SAIND: Scene Animation using (RGB) Images aNd Depth maps

Surya Pratap Singh, Nitin Jotwani, Vinayak Bassi, Kiran Davuluri
{suryasin, njotwani, vbassi, kirandav}@umich.edu

I. Executive Overview

This paper is based on generating animated images with stylized and realistic effects of real-world RGB images fused with its depth map and edge map. We propose a framework that amalgamates the power of existing state-of-the-art architectures and a heuristically designed fusion algorithm. The baseline of our work comes from the work done in Scenimefy [1]. They have proposed a novel semi-supervised image-to-image translation approach to guide the anime style learning using the pseudo-paired data.

Our overall objective can be categorized into three sequential steps. Firstly, we focus on generating edge and depth maps for a given RGB input image. Secondly, we fuse the RGB image with the depth map and its corresponding customised gradient map using our proposed algorithm, EnGD. Lastly, we use a generative adversarial network architecture, StyleGAN2 [2], to train, generate, and evaluate the stylized effects of animated images.

We have demonstrated the results in sequential steps on some real-world images whereas our training was conducted on pseudo pseudo-paired Shinkai-style anime dataset used in the Scenimefy paper.

Our code is publicly available at https://github.com/SuryaPratapSingh37/SAIND.git

II. Background/Impact

Automatic rendering of anime-style images has a high utilitarian value in society. Having a market capture of USD 26 Billion [3], its scope is also beyond the anime industry itself. We felt that its relevance in society can be seen in the following industries:

- **Entertainment Industry:** Its unique art style, storytelling, and cultural nuances have attracted a diverse audience worldwide, leading to substantial revenue generation and job creation. It is used in movies, TV series, and video games as well.
- **Marketing and Advertising:** They resonate particularly well with younger demographics, making them effective for products and services targeted at these age groups.
- **Fashion and Merchandise:** Clothing, accessories, and a wide range of products featuring anime characters and themes are highly popular, contributing to the retail and fashion industries.
- **Education and Training:** It is being used in language learning and as a tool to engage students in various subjects.

Counting all the utility values in society, we also work in this domain because of the challenges and nuanced side of anime image generation reported in the literature. The majority of the work has been conducted using an unsupervised image-to-image translation framework which impedes the idea of semantic segmentation from the rich textures of anime images. Moreover, there is a lack of data that maps the real world to anime images. Using the approach of implementing semi-supervised generation of real-world RGB+Depth+Edge images we expected a more immersive and realistic generation of anime images. This can be further translated to anime video generation using the image frames for real-world video.

III. Method

A. Edge detector

Edge detection plays a crucial role in anime scene generation due to its significance in capturing and emphasizing the essential features and details that define the visual style of anime. From literature review, we found traditional methods and recent Deep Learning approaches [4], especially those based on Convolutional Neural Networks (CNN), have strived for higher metrics but often at the expense of computational intensity, rendering them impractical for low-capacity devices and real-world applications. So in our project, we've used the LDC (Lightweight Dense CNN) model for edge detection. The LDC architecture addresses these challenges by offering a novel lightweight solution. It builds upon the DexiNed [5] model but introduces modifications for a better trade-off between performance and applicability. By employing smaller convolution filter sizes and compact convolution blocks, LDC model comprises less than 1 million parameters—fifty times smaller than [5] and lighter than most state-of-the-art approaches. This reduction in parameters does not compromise performance; in fact, LDC outperforms DexiNed [5] and BDCN [6] in terms of edge detection, producing cleaner edge maps with significantly fewer parameters. For instance, LDC exhibits only 2 % of the parameters compared to DexiNed [5] and 4 % compared to BDCN [6]. The results from LDC, as demonstrated in various datasets, including MDBD, BIPED, and BRIND, consistently indicate superior edge detection performance, surpassing other state-of-the-art approaches. For our project, we've used the LDC pre-trained weights over the MDBD dataset, since it consists of a wide array of challenging natural scenes. In the context of our anime scene generation project, the choice of LDC is strategic. Its lightweight design aligns with the project's need for efficiency

and practicality, enabling the adaptive learning of valuable features, especially high-frequency information, crucial for generating detailed and visually appealing scenes from RGB images. The LDC edge detector emerges as an optimal choice, balancing computational efficiency with state-of-the-art edge detection performance, making it well-suited for the demands of our anime scene generation application.

B. TokenFusion

Following an ablation study, we deduce there is a need to manipulate essential modal features such as edge, RGB, depth, etc for effective scene animation. We have obtained RGB and EdgeMaps from methods as provided in the earlier sections. Various adaptations of transformers have focused on single-modal vision tasks, where self-attention modules are stacked to handle input sources like images. After extensive survey, we found out that feature fusion and generation of multi-modal features is imperative to give a realistic perspective. We utilize TokenFusion [7] to generate DepthMaps by fusing RGB and their corresponding EdgeMaps. TokenFusion is tailored for transformers-based vision tasks to fuse multiple modalities and aggregate inter-modal features. It surpasses state-of-the-art methods in multimodal image-to-image translation, RGB-depth semantic segmentation, etc. It beats best performing CNN based methods and existing transformer based models by a great margin in terms of FID & MAE evaluation metrics. Experimentally, we find DepthMaps generated from this approach add greater value to gradient aspects in critical depth-patches with consistent depth.

C. EnGD algorithm

We propose a novel algorithm to fuse a given RGB image with its corresponding depth map and its corresponding gradient map. The EnGD algorithm is multi-step process that enhances the style of an image. This is an attempt to bring realistic effects to the image with a sense of depth perception.

The resulting image in RGB space exhibits a new sense of style which attempts to encapsulate its depth perception and hence make the image more realistic with emphasized contrasts for nearby subjects and heightened brightness in regions where the depth changes significantly. The algorithm operates on a different color space namely hue, saturation and value so that it can manipulate the saturation and value color space.

Here's a more detailed and expanded breakdown of the steps involved:

Step 1: Preprocessing the Depth Map

1) **Depth Map Generation:** The initial depth map is generated using the token fusion algorithm.
2) **Normalization:** The depth map is normalized to ensure consistent intensity levels across the image.
3) **Gradient Calculation:** Gradients in both the x and y directions are computed from the normalized depth map. A customised $Sobel$ operator is used with a stride of 2 and kernel size of 3. The resolution of the image gradient is maintained similar to the original image.
4) **Combined Gradient Map:** The absolute sum of x and y gradients is obtained to create a unified gradient map representing edge information in the depth.

Step 2: Refining the Gradient Map

1) **Smoothing Operation:** The gradient map undergoes convolution with a smoothing operator to refine the edges of the gradient map before fusing it with the original image.
2) **Normalization:** The refined gradient map is normalized to obtain the final gradient map (referred to as "gradient map star").

Step 3: Processing the RGB Image in HSV Color Space

1) **Conversion to HSV:** The RGB image is converted to the HSV color space.
2) **Manipulation of Saturation and Value Spaces:**
   - **Saturation Manipulation:** The saturation component of the HSV image is multiplied pixel-wise by the depth map. This emphasizes high contrast for nearby subjects, enhancing the perception of depth.
   - **Value Manipulation:** The value component of the HSV image is multiplied pixel-wise by the complement of the gradient map. This emphasizes high brightness in areas where there's a significant change in depth.

Step 4: Returning to RGB Space

1) **Conversion Back to RGB:** The modified HSV image is converted back to the RGB color space, retaining the enhanced features.

Final Output

The resulting image in RGB space exhibits enhanced depth perception, with emphasized contrasts for nearby subjects and heightened brightness in regions where the depth changes significantly. This process effectively creates a stylized image that conveys a sense of depth perception beyond the original RGB representation.

D. Scenimefy

While performing the literature survey, we figured that there is a lack of high quality anime datasets. The same was reported in our base paper on Scenimefy [1]. They mention that a big domain gap in real-world and anime datasets does impact the accuracy of anime image generation. Scenimefy has been typically used for transferring real-scene images into anime styles. Moreover, most of the literature has reported to use StyleGAN [8] using an unsupervised learning framework. Scenimefy introduces an Image-to-Image translation framework that guides the learning with structure-consistent pseudo-paired data that is derived from a semantic-constrained StyleGAN . It learns effective pixel-wise correspondence and generates fine details between the two domains, guided by a

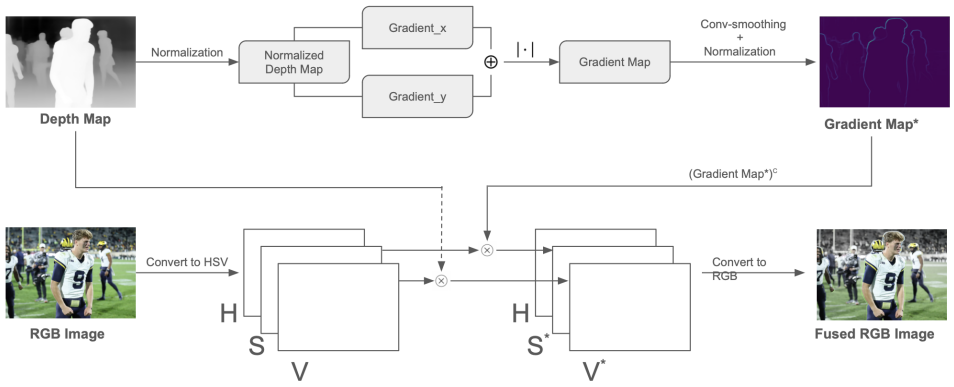

Fig. 1. EnGD: Our Proposed Fusion Approach

novel patch-wise contrastive style loss. In terms of learning, it introduces semi-supervised image-to-image translation by ingesting pseudo-paired anime data while the unsupervised branch learns the true target distribution using original real and anime datasets. In training, the loss function is defined as:

$$L_{i2i} = L_{unsup} + \lambda_{sup} * L_{sup} \quad (1)$$

where both L_{unsup} and L_{sup} are summation of GAN loss and style patch function and λ_{sup} decays following the cosine function as training proceeds.

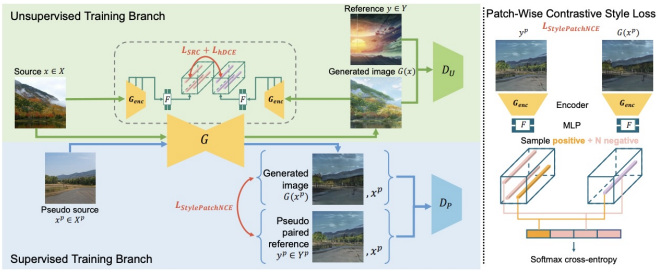

Fig. 2. Scenimefy architecture. Taken from [1].

IV. Prototype

The data set comprises 1,986 image frames extracted from nine Shinkai films, each frame having dimensions of 1920 x 1080. Frames were extracted using a Python script to collect key moments from movies such as 'Weathering with You' (2019), 'Your Name' (2016), 'Children Who Chase Lost Voices' (2011), among others.

The pipeline for the project is segmented into multiple sections as described below.

A. LDC (Lightweight Dense CNN) for Edge Detection:

- Utilizes an off-the-shelf edge detector to generate edge maps for each image.
- We invert the edge map as the initial output represents edges in black pixels for alignment with the subsequent processes.

B. Token Fusion:

- Feeds original RGB images and their corresponding inverted edge maps (from LDC) to vision transformers to produce depth maps while preserving the image resolution.

C. EnGD (Edge and Gradient of Depth):

- Integrates the depth map and a custom gradient map derived from the depth map using EnGD, enhancing the image's style and depth perception.
- Fusion Process:
  - Converts RGB images to HSV color space.
  - Multiplies the saturation component by the depth map and the value component by the complement of the custom gradient map on a pixel-wise basis.

D. Scenimefy:

- A semi-supervised image-to-image translation framework used to generate Shinkai-style anime images.
- Training data used by authors of the Scenimefy.
  - Real-world scene photos and anime scene images for the unsupervised branch. Used 90,000 natural landscape images from the Landscapes High-Quality (LHQ) data set as their training set
  - A pseudo paired dataset for the supervised branch. The study also contributed a high resolution Shinkai-style pure anime scene data set comprising 5,958 images.
- Model Training:
  - We further train the Scenimefy model on 1,986 generated fused images for 20 epochs using model checkpoints provided by the paper's authors. We trained with decaying factor $\lambda_{sup} = 0.05$ as suggested in the paper.
  - We used Nvidia Tesla V100-SXM2 16GB GPU as training resource.
- The final fused image is trained in the generative adversarial network of Scenimefy, resulting in a Shinkai-style anime image with realistic effects and depth

Fig. 3. Results

perception. We evaluate our generated stylized images using FID (Fréchet Inception distance) score. Lower the FID score, the better the output of the GAN.

This refined pipeline leverages multi-step algorithms and semi-supervised learning to transform original frames into Shinkai-style anime images with enhanced depth and visual aesthetics.

V. Results

As mentioned above, we trained 1986 Shinkai-styled RGB fused image dataset on StyleGAN. We report FID score as our evaluation metric. We reported FID score of 66.343 whereas Scenimefy baseline paper reported best results in this domain (FID 44). However our results are at-par with other unsupervised Image-to-Image architectures. Our lesser FID score may be a reflectance of lesser number of training data-points. In the inference we took a mixture of 15 real-world images of natural scenes with different object orientations and lightning. After passing through the system pipeline we were able to generate Shinkai-styled anime images. Two of them have been reported with the intermediate results in Fig. 3. We can see that our pipeline enables to give a more immersive/depth perception to the main object in the images. However, we also found that our pipeline does not generate optimum results for cases where the object is not predominant w.r.t background in the image. It was also sensitive to the lightning in the image. For low-light semi-dark images our anime images generated may not be optimal to the style. In future, we want to adapt newer styles by enhancing the training to more set of images. Although we did mention that there is a significant domain gap between real and anime scenes, and we need to generate high-quality anime scene dataset. We also want to iteratively improve our algorithm to capture features in the images which are light sensitive and rich in texture.

References

[1] 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*, 2023, pp. 7357–7367.

[2] 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*, 2020, pp. 8110–8119.

[3] "Top Market Research Intelligence and Consulting Firm — SkyQuest Technology Consulting Pvt. Ltd. — skyquestt.com," https://www.skyquestt.com/report/anime-market, [Accessed 07-12-2023].

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] X. Soria, A. Sappa, P. Humanante, and A. Akbarinia, "Dense extreme inception network for edge detection," *Pattern Recognition*, vol. 139, p. 109461, 2023.

[6] J. He, S. Zhang, M. Yang, Y. Shan, and T. Huang, "Bdcn: Bi-directional cascade network for perceptual edge detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 1, pp. 100–113, 2020.

[7] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 186–12 195.

[8] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.