Scientific Journal of Informatics
Vol. 13, No. 2, May 2026
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Region-Aware Double-Tail GAN with AdaIN–LADE Adaptive Style Transfer for Cinematic Anime Background Stylization in Makoto Shinkai Style

Agus Purwanto[1*], Kusrini[2], Ema Utami[3], David Agustriawan[4]
[1, 2, 3, 4]Doctoral Informatics Study Program, Universitas Amikom Yogyakarta, Indonesia

Abstract.
Purpose: Background stylization in a specific anime art direction remains challenging because global style transfer often yields inconsistent stylization across semantic regions. Our prior Double-Tail GAN (DTGAN) with Adaptive Instance Normalization (AdaIN) and Linearly Adaptive Denormalization (LADE) can produce Shinkai-like backgrounds, but it still exhibits region-specific failures such as unstable sky gradients, over-stylized vegetation textures, and reduced edge clarity in buildings.
Methods: We propose a region-aware extension of DTGAN by conditioning the generator on semantic masks (sky, vegetation, and buildings) and optimizing with coverage-aware, region-weighted objectives. Semantic masks are generated automatically using a lightweight transformer-based semantic segmentation model and refined via simple morphological filtering to stabilize mask boundaries during training.
Result: Experiments on real photographs and Makoto Shinkai-style background frames show that region-aware conditioning improves both global and region-level quality compared with DTGAN without masks. The proposed method reduces global FID from 74.5 to 65.8 and LPIPS from 0.505 to 0.448, while improving sky-gradient similarity and overall palette consistency.
Novelty: This work contributes (i) a practical mask-conditioned DTGAN formulation for local controllability in cinematic anime background stylization, and (ii) a coverage-aware region-weighting strategy that mitigates over-stylization and style leakage when semantic regions occupy imbalanced areas, without requiring manual mask annotation.

Keywords: : Double-tail Generative Adversarial Network, Region-Aware, Anime Background Production, AdaIN, LADE, Makoto Shinkai
Received January 2026 / **Revised** February 2026 / **Accepted** May 2026

INTRODUCTION

Background stylization for cinematic anime production remains challenging because a single global style constraint often yields uneven results across semantic regions. In our prior stage, Double-Tail GAN (DTGAN) with adaptive style transfer (AdaIN for global statistics and LADE for artifact suppression) produced Shinkai-like backgrounds but still exhibited characteristic failures: unstable sky gradients, over-stylized vegetation textures, and reduced edge clarity on buildings. [1] More broadly, GAN-based image synthesis and neural style transfer have become established foundations for photo stylization, but global stylization objectives still tend to treat a scene too uniformly when semantics differ strongly across regions [2], [3], [4], [5].

These failures indicate that background stylization is not a uniform transformation. The sky requires smooth, globally consistent color transitions; vegetation benefits from textured but controlled brush-like patterns; and buildings demand structure preservation with stable tone mapping. When the generator is not explicitly guided by semantic regions, style leakage across boundaries and region neglect under imbalanced region areas can occur.

*Corresponding author.
Email addresses: agus@amikom.ac.id (Purwanto)*, kusrini@amikom.ac.id (Kusrini), ema.u@amikom.ac.id (Utami), david.agustriawan@i3l.ac.id (Agustriawan)
DOI: 10.15294/sji.v13i2.40665

Region-aware GANs address this by incorporating semantic priors (e.g., segmentation maps) into generation. However, representative approaches such as SPADE and SEAN inject semantic information through layer-wise normalization, which can increase training complexity and depend heavily on high-quality segmentation labels. For photo-to-anime background stylization with limited style data, we adopt a lighter integration: semantic masks are concatenated as additional input channels, while region control is emphasized through an objective-level design (region-specific losses) rather than through heavy layer-wise feature modulation [6], [7], [8], [9], [10].

Recent photo-to-anime and controllable cartoonization studies such as AnimeGAN, CartoonLossGAN, Scenimefy, and Interactive Cartoonization show that visually appealing translation is feasible with lightweight generators and perceptual controls; however, these methods generally emphasize global appearance and do not explicitly enforce region-balanced stylization for cinematic backgrounds [11], [12], [13], [14].

Compared with prior region-aware image synthesis methods such as SPADE and SEAN, the proposed method introduces semantic guidance in a lighter way, namely through mask concatenation and coverage-aware region-weighted losses, while preserving the original DTGAN two-tail refinement pathway. This design reduces parameter overhead and training complexity, making it more suitable for limited photo-to-anime background datasets. In contrast to global stylization methods such as AnimeGAN, CartoonLossGAN, and related photo-to-anime approaches, our method explicitly targets region-balanced stylization for sky, vegetation, and building areas, which are critical for cinematic anime backgrounds. Therefore, the main contribution of this study is not only improving stylization quality, but also improving local controllability and region fidelity without relying on heavy semantic-normalization layers.

This paper proposes a region-aware extension of DTGAN by (1) generating semantic masks for sky, vegetation, and building regions using a lightweight segmentation model with morphological refinement; (2) designing a coverage-aware region-weighted loss that balances optimization across regions under area imbalance; and (3) evaluating improvements using both global metrics (FID, LPIPS) and region-level metrics (Region-FID/RFID, SGS, PSI, SFD) via an ablation study.

METHODS

To make the contribution verifiable (not only visually convincing), this study explicitly tests whether region-aware supervision improves stylization fidelity per semantic area. In particular, we aim to demonstrate that (i) semantic-mask guidance reduces style leakage between sky, vegetation, and building regions, and (ii) coverage-aware weighting stabilizes training by preventing small/rare regions from dominating gradients. These claims are evaluated using both global metrics (FID and LPIPS) and region-level metrics (Region-FID/RFID, SGS, PSI, and Style Feature Distance/SFD), supported by an ablation setting that isolates the effect of masks, AdaIN, LADE, and adaptive weights. In summary, we aim to demonstrate that region-aware supervision via semantic masks and coverage-aware weighting yields measurable improvements in per-region stylization metrics, beyond what can be achieved by global style transfer alone.

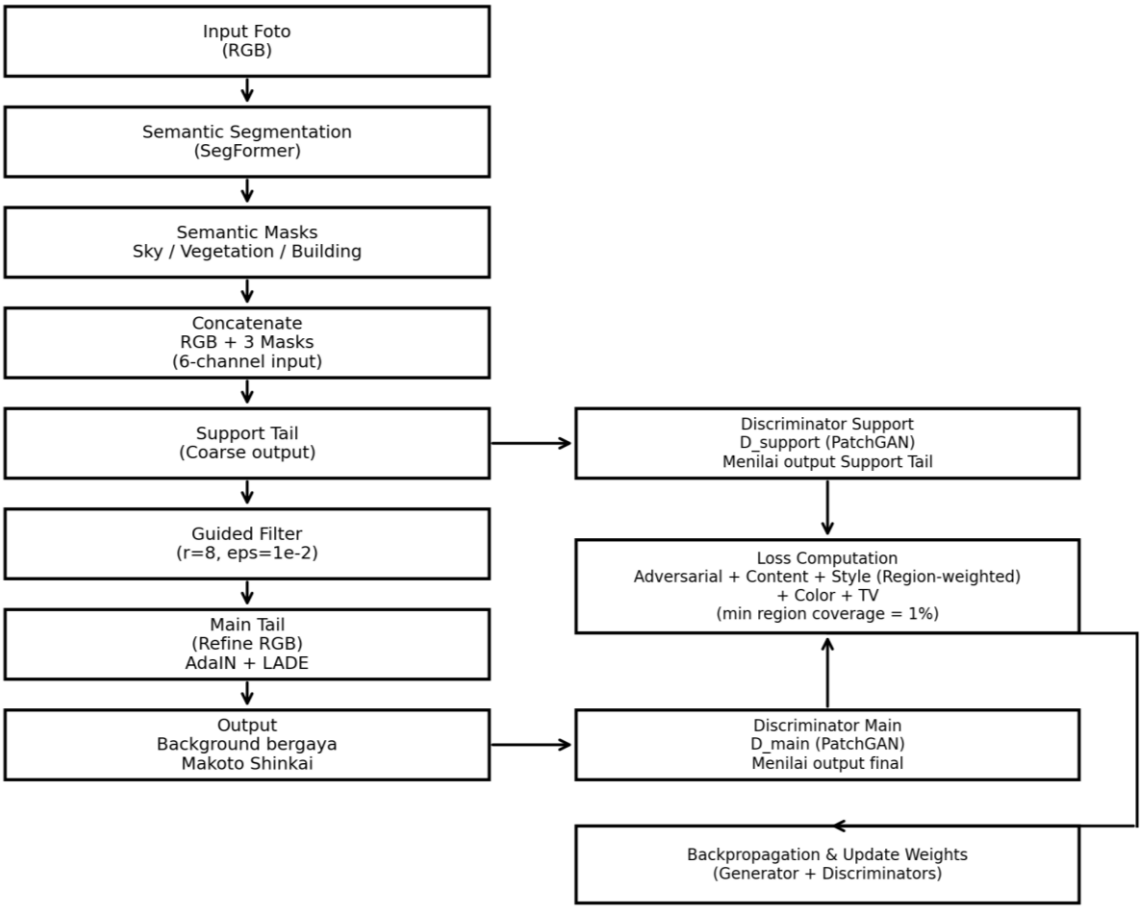

Figure 1. Overview of the Proposed Stage-2 Region-Aware DTGAN: Semantic Mask Generation, Region-Weighted Objectives, and AdaIN–LADE Adaptive Style Transfer For Shinkai-Style Background Stylization

Dataset and Pre-processing. Two image domains were prepared: (1) real-world photographs as content inputs and (2) Makoto Shinkai-style background frames as the target style domain. The photo dataset contains 6,656 outdoor and urban-scene images collected from publicly available sources. The style domain contains 1,650 Shinkai-like background frames and illustrations, then augmented to 6,600 images to reduce imbalance. All images were resized to 256x256, normalized to [-1,1], and augmented using random horizontal flip, random rotation (±10°), and random crop. The 256x256 training resolution was selected as a pragmatic compromise between stylization detail, GPU cost, and adversarial training stability for a moderate-sized dataset [15], [16].

Semantic Mask Generation. Region-aware conditioning requires pixel-level semantic masks. We used an automatic segmentation model (SegFormer-B0 fine-tuned on ADE20K) to predict semantic maps. Three binary masks were derived via class mapping: sky, vegetation (tree/grass/plant/palm), and building (building/house). To stabilize training, each mask was refined by morphological closing and opening using a 5x5 elliptical kernel. Regions with coverage below 1% were ignored to avoid unstable gradients [17]. More advanced edge-aware refinement such as guided filtering could be explored in future preprocessing, but simple morphology was sufficient for the present pipeline [18].

Region-Aware DTGAN Architecture. The proposed model follows the Double-Tail Generator structure: a support tail produces coarse stylization, and a main tail refines details. For region awareness, the input is extended from 3 channels (RGB) to 6 channels by concatenating the three semantic masks. The adaptive style transfer module applies AdaIN to transfer global style statistics and LADE to suppress local artifacts

during refinement. Discriminators follow a PatchGAN design and are trained to discriminate real Shinkai-style backgrounds from generated outputs [1], [19], [20].

Region-Weighted Objective. Let R = {sky, veg, bld} denote semantic regions with binary masks $M_r \in \{0,1\}^{H\times W}$. For each region r, we compute the region coverage c_r (pixel ratio) and use it to derive normalized weights w_r for region-specific losses.

$$c_r = (1/HW) \Sigma_{i=1}^{H} \Sigma_{j=1}^{W} M_r(i,j) \quad (1)$$

$$R^+ = \{ r \in R \mid c_r \geq \tau \} \quad (2)$$

$$w_r = c_r / (\Sigma_{k\in R^+} c_k), \quad \Sigma_{r\in R^+} w_r = 1 \quad (3)$$

$$L_{style} = \Sigma_{r\in R^+} w_r L_{style}^{(r)} \quad (4)$$

$$L_{total} = \lambda_{adv} L_{adv} + \lambda_c L_{content} + \lambda_s L_{style} + \lambda_{id} L_{id} \quad (5)$$

Here, τ is a minimum coverage threshold (we use $\tau = 1\%$ in our experiments) to ignore tiny regions that can produce unstable gradients. $L_{style}^{(r)}$ denotes the style loss computed only within region r, and λ_{adv}, λ_c, λ_s, and λ_{id} are scalar weights for adversarial, content, style, and identity losses, respectively. Coverage-aware weighting prevents dominant regions (e.g., sky) from overwhelming training while reducing region neglect for sparse regions (e.g., building facades). This balancing intuition is related to broader multi-objective weighting strategies in vision systems, where adaptive weighting helps prevent one task or loss term from dominating optimization [21].

Adaptive Instance Normalization (AdaIN) aligns the first- and second-order statistics of content features to those of a style reference. Given content feature f_c and style feature f_s, we use: [22]

$$AdaIN(f_c, f_s) = \sigma(f_s) \odot (f_c - \mu(f_c)) / (\sigma(f_c) + \varepsilon) + \mu(f_s) \quad (6)$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ are the channel-wise mean and standard deviation, $\odot$ denotes element-wise multiplication, and ε is a small constant for numerical stability. In our DTGAN refinement tail, AdaIN provides global style statistics transfer, while LADE complements it by suppressing local artifacts during denormalization.

Evaluation Protocol. We report global Fréchet Inception Distance (FID) and LPIPS, and region-level metrics to quantify style similarity: region-wise FID (computed using masks), Sky Gradient Similarity (SGS), Palette Similarity Index (PSI), and Style Feature Distance (SFD) based on deep feature statistics. Ablation experiments were conducted by selectively removing AdaIN, LADE, and semantic masks, then comparing changes on sky, vegetation, and building regions. To reduce subjectivity, SGS, PSI, and SFD are designed as region-awaremeasures linked to the visual cues of Makoto Shinkai backgrounds: (i) SGS captures the smoothness and directionality of vertical color transitions in the sky, (ii) PSI measures palette proximity in perceptual color space, and (iii) SFD quantifies deep style-feature differences using pretrained perceptual embeddings. However, proposing new metrics requires evidence that they behave consistently and reflect perceived stylization quality, this is beyond our limitation [23], [20]. The use of FID/LPIPS follows common GAN and perceptual-evaluation practice, although these metrics also have known limitations and implementation sensitivities; our SFD embedding is therefore interpreted as a complementary perceptual cue rather than a standalone judgment [24], [25], [26].

RESULTS AND DISCUSSIONS

Training and Mask Quality. Training and Mask Quality. Stage 2 training follows the same optimization setting as Stage 1 (batch size 8, 100 epochs, generator learning rate 1e-4, discriminator learning rate 2e-4). Semantic masks are generated automatically; therefore, mask quality becomes a critical factor. In our dataset, sky coverage is typically dominant, while building areas may be sparse. The proposed coverage-aware weighting mitigates this imbalance by scaling region losses according to the pixel ratio of each region. The Dataset composition and semantic mask statistics are shown at Table 1.

Table 1. Dataset composition and semantic mask statistics

Component	Quantity	Description
Real-photo dataset	6,656 images	Outdoor/urban photos used as content inputs.
Shinkai-style dataset	1,650 images (aug. to 6,600)	Background frames/illustrations as target style domain.
Semantic masks (photo)	19,968 files	Three binary masks per photo: sky, vegetation, building.
Semantic masks (style)	19,800 files	Three binary masks per style image (aligned per dataset).
Region coverage (mean)	Sky 34.3%, Veg 17.0%, Bld 3.1%	Average pixel coverage in training photos; used for adaptive weighting.

Ablation Study. To verify the contribution of each component, we conducted an ablation study across five configurations: (A) baseline DTGAN without AdaIN/LADE and without masks; (B) +AdaIN only; (C) +LADE (with AdaIN); (D) +Semantic masks (fixed region weights); and (E) +Coverage-aware adaptive weights (proposed). Table 2 shows that each component consistently improves global quality. The proposed configuration (E) achieves the best global FID (65.8) and LPIPS (0.448). The Global ablation study using FID and LPIPS are shown in Table 2.

Table 2. Global ablation study using FID and LPIPS (lower is better).

Config.	Description	FID	LPIPS	Note
A	Baseline DTGAN (no AdaIN/LADE, no masks)	92.4	0.590	Global stylization only
B	+ AdaIN	84.7	0.540	Global style stats transfer
C	+ LADE (with AdaIN)	74.5	0.505	Artifact suppression
D	+ Semantic masks (fixed weights)	69.4	0.470	Region-aware loss
E	+ Adaptive region weights (proposed)	65.8	0.448	Coverage-aware weighting

The ablation results in Table 2 indicate that each added component contributes progressively to better stylization quality. Relative to the baseline DTGAN, adding AdaIN reduces FID from 92.4 to 84.7, suggesting that global style-statistics transfer already improves alignment with the target anime domain. Introducing LADE further reduces FID to 74.5 and LPIPS to 0.505, indicating that artifact suppression is important for preserving perceptual quality. The largest improvement appears after adding semantic masks and adaptive region weighting, with the proposed model reaching 65.8 FID and 0.448 LPIPS. This trend confirms that global style transfer alone is insufficient for cinematic backgrounds, where different semantic regions require different stylization behavior.

Figure 2. Illustration of Semantic-Region-Guided Stylization In The Proposed Framework
The scene is first decomposed into semantic regions, and the resulting masks are used to guide stylization so that different areas are stylized more consistently according to their semantic roles

Figure 2 provides a conceptual illustration of the proposed region-aware stylization framework. It shows how an input scene can be decomposed into semantic regions and how region-guided stylization is intended to improve sky smoothness, palette coherence, and structural preservation in building areas. This illustration is included to clarify the expected behavior of the method, particularly in scenes containing visually dominant sky, vegetation, and building regions. The actual quantitative improvements of the proposed method are reported separately in Tables 2 and 3.

Region-Level Evaluation. Global metrics can hide region-specific artifacts; therefore, we evaluated sky, vegetation, and building regions separately. Table 3 reports region-wise FID along with SGS, PSI, and SFD. The proposed method improves all regions, with the largest relative gain on the sky region, reflecting more stable Shinkai-like gradients and palette transitions. Figure 2 visualizes the region-wise FID improvement.

Discussion and Failure Cases. The region-aware design reduces over-stylization by constraining style loss within each semantic region and by discouraging style leakage across boundaries. However, failures still occur when segmentation masks are inaccurate (e.g., thin power lines misclassified as building edges, or distant trees mislabeled as buildings). In such cases, the model may transfer vegetation textures to wall regions or produce unnatural color banding in the sky. These findings indicate that further improvements may come from mask refinement, higher-resolution segmentation, or lightweight manual correction for a small subset of difficult scenes. Table 3 reports the region-wise metrics, showing that the largest gains appear in the sky and vegetation regions when semantic masks and coverage-aware weighting are enabled. [17], [27], [28].

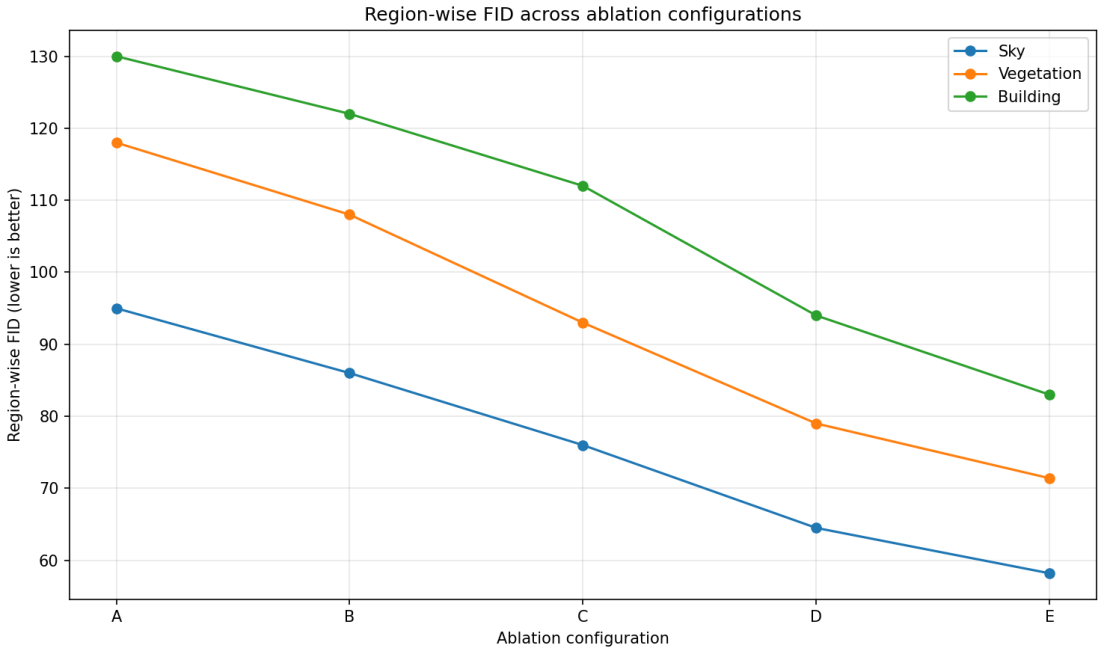

Figure 3. Region-wise FID Across Ablation Configurations (Lower Is Better)

Table 3. Region-wise Evaluation (RFID, SGS, PSI, SFD) Across Ablation Configurations

Config.	AdaIN	LADE	Masks	RFID (Sky/Veg/Bld)	SGS	PSI / SFD
A	No	No	No	95.0 / 118.0 / 130.0	0.52	0.41 / 1.15
B	Yes	No	No	86.0 / 108.0 / 122.0	0.58	0.47 / 1.03
C	Yes	Yes	No	76.0 / 93.0 / 112.0	0.63	0.55 / 0.92
D	Yes	Yes	Yes	64.5 / 79.0 / 94.0	0.70	0.61 / 0.78
E (Proposed)	Yes	Yes	Yes	58.2 / 71.4 / 83.0	0.74	0.66 / 0.70

The region-level results reveal why the proposed method is more suitable for cinematic background stylization. The largest gains appear in sky and vegetation regions, which are visually dominant and highly sensitive to stylization errors. Sky RFID decreases from 95.0 to 58.2, while vegetation RFID decreases from 118.0 to 71.4. At the same time, SGS increases from 0.52 to 0.74, showing that the proposed model better preserves smooth vertical sky gradients.

In contrast, our goal is style translation for real photos into cinematic anime backgrounds while maintaining the original DTGAN refinement design. Therefore, we adopt a lighter region-awareness mechanism: semantic masks act as (1) additional conditioning input channels and (2) region-wise loss selectors with coverage-aware weights. This choice preserves the DTGAN two-tail refinement pathway and avoids over-constraining the generator with hard semantic normalization, which can cause over-stylization or texture "stamping" when masks are imperfect. Nevertheless, our method is not universally superior; if high-fidelity semantic control is required (e.g., fully controllable synthesis from masks), SPADE/SEAN may be preferable. For a compact comparison, Table 4 summarizes key differences [6],[7].

Table 4. Comparison with Semantic-Normalization Region-Aware GANS (SPADE, SEAN) Versus Our Objective-Level Semantic Guidance

Aspect	Our approach	SPADE	SEAN	Implication
semantics enters	Input + loss (region-weighted)	Normalization layers	Normalization layers + style	Affects training stability & compute
Need masks at inference	Yes (for region metrics/loss)	Yes	Yes	Deployment complexity
Parameter overhead	Low	Medium–High	High	Risk of overfitting on small data
Mask sensitivity	Medium	High	High	Failure cases differ
Best suited for	Photo→anime stylization	Semantic synthesis	Semantic style editing	Match method to task

Compared with SPADE and SEAN, our method is less dependent on heavy semantic-normalization layers and rich style codes. This makes it more practical for limited photo-to-anime datasets while preserving the original DTGAN refinement structure. Compared with representative photo-to-anime stylization baselines from Stage 1, the current contribution is not merely another global stylization variant; rather, it addresses a different problem, namely region neglect and style leakage across sky, vegetation, and building regions. Therefore, the novelty of this study lies in adding local semantic control to an already strong stylization backbone.

Discussion and Failure Cases. The region-aware design reduces over-stylization by constraining style loss within each semantic region and by discouraging style leakage across boundaries. However, failures still occur when segmentation masks are inaccurate (e.g., thin power lines misclassified as building edges, or distant trees mislabeled as buildings). In such cases, the model may transfer vegetation textures to wall regions or produce unnatural color banding in the sky. These findings indicate that further improvements may come from mask refinement, higher-resolution segmentation, or lightweight manual correction for a small subset of difficult scenes. As shown in Table 5, the DTGAN baseline with AdaIN-LADE already provides a strong starting point relative to representative photo-to-anime baselines. Building upon this competitive backbone, the current study focuses on a different gap: improving per-region style fidelity and reducing region neglect using semantic masks and coverage-aware weighting (see Table 2 and Table 3). [1], [19], [11], [12], [13], [14].

Table 5. Comparison with Representative Photo-To-Anime Stylization Baselines (Stage-1 Benchmark Setting)

Method	Type	Region-aware	FID ↓	PSNR ↑ (dB)
CartoonGAN	Unpaired GAN	No	65.2	16.3
AnimeGAN	Unpaired GAN	No	59.8	17.2
AnimeGANv2	Unpaired GAN	No	51.6	18.0
White-box	Rule+GAN hybrid	Partial	78.3	15.5
DTGAN only	Double-tail GAN	No	44.1	18.1
DTGAN + AdaIN + LADE	DTGAN + adaptive style	No	38.7	19.4

CONCLUSION

This paper presented a region-aware extension of DTGAN for cinematic anime background stylization in Makoto Shinkai style. The main contributions are threefold: the integration of semantic masks as lightweight conditioning inputs, the introduction of coverage-aware region-weighted losses, and the demonstration that these components improve both global and region-level stylization quality. Experimental results show consistent improvements over the non-region-aware DTGAN in FID, LPIPS, RFID, SGS, PSI, and SFD. Nevertheless, limitations remain when semantic masks are noisy, especially in thin structures or distant objects, which can still lead to style leakage or local artifacts. Future work will focus on improving segmentation robustness, exploring higher-resolution stylization, and incorporating lighting-specific guidance to better match cinematic anime atmosphere.

Limitations remain in scenes where semantic masks are noisy or where small objects dominate the composition. Future work will focus on improving segmentation robustness and exploring lighting-specific guidance (e.g., bloom/lens-flare maps) to further match Shinkai's cinematic atmosphere [17], [27], [28], [29].

This paper presented a region-aware extension of DTGAN for cinematic anime background stylization in Makoto Shinkai style. By adding semantic masks as conditioning inputs and by introducing coverage-aware region-weighted losses, the model gains better control over sky, vegetation, and building regions. Experimental results show consistent improvements over the non-region-aware DTGAN, both globally (FID/LPIPS) and on region-level metrics (RFID, SGS, PSI, SFD).

REFERENCES

[1] A. Purwanto, K. Kusrini, E. Utami, and D. Agustriawan, "Development of Double-Tail Generative Adversarial Network with Adaptive Style Transfer for Anime Background Production with Makoto Shinkai's Stylization," *Scientific Journal of Informatics*, vol. 12, no. 1, pp. 159–170, 2025.

[2] I. G. et al., "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[3] B. Li, Y. Zhu, Y. Wang, C. W. Lin, B. Ghanem, and L. Shen, "AniGAN: Style-Guided Generative Adversarial Networks for Unsupervised Anime Face Generation," *IEEE Trans. Multimedia*, vol. 24, pp. 4077–4091, 2022, doi: 10.1109/TMM.2021.3113786.
[4] L. Gatys, A. Ecker, and M. Bethge, "Image Style Transfer Using Convolutional Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[5] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural Style Transfer: A Review," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 11, pp. 3365–3385, 2020, doi: 10.1109/TVCG.2019.2921336.
[6] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic Image Synthesis with Spatially-Adaptive Normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[7] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, "SEAN: Image Synthesis with Semantic Region-Adaptive Normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
[8] N. Qi, Y. Li, R. Fu, and Q. Zhu, "STRAT: Image style transfer with region-aware transformer," *Neurocomputing*, vol. 617, p. 129039, 2025, doi: 10.1016/j.neucom.2024.129039.
[9] Z. Li *et al.*, "Parsing-Conditioned Anime Translation: A New Dataset and Method," *ACM Trans. Graph.*, vol. 42, no. 3, pp. 1–14, 2023, doi: 10.1145/3585002.
[10] Y.-S. Liao and C.-R. Huang, "Semantic Context-Aware Image Style Transfer," *IEEE Transactions on Image Processing*, vol. 31, pp. 1911–1923, 2022, doi: 10.1109/TIP.2022.3149237.
[11] J. Chen, G. Liu, and X. Chen, "AnimeGAN: A Novel Lightweight GAN for Photo Animation," in *Artificial Intelligence Algorithms and Applications: ISICA 2019*, Springer, 2020, pp. 242–256.
[12] Y. Dong, W. Tan, D. Tao, L. Zheng, and X. Li, "CartoonLossGAN: Learning surface and coloring of images for cartoonization," *IEEE Transactions on Image Processing*, vol. 31, pp. 485–498, 2021.
[13] Y. Jiang, L. Jiang, S. Yang, and C. C. Loy, "Scenimefy: Learning to Craft Anime Scene via Semi-Supervised Image-to-Image Translation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
[14] N. Ahn, J. Seo, and K. Sohn, "Interactive Cartoonization with Controllable Perceptual Factors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
[15] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8110–8119.
[16] Z. Li *et al.*, "A Systematic Survey of Regularization and Normalization in GANs," *ACM Comput. Surv.*, vol. 55, no. 11, pp. 1–37, 2023, doi: 10.1145/3569928.
[17] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
[18] K. He, J. Sun, and X. Tang, "Guided Image Filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, 2013, doi: 10.1109/TPAMI.2012.213.
[19] G. Liu, X. Chen, and Z. Gao, "A Novel Double-Tail Generative Adversarial Network for Fast Photo Animation," *IEICE Trans. Inf. Syst.*, vol. E107.D, no. 1, pp. 72–82, 2024, doi: 10.1587/transinf.2023EDP7061.
[20] R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
[21] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
[22] X. Huang and S. Belongie, "Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.
[23] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[24] A. Borji, "Pros and cons of GAN evaluation measures: New developments," *Computer Vision and Image Understanding*, vol. 215, p. 103329, 2022, doi: 10.1016/J.CVIU.2021.103329.

[25] T. Kynkäänniemi, T. Karras, M. Aittala, T. Aila, and J. Lehtinen, "The role of ImageNet classes in Fréchet Inception Distance," *ArXiv*, 2022.
[26] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *ArXiv*, 2014.
[27] B. Cheng, A. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention Mask Transformer for Universal Image Segmentation (Mask2Former)," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
[28] A. K. et al., "Segment Anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
[29] L. Zhang, A. Rao, and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models (ControlNet)," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.