Electronic Letters on Computer Vision and Image Analysis 25(2):22-44, 2026

seGANmentation: Rethinking Semantic Segmentation as a Generative Framework

Maryam Berijanian[1], Max Fahrenkrug[2], Abdi Kusata[3], Amin Sehati[4], Dirk Colbry[1]

[1] *Michigan State University, East Lansing, MI, USA*
[2] *University of Notre Dame, Notre Dame, IN, USA*
[3] *Lafayette College, Easton, PA, USA*
[4] *Trine University, Allen Park, MI, USA*

Received 7th of September 2024; accepted 3rd of July 2026

Abstract

This research explores the potential of state-of-the-art Generative Adversarial Networks (GANs) for segmentation via image-to-label (I2L) translation. The presented model, seGANmentation, integrates convolutional neural networks and vision transformers to directly generate segmentation labels from input images. The primary goal involves reducing manual labeling constraints and enhancing segmentation performance using synthetic data. Evaluating the framework against a suite of descriminative baselines isolates the impact of adversarial I2L translation across different architectural paradigms. Results are additionally contextualized against GAN frameworks such as Pix2Pix and SegAN. While the fully supervised discriminative models achieve near-perfect pixel-wise overlap, seGANmentation yields an Accuracy of 0.90, a Dice Coefficient of 0.83, and an IoU of 0.71. However, extended evaluation using Fréchet Inception Distance (FID) reveals a trade-off: the adversarial approach successfully learns a distributionally realistic prior over the mask domain, generating smoother silhouettes that outperform lightweight discriminative models prone to jagged upsampling artifacts. Although seGANmentation trails in pixel-perfect boundary adherence, it establishes the viability of a purely generative approach to segmentation. The proposed framework offers particular value for data-scarce scenarios, potentially enhancing downstream task accuracy through synthetic data augmentation. Furthermore, utilizing GAN-generated masks as initial priors for traditional segmentors could significantly reduce the requirement for extensive training pairs and open pathways to unsupervised segmentation. These hypotheses remain open for future studies. The source code is publicly available at: https://github.com/maryambrj/seGANmentation.

Key Words: Generative Adversarial Network (GAN), Vision Transformer (ViT), Image Segmentation, Image-to-Image (I2I) Translation, Synthetic Data Augmentation, Data Scarcity.

Correspondence to: berijani@msu.edu

Recommended for acceptance by Angel D. Sappa
https://doi.org/10.5565/rev/elcvia.1995
ELCVIA ISSN:15108-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

1 Introduction

Unpaired image-to-image (I2I) translation is a fundamental problem in computer vision. It enables the mapping between two data domains without requiring paired examples. Generative Adversarial Networks (GANs) have advanced this field by enabling the generation of high-quality, realistic images. GANs have been applied successfully in various tasks such as image generation, I2I translation, and facial attribute manipulation [1].

The primary motivation for this work stems from recent advancements in deep learning which show that combining different architectural paradigms can lead to significant improvements. We integrate convolutional neural networks (CNNs) with vision transformers (ViTs) [2] within a GAN-based framework. CNNs excel at capturing local features and textures, while ViTs are adept at modeling global relationships. This hybrid approach is particularly beneficial for segmentation tasks, where both fine details and overall context are essential.

The task of I2I translation has been revolutionized by GANs and Diffusion Models (DMs). GANs enable the generation of diverse probability densities and support a variety of generator architectures. Their flexibility has resulted in state-of-the-art performance for synthesizing image data [1]. In contrast, traditional DMs often require high computational resources and multiple evaluations during inference [3].

Semantic segmentation has also advanced through the use of CNNs. Models such as U-Net, SegNet [4], FCN [5], and Mask R-CNN [6] have become cornerstones in applications from driverless cars to medical imaging [7]. However, a major challenge remains: deep NNs require a large number of manual labels for training.

A central obstacle to broader deployment of semantic segmentation is the *labeling bottleneck*: pixel-level annotation is expensive, slow, and difficult to scale across domains and sensing conditions [**?**]. This creates a gap between the practical availability of labeled data and the data requirements of modern segmentation systems. In addition, purely discriminative segmentation models are typically optimized with per-pixel supervision and therefore struggle when labels are scarce [8]; moreover, they do not explicitly model the *distribution* of valid masks, limiting their ability to enforce global structural priors (e.g., coherent shapes and plausible region topology) under weak supervision [9, 10].

To address these limitations, we formulate segmentation as a domain-translation problem, learning a mask-domain prior via adversarial training rather than relying solely on dense pixel-wise supervision. We pursue a *generative* perspective where a GAN is used to learn the style and global structure of segmentation masks (i.e., what a valid mask should look like) and leverages this learned prior during training. Unlike traditional GAN-based segmentation methods that use adversarial critics as auxiliary refiners or rely on indirect adaptation pipelines, our framework, *seGANmentation*, treats segmentation as a direct, standalone Image-to-Label (I2L) engine. As detailed in Table 1, existing adversarial frameworks often face methodological gaps, such as requiring fully paired supervision, functioning only as secondary refiners, or being restricted to simple geometric priors. We leverage style modulation and a hybrid ViT-CNN generator to ensure that generated masks are not only spatially aligned but also distributionally realistic. The novelty of seGANmentation lies in its ability to explicitly learn a global structural prior over the mask domain, enabling segmentation with substantially reduced supervision by complementing overlap-based objectives with adversarial learning.

While GANs have been used for I2I translation [13], their application as a direct I2L engine that leverages global contextual relationships (via Transformers) remains under-explored [15]. Current GAN-based segmentation often relies on simple geometric priors or serves merely as a pre-processing step [11]. These suggest a gap between unpaired translation methods that excel at image-domain mapping and segmentation systems that require dense supervision or use adversarial learning only as an auxiliary mechanism. This motivates the need for a framework that can generate both realistic synthetic images and their corresponding masks simultaneously [16], avoiding the high-frequency checkerboard artifacts common in lightweight discriminative models.

This work proposes a novel GAN-based framework to address this challenge. Our approach directly translates images into their corresponding segmentation labels. In doing so, it not only generates new images suitable for data augmentation but also creates their labels. This reduces the reliance on manual annotation. The dual benefits of data augmentation and segmentation are expected to enhance the accuracy of networks designed for semantic segmentation tasks.

Table 1: Analysis of related adversarial segmentation frameworks and identified methodological limitations.

Method	Application	Key Techniques	Limitations
Gadermayr et al. [11]	Domain-Independent Segmentation	Dual translation pipelines	Relies on indirect domain adaptation pipelines; lacks direct I2L translation capability
Gadermayr et al. [12]	Digital Pathology Segmentation	GAN with adversarial loss and cycle consistency; artificial annotation generation	Restricted to simple, handcrafted geometric priors (circles/ellipses); insufficient for complex object segmentation
Pix2Pix [13]	Paired I2I Translation	Conditional GAN combined with traditional L1 loss	Requires fully paired image–label supervision; adversarial loss shown to reduce segmentation IoU compared to regression baselines
SegAN [14]	Medical Image Segmentation	Multi-scale adversarial training	Functions primarily as an auxiliary refiner

We address these limitations by implementing a hybrid ViT-CNN architecture within an adversarial framework. By leveraging both convolutional layers for local features and a Transformer bottleneck for global context, seGANmentation treats segmentation as a direct domain-translation task. Technical specifics of the architectural integration and style modulation are detailed in Section 3.2.

This combination enables UVCGAN to produce high-quality translations that maintain strong correlations between the original and translated images, a feature crucial for scientific simulations. This ensures a better starting state for the generators and contributes significantly to the model's efficacy. The architecture change, combined with these training enhancements, is critical for UVCGAN's improved performance.

The model's effectiveness is quantified using metrics like Fréchet Inception Distance (FID) and Kernel Inception Distance (KID), where lower scores denote higher similarity to the target domain images. UVCGAN excels in these metrics, demonstrating its superior performance both quantitatively and qualitatively in I2I translation tasks. It effectively handles long-range spatial dependencies, a challenge in scientific domain research, while maintaining crucial features of the source images.

Building upon the success of UVCGAN, the advanced UVCGAN-v2 introduces further enhancements. This revised model features a redesigned generator with style modulation, in which images are encoded as a sequence of tokens and processed by the Transformer network [17]. A learnable style token, S, is introduced, which upon processing by the Transformer, serves as a latent image style. This style is then applied across the U-Net's decoding branch through specific style vectors derived from S, enhancing the generator's capability. Such architectural nuances enable UVCGAN-v2 to directly deduce the target style from the source image.

These innovations directly support our approach and the primary reason for selecting this model as the backbone of our work. By leveraging UVCGAN-v2's ability to generate detailed and stylistically consistent translations, our framework can directly convert images into segmentation labels. This capability reduces the need for extensive manual labeling and makes our method particularly attractive in data-scarce scenarios. The integration of UVCGAN-v2 provides both the fine-detail preservation of U-Net architectures and the global context understanding of Transformers, which are essential for generating accurate segmentation masks.

To ensure the integrity of our implementation, we successfully reproduced the original I2I translation results reported in [18]; qualitative samples of this reproduction are provided in Appendix A.

From an evaluation perspective, this work distinguishes between two types of baselines. First, we conduct a controlled architectural comparison against a comprehensive suite of discriminative segmentation models, including U-Net [19], DeepLabV3+ [20], LinkNet [21], and SegFormer [22]. This evaluates the generative framework across established architectural paradigms, ranging from lightweight convolutional networks to vision transformers. Second, we contextualize our findings against representative GAN-based segmentation approaches such as Pix2Pix [13] and SegAN [14], which employ adversarial learning in fundamentally different roles. This distinction isolates the effect of treating segmentation as a direct I2L translation problem, rather than

Table 2: Summary of foundational GAN architectures for unsupervised I2I translation.

Method	Application	Key Techniques	Notes
CycleGAN [23]	Unpaired I2I Translation	Cycle-consistency	Baseline method for unpaired translation
Berijanian et al. [24]	Medical I2I Translation	Unsupervised many-to-many translation	Streamlines translation between multiple domains
UVCGAN [25]	Unpaired I2I Translation	Hybrid U-Net + ViT; self-supervised pre-training; gradient penalty	Preserves both fine details and global context
UVCGAN-v2 [18]	Unpaired I2I Translation	Style modulation; token-based encoding; Transformer	Superior FID and KID performance

as localized pixel-wise regression or adversarial refinement. The key contributions of this work are twofold:

- Development of a hybrid UNet-ViT framework for direct I2L translation, based on UVCGAN-v2 [18], converting images into their corresponding segmentation labels.
- Evaluation of the proposed framework against a diverse suite of discriminative baselines (U-Net, LinkNet, SegFormer, DeepLabV3+) to isolate the impact of generative adversarial learning.

2 Related Work

In the field of synthetic image generation and unpaired I2I translation, GANs have provided a foundational framework for learning mappings between domains without paired supervision. Table 2 summarizes representative translation backbones that are most relevant to this work. CycleGAN [23] established cycle consistency as a practical constraint for unpaired translation, enabling one-to-one mappings that are especially important when semantic content must be preserved. Berijanian et al. [24] introduced an unsupervised many-to-many image translation framework for medical applications, enabling translation between any two image domains using a single GAN, thus eliminating the need for training separate GANs for each domain pair and emphasizing scalability across multiple domains. However, adapting these general-purpose translation models for segmentation requires moving beyond mere texture mapping to explicitly preserving global object topologies.

While alternative generative paradigms such as Diffusion models (DMs) show impressive performance, they present analytical limitations for our specific objective, such as neglecting source domain data during training, high computational demands, and reliance on pixel-wise errors for image consistency. Because segmentation relies heavily on spatial correspondence between the source image and target mask, UVCGAN addresses these challenges by building upon the CycleGAN framework [23, 25] and marks a pivotal advancement.

CycleGAN was one of the first methods to enable unpaired I2I translation by enforcing cycle-consistency, which is crucial for scientific applications requiring one-to-one mappings. However, many subsequent models have relaxed this constraint to boost image generation diversity, which sometimes altered crucial image characteristics [25]. To bridge the gap between localized pixel translation and structural coherence, UVCGAN [25] preserves cycle consistency while incorporating modern stabilization strategies (e.g., self-supervised pre-training and gradient regularization) and a hybrid U-Net+ViT generator to capture both local detail and global structure. This combination sets UVCGAN apart from predecessors like CycleGAN, UNIT [26], and MUNIT [27]. UVCGAN-v2 [18] further improves translation quality via style modulation and token-based encoding, demonstrating that hybrid CNN–Transformer generators can produce spatially consistent outputs under challenging domain shifts. These properties make UVCGAN-v2 an appropriate backbone for our setting, where mask realism and global structural coherence are central concerns.

Transitioning from general-purpose translation to segmentation, an analysis of prior adversarial approaches reveals that they differ primarily in how GANs are integrated into the learning pipeline. They can be grouped

Table 3: Comparison between discriminative segmenters and the proposed generative I2L framework.

Feature	Discriminative (e.g. U-Net)	Generative (seGANmentation)
Objective	Minimize per-pixel error	Model realistic mask distributions
Strengths	Exact boundary adherence	Shape coherence & silhouette smoothness
Failure Mode	Jagged edges & pixelated artifacts	Spatial misalignment
Utility	Direct segmentation only	Simultaneous data augmentation

into three categories, each with limitations that motivate our approach. Some methods rely on paired supervision (e.g., Pix2Pix [13]), others use adversarial critics as auxiliary refiners (e.g., SegAN [14]), and others (e.g., [11]) employ indirect domain adaptation pipelines. These are central to understanding how seGANmentation differs from prior work. Table 1 summarizes these representative examples and their key methodological gaps.

First, paired translation frameworks such as Pix2Pix [13] show that conditional GANs can learn accurate I2L mappings when fully paired image–mask supervision is available. However, this requirement directly conflicts with the labeling bottleneck that motivates our work: Pix2Pix assumes dense paired data and therefore does not reduce annotation effort. Moreover, in segmentation settings, adversarial objectives alone do not guarantee improved overlap metrics compared with regression-based training, which reinforces the need for segmentation-aware objectives rather than purely realism-driven supervision.

Second, several adversarial segmentation approaches (e.g., SegAN [14]) incorporate a discriminator to improve mask quality, but the GAN component typically serves as an auxiliary refiner on top of a conventional supervised segmentor. In other words, the segmentation network remains the primary engine and still depends on dense labeled training, while the adversarial loss is used mainly to sharpen or regularize predictions.

Third, domain adaptation pipelines such as Gadermayr et al. [11] aim for domain-independent segmentation by translating images between domains and then applying a supervised segmentor in the target domain. While effective for cross-domain transfer, this strategy is indirect: it relies on a multi-stage translation-and-segment procedure and does not provide an end-to-end I2L translation model. Related work in digital pathology [12] further reduces supervision by generating artificial annotations (e.g., circles/ellipses) and translating them into realistic masks. However, such geometric priors are intrinsically restricted and do not generalize to complex object boundaries (e.g., cars), making them unsuitable for the segmentation regimes considered here.

Taken together, Table 1 highlights a common gap in prior work: existing GAN-based segmentation approaches either require fully paired image–label supervision (e.g., Pix2Pix [13]), employ adversarial learning only as a secondary refinement on top of a conventional segmentor (e.g., SegAN [14]), or rely on indirect domain-adaptation or simple label priors that do not generalize to complex shapes [11, 12]. While these formulations (paired, unpaired, refinement-based, and domain-adaptation pipelines) have demonstrated success within their problem settings, they do not treat segmentation itself as a first-class generative translation task.

In contrast, our objective is to formulate segmentation explicitly as an I2L translation problem, in which a generative model directly produces masks, learns a prior over valid structures via adversarial training, and optimizes segmentation-aware losses that align generated masks with ground truth. seGANmentation addresses existing gaps by adapting the UVCGAN-v2 framework due to its unique handling of long-range dependencies through a Transformer bottleneck, which is critical for maintaining the global topological coherence of objects (e.g., the silhouette of a car). While models like Pix2Pix require dense paired data, our adaptation moves toward a framework capable of enforcing global structural realism via adversarial learning in the target domain, providing a pathway toward segmentation in data-scarce regimes where per-pixel supervision is limited. Based on the insights gained from our results, Table 3 summarizes these fundamental differences, contrasting the proposed generative approach with traditional discriminative models across key operational dimensions.

From an evaluation standpoint, a comparison is established against a comprehensive suite of discriminative baselines (U-Net, DeepLabV3+, LinkNet, and SegFormer). Evaluating against these diverse architectures isolates the effect of adversarial mask-domain learning and generative training objectives from architectural differences. Furthermore, comparisons with Pix2Pix and SegAN serve to contextualize the findings across dif-

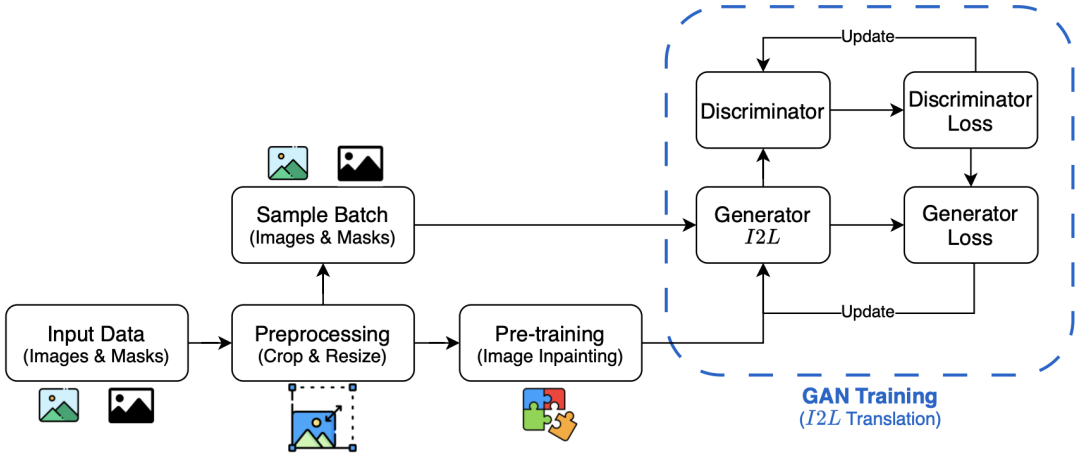


Figure 1: Flowchart of the seGANmentation training process, illustrating the combined influence of adversarial learning and overlap-based segmentation objectives.

ferent GAN-based segmentation paradigms, rather than acting as direct architectural baselines. This framing motivates the methodology described in the next section.

3 Methodology

3.1 Overview

This work introduces *seGANmentation*, a GAN-based framework that formulates semantic segmentation as a *domain translation* problem from the RGB image domain I to the mask domain M. Inspired by the hybrid U-Net–ViT architecture of UVCGAN-v2 [18], the objective is to learn a generator G that produces a segmentation mask $\hat{y} = G(x)$ that is *structurally aligned* with the ground-truth mask $y \in M$ and *realistic* under the mask domain distribution. This design reduces reliance on dense manual annotation and simultaneously enables synthetic image–mask data augmentation for data-scarce settings.

The end-to-end training pipeline, illustrated in Figure 1 consists of the following core stages:

- Self-supervised pre-training: Following standard preprocessing, the generator G is pre-trained via an image inpainting task to learn visual representations without annotations.
- Feature Extraction: Images are processed by the hybrid generator, where the U-Net backbone captures local spatial details (e.g., object boundaries) and the ViT bottleneck models global context.
- Adversarial I2L Training: The generator is fine-tuned iteratively using a composite loss function (Eq. 3). This combines an adversarial discriminator D (enforcing mask realism), overlap-based segmentation objectives (IoU and Dice), and an optional cycle-consistency path to preserve structural integrity.

To isolate the effects of generative mask-domain learning from architectural biases, the primary comparison is conducted against a comprehensive suite of discriminative baselines. This suite includes U-Net [19] as the standard encoder-decoder model. We selected DeepLabV3+ [20] as an additional baseline because its use of Atrous Spatial Pyramid Pooling (ASPP) [28] provides a robust, modern contrast to U-Net's traditional skip-connection design. Furthermore, to compare our generative approach against a supervised model optimized for inference efficiency rather than pixel-perfection, we included LinkNet [21], which utilizes a lightweight decoder designed for real-time edge device inference. Finally, the inclusion of SegFormer [22], a purely Transformer-based architecture, extends the evaluation to ViTs and rigorously contrasts with our hybrid ViT-CNN approach. Together, this approach enables an evaluation across diverse structural paradigms.

At test time, this workflow enables direct I2L inference via a single forward pass, providing a flexible foundation for data augmentation, semi-supervised learning, and future extensions to unsupervised segmentation.

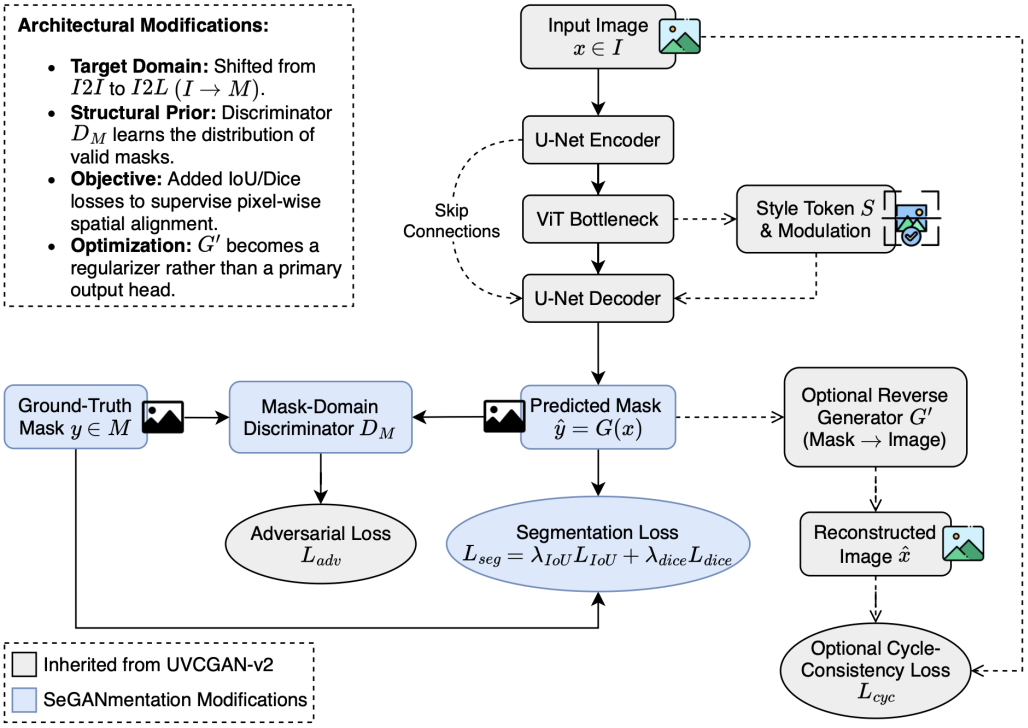


Figure 2: The seGANmentation architecture, featuring a hybrid ViT-CNN generator and style modulation based on UVCGAN-v2 [18]. Blue components indicate task-specific elements designed for direct I2L translation and mask-domain adversarial learning. The output is a segmentation mask $\hat{y}$, a mask-domain discriminator is used, and segmentation-aware losses L_{seg} are integrated into the generator objective.

3.2 Hybrid Architecture Design

The core of seGANmentation is a hybrid UNet-ViT generator based on UVCGAN [25] and its successor, UVCGAN-v2 [18]. It builds on the cycle-consistency principle of CycleGAN [23] by combining a U-Net encoder–decoder with a ViT bottleneck and employing modern GAN stabilization techniques, including spectral normalization, gradient penalty, and exponential moving average of generator weights. This specific combination is utilized because the U-Net encoding path extracts input features and uses skip connections to pass high-frequency, fine spatial details to the decoder, which is crucial for preserving object boundaries, while the ViT bottleneck learns pairwise relationships among low-frequency features and captures long-range dependencies and global context, which are crucial for coherent object shape. In addition, UVCGAN-v2 introduces a modular discriminator architecture, comprising a traditional body and a specialized head, which helps prevent mode collapse and improves feature comparison between target and translated images.

To formulate this explicitly for segmentation rather than general image translation, the architecture incorporates the following characteristics:

- I2L translation: the target domain is a segmentation mask rather than an RGB image.
- Segmentation-aware supervision: an explicit segmentation loss L_{seg} based on IoU and Dice (Eq. 6) is added to directly optimize overlap with ground truth.
- Optional cycle regularization: since mask→image translation is not required at inference, we evaluate both a one-way I2L model and an optional two-way variant with cycle consistency (Eq. 3–7).

Figure 2 provides an overview of these modifications, highlighting how the UVCGAN-v2 backbone is adapted for direct I2L segmentation while preserving the hybrid U-Net–ViT generator and style modulation.

3.3 Model Components

Generator. The generator G follows the UVCGAN-v2 [18] hybrid design, combining a U-Net encoder–decoder with a ViT bottleneck. The U-Net path preserves spatial detail via skip connections (useful for boundaries and thin structures), while the transformer bottleneck captures global context (useful for coherent object shape). In the I2L setting, G maps an RGB image x directly to a mask $\hat{y}$.

Discriminator. A discriminator D operates in the mask domain and is trained to distinguish real masks y from generated masks $\hat{y} = G(x)$. This adversarial pressure encourages outputs that are globally mask-like (e.g., coherent regions, plausible shapes), complementing overlap-based segmentation losses.

Optional reverse direction. In the default I2L setting, the reverse mapping (mask $\rightarrow$ image) is unnecessary and can be removed. However, we also evaluate an optional reverse generator G' and a second discriminator D' to enable cycle consistency, as a regularizer that can improve stability and mask quality in some regimes.

3.4 Training Procedure

Training is conducted in two stages, matching the flow in Figure 1.

Stage 1: Self-supervised pre-training via self-supervision. Following UVCGAN-v2 [18], we pre-train G using an image inpainting task to learn visual representations before adversarial training. This step, as highlighted in the paper, significantly enhances performance when the same dataset is used for both pre-training and the translation task. Each input image is divided into 32×32 patches, and 40% of patches are randomly masked by zeroing their pixel values. The generator is trained to reconstruct the original image by minimizing pixel-wise L_1 reconstruction loss. The resulting weights initialize G for the I2L task (Figure 1, "Pre-training").

Stage 2: I2L adversarial training with optional cycle consistency. In each iteration, we sample a batch of paired examples (x, y) (image and corresponding ground-truth mask). The generator predicts a mask $\hat{y} = G(x)$. The discriminator is updated to classify y as real and $\hat{y}$ as fake, and the generator is updated to fool the discriminator and minimize segmentation error relative to y. This corresponds to the "Generator/Discriminator $\rightarrow$ Loss $\rightarrow$ Update" loop in Figure 1. When cycle consistency is enabled, we additionally compute an image reconstruction $\hat{x} = G'(\hat{y})$ and enforce a reconstruction constraint between x and $\hat{x}$. This adds a structural regularizer that can reduce degenerate mappings and improve alignment, at the cost of additional compute.

3.5 Loss Functions

We retain the loss structure from UVCGAN-v2 where applicable, while adding segmentation-aware terms to explicitly optimize mask quality.

Discriminator objectives. The discriminators D and D' function as binary classifiers, distinguishing between real images from the original training set and synthetic outputs from the generators G and G'. The discriminators output, $D(x)$ or $D'(x)$, is either 0 or 1 depending on whether the input x is classified as fake or real, respectively. The discriminators are updated during training using the discriminator losses L_D and $L_{D'}$ for the forward and reverse directions defined as:

$$\begin{aligned} L_D &= E_{x \in I} L_c(D(G(x)), 0) + E_{y \in M} L_c(D(y), 1) + \lambda_{GP} L_{GP} \\ L_{D'} &= E_{y \in M} L_c(D'(G'(y)), 0) + E_{x \in I} L_c(D'(x), 1) + \lambda_{GP} L'_{GP} \end{aligned} \tag{1}$$

where I represents the set of input images and M denotes the ground-truth segmentation masks. $G(x)$ and $G'(x)$ are the outputs generated by the generators, and L_c can be any suitable binary classification loss function. The gradient penalty losses L_{GP} and L'_{GP}, also named as γ-centered GP regularization are defined as:

$$L_{GP} = E\left[\frac{(\|\nabla_x D(x)\|_2 - \gamma)^2}{\gamma^2}\right] \quad , \quad L'_{GP} = E\left[\frac{(\|\nabla_x D'(x)\|_2 - \gamma)^2}{\gamma^2}\right]. \tag{2}$$

The gradient penalty loss enhances training stability and reduces hyperparameter sensitivity, as evidenced in [18]. The study confirms that using GP with pre-trained generators yields superior performance. The coefficient λ_{GP} controls the weight of the gradient penalty terms.

Generator objectives. The generator losses L_G and $L_{G'}$ combine adversarial, cycle, identity, consistency, and segmentation terms:

$$\begin{aligned} L_G &= L_{adv} + \lambda_{cyc} L_{cyc} + \lambda_{idt} L_{idt} + \lambda_{cons} L_{cons} + L_{seg} \\ L_{G'} &= L'_{adv} + \lambda'_{cyc} L'_{cyc} + \lambda_{idt} L'_{idt} + \lambda_{cons} L'_{cons} + L'_{seg} \end{aligned} \tag{3}$$

The coefficients λ_{cyc}, λ'_{cyc}, λ_{idt} and λ_{cons} are hyperparameters. Note that λ_{cyc} and λ'_{cyc} are independent coefficients for the two directions of translation. This is the only loss term that is asymmetric because one is related to the reconstruction of images and the other to the reconstruction of masks, with their loss functions differing accordingly. For image reconstruction, the loss L_{cyc} is based on the pixel-wise difference between the initial and reconstructed image, while for L'_{cyc} this paper used a different loss as explained later.

The individual losses, namely L_{adv} (adversarial loss), L_{cyc} (cycle consistency loss), L_{idt} (identity loss), L_{cons} (consistency loss), and L_{seg} (segmentation loss), are each defined in the subsequent paragraphs.

Adversarial, identity, and cycle losses.

$$\begin{aligned} L_{adv} &= E_{x \in I} L_c(D(G(x)), 1) \\ L_{cyc} &= E_{x \in I} L_{pix}(G'(G(x)), x) \\ L_{idt} &= E_{y \in M} L_{pix}(G(y), y) \end{aligned} \tag{4}$$

Here, L_{adv} is defined in opposition to the discriminator's objectives. The adversarial loss is a fundamental component of all GANs, fostering a competitive dynamic between the generator and discriminator. The cycle consistency loss, used in all CycleGAN-like models, ensures that translating an image to the label domain and back to the image domain yields the original image, thereby maintaining consistency and reducing artifacts. As shown in [18], the combination of identity loss with GP loss enhances overall performance.

In our experiments, the L_1 loss function was chosen for L_{idt}, but experiments were condcucted with both L_1 and L_2 for L_{cyc} in the I2L direction. L_{pix} represents any pixel-wise loss function capable of quantifying the difference between its inputs, such as L_1 or L_2 loss.

Consistency loss. To encourage structural agreement between input images and generated masks at a coarse scale, we use:

$$L_{cons} = E_{x \in I} L_1(R(G(x)), R(x)) \tag{5}$$

where R is a downsizing filter that reduces input resolution to 32×32 pixels, and L_1 computes the difference between the resized input image x and the resized synthetic mask $G(x)$. This loss can be instrumental in enhancing the consistency between the input image and the generated mask.

Segmentation loss. In this framework, the segmentation loss L_{seg} is computed as a linear combination of two distinct losses: one based on IoU and the other on Dice Coefficient to optimize overlap with ground truth.

$$\begin{aligned} L_{seg} &= \lambda_{IoU} L_{IoU} + \lambda_{dice} L_{dice} \\ L_{IoU} &= 1 - IoU = 1 - E_{x \in I,\, y \in M} \left(\frac{|y \cap G(x)| + \varepsilon}{|y \cup G(x)| + \varepsilon} \right) \\ L_{dice} &= 1 - C_{dice} = 1 - E_{x \in I,\, y \in M} \left(\frac{2\,|y \cap G(x)| + \varepsilon}{|y| + |G(x)| + \varepsilon} \right) \end{aligned} \tag{6}$$

Here, y represents the ground-truth mask corresponding to the image x, and $G(x)$ is the mask generated by the model. In these equations, ε, a small constant ($\varepsilon = 10^{-6}$), is added to prevent division by zero.

The contributions of these two metrics to the total loss are determined by their coefficients, λ_{dice} and λ_{IoU}.

Reverse-direction losses. When the cycle pathway is enabled, the reverse-direction objectives are similarly defined as in Equations 4 through 6, with a different definition for L'_{cyc} as follows:

$$\begin{aligned} L'_{adv} &= E_{y\in M} L_c(D'(G'(y)), 1), & L'_{seg} &= \lambda_{IoU} L'_{IoU} + \lambda_{dice} L'_{dice}, \\ L'_{cyc} &= E_{y\in M} L_s(G(G'(y)), y), & L'_{IoU} &= 1 - E_{y\in M,\, x\in I}\left(\frac{|x \cap G'(y)| + \varepsilon}{|x \cup G'(y)| + \varepsilon}\right), \\ L'_{idt} &= E_{x\in I} L_{pix}(G'(x), x), & & \\ L'_{cons} &= E_{y\in M} L_1(R(G'(y)), R(y)), & L'_{dice} &= 1 - E_{y\in M,\, x\in I}\left(\frac{2|x \cap G'(y)| + \varepsilon}{|x| + |G'(y)| + \varepsilon}\right). \end{aligned} \tag{7}$$

Here, L_s is a loss that compares an input segmentation mask y with its reconstruction mask $G(G'(y)$. The loss was selected to be either L'_{IoU} or L'_{dice}, and we experimented with both for comparison.

3.6 Inference

At test time, only the forward generator G is used. Given an input image x, the predicted segmentation is obtained in a single forward pass as $\hat{y} = G(x)$, without post-processing. This matches the forward path in Figure 1 and enables direct I2L prediction.

Algorithm 1 summarizes the complete training procedure described above, detailing the interaction between generator, discriminator, and segmentation losses during training. The algorithm begins with a self-supervised pre-training for the generator using image inpainting. Then, in each iteration, a batch of images and their corresponding segmentation masks are sampled. The generator produces synthetic masks, and the discriminator is updated based on adversarial loss. The generator is updated using a combination of losses (Eq. 3).

Algorithm 1 seGANmentation Training

Require: Dataset $\mathcal{D} = \{(I, M)\}$ where I is an image and M is its segmentation mask
Require: Generator G, Discriminator D
Require: Learning rate α, Batch size B, Loss weights $\lambda_{adv}, \lambda_{seg}, \lambda_{cyc}, \lambda_{idt}, \lambda_{cons}, \lambda_{GP}$
Pre-train Generator: Train G with self-supervised image inpainting
for each training iteration **do**
  Sample mini-batch $\{(I_b, M_b)\}_{b=1}^{B}$ from dataset
  Generator Forward Pass: $G(I) \rightarrow \hat{M}$ (Generate synthetic mask)
  Compute Discriminator Loss: Eq. 1
  Update Discriminator: $\theta_D \leftarrow \theta_D - \alpha \nabla_{\theta_D} L_D$
  Compute Generator Loss: Eq. 3
  Update Generator: $\theta_G \leftarrow \theta_G - \alpha \nabla_{\theta_G} L_G$
end for
return Trained Generator G

4 Experimental Setup

4.1 Dataset Description and Pre-processing

For this study, the Carvana dataset* was selected from Kaggle's Image Masking Challenge (2017). Given the novel formulation of segmentation as an adversarial I2L translation task, we utilize the Carvana dataset as a controlled, single-object proof-of-concept. This allows us to isolate the behavior of mask-domain adversarial learning before scaling to complex, multi-class environments commonly encountered in autonomous driving applications. This dataset is relatively small, making it ideal for faster experimental runs and preliminary studies. It comprises approximately 5,000 images of cars, each accompanied by a highly accurate, manual

*https://www.kaggle.com/competitions/carvana-image-masking-challenge/data

Table 4: The hyperparameters yielding the optimal I2L translation results.

$\lambda_{\mathbf{GP}}$	$\lambda_{\mathbf{cyc}}$	$\lambda'_{\mathbf{cyc}}$	$\lambda_{\mathbf{idt}}$	$\lambda_{\mathbf{cons}}$	$\lambda_{\mathbf{dice}}$	$\lambda_{\mathbf{IoU}}$	**lr**
1	3	3	0	0	0	5	1e-8

binary segmentation mask. Representative examples of the input images and their corresponding ground-truth masks are illustrated in Figure 3.

The annotated images were divided into training and test sets, with 20% of the images randomly selected for testing. Notably, each car in the dataset is represented by a set of 16 images captured from various angles. To maintain consistency, if one image of a car is chosen for the test set, all 16 images of that car are included. This approach ensures no overlap between the training and test datasets, effectively preventing data leakage and guaranteeing a fair and unbiased evaluation. The high-resolution 1918×1280 images were cropped to a 1280×1280 square and downsized to 256×256 pixels for computational efficiency.

4.2 Training Process

The training process follows the two-step methodology of UVCGAN-v2 [18]: self-supervised generator pre-training via image inpainting, followed by adversarial I2L fine-tuning.

During pre-training, inspired by BERT [29], the 256×256 images were divided into 32×32 patches, with 40% randomly masked. The generator was trained to reconstruct the original image using an AdamW optimizer (weight decay 0.05, $\beta_1 = 0.9$, $\beta_2 = 0.99$) with cosine learning rate annealing over 5 cycles. The initial learning rate was scaled to $5 \times 10^{-3} \times (\text{batch size}/512)$, and generator weights were initialized using a random normal distribution (standard deviation 0.02). Data augmentations included random rotation ($\pm 10^{\circ}$), horizontal flipping ($p = 0.5$), and color jitter (± 0.2). Pre-training ran for 2000 epochs with a batch size of 16 on an NVIDIA A100 GPU, completing in approximately one week.

For the subsequent I2L translation, the generator was optimized using Adam (no weight decay), with data augmentation restricted to random horizontal flips. Training stability was enforced via spectral normalization on the discriminator weights, a zero-centered gradient penalty for the discriminator loss, and exponential moving average for the generator weights (momentum 0.9999). To mitigate mode collapse despite a training batch size of 1, the discriminator's batch head utilized batch normalization with an image cache size of 3, simulating an effective batch size of 4 to accurately estimate batch statistics.

4.3 Hyperparameter Tuning

An extensive grid search was conducted to optimize the loss coefficients (λ_{GP}, λ_{cyc}, λ'_{cyc}, λ_{idt}, λ_{cons}, λ_{dice}, λ_{IoU}) and learning rates. Each configuration was trained for 1000 epochs (requiring approximately 5 days per run on an A100 GPU), as preliminary tests demonstrated no significant performance gains from extending training to 2000 epochs. Comprehensive results for all configurations, with and without cycle consistency, are documented in Tables 7 and 8 in Appendix B.

The grid search yielded several key insights:

- A uniform learning rate of 1×10^{-8} and a discriminator β_1 of 0.5 provided the most stable convergence.
- Setting both the identity loss (λ_{idt}) and consistency loss (λ_{cons}) to 0 yielded the best results. A gradient penalty coefficient (λ_{GP}) of 1 was sufficient, with higher values yielding no further improvements.
- Optimizing purely for IoU ($\lambda_{IoU} = 5$, $\lambda_{dice} = 0$) outperformed combinations utilizing the Dice loss.
- Incorporating bidirectional cycle-consistency constraints improved overlap metrics by approximately 2–3%. The optimal configuration utilized the L_2 norm for image reconstruction ($\lambda_{cyc} = 3$) and the IoU-based loss for mask reconstruction ($\lambda'_{cyc} = 3$).

The final optimal hyperparameters utilized for the core experiments are summarized in Table 4.

Table 5: Comparative performance of seGANmentation and baselines on the Carvana dataset.

Model	Paradigm	Architecture	Accuracy	IoU	Dice	Precision	Recall	AP@50	FID
U-Net	Discriminative	CNN (Encoder-Decoder)	0.9933	0.9781	0.9889	0.9887	0.9891	0.4522	83.41
DeepLabV3+	Discriminative	CNN (ASPP)	0.9926	0.9754	0.9875	0.9895	0.9856	0.4545	104.21
LinkNet	Discriminative	Lightweight CNN	0.9917	0.9727	0.9861	0.9831	0.9892	0.0326	208.58
SegFormer	Discriminative	Vision Transformer (ViT)	0.9924	0.9749	0.9873	0.9853	0.9893	0.4545	134.39
seGANmentation	Generative	Hybrid ViT-CNN	0.8990	0.7091	0.8292	0.8145	0.8319	0.3016	160.79

4.4 Baselines

The framework is evaluated against a diverse suite of SOTA discriminative segmentation architectures: U-Net [19], which shares a similar encoder-decoder structure; DeepLabV3+ [20], which utilizes dilated convolutions for multi-scale context; LinkNet [21], optimized for efficient feature mapping; and SegFormer [22], a purely transformer-based architecture to contrast with our hybrid ViT-CNN approach. This selection provides a rigorous benchmark across dominant modern segmentation paradigms, including CNN-based ASPP [28] and ViTs, while isolating the impact of adversarial learning. For a controlled comparison, all convolutional baselines (U-Net, DeepLabV3+, and LinkNet) utilize a ResNet-34 [30] backbone, while SegFormer employs the mit-b0 encoder [22]. All baselines are trained and evaluated on the Carvana dataset under identical data splits and preprocessing conditions. GAN-based segmentation approaches, including Pix2Pix [13] and SegAN [14], are used for contextual comparison based on reported results from the literature (Section 5) and they rely on different task formulations such as auxiliary refinement or fully paired supervision.

5 Results and Discussion

Some of the qualitative results obtained with the configuration in Table 4 are depicted in Figure 3. This figure displays a comparison of several images from the dataset, including the ground truth, the masks generated by the presented network, and those produced by the baselines. It is important to note that, as previously mentioned, the Carvana dataset features 16 images of each car taken from different angles. For illustrative purposes 4 images of the same car were selected. The results presented here are representative and have not been selectively "cherry-picked". Although the results from this model are not perfect, they intriguingly demonstrate the capability to capture the general shape of cars.

To assess the outcomes of the I2L translation, evaluation was divided into local and global metrics. Local metrics (including Accuracy, IoU, Dice Coefficient, Precision, and Recall) evaluate pixel-wise spatial alignment and boundary precision. Global metrics, specifically Average Precision at a 50% threshold (AP@50) and Fréchet Inception Distance (FID), assess overall object localization and the distributional realism of the generated mask shapes. Unlike Dice or pixel-wise IoU which are highly sensitive to edge misalignments, AP@50 evaluates whether the model identifies the object's global position and approximate extent. This distinction is critical for evaluating generative models, which prioritize holistic shape over high-frequency edge adherence.

The training parameters included a batch size of 16 and a learning rate of $1e-5$. Bilinear upsampling was used for its potential suitability for smaller images. The training was conducted over 10 epochs using an NVIDIA A100 GPU, completing the process in a few hours.

The average values of the metrics were calculated for the results obtained from the discriminative baselines and the seGANmentation framework. All models were trained and evaluated under the same experimental protocol. Performance calculations were made in comparison to the ground truth masks, and the outcomes of these measurements are comprehensively presented in Table 5. Furthermore, detailed tables containing all experimental runs, including hyperparameter configurations and the resulting evaluation metrics for the generative framework, are provided in Tables 7 and 8 in Appendix B.

Comparing the performance across the evaluated architectures, the discriminative baselines achieve the strongest local, pixel-level overlap scores, with models reaching IoU values above 0.97. This is expected,

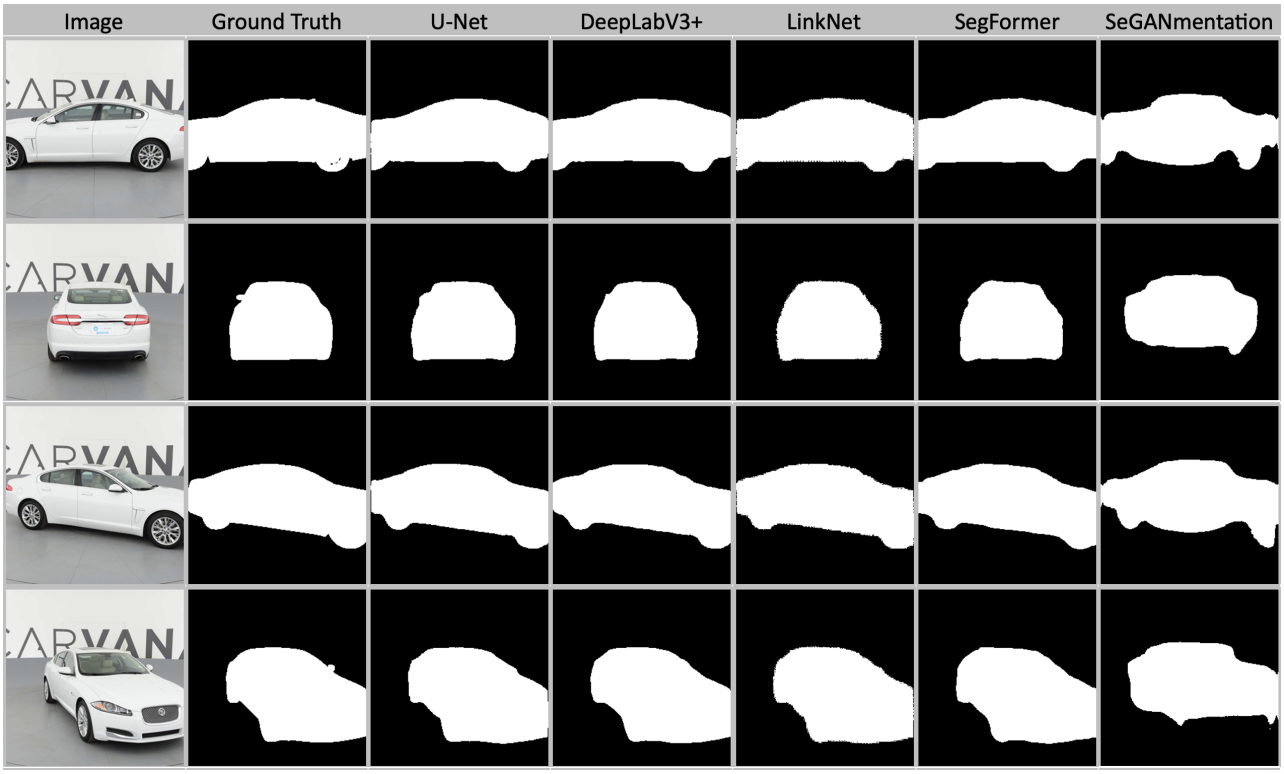


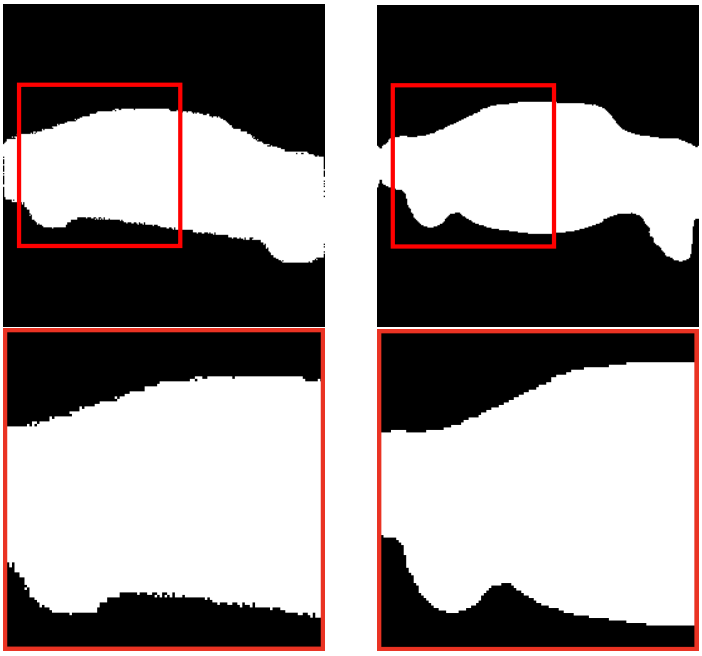
Figure 3: Qualitative comparison of segmentation results, including the ground truth masks alongside the corresponding predictions of baselines and seGANmentation.

since these methods are explicitly optimized for localized pixel-wise alignment and fine boundary recovery.

In contrast, seGANmentation shows a different behavior. While its local metrics are lower, the relative balance between Precision (0.81) and Recall (0.83) indicates that the framework successfully identifies the main object region without severe directional bias. The slightly higher recall suggests a mild tendency to produce larger, smoother mask regions. This is consistent with adversarial I2L translation: the generator favors globally coherent object shapes but does not always recover high-frequency, complex boundary details.

This distinction highlights that the proposed generative framework functions effectively as an object-level localizer rather than a pixel-perfect segmentor. This is further contextualized by global metrics such as AP@50, which evaluate whether the predicted mask sufficiently overlaps the target object at an object-detection level rather than requiring near-perfect boundary alignment. A qualitative inspection of the predicted masks reveals crucial distinctions between the discriminative and generative paradigms. Upon close examination of the discriminative baselines, particularly LinkNet, distinct boundary artifacts become apparent, as highlighted in Figure 4–left. While any digital image inherently displays pixelation under high magnification due to finite grid resolution (an effect naturally present even in seGANmentation), the boundaries of the LinkNet predictions exhibit more structural jaggedness and pixelation. This effect is characteristic of upsampling or checkerboard artifacts generated by the lightweight decoder network. Because these high-frequency boundary errors involve a relatively small number of pixels, they do not severely penalize the overlap-based metrics, allowing LinkNet to maintain a high local IoU (0.97). However, these jagged boundaries produce unnatural mask silhouettes, which directly explains LinkNet's surprisingly poor global FID score (208.6).

Conversely, the adversarial discriminator in the seGANmentation framework explicitly penalizes such unnatural, generated artifacts. Consequently, while the GAN may trade off pixel-perfect boundary adherence, it suppresses these distortions. Even with normal finite-resolution pixelation present, it enforces a smoother and distributionally realistic global mask structure (Figure 4–right), directly resulting in its superior FID perfor-



(a) LinkNet (b) seGANmentation

Figure 4: Comparison of mask boundary fidelity. The top row shows full predictions with the magnified region highlighted in red. The bottom row reveals that while both masks inherently possess finite grid resolution, the discriminative LinkNet baseline (a) exhibits jaggedness and checkerboard upsampling artifacts. In contrast, the generative seGANmentation framework (b) suppresses these artifacts, enforcing a smoother, distributionally realistic silhouette.

Table 6: Performance of representative GAN-based segmentation frameworks. Results are shown for context only and are not directly comparable due to differences in datasets, task formulations, and training protocols.

Method	**Dataset**	**Role of GAN**	**Metric**	**Score**
Pix2Pix [13]	Cityscapes [31]	Primary I2L engine	IoU	0.22
Pix2Pix (L_1 baseline)	Cityscapes	Regression (no GAN)	IoU	0.35
SegAN [14]	BRATS 2015 [32]	Adversarial refiner	Dice	0.85
U-Net	BRATS 2015	Supervised baseline	Dice	0.80

mance (160.8). These findings underscore the fundamental trade-off of direct I2L segmentation: prioritizing holistic shape distribution and domain realism over localized pixel regression. Importantly, because seGANmentation can potentially operate without paired pixel-level annotations, it presents a valuable alternative in data-scarce or unsupervised environments where coarse localization is sufficient.

For broader context, Table 6 summarizes reported results from other GAN-based segmentation frameworks such as Pix2Pix [13] and domain adaptation–oriented pipelines like SegAN [14]. These results show a clear distinction between generative translation and supervised segmentation. Although these methods address different segmentation formulations, their reported performance provides useful context for understanding the behavior of adversarial learning in segmentation. For instance, Pix2Pix achieved a class IoU of only 0.22 on the Cityscapes dataset [31], underperforming a standard L_1 loss baseline of 0.35 [13].

SegAN leveraged a multi-scale L_1 adversarial critic to achieve a Dice coefficient of 0.85 on whole tumors in the MICCAI BRATS 2015 brain tumor dataset [32], surpassing its 0.80 U-Net baseline [14]. However, our seGANmentation model reached a Dice score of 0.83 on the Carvana dataset, falling short of its 0.99 U-Net benchmark. This suggests that SegAN's strategy of using an adversarial critic to regularize a dedicated

segmentor is significantly more effective for ensuring spatial contiguity than a pure I2L translation approach.

These findings reveal that a hybrid framework (one that retains a primary segmentation engine while utilizing GAN components for auxiliary refinement) may be superior for achieving the fine-grained precision required in complex segmentation tasks. Importantly, this observation does not indicate a deficiency of adversarial learning itself, but rather highlights the limitations of using pure I2L translation as a standalone segmentation strategy in fully supervised settings.

5.1 Failure Analysis and Limitations

As shown in Figure 3 and Table 5, the supervised discriminative baselines achieve near-perfect pixel-level segmentation performance, benefiting from architectures specifically designed for explicit boundary recovery. In contrast, seGANmentation, while offering a novel approach by generating both images and segmentation labels within a single network, falls short in localized segmentation accuracy. Despite these limitations, several important observations emerge that highlight both current failure modes and opportunities for improvement.

Performance and Metric Disparities: One notable pattern in the results is the disparity between local and global metrics. While the IoU is relatively low (0.71), the Accuracy (0.90), Dice coefficient (0.83), and Recall (0.83) remain notably higher. This disparity occurs because metrics like Accuracy and Recall are more forgiving of small boundary inconsistencies as long as the object's core is identified. The relatively higher Recall indicates that seGANmentation successfully captures the general structure of the target, but struggles with the fine-grained, high-frequency details required for strict IoU optimization.

Common Failure Cases:

- Boundary Inconsistencies: The primary failure mode of seGANmentation is its difficulty in capturing sharp boundaries. The model often produces segmentation masks that are slightly misaligned with object edges, leading to lower IoU and AP@50 scores. This issue is particularly pronounced in cases involving occlusions or highly irregular contours.
- Over-Smoothing: Another issue is the tendency of the model to generate overly smooth segmentation masks, particularly in areas with complex textures. This likely stems from the adversarial loss optimization, which prioritizes global shape coherence over local precision. Notably, while this over-smoothing penalizes pixel-wise metrics, it simultaneously prevents the unnatural jagged upsampling artifacts seen in a lightweight discriminative model like LinkNet, thereby contributing to seGANmentation's competitive FID score.
- Sensitivity to Hyperparameters: The ablation studies indicate that seGANmentation's performance is highly sensitive to the weighting of different loss terms. Small variations in hyperparameters can lead to fluctuations in results, suggesting that careful tuning is required to balance generative realism with spatial accuracy.

Impact of Cycle Consistency: Adding cycle consistency loss improved segmentation accuracy by approximately 2-3%, but at the cost of increased computational overhead. This suggests that maintaining a cycle consistency constraint is beneficial, though not sufficient on its own to bridge the performance gap with supervised baselines. Future work should explore whether alternative constraints, such as boundary-focused loss terms, could further improve accuracy.

Hybrid Strategies: Given its current limitations, seGANmentation could be effective as part of a hybrid segmentation pipeline. One promising direction is using GAN-generated segmentation labels as an initial structural prior, which could then be refined using traditional discriminative methods like U-Net or DeepLabV3+. This strategy leverages the generative power of GANs for global shape distribution while maintaining the boundary reliability of supervised learning-based approaches. This hybrid strategy has the potential to serve as a complementary tool to existing segmentation networks.

Advantages in Data-Scarce Environments: While standard discriminative models rely heavily on large, dense, and fully annotated datasets, seGANmentation presents a possible alternative for semi-supervised or unsupervised segmentation (though unsupervised implementation remains for future work). By generating

labels directly alongside images, the generative framework could serve as a pre-training step or an augmentation method for domains where labeled data is prohibitively expensive or scarce.

Dataset and Generalizability Limitations. A primary limitation of this study is its reliance on a single dataset, which features binary, single-object segmentation with relatively clean backgrounds. While this controlled environment is ideal for isolating the baseline behavior of adversarial I2L translation and proving the underlying concept, it limits the immediate generalizability of our findings. Real-world applications, such as autonomous driving or scene understanding, involve complex scenes with multi-class overlapping objects and severe occlusions. Therefore, the reported results represent a foundational proof-of-concept rather than a definitive benchmark across diverse tasks. Scaling this framework to multi-class datasets to evaluate the learned mask-priors under complex scene distributions remains a critical direction for future work.

In summary, while seGANmentation does not surpass discriminative baselines in pixel-wise accuracy, it highlights the feasibility of direct GAN-based segmentation and opens avenues for future research. The results emphasize the importance of fine-tuning loss functions, exploring boundary-focused regularization techniques to prevent over-smoothing, and developing hybrid approaches that integrate adversarial mask-domain learning with conventional deep learning models. Future efforts should focus on refining segmentation loss terms, incorporating additional regularization techniques to prevent mode collapse and over-smoothing, and additional adversarial training strategies to boost boundary accuracy and overall performance.

5.2 Training Dynamics and Stability

To assess training stability, we tracked the evolution of the adversarial objectives (generator and discriminator losses in both directions), cycle-consistency losses, the IoU-based objective, an edge-sensitive metric, and the discriminator gradient-penalty terms over 1000 epochs (Fig. 5). Overall, the curves suggest a stable adversarial game with a clear two-phase learning behavior.

Two-phase convergence. The training dynamics exhibit a distinct transition around epoch ≈ 200. In the early stage (roughly epochs 0–200), the model is still learning a coarse alignment between the image domain and the mask domain. During this period, the generator adversarial losses remain high (near 1.0), indicating that the discriminators can reliably distinguish generated outputs from real samples. In parallel, discriminator losses rapidly decrease to near-zero at the very beginning and remain low, which is expected when the discriminator initially dominates before the generator learns meaningful structure.

After the transition (epochs $\gtrsim 200$), the generator losses drop and settle into a narrow band (0.3–0.4), while discriminator losses rise and stabilize around 0.18–0.22. This is consistent with the generator improving enough that discriminators no longer separate real and fake samples, leading to an equilibrium. None of the losses diverge or collapse; instead, they remain bounded with small oscillations, suggesting stable training.

Cycle-consistency behavior. The cycle-consistency losses peak around the same transition region (with a spike near epoch 190), and then decay steadily. This pattern is typical when the generators begin to substantially change their outputs: as the mapping improves and becomes less noisy, reconstruction errors decrease and the cycle losses stabilize. After epoch ≈ 600, both cycle losses remain low and relatively flat (λ_{cyc} or `cycle_a` ≈ 0.10 and λ'_{cyc} or `cycle_b` ≈ 0.02), indicating that bidirectional mappings preserve a consistent structure between domains. The asymmetry (`cycle_a` consistently higher than `cycle_b`) is expected because the two directions correspond to different reconstruction difficulties (image reconstruction versus mask reconstruction) and are optimized under different effective loss landscapes.

Segmentation objective and boundary sensitivity. The IoU-related training curve in Figure 5 decreases sharply during the first 200 epochs and then plateaus, which indicates that most improvements in overlap-based quality occur early, followed by a long refinement regime. The edge metric shows a rapid rise early on (peaking around epochs 100–200), followed by a decline and stabilization. Qualitatively, this is consistent with the generator first learning to place high-frequency transitions (i.e., discovering object boundaries), and then smoothing and refining them as adversarial and reconstruction terms encourage globally coherent masks.

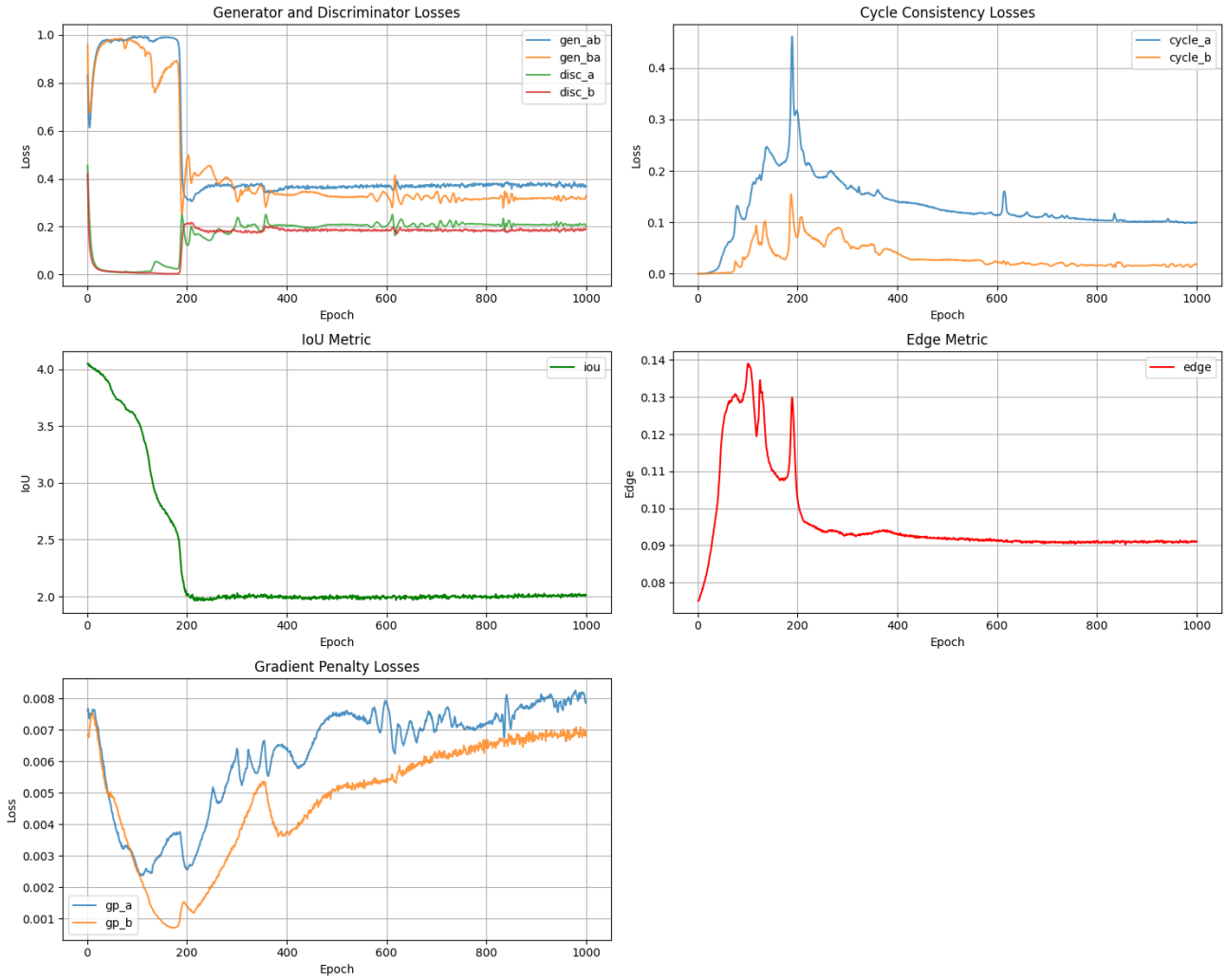


Figure 5: Training curves for seGANmentation over 1000 epochs, showing generator/discriminator losses, cycle-consistency losses, IoU and edge metrics, and gradient penalty terms.

The coincidence of the edge spikes with the cycle-consistency spike further supports the interpretation that this phase corresponds to a rapid reorganization of the learned mapping rather than an instability event.

Gradient penalty and adversarial stability. Finally, the gradient-penalty losses remain small throughout training ($\approx 10^{-3}$–10^{-2}) and do not exhibit runaway growth. This behavior supports the role of the zero-centered GP regularizer in preventing exploding discriminator gradients and reducing sensitivity to hyperparameters. Together with the bounded discriminator/generator losses, these trends suggest that the run avoids common GAN failure modes such as catastrophic divergence or sustained mode collapse.

Practical implication. Since the IoU-related objective plateaus shortly after the transition point, the best validation checkpoints are expected to occur relatively early (soon after the stabilization regime begins), while later epochs primarily contribute to incremental boundary refinement rather than large gains in region overlap. This motivates using early stopping (or model selection based on validation IoU/Dice) and focusing subsequent experimentation on loss re-weighting or boundary-aware terms to improve fine-grained mask quality.

6 Societal Impacts

As AI continues to advance, it is important to consider potential negative impacts. This research on GAN-based segmentation raises several ethical, social, and environmental concerns that merit careful consideration.

One key concern is the potential for malicious, unintended, or deceptive use of generative AI. Generative models are capable of producing realistic synthetic images, which may be exploited for disinformation, the creation of fake identities, or other deceptive practices that undermine trust in digital media and automated decision-making. Closely related are security risks, as synthetic images may be used to bypass biometric authentication mechanisms such as facial recognition, facilitating identity theft or fraud. Mitigating these risks requires the establishment of ethical use guidelines, the development of robust detection and authentication systems, and the integration of synthetic-data-aware defenses. Ongoing collaboration between researchers, practitioners, and regulatory bodies is essential to promote accountability and responsible deployment.

Fairness and bias also present significant challenges. Models trained on synthetic data may inadvertently amplify existing biases if the generated samples do not adequately represent real-world diversity, potentially leading to unfair outcomes for underrepresented or disadvantaged groups. This risk can be reduced by training models on diverse and representative datasets, regular bias audits, and incorporating bias detection and mitigation strategies throughout the model development lifecycle.

Privacy considerations are particularly important for generative AI systems. Although generated outputs are synthetic, there is a risk that sensitive information from the training data could be implicitly memorized, enabling model inversion or membership inference attacks that reconstruct private images. Privacy risks can be mitigated through the use of techniques such as differential privacy, federated learning, and secure training protocols that limit the exposure of individual data points during model optimization.

The environmental impact of training large models is another concern. Training deep generative architectures requires substantial computational resources, resulting in high energy consumption and increased carbon emissions. These effects can be alleviated by optimizing model architectures, reducing unnecessary experimentation, improving training efficiency, and prioritizing energy-efficient hardware or renewable energy sources.

In conclusion, while GAN-based segmentation offers significant promise (particularly for reducing annotation costs and enabling learning in data-scarce scenarios), its societal implications must be carefully addressed. By proactively integrating mitigation strategies, the responsible development and deployment of generative AI systems can help ensure that these technologies provide meaningful and equitable benefits to society as a whole.

7 Conclusion

This research introduced seGANmentation, a novel framework that formulates segmentation as a direct I2L translation problem using adversarial learning. While seGANmentation does not outperform discriminative baselines in terms of pixel-level accuracy in fully supervised settings, the results demonstrate that treating segmentation as a generative translation task is feasible. This approach highlights a trade-off between global mask realism (evidenced by competitive FID scores) and fine-grained boundary precision. In contrast to localized pixel-wise regression, adversarial training encourages globally plausible mask distributions and can successfully capture the global topological distribution of segmentation masks. However, the framework remains sensitive to hyperparameter tuning and architectural complexity, particularly given the integration of ViT components that require careful optimization to balance global context with local detail.

These findings suggest that generative segmentation offers unique advantages for data augmentation and semi-supervised learning. By synthesizing image–label pairs, this approach can expand training datasets and reduce reliance on manual annotation. In addition, GAN-generated masks may serve as structural priors or coarse initial labels that can be refined using off-the-shelf discriminative segmentors, thereby reducing the need for extensive paired supervision and helping bypass the labeling bottleneck in data-scarce environments. Future work will investigate the effectiveness of adversarial I2L translation in data-scarcity regimes, including hybrid pipelines where GAN-generated masks serve as coarse initializations for discriminative models. The feasibility

of fully unsupervised segmentation through cycle-consistency constraints, and the potential of a single network producing joint visual and segmentation outputs, remain promising directions for further exploration.

Overall, this work positions adversarial I2L translation not as a replacement for discriminative segmentation, but as a complementary paradigm. Compared to fully supervised CNNs and Transformers, as well as prior GAN-based approaches such as Pix2Pix and SegAN that rely heavily on paired supervision or use adversarial networks merely as auxiliary refiners, seGANmentation explores a fundamentally different operating point where the generative model itself serves as the primary segmentation engine. While this formulation sacrifices precise boundary adherence in fully supervised settings, it successfully learns meaningful mask priors and generates naturally cohesive, distributionally realistic silhouettes. As such, seGANmentation opens new opportunities for reducing annotation costs and enabling coarse segmentation under weak, limited, or absent supervision, providing a foundation for future research at the intersection of generative modeling and semantic understanding.

References

[1] Z. Wang, Q. She, and T. E. Ward, "Generative adversarial networks in computer vision: A survey and taxonomy," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021. [Online]. Available: https://doi.org/10.1145/3439723

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2010.11929

[3] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023. [Online]. Available: https://doi.org/10.1109/TPAMI.2023.3261988

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2644615

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015, pp. 3431–3440. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298965

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988. [Online]. Available: https://doi.org/10.1109/TPAMI.2018.2844175

[7] I. Ulku and E. Akagündüz, "A survey on deep learning-based architectures for semantic segmentation on 2D images," *Applied Artificial Intelligence*, vol. 36, no. 1, 2022. [Online]. Available: https://doi.org/10.1080/08839514.2022.2032924

[8] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 671–12 681. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.01269

[9] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," *ArXiv*, vol. abs/1611.08408, 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1611.08408

[10] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1802.07934

[11] M. Gadermayr, V. Appel, B. M. Klinkhammer, P. Boor, and D. Merhof, "Which way round? a study on the performance of stain-translation for segmenting arbitrarily dyed histological images," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2018, pp. 165–173. [Online]. Available: https://doi.org/10.1007/978-3-030-00934-2_19

[12] M. Gadermayr, L. Gupta, B. M. Klinkhammer, P. Boor, and D. Merhof, "Unsupervisedly training GANs for segmenting digital pathology with automatically generated annotations," in *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, ser. Proceedings of Machine Learning Research, vol. 102, July 2019, pp. 175–184. [Online]. Available: https://doi.org/10.48550/arXiv.1805.10059

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134. [Online]. Available: https://doi.org/10.1109/CVPR.2017.632

[14] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang, "Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation," *Neuroinformatics*, vol. 16, pp. 383–392, 2018. [Online]. Available: https://doi.org/10.1007/s12021-018-9377-x

[15] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M. P. Lungren, S. Zhang, L. Xing, L. Lu, A. Yuille, and Y. Zhou, "TransUNet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers," *Medical Image Analysis*, vol. 97, p. 103280, 2024. [Online]. Available: https://doi.org/10.1016/j.media.2024.103280

[16] Y. Zhang, H. Ling, J. Gao, K. Yin, J.-F. Lafleche, A. Barriuso, A. Torralba, and S. Fidler, "DatasetGAN: Efficient Labeled Data Factory with Minimal Human Effort," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2021, pp. 10 140–10 150. [Online]. Available: https://doi.org/10.1109/CVPR46437.2021.01001

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Neural Information Processing Systems*, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1706.03762

[18] D. Torbunov, Y. Huang, H.-H. Tseng, H. Yu, J. Huang, S. Yoo, M. Lin, B. Viren, and Y. Ren, "UVCGAN v2: An improved cycle-consistent GAN for unpaired image-to-image translation," *arXiv preprint arXiv:2303.16280*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.16280

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention (MICCAI)*, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-319-24574-4_28r

[20] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818. [Online]. Available: https://doi.org/10.1007/978-3-030-01234-2_49

[21] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE visual communications and image processing (VCIP)*. IEEE, 2017, pp. 1–4. [Online]. Available: https://doi.org/10.1109/VCIP.2017.8305148

[22] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2105.15203

[23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251. [Online]. Available: https://doi.org/10.1109/ICCV.2017.244

[24] M. Berijanian, N. S. Schaadt, B. Huang, J. Lotz, F. Feuerhake, and D. Merhof, "Unsupervised many-to-many stain translation for histological image augmentation to improve classification accuracy," *Journal of pathology informatics*, vol. 14, 2023. [Online]. Available: https://doi.org/10.1016/j.jpi.2023.100195

[25] D. Torbunov, Y. Huang, H. Yu, J. Huang, S. Yoo, M. Lin, B. Viren, and Y. Ren, "UVCGAN: UNet vision transformer cycle-consistent GAN for unpaired image-to-image translation," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 702–712. [Online]. Available: https://doi.org/10.1109/WACV56688.2023.00077

[26] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *Advances in neural information processing systems*, vol. 30, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1703.00848

[27] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189. [Online]. Available: https://doi.org/10.1007/978-3-030-01219-9_11

[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018. [Online]. Available: https://doi.org/10.1109/TPAMI.2017.2699184

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, June 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/N19-1423

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[31] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223. [Online]. Available: https://doi.org/10.1109/CVPR.2016.350

[32] B. H. Menze, A. Jakab, S. Bauer *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, pp. 1993–2024, 2015. [Online]. Available: https://doi.org/10.1109/TMI.2014.2377694

[33] J. Kim, M. Kim, H. Kang, and K. H. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in *International Conference on Learning Representations (ICLR)*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2020, p. 6000–6010. [Online]. Available: https://doi.org/10.48550/arXiv.1907.10830

[34] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1710.10196

[35] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.00821

A Additional Validation and Reproduction of UVCGAN-v2 Results

To validate our implementation before adapting it to I2L translation, we reproduced the results from UVCGAN-v2 [18] using the same datasets (Anime [33], CelebA-HQ [34], and AFHQ [35]) and training protocol. The experiments were conducted over 1000 epochs on an NVIDIA A100 GPU.

Translation tasks included pairs such as anime-selfie, male-female, cat-dog, wild-cat, and wild-dog. Figure 6 displays a selection of these results, which are not included in the original paper.

We observed FID/KID values consistent with those reported in [18], providing confidence that our codebase matches the intended backbone behavior prior to introducing segmentation-specific modifications.

B Hyperparameter Tuning

Here, we provide comprehensive tables summarizing the results from our hyperparameter grid search experiments for I2L translation training. These tables include experiments conducted both without (Table 7) and with (Table 8) cycle consistency loss. They offer insights into the performance metrics under different configurations.

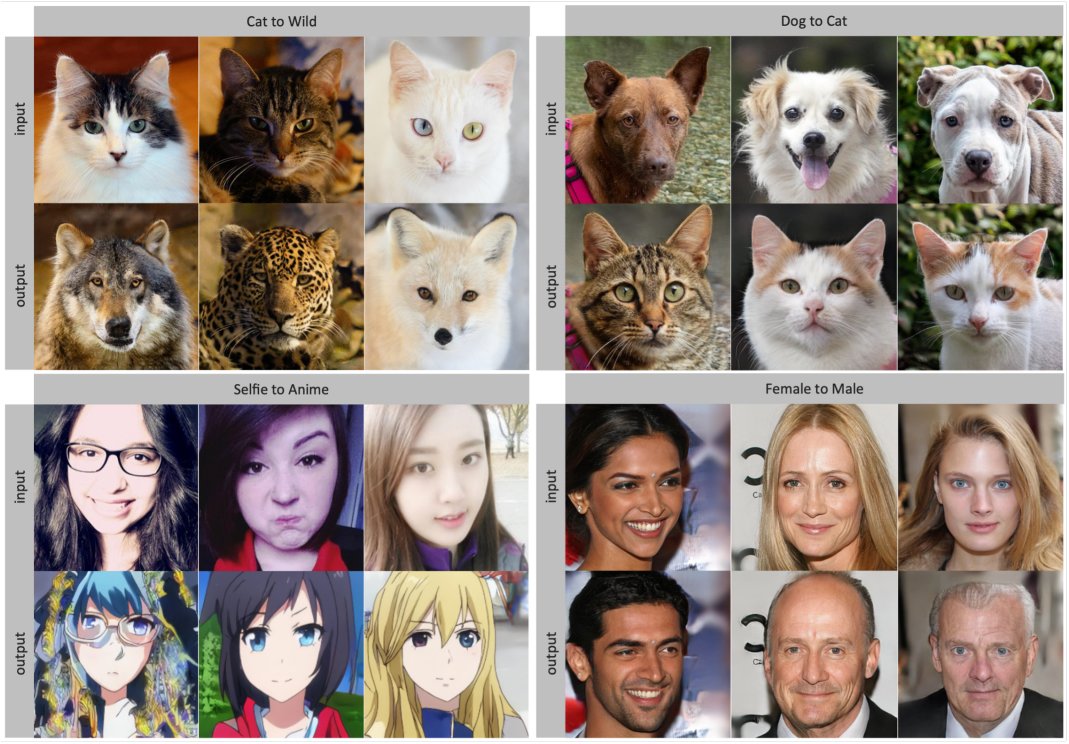


Figure 6: Sample translations from Anime, CelebA-HQ, and AFHQ datasets, featuring four transformations: cat2wild, dog2cat, selfie2anime, and female2male. These results reproduce findings from UVCGAN-v2 [18].

Table 7: Performance metrics for experiments without cycle consistency loss using various hyperparameters.

No. of epochs	lr	λ_{GP}	λ_{idt}	λ_{cons}	λ_{dice}	λ_{IoU}	β_1	Accuracy	Dice	IoU
2000	**1e-8**	**1**	**0.5**	**0.5**	**0**	**5**	**0.5**	**0.88**	**0.80**	**0.68**
1000	**1e-8**	**1**	**0.5**	**0.5**	**0**	**5**	**0.5**	**0.88**	**0.80**	**0.67**
1000	1e-8	1	0.5	0.5	5	0	0.5	0.86	0.77	0.64
2000	1e-8	1	0.5	0.5	5	0	0.5	0.86	0.77	0.63
2000	1e-9	1	0.5	0.5	0	5	0.5	0.59	0.59	0.42
2000	1e-8	1	0.5	5	0	5	0.5	0.77	0.70	0.54
2000	1e-7	1	0.5	0.5	0	5	0.5	0.82	0.71	0.56
2000	1e-8	1	0.5	0.5	0	10	0.5	0.84	0.73	0.58
2000	**1e-8**	**1**	**0.5**	**0.5**	**0**	**5**	**0.9**	**0.88**	**0.80**	**0.67**
2000	1e-8	1	2	2	2	2	0.5	0.81	0.69	0.54
2000	**1e-8**	**1**	**0**	**0**	**0**	**5**	**0.5**	**0.88**	**0.80**	**0.68**
2000	**1e-8**	**3**	**0.5**	**0.5**	**0**	**5**	**0.5**	**0.88**	**0.80**	**0.67**

Table 8: Performance metrics for experiments with cycle consistency loss using various hyperparameters.

lr	λ_{dice}	λ_{IoU}	λ_{cyc,L_1}	λ_{cyc,L_2}	$\lambda'_{cyc,dice}$	$\lambda'_{cyc,IoU}$	Accuracy	Dice	IoU
1e-8	**0**	**5**	**0**	**0**	**0**	**0**	**0.89**	**0.82**	**0.69**
1e-8	0	5	0	0	0	5	0.87	0.79	0.66
1e-8	0	5	0	0	5	0	0.88	0.80	0.67
1e-8	**0**	**5**	**5**	**0**	**0**	**0**	**0.89**	**0.82**	**0.70**
1e-8	0	5	0	5	0	0	0.81	0.74	0.59
1e-8	5	5	0	0	0	5	0.87	0.80	0.66
1e-8	0	2	0	0	0	5	0.85	0.77	0.63
5e-8	0	5	0	0	0	5	0.88	0.81	0.68
1e-8	0	5	0	0	0	9	0.87	0.79	0.66
1e-8	0	5	0	0	0	1	0.87	0.80	0.67
1e-8	**0**	**5**	**0**	**3**	**3**	**0**	**0.89**	**0.81**	**0.69**
1e-8	**0**	**5**	**0**	**3**	**0**	**3**	**0.90**	**0.83**	**0.71**