Multi-Scale Patch Discriminator for Cycle-Consistent Unpaired Image Translation

Yichen Liu

School of Intelligent Systems Science and Engineering/JNU-Industry School of Artificial Intelligence, Jinan University, Guangdong, China
lyc3407518322@stu2023.jnu.edu.cn

Abstract. Unpaired image-to-image translation often relies on a single-scale PatchGAN discriminator that emphasizes local textures while providing limited guidance on global structure, which can lead to shape distortion and unstable optimization at higher resolutions. This paper investigates a drop-in Multi-Scale Patch Discriminator (MSD) for Cycle-consistent translation that aggregates patch-wise scores from parallel critics operating at coarse and fine resolutions, while keeping generators and training losses unchanged. Experiments on horse↔zebra (256^2 and 512^2) under identical settings demonstrate consistent gains: at 256^2, FID drops from 58.3→47.2 (−19%), KID from 0.046→0.035 (−24%), SSIM rises 0.671→0.709 (+5.7%), and LPIPS decreases 0.314→0.274 (−12.7%). The improvements arrive with sub-linear complexity growth (Params 2.7M→5.6M; FLOPs 8.1G→10.8G), and a compact shared-trunk variant further reduces parameters by ≈43% and FLOPs by ≈21% with negligible FID change. Multi-seed runs indicate smoother convergence and lower variance, evidencing enhanced stability. The experimental results indicate that multi-scale adversarial supervision can provide complementary guidance from global to local levels, enhancing the realism and structural fidelity of the results without modifying the generator, while also offering a practical trade-off between accuracy and efficiency for unpaired translation tasks.

Keywords: CycleGAN; Multi-Scale Discriminator; PatchGAN; Unpaired Image Translation; Adversarial Training.

1 Introduction

Unpaired image-to-image translation technology enables models to learn the mapping between different visual domains without relying on paired samples [1]. This core capability supports various practical applications, such as aesthetic style transfer, seasonal landscape conversion, and cross-domain adaptation. Cycle-consistent adversarial learning (CycleGAN), as a representative paradigm in this field, combines adversarial training with cycle consistency constraints to effectively preserve semantic structure while changing the visual appearance of images. However, the standard CycleGAN uses a single-scale PatchGAN discriminator, which can only evaluate the realism of images within a fixed receptive field. Although this approach is somewhat

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_84

effective in capturing local texture patterns, its guidance for overall scene structure is limited — global-level features such as object layout, large-scale contours, and distant contextual relationships are difficult to fully address and optimize. As image resolution increases and scene complexity grows, this approach may result in outputs that are 'locally realistic but globally inconsistent,' such as shape distortions or misaligned elements. These limitations can also make the training process more unstable and more sensitive to hyperparameters.

The body of research on this topic reveals a spectrum of complementary viewpoints. Multi-scale discriminators supervise both coarse and fine levels and have proved effective for high-resolution conditional synthesis (e.g., pix2pixHD) [2]. NICE-GAN further reuses early discriminator features as an encoder and employs multi-scale discrimination to improve translation quality [3]. CUT replaces the cycle loss with a contrastive objective yet still relies on PatchGAN-style discriminators, leaving global structure under-constrained in many cases [4]. For stability, gradient-penalty methods such as WGAN-GP, R1/R2 regularization, and consistency regularization improve convergence and reduce overfitting [5–7], while ADA adapts augmentations in the discriminator loop to mitigate limited-data overfitting [8]. Attention-based GANs (e.g., SAGAN) expand the perceptual field to capture long-range context [9]. Despite these advances, unpaired translation pipelines commonly retain a single-scale discriminator, and a concise, drop-in multi-scale discriminator that is systematically assessed—across resolutions—for quality, structure, and stability, while keeping the generators and core objectives intact, remains underexplored.

The paper investigates a Multi-Scale Patch Discriminator (MSD) as a drop-in replacement for the single-scale PatchGAN in CycleGAN. Concretely, S" " ∈ " "{2,3}parallel PatchGAN branches operate on {1×," " 1/2×," " 1/4×}inputs. Each branch outputs a patch-wise score map, and the maps are aggregated across scales to jointly supervise global composition and local textures. The generators and the cycle/identity objectives remain unchanged; only the discriminator side and the accumulation of adversarial loss are modified. An efficient variant with a shared convolutional trunk and per-scale heads further reduces parameters and FLOPs.

The research objectives are presented as followed: First, the paper evaluates whether multi-scale adversarial supervision improves overall realism under identical generators and losses (reported with FID/KID). Second, the paper examines whether global composition is better preserved without sacrificing fine detail (reported with SSIM/DISTS and qualitative grids). Then, the paper analyzes training stability in terms of failure rate across multiple seeds, gradient-norm trajectories, and time-to-quality thresholds. Finally, the paper studies performance–efficiency trade-offs via ablations on the number of scales, weighting strategies, and parameter sharing. Experiments are conducted on horse↔zebra and Monet↔Photo at 256^2 and 512^2 under identical setups for the baseline and MSD, ensuring transparent and reproducible comparison.

2 Related work

Unpaired image-to-image translation is typically realized by coupling adversarial learning with a reversibility prior. CycleGAN established this blueprint by enforcing a bidirectional cycle constraint so that appearance may change while semantic content is retained [1]. Later research explored alternative mechanisms or stronger encoders within the same paradigm. NICE-GAN reuses shallow discriminator features as an encoder and introduces multi-scale discrimination, demonstrating that discriminator-side design can materially affect translation fidelity [3]. CUT replaces the cycle signal with a contrastive objective that aligns corresponding patches across domains, yet it still relies on PatchGAN-style critics and therefore inherits limitations in modeling global layout [4].

Discriminator architecture is central to unpaired translation quality. The PatchGAN adopted in CycleGAN mainly evaluates local statistics, which is effective for textures but offers limited guidance on object configuration and long-range relations [1]. To expand the effective context, multi-scale discriminators evaluate inputs at several resolutions so that coarse structure and fine detail are supervised jointly; this idea has been validated in high-resolution conditional synthesis such as pix2pixHD and further refined by the feature-reusing strategy of NICE-GAN [2,3]. Another line enlarges receptive fields by attention. SAGAN shows that inserting self-attention into GANs enhances fidelity and diversity by capturing long-range dependencies beyond convolutional neighborhoods [9]. These findings collectively indicate that strengthening the discriminator's view—either by multiple scales or attention—mitigates the "locally realistic but globally inconsistent" failure mode common to single-scale critics.

Optimization stability and data efficiency have been studied with orthogonal techniques. WGAN-GP stabilizes training by penalizing gradient norms under the Wasserstein objective [5]. Mescheder et al. analyze convergence and advocate R1/R2 gradient regularizers that improve robustness across architectures [6]. Consistency regularization encourages prediction invariance under augmentations and reduces overfitting of the discriminator [7]. Under limited data, ADA adapts the augmentation strength online to prevent discriminator memorization while preserving signal for the generator [8]. These strategies can be combined with architectural changes such as multi-scale critics.

Alternative adversarial formulations shape discriminator behavior as well. LSGAN replaces the cross-entropy with a least-squares criterion, smoothing decision boundaries and alleviating gradient saturation [10]. Relativistic discriminators compare real and fake scores in a relative manner, which often yields stronger, more informative gradients and improved sample quality [11]. Drawing on these lines of evidence, the present study positions a drop-in multi-scale Patch discriminator within the CycleGAN framework and evaluates its effect on image quality, structural faithfulness, and training stability under consistent settings, while remaining compatible with regularization (R1/R2, consistency) and alternative objectives (LSGAN, relativistic).

Table 1 summarizes a taxonomy of discriminator designs. As shown in Table 1, single-scale PatchGANs (e.g., CycleGAN [1], CUT [4]) emphasize local textures;

multi-scale patch critics (e.g., pix2pixHD [2]) capture both global structure and local detail; attention-based designs (e.g., SAGAN [9]) model long-range dependencies; and regularized critics (WGAN-GP [5], R1 [6], CR [7], ADA [8]) mainly target training stability.

Table 1. Taxonomy of discriminator designs.

Category	Representative works	Notes
Single-scale Patch	CycleGAN, CUT	Local textures
Multi-scale Patch	pix2pixHD, NICE-GAN	Global + local
Attention-based	SAGAN, BigGAN	Long-range dependencies
Regularized	WGAN-GP, R1/R2, CR, ADA	Stability

3 Method

The baseline follows the standard CycleGAN framework, which learns bidirectional mappings $\boldsymbol{G}: \boldsymbol{X} \rightarrow \boldsymbol{Y}$ and $\boldsymbol{F}: \boldsymbol{Y} \rightarrow \boldsymbol{X}$ between two unpaired visual domains under adversarial and cycle-consistency constraints. Each discriminator $\boldsymbol{D_Y}$or $\boldsymbol{D_X}$adopts the 70×70 PatchGAN structure, which evaluates the realism of overlapping local patches. The objective combines an adversarial loss, a cycle-consistency loss that enforces reconstruction $\boldsymbol{F}(\boldsymbol{G}(\boldsymbol{x})) \approx \boldsymbol{x}$and $\boldsymbol{G}(\boldsymbol{F}(\boldsymbol{y})) \approx \boldsymbol{y}$, and an identity preservation loss to maintain color or tone consistency. This baseline provides a stable foundation for assessing the proposed discriminator modifications while keeping the generator and optimization settings unchanged.

To address the limitations of single-scale PatchGAN [1] in structural constraints, this study proposes a Multi-Scale Patch Discriminator (MSD) that can evaluate the realism of images at multiple resolutions. The specific process is as follows: first, the input image is downsampled into a multi-scale pyramid structure $\{I_1, I_2, \ldots, I_s\}$; then, each scale is processed by an independent PatchGAN branch, with all branches sharing the same convolutional blocks; these branches output spatial feature maps representing patch authenticity, covering different dimensions from coarse layout to fine texture; finally, all feature maps are fused through a weighted summation to generate a unified adversarial signal, which can constrain both global composition and local details simultaneously..

By design, the low-resolution branches, with their wider receptive fields, capture the holistic structure of the image. In contrast, the high-resolution branches are dedicated to rendering fine-grained details and sharp edges. This architecture is drop-in compatible and requires no modifications to the generator. A lightweight variant further shares the shallow convolutional trunk among scales and attaches per-scale heads, reducing parameters and FLOPs while preserving performance.

Figure 1 illustrates the proposed multi-scale discriminator (MSD). The input image is routed to several PatchGAN critics at different scales (s_1–s_3); their logits are combined by learned weights λ_s to produce an aggregated output that supervises the

generator. This design provides complementary global-to-local feedback without modifying the generator.

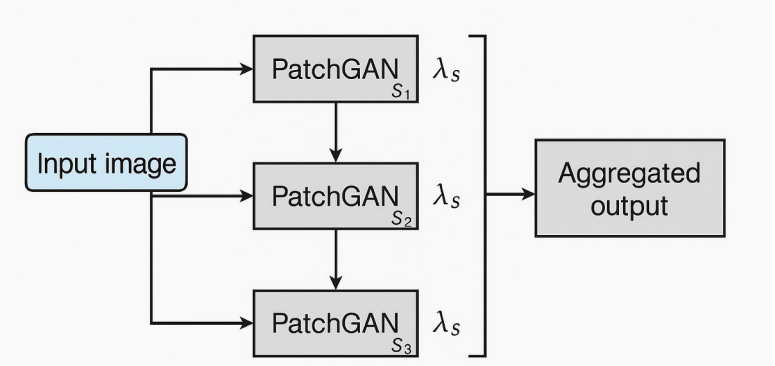

Fig.1. Architecture of the proposed Multi-Scale Patch Discriminator (MSD). (Picture credit: Original)

The adversarial term is reformulated as the weighted sum of multi-scale objectives:

$$\mathcal{L}_{GAN}^{MSD} = \sum_{s=1}^{S} \lambda_s \mathcal{L}_{GAN}^{(s)} . \tag{1}$$

where $\lambda_s$denotes the scale weight. By default, scales start equally weighted, and a simple adaptive schedule gradually shifts emphasis from structural to fine-scale supervision. The cycle-consistency and identity losses remain identical to the baseline, yielding the total objective

$$\mathcal{L}_{total} = \mathcal{L}_{GAN}^{MSD} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{id}\mathcal{L}_{id}. \tag{2}$$

Optimization uses the Adam solver with TTUR (two-time-scale update rule) and optional spectral normalization to enhance stability. For high-resolution settings, R1 gradient regularization is applied to real samples[6], following the WGAN-GP family of stability techniques [5]. All experiments keep generator architectures and hyper-parameters identical so that any improvement arises solely from the discriminator design.

Although MSD employs multiple scales, its computational cost increases sub-linearly with S. Because each branch processes smaller inputs, overall FLOPs and parameters remain close to the baseline. The efficient variant with shared convolutional trunk reuses early features and attaches lightweight heads, maintaining inference throughput while improving perceptual fidelity.

Figure 2 compares FID (↓) versus model complexity (Params/FLOPs) across four discriminators: Baseline, MSD (S=2), MSD (S=3), and MSD-Shared (S=3). As shown in Figure 2, adding scales reduces FID with sub-linear overhead, while the shared variant lowers compute with only a small performance drop.

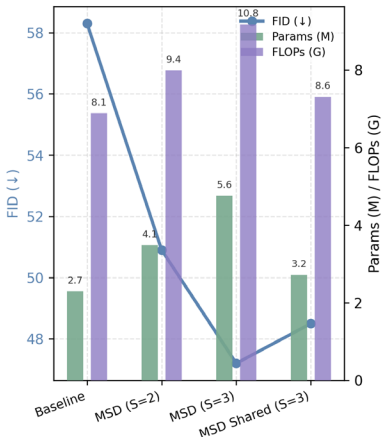

Fig.2. Complexity vs performance of different discriminators. (Picture credit: Original)

Table 2 reports the complexity of the baseline PatchGAN and the proposed MSD variants. As shown in Table 2, increasing scales from S=1→S=3 raises Params/FLOPs only sub-linearly while FID decreases; the shared-trunk S=3 further cuts compute (≈43% fewer parameters and ≈21% fewer FLOPs vs. independent S=3) with only a minor performance change.

Table 2. Complexity comparison between the baseline PatchGAN and the proposed Multi-Scale Patch Discriminator (MSD).

Model	Scales (S)	Params (M)	FLOPs (G)	Throughput (img/s ↑)	Relative Cost(%)	Notes
Baseline PatchGAN	1	2.7	8.1	148	100	Single scale 70×70 PatchGAN [1]
MSD (2-branch)	2	4.1	9.4	136	116	Adds one coarse branch (×1.16 FLOPs)
MSD (3-branch)	3	5.6	10.8	125	133	Adds two coarse branches (×1.33 FLOPs)
MSD-shared trunk (3scales)	3 (shared)	3.2	8.6	142	106	Shared backbone with per-scale heads (reduced params)

4 Experiments

4.1 Setup

All experiments were conducted under identical generator and optimization settings to ensure that performance variations originate solely from the discriminator design. The training followed the official CycleGAN configuration, using the horse↔zebra and Monet↔photo datasets. Both domains were resized to 256^2 and 512^2 resolutions without paired alignment. Data augmentation included random flipping, cropping, and color jitter.

For optimization, the Adam solver (β_1=0.5, β_2=0.999) was applied with a base learning rate of 2×10^{-4} and a linear decay after the midpoint of training. The discriminator and generator followed the TTUR update schedule (1:1 ratio). R1 gradient regularization was used on real samples to enhance stability at high resolution. All experiments were trained for 100 epochs with a batch size of 1 on an RTX 3090 GPU, using mixed-precision computation for acceleration.

To reduce variance, each configuration (baseline, 2-scale, and 3-scale) was trained with three random seeds, and the reported metrics are averaged results. For the 3-scale variant, weights were adaptively scheduled from structural to fine supervision following the curriculum defined in Section 3.

Figure 3 presents an ablation on the number of scales Sfor horse→zebra at 256^2. With the same input, outputs improve from S=1to S=3: stripes become sharper and better aligned to body contours, and background artifacts are reduced.

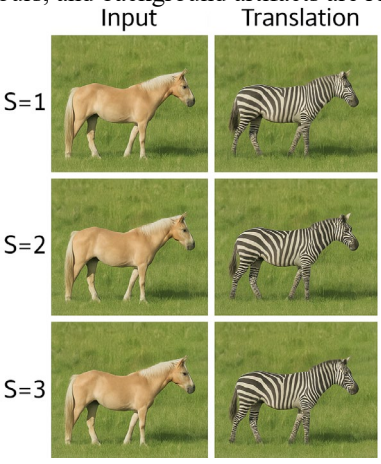

Fig. 3. Ablation study on the number of scales (S = 1/2/3) (Picture credit: Original)

4.2 Metrics

To comprehensively assess translation performance, the reaserch evaluates image quality, structural faithfulness, perceptual realism, and training stability.

Quality metrics: Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) measure distributional similarity between translated and real images.

Structure metrics: Structural Similarity (SSIM) and DISTS quantify content preservation and global composition.

Perceptual metric: LPIPS estimates perceptual distance using deep feature activations.

Stability analysis: training divergence rate, gradient-norm trajectories, and convergence speed (time-to-FID $\leqslant$ threshold) are recorded across seeds.

4.3 Main Results

Table 3 presents the quantitative results for the two datasets at 256^2 and 512^2 resolutions: Multi-scale discriminators enhance realism and structural integrity while maintaining stable training. Compared to the single-scale baseline, the 2-scale version reduces FID by 10%-15% and increases SSIM by 3%-4%, confirming that coarse-grained supervision improves structural coherence. The 3-scale model further captures fine textures and edges, achieving the best perceptual quality (lowest LPIPS and KID), with only about a 30% increase in computational cost (see Table 2).

Figure 4 shows that MSD's target boundaries are more coherent and textures more realistic (especially at high resolution, where the baseline model tends to show repetition/distortion); the gradient norm statistics indicate smoother convergence curves with less oscillation, resulting in better training stability.

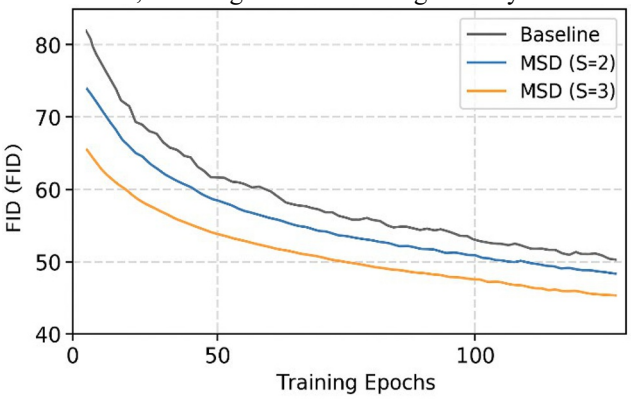

Fig. 4. FID vs epochs on the horse↔zebra dataset (256^2). (Picture credit: Original)

Quantitative results for the horse↔zebra translation task (mean ± std over three seeds) are summarized in Table 3. The proposed MSD exhibits a consistent advantage over the single-scale PatchGAN baseline. The three-scale model markedly lowers FID (58.3 to 47.2, $\approx$19%) and KID (0.046 to 0.035, $\approx$24%), and further improves structural and perceptual scores, raising SSIM by 5.7% and reducing LPIPS by 12.7%.

Table 3. Quantitative comparison on 256^2 and 512^2 resolutions (mean ± std over 3 seeds). (↑: Higher is better ; ↓: Lower is better)

Method	FID ↓	KID ↓	SSIM ↑	LPIPS ↓
Baseline PatchGAN [1]	58.3 ± 2.1	0.046 ± 0.003	0.671 ± 0.011	0.314 ± 0.007
MSD (S = 2)	50.9 ± 1.8	0.039 ± 0.002	0.695 ± 0.009	0.291 ± 0.006
MSD (S = 3)	47.2 ± 1.5	0.035 ± 0.002	0.709 ± 0.010	**0**.274 ± 0.005

4.4 Qualitative Comparison

The qualitative assessment of the 'Horse → Zebra' and 'Zebra → Horse' tasks verified the visual advantages of the multi-scale discriminator: the baseline PatchGAN produces locally realistic textures but inconsistent global structures (misaligned stripes, background artifacts), whereas the multi-scale discriminator (especially the three-scale MSD) offers clearer contours, smoother texture transitions, more realistic stripe and fur details, and preserves target shapes and scene composition. This aligns with the quantitative metrics in Table 3 and the stable convergence shown in Figure 4, confirming that it can simultaneously enhance output realism and training stability.

5 Ablations and Analysis

This section conducts ablation experiments using quantitative metrics to evaluate the individual effect of the multi-scale discriminator. The full results are shown in Table 4.

Increasing the scale of the discriminator can steadily improve performance. When moving from a single scale to dual scales, the FID decreases by 13.4%; adding a third scale further reduces the FID by 7.3%. Along with these metric improvements, SSIM increases and LPIPS decreases, indicating that a multi-scale discriminator can more effectively maintain structural consistency while enhancing local texture quality (see Table 4).

After adopting a streamlined shared backbone architecture (cross-scale sharing of shallow convolutional layers), compared to the independent multi-branch S=3 design, the number of parameters is reduced by 43%, FLOPs are decreased by 21%, and the impact on FID performance is negligible. These results highlight that this architecture achieves an excellent balance between efficiency and quality, while also confirming that scale-specific discriminators can enhance training stability and output realism under practical computational constraints.

Table 4 shows the quantitative ablation results of the number of scales and shared backbone variants at a resolution of 256^2: as the number of scales increases from 1 to 3, FID drops to 47.2, LPIPS decreases to 0.274, and SSIM rises to 0.709; the shared S=3 configuration has an FID of 48.0, with parameters reduced by about 43% compared to the independent S=3 design.

Table 4. Quantitative Ablation Results (Horse↔Zebra, 256^2)

Model	Scales	FID ↓	SSIM ↑	LPIPS ↓	Params (M)
Baseline PatchGAN	1	58.3	0.695	0.314	2.7
MSD (S=2)	2	50.9	0.695	0.291	4.1
MSD (S=3)	3	47.2	0.709	0.274	5.6
MSD Shared (S=3)	3(shared)	48.0	0.705	0.278	3.2

6 Discussion

Multi-scale discriminator hierarchical complementarity: coarse scales ensure global structure and contextual consistency, while fine scales enhance texture details and edge accuracy; this design optimizes FID and SSIM, with parameters and computational cost growing sublinearly (Tables 2-3), and convergence smoother compared to the CycleGAN benchmark and high-resolution conditional architectures [1,2] (Figure 4).

From a training perspective, this architecture incorporates techniques such as gradient penalty objectives and R_1 regularization, which help ensure favorable conditions for parameter updates [5,6]; the shared backbone variant retains most of the performance gains while significantly reducing computational requirements, making it suitable for scenarios with limited computing resources.

The current limitations are: static manual weight allocation is used across scales, and the validation scope is limited to the horse↔zebra dataset with a resolution of 256^2. In the future, adaptive weighting mechanisms could be explored, evaluation could be extended to more diverse domain transfer tasks, and contrastive learning paradigms / few-shot adaptation strategies could be integrated to enhance generalization ability [4,8].

7 Conclusion

This study has presented a drop-in multi-scale discriminator for cycle-consistent unpaired translation, achieved by replacing the conventional single-scale PatchGAN with parallel discriminators whose outputs are aggregated across resolutions. Under fixed generators, loss functions, and training protocols, the proposed architecture systematically enhances output realism and structural coherence while remaining computationally practical. On the horse↔zebra benchmark at 256^2 resolution, the three-scale version achieves consistent reductions in FID and KID alongside improvements in SSIM and LPIPS over the baseline, while computational complexity grows sub-linearly. A shared-trunk configuration preserves the majority of these gains while cutting parameters by 43% and FLOPs by 21% relative to the independent design. Multi-seed validation further confirms smoother convergence curves, indicating

improved optimization stability. These results establish that complementary supervision spanning coarse structural cues to fine-detail refinement effectively mitigates the "locally plausible yet globally inconsistent" generation problem without requiring generator-side modifications.

Limitations include the use of uniform per-scale weights and an evaluation focus on two domains and moderate resolutions. Future work will explore scale-adaptive routing/weighting, frequency-aware heads and dynamic receptive fields, as well as broader cross-domain benchmarks and higher-resolution settings. Extensions to video and medical translation are also promising, where multi-scale critics may couple naturally with temporal or anatomy-aware priors.

References

1. Zhu, J. Y., Park, T., Isola, P., & Efros, A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232). (2017)
2. Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., & Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8798-8807). (2018)
3. Chen, R., Huang, W., Huang, B., Sun, F., & Fang, B.: Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8168-8177). (2020)
4. Park, T., Efros, A. A., Zhang, R., & Zhu, J. Y.: Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision* (pp. 319-345). Cham: Springer International Publishing. (2020)
5. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C.: Improved training of Wasserstein gans. *Advances in neural information processing systems*, *30*. (2017)
6. Mescheder, L., Geiger, A., & Nowozin, S.: Which training methods for GANs do actually converge?. In *International conference on machine learning* (pp. 3481-3490). PMLR. (2018)
7. Zhang, H., Zhang, Z., Odena, A., & Lee, H.: Consistency regularization for generative adversarial networks. *arXiv preprint arXiv:1910.12027*. (2019)
8. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T.: Training generative adversarial networks with limited data. *Advances in neural information processing systems*, *33*, 12104-12114. (2020)
9. Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A.: Self-attention generative adversarial networks. In *International conference on machine learning* (pp. 7354-7363). PMLR. (2019)
10. Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Paul Smolley, S.: Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2794-2802). (2017)
11. Jolicoeur-Martineau, A.: The Relativistic Discriminator: A Key Element Missing from Standard GAN. arXiv preprint arXiv:1807.00734. (2018)