structural adherence as a diagnostic probe, the re-segmented vessel maps extracted from synthesized images are compared against the corresponding input vessel masks on the CHASE subset (N = 28). Re-segmentation consistency improves from Dice 0.5138 ± 0.0461 (IoU 0.3470 ± 0.0414) at epoch 70 to Dice 0.6015 ± 0.0355 (IoU 0.4301 ± 0.0363) at epoch 190. The auxiliary U-Net segmenter was trained offline on real fundus images from DRIVE and CHASE_DB1 with a held-out split (seed 42; val 0.15, test 0.15) and is kept frozen for all analyses. Because the segmenter is trained on the same public datasets and the re-segmentation subset may overlap with its training split, re-segmentation Dice/IoU are reported as within-protocol diagnostic probes rather than an unbiased downstream evaluation.

3 Experiments

3.1 Datasets and Paired Sample Construction

The experiment used paired retinal fundus datasets (DRIVE, CHASE DB1) with vessel annotations. Each sample is constructed as a paired tuple (x, m), where x is an RGB fundus image and m is its corresponding single-channel vessel mask. Fundus images are loaded in RGB format and vessel annotations are loaded as single-channel masks. All checkpoints and methods use the same evaluation protocol, so the results can be compared directly.

For DRIVE, vessel ground-truth masks are taken from the first manual annotations and paired with the corresponding fundus images [14]. The files under training/mask and test/mask represent field-of-view (FOV) region masks rather than vessel-annotation masks and are therefore excluded from both training targets and evaluation. For CHASE DB1, the first observer annotation (1stHO) is used as the vessel mask [15].

3.2 Preprocessing

For each paired sample, the RGB fundus image and grayscale vessel mask is processed with same geometric transforms (center crop, resize to 512×512, random horizontal flip, and random vertical flip) to preserve pixel alignment. Bicubic interpolation is used for the fundus image, whereas nearest-neighbor interpolation is used for the mask to avoid boundary smoothing. Both inputs are converted to tensors and normalized from [0, 1] to [-1, 1], matching the generator's tanh output range.

3.3 Training Settings (Pix2Pix Baseline)

All experiments reported in Section 4 primarily correspond to the Pix2Pix baseline trained under a fixed preprocessing and evaluation protocol at. Images and masks are normalized to [-1,1] as described in Section 2.1. To reduce reporting bias under non-monotonic adversarial dynamics, multiple late-stage checkpoints are summarized rather than relying solely on the final epoch.[9] All parameters are described in Table 1.

Table 1. Training settings (512x512)

Item	Value
Resolution	512×512
Batch size	2
Epochs	200
Optimizer	Adam
Learning rate	0.0002
Adam betas	(0.5,0.999)
Loss weights	$\lambda_1 = 120, \lambda_{tv} = 0.05, \lambda_{\text{edge}} = 0$
Checkpoint reporting	Epochs 160, 180, 190, 200 (Table 2)
Metric evaluation	Every 10 epochs (Fig. 1)

4 Results

4.1 Quantitative Results

This part presents findings for mask-to-fundus generation using the combined DRIVE+CHASE DB1 paired dataset (N = 48). It is worth noting that this study does not employ a validation/test split in order to simulate the limited-sample setting under privacy-preserving constraints for patients, and to analyze non-monotonic training behaviors under a consistent protocol. All metrics are calculated in-sample on the training pairs. Therefore, the reported values are internal epoch-to-epoch comparisons of realism and paired fidelity, rather than rigorous estimates of out-of-sample generalization ability.

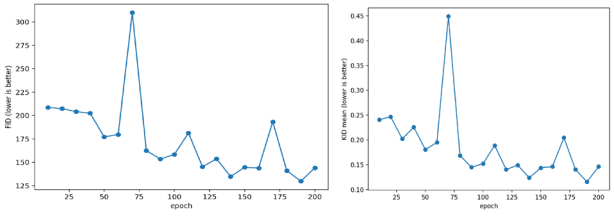
Table 2. Quantitative metrics on DRIVE+CHASE DB1(N=48)

Checkpoint	FID ↓	KID mean ↓	LPIPS ↓	MS-SSIM ↑
Epoch 160 (best LPIPS)	143.8449	0.146306	**0.2988**	0.8687
Epoch 180 (best MS-SSIM)	141.1606	0.140359	0.3093	**0.8761**
Epoch 190 (best FID/KID)	**129.7321**	**0.115495**	0.3289	0.8481
Epoch 200 (final)	144.1005	0.146273	0.3147	0.8315

To emphasize the non-monotonic behavior of adversarial training, several late-stage checkpoints are reported in Table 2. The best distribution-level realism occurs at epoch 190 (lowest FID and lowest KID mean), whereas the strongest paired-fidelity scores (lowest LPIPS and highest MS-SSIM) appear at nearby epochs. This mismatch shows that the best values for realism and fidelity may not occur at the same epoch. This situation reveals a realism–fidelity trade-off during training.

4.2 Metric Curves Across Training

Figure 1 visualizes the four metrics using results measured every 10 epochs. Figure 1 clearly shows that the non-monotonic behavior in this experiment was mainly manifested in the early and late stages, while the behavior in the middle was relatively flat compared to the early and late stages. This may occur because the generator and discriminator update against each other and progress can fluctuate rather than improve steadily across epochs. Besides, the mixed DRIVE+CHASE DB1 setting also introduces residual appearance differences between domains, which can amplify the sensitivity of distribution-level metrics such as FID and KID to the chosen checkpoint. These Dice/IoU values are reported as within-protocol diagnostics under a fixed segmenter.



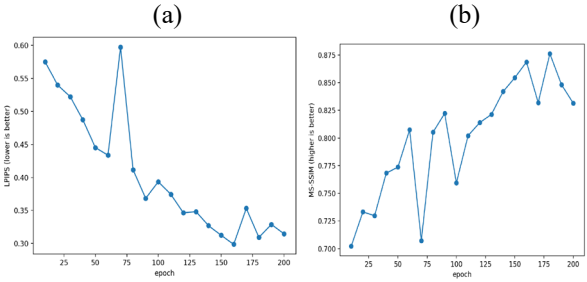


Fig. 1. Metric curves on the mixed paired set (N = 48) ((a) FID, (b) KID mean, (c) LPIPS, (d) MS-SSIM.) (Picture credit: Original)

4.3 Qualitative Results

To complement the quantitative results, qualitative visualizations are added to judge overall appearance, color/illumination plausibility, and vessel-structure consistency. Figure 2 illustrates the difference between the "best" epoch and the "worst" among 200 epochs (190 and 70, respectively). From the Figure 2, it is clear to see that the vessel-structure of epoch 190 is much better than epoch 70, and the specific directions of tiny blood vessels can be clearly distinguished at epoch 190.

Fig. 2. Epoch 190 and 70 comparison ((a) 190 real, (b) 190 synthesized, (c) 70 real, (d) 70 synthesized.) (Picture credit: Original)

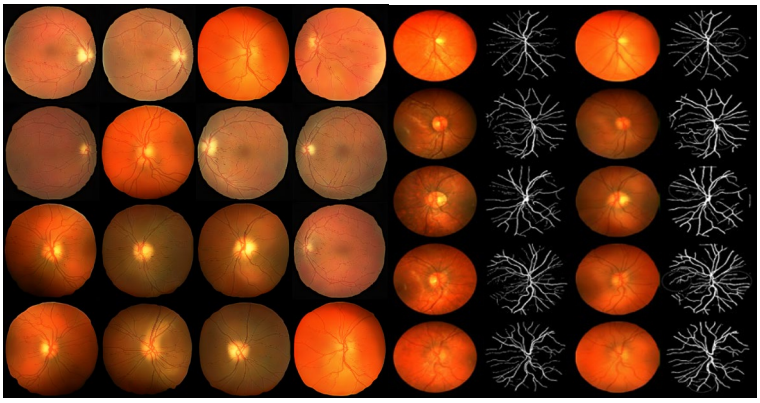


Fig. 3. Qualitative results at epoch 190. ((a) Synthesized samples. (b) Four-column visualization $(x, m, \hat{x}, \tilde{m})$) (Picture credit: Original)

To demonstrate the vessel-structure consistency achieved in this experiment. Figure 3 presents qualitative results at epoch 190 (Note: The clearly visible improvements in Figure 3(b) are highlighted (circled) in the figure). while (a) shows the synthesized samples, and (b) shows the comparison in the four-column visualization. Each visualization follows a four-column layout: real fundus image x, vessel ground truth m, synthesized fundus $\hat{x}$, and predicted vessel mask $\tilde{m}$. The mask $\tilde{m}$ is produced by the auxiliary segmenter for visualization only.

In summary, the above results demonstrate that the quantitative curves are non-monotonic but show overall improvement over training. Distribution-level realism (FID/KID) is best at epoch 190 in this experiment (when validated across multiple runs, the best epoch also tends to occur in the later stages of training), while paired perceptual and structural similarity peaks at nearby epochs (with a low probability of occurring at the same epoch). Given this trade-off, epoch 190 is used as the representative checkpoint in the presentation because it provides the strongest distribution alignment and yields stable qualitative results.

5 Discussion & Limitations

5.1 Discussion

The core of this experiment is that the Pix2Pix framework can generate a reasonable overall appearance at a high resolution (512×512) in a mask-to-fundus setting and can maintain the vascular geometry in most samples. In the actual checkpoints reported here, epoch 190 gives the best distribution-level realism (lowest FID/KID), while LPIPS and MS-SSIM reach their best values at nearby epochs rather than the same one. Adversarial training in GANs can cause noticeable fluctuations in the model's learning of overall texture and illumination, while the paired loss (such as L1) forces the model to fit each target image. When the image resolution is high and the training data are scarce, it becomes difficult for the model to both learn background detail textures well and accurately reproduce blood vessel edges, and the conflict is most visible around thin vessels and at vessel boundaries. Differences in data collection across datasets (such as camera model, field of view, and illumination conditions) can disrupt the overall style distribution of images. These inter-domain differences can affect metrics like FID/KID, which rely on feature statistics computed over a batch, so the values may vary with the specific batch or the DRIVE–CHASE mix at a given checkpoint. As a result, inconsistencies are more likely to arise between the overall realism reflected by FID/KID and the single-image pairing quality reflected by LPIPS/MS-SSIM.

5.2 Limitations

To simulate real-world patient privacy scenarios, large-scale datasets commonly used in other studies were not used. Furthermore, due to budget constraints, access restrictions, and limited compute resources, this study used only two publicly available datasets: CHASE_DB1 and DRIVE. As a result, the estimated metrics exhibit higher

variance than in many other GAN studies, especially for feature-based distribution metrics such as FID and KID. Also, only four image-level metrics are reported in this study. They do not directly quantify clinical utility or the impact of synthesized images on downstream tasks. Re-segmentation Dice/IoU depend on the auxiliary segmenter (trained on DRIVE+CHASE), so the paper treats them as diagnostic probes rather than unbiased downstream metrics. This study focuses on analyzing the non-monotonic dynamics of adversarial training and employs a fixed protocol for in-sample evaluation on a hybrid DRIVE+CHASE DB1 paired dataset (N=48). The current implementation does not include explicit held-out partitioning. Therefore, absolute values of the metrics should be interpreted with caution, while relative comparisons across checkpoints are more meaningful.

6 Conclusion

In conclusion, this study assesses vessel-mask-to-fundus translation at a resolution of 512×512 using a Pix2Pix-style GAN baseline on the DRIVE+CHASE_DB1 paired dataset (N = 48). Realism, at the distribution level is measured via FID/KID while paired fidelity is evaluated using LPIPS/MS-SSIM. Throughout later-stage checkpoints these metrics may. Might not achieve their best values simultaneously at the same epoch. Qualitatively, vessel geometry is generally preserved, while global appearance varies across training; epoch 190 is typically cleaner and more consistent than epoch 70, which shows stronger color and texture distortions. Future work should include matched Pix2Pix vs conditional WGAN-GP comparisons, dataset-wise reporting with explicit cross-domain evaluation, and downstream task tests (e.g., vessel segmentation) to assess practical utility.

References

1. Goodfellow, I. J. Pouget-Abadie, J. Mirza, M. et al.: Generative Adversarial Nets," in Advances in Neural Information Processing Systems (NeurIPS), (2014)
2. Coyner, A. S., Chen, J. S., Chang, K. et al.: Synthetic Medical Images for Robust, Privacy-Preserving Training of Artificial Intelligence: Application to Retinopathy of Prematurity Diagnosis, Ophthalmology Science, vol. 2, no. 2, p. 100126, (2022)
3. Kim, M., Kim, Y. N., Jang, M. et al.: Synthesizing realistic high-resolution retina image by style-based generative adversarial network and its utilization, Scientific Reports, vol. 12, p. 17307, (2022)
4. Mirza, M. and Osindero, S.: Conditional Generative Adversarial Nets, arXiv preprint arXiv:1411.1784, (2014)
5. Isola, P., Zhu, J. Y., Zhou, T. and Efros, A. A.: Image-to-Image Translation with Conditional Adversarial Networks, in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), (2017)
6. Zhu, J. Y., Park, T., Isola, P. and Efros, A. A.: Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, in Proc. IEEE Int. Conf. on Computer Vision (ICCV), (2017)

7. Arjovsky, M., Chintala, S. and Bottou, L.: Wasserstein Generative Adversarial Networks, in Proc. 34th Int. Conf. on Machine Learning (ICML), Proc. Mach. Learn. Res. (PMLR), vol. 70, pp. 214–223, (2017)
8. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. and Courville, A.: Improved Training of Wasserstein GANs, in Advances in Neural Information Processing Systems (NeurIPS), (2017)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. and Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, in Advances in Neural Information Processing Systems (NeurIPS), (2017)
10. Bińkowski, M., Sutherland, D. J., Arbel, M. and Gretton, A.: Demystifying MMD GANs, arXiv:1801.01401, (2018)
11. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. and Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), (2018)
12. Wang, Z., Simoncelli, E. P. and Bovik, A. C.: Multiscale Structural Similarity for Image Quality Assessment, in Proc. Asilomar Conf. on Signals, Systems and Computers, (2003)
13. Ronneberger, O., Fischer, P. and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in Medical Image Computing and Computer-Assisted Intervention (MICCAI), (2015)
14. Staal, J., Abràmoff, M. D., Niemeijer, M., Viergever, M. A. and van Ginneken, B.: Ridge-Based Vessel Segmentation in Color Images of the Retina, IEEE Transactions on Medical Imaging, vol. 23, no. 4, pp. 501–509, (2004)
15. Fraz, M. M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A. R., Owen, C. G. and Barman, S. A.: An ensemble classification-based approach applied to retinal blood vessel segmentation," IEEE Trans. Biomed. Eng., vol. 59, no. 9, pp. 2538–2548, (2012) doi: 10.1109/TBME.2012.2205687.