변 혜 원 교수 지도
석사학위 청구논문

동적 최적화 기반 한국어-영어 장면 텍스트 스타일 전이 시스템

2024

성신여자대학교 대학원
미래융합기술공학과
김 예 림

동적 최적화 기반 한국어-영어 장면 텍스트 스타일 전이 시스템

변 혜 원 교수 지도

이 논문을 석사 학위 논문으로 제출함

2024년 5월

성신여자대학교 대학원

미래융합기술공학과

김 예 림

인 준 서

김예림의 석사학위 논문으로 인준함

2024년 6월

심사위원장 오 장 민 (서명 또는 인)

심 사 위 원 고 원 준 (서명 또는 인)

심 사 위 원 변 혜 원 (서명 또는 인)

성신여자대학교 대학원

논 문 개 요

본 연구는 외형 차이가 큰 한국어-영어 간 장면 텍스트 스타일 전이 시스템을 제안한다. 본 시스템은 StyleGAN 기반의 단일 언어 간 텍스트 스타일 전이 모델인 TSB(Text Style Brush)를 이중 언어 간 텍스트 스타일 전이 모델로 확장한다. 첫 번째 단계에서는 영어-영어 텍스트 스타일 전이 학습을 진행하고, 두 번째 단계에서는 한국어-영어 텍스트 스타일 전이 학습을 수행하는 2단계 학습 과정을 통해, 스타일 전이 결과에 대한 정답 이미지 없이 이중 언어 간 스타일 전이를 가능하게 한다. 텍스트 인식 분야에서 최고 성능을 달성한 TRBA 모델 구조를 기반으로 한국어 인식기를 학습하여 텍스트 스타일 전이 학습 시 생성된 이미지의 글자를 인식해 올바르게 생성되었는지 평가하는 데 사용한다. VGG16 기반의 구조를 사용하여 한국어 및 영어 글꼴 분류기를 학습하고 이를 통해 텍스트 스타일 전이 학습 시 입력 이미지와 생성된 이미지가 동일한 글꼴을 표현하는지 평가한다. 기존 최적화 방법론인 GL-GAN에 동적 임곗값 설정 방법을 도입한 동적 최적화를 수행함으로써 텍스트 스타일 전이 학습 시 한국어와 영어의 외형 차이로 인해 발생하는 아티팩트를 제거하고 품질을 향상시킨다. 기존 텍스트 스타일 전이 및 최적화 연구와의 성능 비교와 자체적인 절제 연구를 통해 제안하는 시스템의 성능을 입증한다.

목 차

그 림 목 차

표 목 차

Ⅰ. 서 론

'장면 텍스트(Scene text)'란 촬영된 이미지나 비디오에 나타나는 텍스트를 의미한다. 일본 드라마를 한국에 수출하는 플랫폼 '도라마 코리아'에서는 [그림 1-1]과 같이 일본어 장면 텍스트를 번역된 한국어로 변환하는 서비스를 제공하고 있다. 이 과정에서 중요한 점은 단순히 텍스트를 번역하는 것에 그치지 않고, 글꼴, 색상, 스타일을 전이함으로써 영상의 전반적인 시각적 일관성을 유지하고 영상 시청 과정에서 혼란을 방지하는 것이다. 그러나 이러한 과정은 수작업으로 이루어지기 때문에 많은 시간과 비용이 소모된다. 따라서 이러한 작업의 자동화가 필요하다. 이를 한국 콘텐츠 해외 수출에 적용하기 위해서는 한국어의 스타일을 번역된 영어로 전이할 수 있도록 텍스트 간 스타일 전이에 관한 연구가 필요하다.

[그림 1-1] '도라마 코리아' 장면 텍스트 번역 예시

TSB(Text Style Brush)[1]는 StyleGAN[2] 기반의 약한 자기 지도 학습 방식을

1) Krishnan, P., Kovvuri, R., Pang, G., Vassilev, B., & Hassner, T. (2023). Textstylebrush: transfer of text aesthetics from a single example. IEEE Transactions on Pattern Analysis and Machine Intelligence.

사용한 영어-영어 텍스트 스타일 전이 모델이다. TSB는 기존 연구와 비교하여 다음과 같은 장점을 보인다. 기존에는 동일한 스타일의 여러 다른 텍스트가 필요하여 퓨-샷 학습(Few-shot learning)[3] 환경에서 스타일 전이 학습을 수행했지만[4][5][6], TSB는 단일 텍스트만으로도 스타일 전이가 가능하여 원-샷 학습(One-shot learning)[7] 환경에서 스타일 전이 학습이 가능하다. 또한, 기존에는 한 글자씩 스타일 전이를 수행했지만[3][4][5][8], TSB는 단어 수준에서도 스타일 전이가 가능하다. 마지막으로, 기존에는 텍스트 스타일 전이 결과에 대한 정답 이미지가 필요했지만[3][4][5][9], TSB는 입력 이미지를 정답 이미지로 사용하는 새로운 학습 방법을 제시하였다. 그러나 TSB는 단일 언어 간 스타일 전이에만 적용된다는 한계가 있다.

Xie, Yangchen, et al.[10]은 TSB를 기반으로 한 이중 언어 간 텍스트 스타일

---

2) Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4401-4410).

3) Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. Advances in neural information processing systems, 30.

4) Li, C., Taniguchi, Y., Lu, M., & Konomi, S. I. (2021). Few-shot font style transfer between different languages. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 433-442).

5) Pan, W., Zhu, A., Zhou, X., Iwana, B. K., & Li, S. (2023). Few shot font generation via transferring similarity guided global style and quantization local style. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 19506-19516).

6) Wang, C., Zhou, M., Ge, T., Jiang, Y., Bao, H., & Xu, W. (2023). Cf-font: Content fusion for few-shot font generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1858-1867).

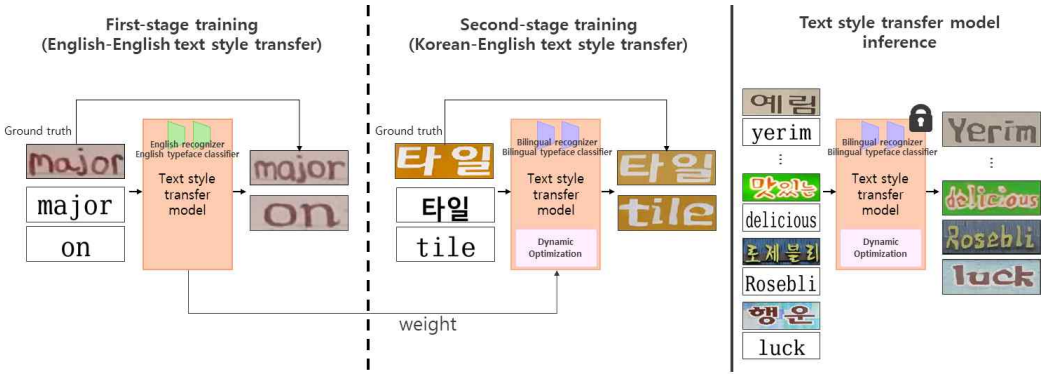
7) Vinyals, O., Blundell, C., Lillicrap, T., & Wierstra, D. (2016). Matching networks for one shot learning. Advances in neural information processing systems, 29.

8) Azadi, S., Fisher, M., Kim, V. G., Wang, Z., Shechtman, E., & Darrell, T. (2018). Multi-content gan for few-shot font style transfer. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7564-7573).

9) Kong, Y., Luo, C., Ma, W., Zhu, Q., Zhu, S., Yuan, N., & Jin, L. (2022). Look closer to supervise better: One-shot font generation via component-based discriminator. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13482-13491).

10) Xie, Y., Chen, X., Zhan, H., Shivakumara, P., Yin, B., Liu, C., & Lu, Y. (2024). Weakly supervised scene text generation for low-resource languages. Expert

전이 모델을 제안하였다. 이 모델은 '통합 어텐션 모듈(Integrated attention module)'을 도입하여 각 언어의 구도 차이 및 획 차이를 학습함으로써 이중 언어 간 텍스트 스타일 전이를 가능하게 하였다. 그러나 XIE, Yangchen, et al.의 연구는 특징 차이가 적은 비교적 유사한 형태를 가진 언어 간(영어-카자흐어, 중국어-한국어) 스타일 전이가 가능하다는 한계를 지닌다.



[그림 1-2] 한국어-영어 텍스트 스타일 전이 시스템 파이프라인

본 논문에서는 TSB를 활용하여 외형 차이가 큰 이중 언어 간 텍스트 스타일 전이 시스템을 제안한다. TSB의 입력 이미지를 정답 이미지로 활용하는 학습 방식은 한국어-영어 텍스트 스타일 전이가 불가능하다. 이는 한국어와 영어의 임베딩 공간 사이의 거리 차이가 크기 때문이다. 따라서 [그림 1-2]와 같이 2단계 학습으로 구성하였다. 첫 번째 단계에서는 영어-영어 텍스트 스타일 전이 학습을 수행하고, 두 번째 단계에서는 한국어-영어 텍스트 스타일 전이 학습을 진행한다. 이때 첫 번째 단계에서 학습된 가중치를 활용하여 StyleGAN의 매핑 네트워크에서 영어의 잠재 공간을 사용함으로써 한국어의 스타일을 영어에 전이한다.

이 시스템은 동일 언어 간(영어-영어) 텍스트 스타일 전이 모델인 TSB를 이

Systems with Applications, 237, 121622.

중 언어 간(한국어-영어) 텍스트 스타일 전이 모델로 확장한다. 첫째, 한국어 인식기를 적용한다. 이 인식기는 최고 성능(SOTA:State-of-the-Art)을 달성한 TRBA[11] 모델 구조를 기반으로 학습되었으며, 생성된 이미지의 글자를 인식하여 올바르게 생성되었는지 평가한다. 둘째, 한국어 및 영어 글꼴 분류기를 적용한다. 이 분류기는 VGG16(Visual Geometry Group - 16 Layers)[12] 구조로 학습되었으며, 생성된 이미지가 입력 이미지와 동일한 글꼴을 표현하는지 평가한다. 셋째, 동적인 전역적 및 지역적 이중 수준 최적화를 수행한다. 이 최적화는 스타일 전이 결과의 품질을 향상시키기 위해 GL-GAN[13]의 최적화 방법론에 동적인 임곗값 설정 방법을 도입하였다. 생성된 이미지의 품질에 따라 전역 최적화와 지역 최적화 중 하나를 선택하여 수행하고, 동적인 임곗값을 사용하여 최적화 학습의 방향을 조절한다.

제안하는 시스템의 우수성은 기존의 텍스트 스타일 전이 모델 및 최적화 모델과의 성능 비교와 자체적인 절제 연구를 통해 증명한다. 본 연구의 기여는 다음과 같다.

1. 외형 차이가 큰 이중 언어 간 텍스트 스타일 전이 시스템을 제안한다.

2. 2단계 학습 과정을 도입하여 스타일 전이 결과에 대한 정답 이미지 없이 이중 언어 간 텍스트 스타일 전이를 구현한다.

3. 단일 언어 간 텍스트 스타일 전이 모델을 이중 언어 간 텍스트 스타일 전이 모델로 확장한다.

4. 스타일 전이 결과의 품질 향상을 위해 동적인 전역적 및 지역적 이중 수준

---

11) Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., ... & Lee, H. (2019). What is wrong with scene text recognition model comparisons? dataset and model analysis. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4715-4723).

12) Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

13) Liu, Y., Fan, H., Yuan, X., & Xiang, J. (2022). GL-GAN: Adaptive global and local bilevel optimization for generative adversarial network. Pattern Recognition, 123, 108375.

최적화를 수행한다.

5. 기존의 최적화 모델 GL-GAN을 개선하여 학습 시간을 단축하고 성능을 향상시켰다.

Ⅱ. 관련 연구

1. 스타일 전이

이미지에서 이미지로 스타일을 전이하는 연구는 이전부터 계속해서 진행되어 왔다. 초기에는 CNN(Convolutional Neural Networks)의 다양한 레이어를 활용하여 스타일 이미지로부터 스타일을 추출하고, 이를 콘텐츠 이미지에 전이하는 방식이 사용되었다[14]. 그러나 이 방법은 콘텐츠 이미지가 변경될 때마다 학습을 다시 진행해야 하므로 학습 속도가 느리다는 단점이 있다. 이를 해결하기 위해 Johnson, Justin, et al.[15]은 지각 손실 값을 이용한 피드 포워드(Feed-Forward) CNN을 도입하여 실시간 스타일 전이를 가능하게 했다. 이 방법은 학습 시 스타일 이미지만을 사용하여 스타일을 추출하고, 추론 시에 콘텐츠 이미지에 스타일을 전이하기 때문에 다시 학습할 필요가 없어져 실시간 전이가 가능해졌다. 하지만 이미지에서 스타일을 추출하는 방식을 학습하기 때문에 스타일 이미지가 변경된다면 다시 학습을 진행해야 한다는 문제는 여전히 존재했다. 이후에는 CycleGAN[16]과 StarGAN[17] 기반의 스타일 전이 모델[18][19]이 등장하면서 재학습

14) Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2414-2423).

15) Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14 (pp. 694-711). Springer International Publishing.

16) Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2223-2232).

17) Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8789-8797).

18) Liu, M. Y., Breuel, T., & Kautz, J. (2017). Unsupervised image-to-image translation

할 필요가 없어졌다. 각 도메인에서 동일한 스타일을 나타내는 이미지를 충분히 많이 확보한 후 도메인 간의 이미지를 변환하는 방식을 학습했기 때문이다.

그러나 방대한 데이터를 수집하는 것은 어려운 일이다. FUNIT[20]은 적은 양의 데이터를 사용해 퓨-샷 학습을 가능하게 하였다. 이 모델은 고수준의 시맨틱 정보를 활용하여 이미지 간의 매핑을 수행한다. 이는 입력 이미지와 출력 이미지 사이의 고수준 특징을 추출하고, 이를 기반으로 이미지를 변환한다. 이러한 방식은 작은 데이터 세트에서도 강력한 일반화 성능을 제공하였다.

이후 TSB[1]는 StyleGAN2[21]의 매핑 네트워크를 통해 각 스타일을 잠재 공간으로 매핑함으로써 손쉬운 스타일 조작을 가능케 하였다. 이를 다양한 손실값을 사용한 최적화를 통해 한 장의 이미지만으로도 스타일 전이가 가능한 원-샷 모델을 제안하였다. 본 논문에서는 TSB를 확장하여 하나의 한국어 텍스트의 스타일을 번역한 영어에 전이하는 시스템을 구성하였다.

2. 텍스트 스타일 전이

텍스트의 글꼴, 색깔, 두께 등의 스타일을 새로운 텍스트에 전이하는 연구는 다양하게 진행되고 있다. Li, Chenhao, et al.[4]은 847개의 글꼴 데이터를 사용하여 영어와 중국어 간의 스타일을 전이하는 모델을 제안하였다. 847개의 글꼴은 영어와 중국어를 모두 표현이 가능하여 스타일 전이에 있어 각 글꼴은 서로 정답

networks. Advances in neural information processing systems, 30.

19) Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In Proceedings of the European conference on computer vision (ECCV) (pp. 172-189).

20) Liu, M. Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., & Kautz, J. (2019). Few-shot unsupervised image-to-image translation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10551-10560).

21) Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8110-8119).

으로 활용할 수 있었다. 그러나 이 모델은 글꼴 데이터를 사용하여 한 글자씩 스타일 전이가 가능한 모델로, 단어의 스타일 전이는 불가능하였다. 그뿐만 아니라, Li, Chenhao, et al.의 중국어와 영어를 제외하면, 두 언어를 모두 지원하는 글꼴이 제한적인 언어들이 다수 존재한다. 따라서 스타일 전이에 대한 정확한 정답 데이터를 확보하는 것은 어렵다. 또한, 정답 데이터가 부재하면 텍스트에 스타일을 효과적으로 전이하는 것은 어려운 과제이다.

다음 연구들에서는 스타일 전이 결과에 대한 정답이 없는 데이터를 활용하여 텍스트에 효과적으로 스타일을 전이하도록 시도했다. Park, Song, et al.[22]은 중국어에서 한국어로의 스타일 전이 방법을 제안하였다. Park, Song, et al.은 글자의 구성 요소 스타일과 콘텐츠 특징을 분리하여 학습하는 'MX-Font(Multiple Localized eXperts Fewshot Font Generation Network)'를 제안했다. MX-Font는 다중 지역 전문가를 활용하여 지역적인 스타일을 포착하고, 약한 지도를 통해 각 전문가가 특정 로컬 개념에 특화되도록 유도한다. 이러한 방법을 통해 스타일 전이 결과에 대한 정답이 없는 데이터를 기반으로 한 효과적인 스타일 전이가 가능해졌다. 하지만 한 글자씩의 스타일 전이가 가능했으며, '充': ['亠', '厶', '川']와 같은 각 글자의 구성 요소에 대한 라벨 정보가 추가로 필요했다.

Kong, Yuxin, et al.[9]은 중국어-한국어 간 스타일 전이 방법을 제안하였다. Kong, Yuxin, et al.은 글자의 구성 요소를 판별하는 '구성 요소 기반 판별기(Component-Based Discriminator)'를 제안하여 스타일 전이 과정에서 더욱 정교한 판별을 실현하고자 하였다. 기존 연구들이 주로 전체적인 스타일을 고려하는 데에 그쳤던 반면에, 각 글자의 부분적인 특성에 주목함으로써 보다 세밀한 스타일 변화를 달성하였다. 이러한 방식은 한국어, 중국어와 같이 구성 요소가 복잡한 언어의 특성을 고려한 것이다. 하지만 앞선 연구와 마찬가지로 '充': ['亠', '厶', '川

22) Park, S., Chun, S., Cha, J., Lee, B., & Shim, H. (2021). Multiple heads are better than one: Few-shot font generation with multiple localized experts. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 13900-13909).

']와 같은 라벨 정보가 추가로 필요하였다.

Li, Chenhao, et al.[4], Pan, Wei, et al.[5], Wang, Chi, et al.[6], Azadi, Samaneh, et al.[8]은 개별 글자별로 텍스트의 스타일을 전이하는 모델을 제안했다. 그러나 이러한 방법은 각 글자를 개별적으로 처리하기 때문에 생성하고자 하는 글자의 수와 모델 실행 횟수가 비례하여 증가한다. 이는 효율성 측면에서 한계를 가지고 있다. 이러한 한계를 극복하기 위해 텍스트를 개별 문자가 아닌 단어 수준으로 스타일을 전이하는 장면 텍스트 스타일 전이 연구가 이루어지고 있다.

SRNet[23]은 텍스트와 배경 이미지를 분리한 뒤, 각각의 스타일을 전이하여 혼합하는 방식의 장면 텍스트의 스타일 전이 방식을 제안한다. 이러한 방식은 학습 시 텍스트와 배경 이미지가 분리가 가능한 데이터가 필요하므로 합성 데이터에서만 학습이 가능하다는 한계가 존재한다. TSB는 스타일 전이 결과에 대한 정답 이미지와 추가적인 스타일 정보 없이 장면 텍스트의 스타일 전이가 가능한 모델을 제안한다. 이 모델은 장면 텍스트 자체를 스타일 전이 결과에 대한 정답 이미지로 활용하여 약한 자기 지도 학습을 진행하였다. 그러나 영어에서 영어로 스타일을 전이하는 단일 언어 간 스타일 전이 모델이기 때문에 이중 언어 간 스타일 전이에는 어려움이 존재했다.

3. TSB(Text Style Brush)

TSB(Text Style Brush)[1]는 StyleGAN[2] 기반의 약한 자기 지도 학습 방식을 사용한 영어-영어 텍스트 스타일 전이 모델로, 기존 연구들과 비교하여 여러 가지 장점을 보인다. 기존의 텍스트 스타일 전이 연구들에서는 동일한 스타일을 가진 여러 텍스트가 필요하여 퓨-샷(few-shot) 학습 환경에서 스타일 전이 학습을

23) Wu, L., Zhang, C., Liu, J., Han, J., Liu, J., Ding, E., & Bai, X. (2019, October). Editing text in the wild. In Proceedings of the 27th ACM international conference on multimedia (pp. 1500-1508).

수행하였다[4)5)6)]. 그러나 TSB는 단일 이미지로 텍스트 스타일을 추출하고 전이할 수 있어 원-샷(One-shot) 학습 환경에서 스타일 전이 학습을 수행할 수 있다. 이는 StyleGAN의 내부 매핑 네트워크를 통해 입력 데이터의 스타일을 손쉽게 제어할 수 있기 때문이다.

기존 연구들에서는 글꼴 데이터를 사용하여 스타일 전이를 시도하였다[4)5)6)8)]. 글꼴 데이터를 사용하면 모든 글자 생성을 가능하게 하여 스타일 전이 결과에 대한 정답을 생성할 수 있다. 또한 글꼴 데이터는 배경이 없는 경우가 많아, 글자의 스타일 요소를 추출할 때 배경의 영향을 받지 않는다. 따라서 각 글자를 독립적인 이미지로 다룰 수 있으므로 한 글자 단위로 스타일 전이가 용이하며, 글자의 순수한 스타일 요소만을 다룰 수 있다. 그러나 이러한 방식은 단어 전체의 스타일 전이에 한계가 있다. 단어는 개별 문자 간의 배치와 조합, 그리고 상호작용하는 스타일 요소들이 존재하기 때문이다. TSB는 장면 텍스트 데이터를 사용하여 단어 수준에서 스타일 전이를 가능하게 하여, 보다 자연스럽고 일관된 스타일 전이를 제공한다.

그러나 장면 텍스트는 다양한 배경과 각기 다른 스타일을 가진 텍스트를 포함하고 있으므로 정답 데이터를 수집하는 것은 어려운 일이다. 이를 해결하기 위해 TSB는 스타일 전이 결과에 대한 정답 데이터 없이도 장면 텍스트를 사용하여 스타일 전이를 수행할 수 있도록 약한 자기 지도 학습을 활용한 새로운 학습 방안을 제안하였다. TSB는 입력된 장면 데이터와 동일한 텍스트에도 스타일을 전이한 결과를 생성한다. 그 후 장면 텍스트를 그 결과의 정답 이미지로 사용하여 학습을 진행한다. 즉, 입력 데이터 자체를 정답 이미지로 활용하는 방식이다. 이를 통해 정답 데이터가 없는 상황에서도 스타일 전이를 효과적으로 수행할 수 있게 된다. 그뿐만 아니라, TSB는 생성된 이미지의 진위를 판별하는 판별자뿐만 아니라 입력 이미지와 생성된 이미지의 스타일을 비교하여 평가하는 다양한 모델을 활용한다. 이러한 다양한 평가 모델들을 사용하여 스타일 전이에 필요한 손실 함수를

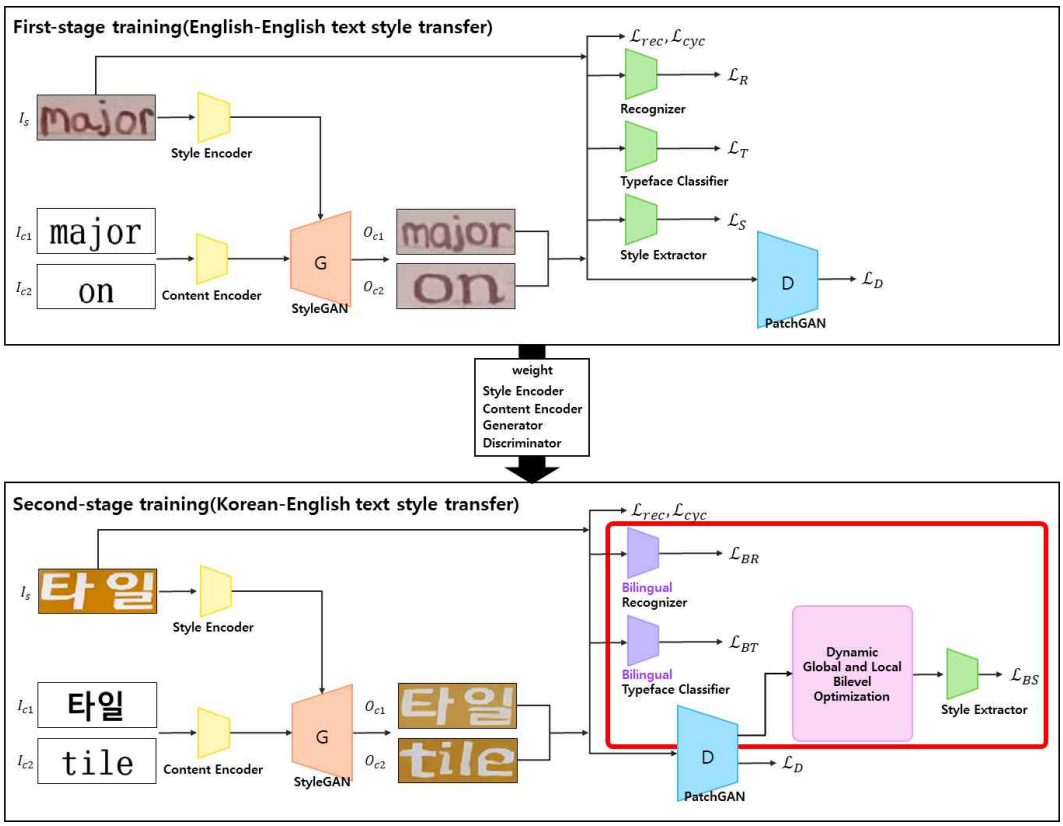
계산하고, 이를 통해 텍스트 스타일 전이 모델을 학습시킨다. 이와 같은 접근법을 통해 TSB는 다양한 손실 함수를 최적화하여 스타일 전이를 가능하게 하였다.

그러나 TSB는 영어-영어 텍스트 스타일 전이 모델로 설계되어 있으며, 모델 내부 구조가 영어에 특화되어 있다. 또한, 언어에 따라 임베딩 공간이 상이한데, 특히 한국어와 영어의 임베딩 공간 거리의 차가 크기 때문에 한국어-영어 텍스트 스타일 전이는 어려움을 겪게 된다. 이러한 차이로 인해 TSB는 한국어와 영어 간의 텍스트 스타일 전이를 효과적으로 수행할 수 없었다.

Xie, Yangchen, et al.[10]은 TSB를 기반으로 한 모델을 제안하였으며, '통합 어텐션 모듈(Integrated Attention module)'을 도입하여 이중 언어 간 스타일 전이를 시도하였다. 이 통합 어텐션 모듈은 텍스트의 전역적 특징인 구도 차이와 지역적 특징인 획 차이를 학습함으로써 TSB를 이중 언어 간 스타일 전이에도 적용할 수 있도록 개선하였다. 그러나 이 모델은 문자를 나열한 형태의 카자흐어와 영어 간, 문자를 조합한 형태의 한국어와 중국어 간 스타일 전이를 시도하였다. 이러한 접근 방식은 특징 차이가 적은 비교적 유사한 형태를 가진 언어 간의 스타일 전이에는 성공적이었으나, 형태가 크게 다른 언어 간의 스타일 전이에는 제한적일 수 있다는 한계가 있다.

Ⅲ. 한국어-영어 장면 텍스트 스타일 전이 시스템

1. 시스템 구조



[그림 3-1] 시스템 구성도

제안하는 한국어-영어 장면 텍스트 스타일 전이 시스템은 [그림 3-1]과 같이 2단계의 학습 과정으로 구성된다. 첫 번째 단계에서는 StyleGAN[2] 기반의 TSB[1]의 모델 구조를 기반으로 한 영어-영어 텍스트 스타일 전이 학습을 수행한다. 두 번째 단계에서는 TSB를 이중 언어 간 스타일 전이 모델로 확장한 구

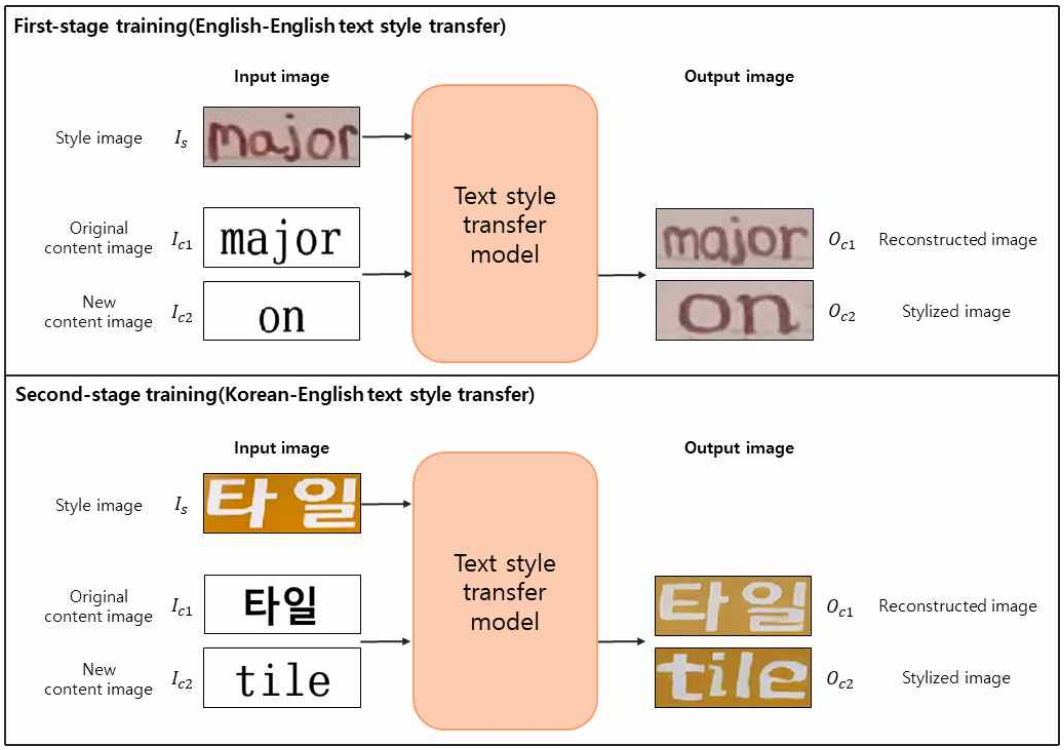
조의 한국어-영어 텍스트 스타일 전이 학습을 진행한다. 두 모델은 스타일 전이 결과에 대한 정답 이미지가 없는 (장면 텍스트 – 텍스트) 쌍을 사용하여 약한 자기 지도 학습을 수행한다.

TSB는 입력 이미지를 정답 이미지로 활용하는 학습 방식을 채택한다. 예를 들어, [그림 3-1]에서 입력 이미지의 스타일을 “on”에 전이하는 것이 목적이지만, “on”에 스타일 전이한 결과를 평가할 수 있는 명확한 정답 이미지가 존재하지 않는다. 따라서 입력 이미지와 동일한 “major”에 스타일 전이한 결과를 입력 이미지와 비교하는 방식으로 입력 이미지를 정답처럼 사용한다. 이 과정에서 판별자뿐만 아니라 다양한 모델을 사용해 스타일 손실값을 계산함으로써 스타일 전이가 가능해지고, 이 가중치를 사용해 “on”에도 스타일 전이가 가능해진다.

하지만 이러한 TSB의 학습 방식은 한국어-영어 스타일 전이에 적용하기 어렵다. 입력 이미지를 정답 이미지처럼 사용한다면 한국어에 대해서 효과적으로 스타일 전이가 가능해지지만, 영어에 대해서는 불가능하다. 이는 한국어와 영어의 임베딩 공간의 거리 차이가 크기 때문이다. 이러한 문제를 해결하기 위해, 2단계 학습 과정을 도입하였다. 두 번째 단계에서 첫 번째 단계의 가중치를 사용하여 StyleGAN의 매핑 네트워크에서 영어의 잠재 공간을 사용함으로써 한국어의 스타일을 영어에 전이가 가능해진다.

또한, 2단계 학습에서는 TSB를 이중 언어 간 스타일 전이 모델로 확장하기 위해 이중 언어 인식기, 이중 언어 글꼴 분류기, 동적 최적화 3가지를 도입하였다.

2. 2단계 학습



[그림 3-2] 입출력 이미지 용어 설명

- 스타일 이미지(I_s, Style image) : 추출하고자 하는 스타일의 장면 텍스트
- 원본 콘텐츠 이미지(I_{c1}, Original content image) : 스타일 이미지와 동일한 원본 텍스트($c1$, Original text)를 이미지로 변환한 것
- 신규 콘텐츠 이미지(I_{c2}, New content image) : 임의로 선택된 신규 텍스트($c2$, New text)를 이미지로 변환한 것
- 재건 이미지(O_{c1}, Reconstructed image) : 원본 텍스트에 스타일을 전이한 이미지

- 스타일 적용 이미지(O_{c2}, Stylized image) : 신규 텍스트에 스타일을 전이한 이미지

1) 영어-영어 텍스트 스타일 전이 학습

영어-영어 텍스트 스타일 전이 모델은 총 7개의 네트워크로 구성되어 있으며, 크게 이미지 생성에 관여하는 네트워크와 평가에 관여하는 네트워크로 구분할 수 있다. 이미지 생성에 관여하는 네트워크는 (1) 텍스트의 스타일을 추출하는 스타일 인코더(Style Encoder), (2) 텍스트의 글자 형태를 추출하는 콘텐츠 인코더(Content Encoder), (3) 원본 텍스트의 스타일을 새로운 텍스트에 전이한 이미지를 생성하는 StyleGAN 기반의 생성자(Generator)로 구성되어 있다. 이미지 평가에 관여하는 네트워크는 (4) 생성된 텍스트를 인식하여 올바른 형태의 글자를 생성했는지 평가하는 텍스트 인식기(Recognizer), (5) 생성된 텍스트의 글꼴이 원본 텍스트의 글꼴과 동일한지 평가하는 글꼴 분류기(Typeface classifier), (6) 생성된 텍스트의 시각적 특징과 질감 표현이 원본 텍스트와 동일한지 평가하는 VGG16(Visual Geometry Group - 16 Layers)[12] 기반의 스타일 추출기(Style Extractor), (7) 텍스트가 생성된 텍스트인지 실제 텍스트인지 판별하는 PatchGAN[24] 기반의 판별자(Discriminator)로 구성되어 있다.

영어-영어 스타일 전이 학습 모델의 입력은 스타일 이미지, 원본 콘텐츠 이미지, 그리고 신규 콘텐츠 이미지로 이루어져 있으며, 출력은 재건 이미지, 스타일 적용 이미지로 구성된다. 모든 입력과 출력은 영어로 이루어져 있다.

원본 텍스트 및 신규 텍스트에 스타일 이미지의 스타일을 전이하는 것을 목표로 하며, 학습 과정은 다음과 같다. 스타일 인코더와 콘텐츠 인코더는 각각

24) Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-1134).

스타일 이미지와 콘텐츠 이미지에서 특징값을 추출한다. 생성자는 스타일 매핑 네트워크를 사용해 스타일 특징값을 스타일 벡터로 변환한 후 콘텐츠 특징값과 연산하여 재건 이미지와 스타일 적용 이미지를 생성한다. 텍스트 인식기, 글꼴 분류기, 스타일 추출기 그리고 판별자는 스타일 이미지를 재건 이미지 및 스타일 적용 이미지와 비교하여 평가하고 다양한 손실값을 계산한다. 계산된 손실값을 사용하여 스타일 및 콘텐츠 인코더, 생성자 그리고 판별자를 학습시킨다.

2) 한국어-영어 텍스트 스타일 전이 학습

한국어-영어 텍스트 스타일 전이 모델은 영어-영어 텍스트 스타일 전이 학습 모델과 유사하지만, 입력 데이터에 차이가 있다. 영어-영어 텍스트 스타일 전이 학습 모델은 모든 입출력이 영어로 이루어져 있지만, 한국어-영어 텍스트 스타일 전이 모델은 스타일 이미지와 원본 콘텐츠 이미지는 한국어로 이루어져 있고, 신규 콘텐츠 이미지는 영어로 이루어져 있다.

학습 과정은 영어-영어 텍스트 스타일 전이 학습과 동일하지만, 생성된 이미지를 평가하는 과정에서 사용되는 네트워크와 해당 네트워크를 사용한 손실값 계산 방법에 차이가 있다. 글자가 올바른 형태로 생성되었는지 평가하기 위해 영어만 인식할 수 있는 텍스트 인식기를 한국어와 영어를 모두 인식할 수 있도록 이중 언어 인식기(Bilingual Recognizer)로 교체하고, 생성된 이미지의 한국어와 영어의 글꼴이 스타일 이미지의 글꼴과 동일한지 평가하기 위해 영어 글꼴만 분류할 수 있는 글꼴 분류기를 한국어와 영어의 글꼴을 모두 분류할 수 있는 이중 언어 글꼴 분류기(Bilingual Typeface Classifier)로 교체하였다. 또한 한국어와 영어의 외형적인 차이로 인해 발생한 아티팩트를 제거하여 품질을 향상시키기 위해 동적인 전역적 및 지역적 이중 수준 최적화를 수행하였다.

3. 이중 언어 인식기

이중 언어 인식기는 한국어-영어 텍스트 스타일 전이 학습에서 재건 이미지와 스타일 적용 이미지의 텍스트를 인식하여 올바른 형태의 글자를 생성했는지 평가하는 데 사용된다. 재건 이미지의 경우 한국어 인식기를 사용하여 원본 텍스트와의 일치 여부를 검증하고, 스타일 적용 이미지의 경우 영어 인식기를 사용하여 신규 텍스트와의 일치 여부를 검증한다.

한국어 인식기는 텍스트 인식 분야에서 최고 성능(SOTA: State-of-the-Art)을 달성한 TRBA[11] 모델의 구조를 기반으로 학습되었다. 학습 과정은 다음의 네 단계로 이루어진다. 첫 번째, 변환 단계(Transformation Stage)에서는 TPS(Thin-Plate Spline)[25]를 사용하여 장면 텍스트에서 텍스트 부분을 확대하고 인식하기 쉬운 형태로 변환하는 전처리를 수행한다. 두 번째, 특징 추출 단계(Feature Extraction Stage)에서는 ResNet[26]을 사용하여 이미지에서 특징값을 추출한다. 세 번째, 시퀀스 모델링 단계(Sequence modeling stage)에서는 BiLSTM(Bidirectional LSTM)을 사용하여 추출된 특징값들의 순차적인 상관관계를 파악한다. 네 번째, 예측 단계(Prediction stage)에서는 Attn(Attention-based sequence prediction)을 사용하여 텍스트 인식 결과를 예측한다. 이 모델은 어텐션 기반의 예측 모델이므로 가변적인 텍스트도 예측할 수 있다.

학습 데이터는 한국어-영어 텍스트 스타일 전이 학습과 동일한 데이터인 AI 허브[27]의 “야외 실제 촬영 한글 이미지”에서 202,113개의 (장면 텍스트 -

25) Zhao, J., & Zhang, H. (2022). Thin-plate spline motion model for image animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3657-3666).
26) He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
27) AIHub, Available at: https://www.aihub.or.kr/

텍스트) 쌍을 사용하였다. 한국어는 자음과 모음을 조합하여 글자를 생성하는데, 모든 조합의 경우의 수는 총 11,172개에 이른다. 이에 모든 글자를 인식하는 모델을 구현하는 것은 어려움이 있어, 학습 데이터에 포함된 579개의 글자를 인식할 수 있도록 학습하여 사용하였다. 한국어 인식기의 인식 정확도는 94.39%를 달성하였다.

영어 인식기는 Olga Kozlova[28]가 제공하는 TRBA-PR 모델을 사용하였다. 이 모델은 알파벳 소문자, 대문자, 숫자, 특수기호 등 97개의 글자를 인식할 수 있다.

4. 이중 언어 글꼴 분류기

이중 언어 글꼴 분류기는 한국어-영어 텍스트 스타일 전이 학습에서 재건 이미지 및 스타일 적용 이미지가 스타일 이미지와 동일한 글꼴을 표현하는지 평가하는 데 사용된다. 이 분류기는 VGG16 구조를 기반으로 학습하였으며, 한국어 및 영어 장면 텍스트의 글꼴을 분류하는 모델이다. 학습 데이터는 Gupta, Ankush, et al.[29]이 제안한 방법을 사용하여 직접 생성하였다. 다양한 배경 이미지에 'Google Font[30]'에서 수집한 37개의 한국어 및 영어 글꼴을 사용해 텍스트를 합성하였다. 합성 데이터 세트를 생성하는 구체적인 방법은 6장에서 상세히 다루었다.

이중 언어 글꼴 분류기의 분류 정확도는 61.74%를 달성하였다. 글꼴 분류기의 마지막 분류 레이어를 제거하고 글꼴 특징값 자체를 비교하는 방식을 채택하였다. 이는 분류 라벨을 사용하게 되면 각 글꼴이 독립적인 클래스로 간주

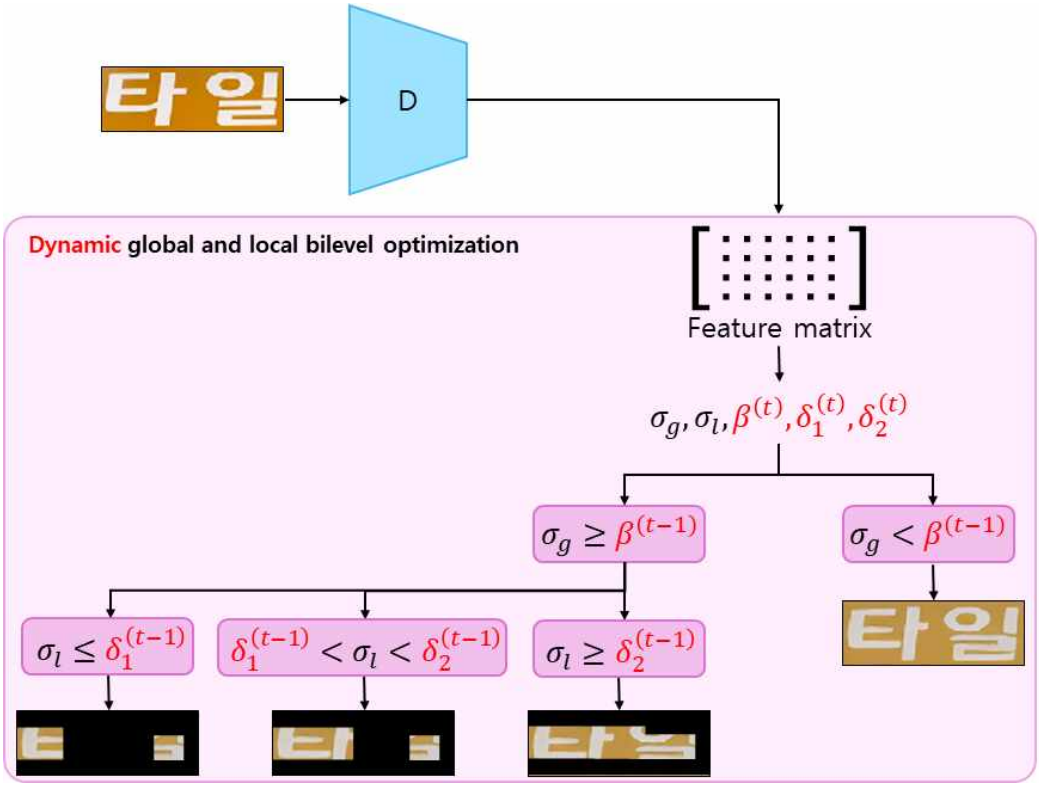
28) deep-text-edit, Github repository, https://github.com/grenlayk/deep-text-edit
29) Gupta, A., Vedaldi, A., & Zisserman, A. (2016). Synthetic data for text localisation in natural images. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2315-2324).
30) Google Fonts, Available at : https://fonts.google.com/

되어 유사한 글꼴 간의 차이를 무시하게 되는 문제점을 해결하기 위함이다. 글꼴 특징값을 비교하면 서로 유사한 글꼴들은 유사한 특징 벡터를 가지므로, 유사성 평가가 더욱 효과적으로 이루어진다. 또한, 이 방법은 글꼴의 미세한 차이도 포착할 수 있어 글꼴의 유사성을 평가하는 데 더 효과적이라고 판단하였다.

5. 동적인 전역적 및 지역적 이중 수준 최적화

한국어는 자음과 모음을 결합하여 글자를 형성하며, 이와 대조적으로 영어는 알파벳을 순차적으로 나열하여 글자를 생성한다. 이런 언어적 특성과 외형의 차이로 인해 한국어-영어 텍스트 스타일 전이 학습에서 생성된 이미지의 일부 배경에 아티팩트가 발생하였다. 이를 제거하고 품질을 향상시키기 위해, 동적인 전역적 및 지역적 이중 수준 최적화를 수행하였다. 이 최적화는 GL-GAN[13]의 최적화 방법론을 개선한 것으로, 그 방법은 다음과 같다.

생성된 이미지의 품질에 따라 전역 최적화 또는 지역 최적화 중 하나를 선택하여 수행한다. 생성된 이미지의 품질이 전체적으로 좋지 않다면, 이미지 전체를 사용해 최적화하는 전역 최적화를 수행하고, 이미지의 일부분만 품질이 좋지 않다면 그 부분만을 사용해 최적화하는 지역 최적화를 수행한다. 이때 최적화 방법을 결정하는 기준이 되는 임곗값인 베타(β)와 지역 최적화 시 저품질 지역을 선정하는 기준이 되는 임곗값 델타(δ_1,δ_2)가 필요하다. GL-GAN의 경우 모든 최적화 학습에서 동일한 임곗값을 사용하지만, 본 연구에서는 에포크마다 업데이트되는 동적인 임곗값을 도입하였다.



[그림 3-3] 동적인 전역적 및 지역적 이중 수준 최적화

$$\mu_k = \frac{\Sigma_{i,j} y_{i,j}}{h \cdot w}, \mu = \frac{\Sigma_{k=1}^{K} \mu_k}{K}$$
$$\sigma_g = \sqrt{\frac{\Sigma_{k=1}^{K} (\mu_k - \mu)^2}{K}}$$

[수식 3-1] 전역 품질 지표

- $y_{i,j}$: 특성 행렬의 요소
- h : 특성 행렬의 열 개수
- w : 특성 행렬의 행 개수
- K : 배치 크기

- σ_g(전역 표준 편차) : 전역 품질 지표를 나타내며, 배치 내에서 이미지의 품질 간 표준 편차로 계산된다.

$$\sigma_k = \frac{\Sigma_{i,j}(y_{i,j} - \mu_k)^2}{h \cdot w}, \sigma_l = \sqrt{\frac{\Sigma_{k=1}^{K} \sigma_k}{K}}$$

[수식 3-2] 지역 품질 지표

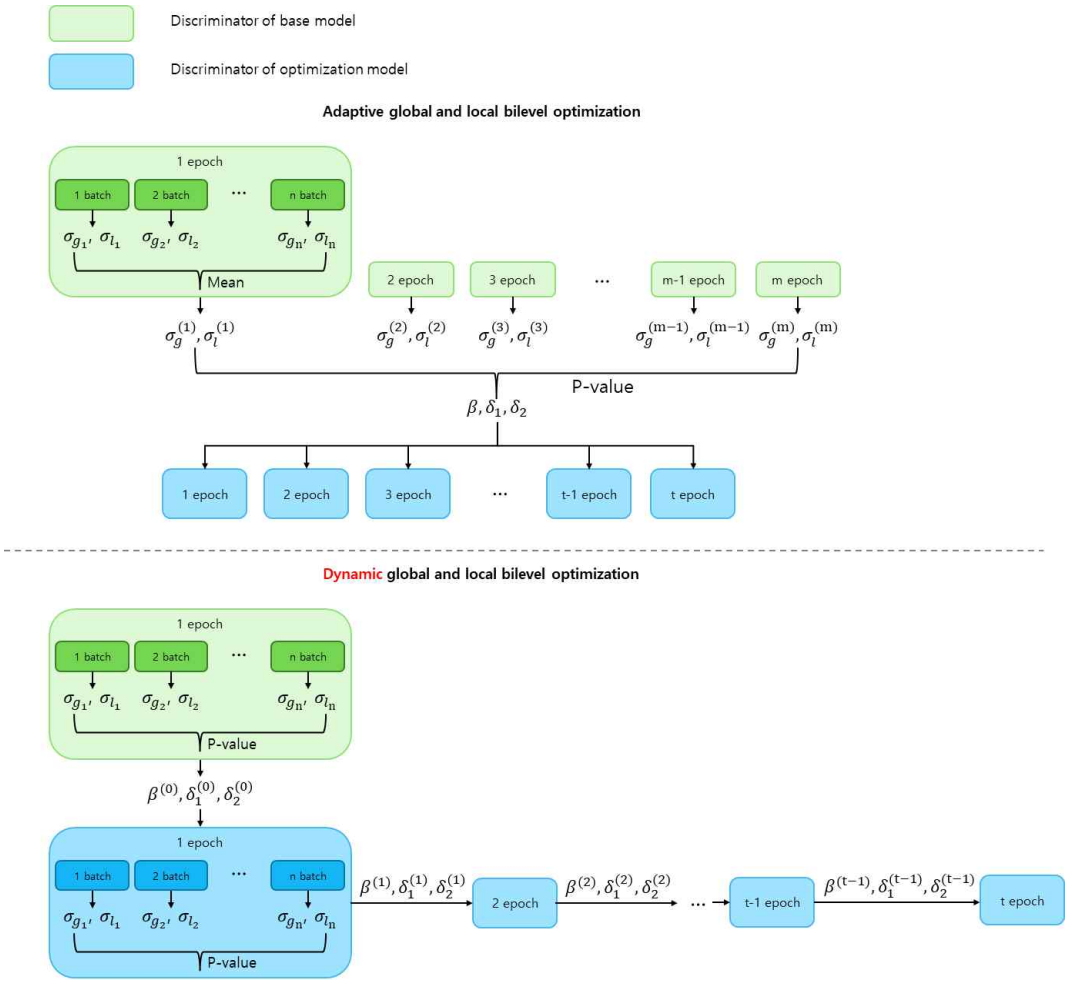
- σ_l(지역 표준 편차) : 지역 품질 지표를 나타내며, 배치 내에서 특징 행렬의 요소 간 표준 편차로 계산된다.
- δ_1, δ_2 : 지역 품질의 기준이 되는 임곗값
- β : 전역 품질의 기준이 되는 임곗값

최적화는 [그림 3-3]과 같은 방식으로 진행된다. 텍스트 스타일 전이 결과로 생성된 이미지를 PatchGAN[24] 기반의 판별자에 입력한 후 출력되는 특성 행렬(Feature matrix)을 사용해 이미지의 품질을 평가한다. 특성 행렬을 통해 총 5가지의 ($\sigma_g, \sigma_l, \beta, \delta_1, \delta_2$)값이 계산된다. 전역 표준 편차가 베타($\beta$)를 초과하는 경우, 이는 이미지의 품질이 전체적으로 좋지 않고 이미지의 내부에 넓은 범위의 불균형이 있다는 것을 의미하므로 전체 이미지를 사용해 최적화하는 전역 최적화를 수행한다. 그렇지 않은 경우, 저품질 지역이 존재하고 좁은 범위의 불균형이 있음을 의미하므로 저품질 지역만을 최적화하는 지역 최적화를 수행한다.

지역 최적화의 경우, 생성된 이미지에서 저품질을 제외한 부분에 마스크 연산을 수행하는데, 저품질 영역의 크기에 따라 마스크 행렬의 크기와 모양이 달라진다. 지역 표준 편차가 델타 원(δ_1)보다 작은 경우, 이는 저품질 지역이 작은 것을 의미하므로 크기가 큰 마스크 행렬과 연산하여 최적화를 수행하고, 지

역 표준 편차가 델타 투(δ_2)보다 큰 경우, 저품질 지역이 큰 것을 의미하므로 크기가 작은 마스크 행렬과 연산하여 최적화를 수행한다.

이때, 특성 행렬을 통해 계산되는 임곗값은 t번 에포크에서 계산되고, 비교할 때 사용되는 임곗값은 t-1 번 에포크에서 계산된 것이다. 즉, 이전 에포크에서 계산된 임곗값을 현재 최적화 학습에 사용하고, 현재 에포크에서 계산된 임곗값은 다음 에포크의 최적화 학습에 사용된다.



[그림 3-4] 최적화 알고리즘 비교

[그림 3-4]는 GL-GAN과 본 논문의 최적화 학습 시 사용되는 임곗값 설정 방법의 차이를 시각적으로 나타낸다. GL-GAN은 최적화를 적용하지 않은 기본 모델을 먼저 학습한다. 기본 모델 학습의 모든 에포크에서는 각 배치에서 출력된 특성 행렬을 사용해 전역 표준 편차와 지역 표준 편차를 계산한 뒤 평균값을 도출한다. 만약 m 에포크까지 학습했다면, m개의 전역 표준 편차 및 지역 표준 편차가 산출된다. 전역 표준 편차의 경우 P-value가 0.7인 값을 베

타로 선정하고, 지역 표준 편차의 경우 P-value가 0.4와 0.7인 값을 델타 원, 델타 투로 선정한다. 이러한 방법은 기본 모델과 최적화 모델을 모두 학습해야 하므로 많은 시간이 소요된다.

본 논문에서는 이를 개선하기 위해 동적인 임곗값을 도입하였다. 베타와 델타값을 이전 에포크에서 출력된 특성 행렬을 사용해 지역 표준 편차 및 전역 표준 편차의 P-value를 기반으로 설정하였으며, 최적화 학습과 동시에 베타와 델타를 에포크마다 업데이트하였다. 이로써 기본 모델을 수렴할 때까지 학습해야 하는 GL-GAN의 방법과 달리 단 1 에포크만 학습해도 최적화 학습이 가능해져 모델 학습 시간이 대폭 감소하였다.

또한, 최적화를 적용하지 않은 기본 모델이 생성한 이미지와 최적화를 적용한 모델이 생성한 이미지 사이에는 차이가 존재한다. 즉, GL-GAN의 경우 기본 모델의 여러 에포크에서 여러 개의 표본을 사용해 임곗값을 설정했지만, 이는 최적화 학습 과정에서 생성한 이미지와는 차이가 존재한다. 그러나 본 논문에서 제안한 방법은 가장 유사하게 이미지를 생성하는 이전 에포크에서 계산된 임곗값을 사용하기 때문에 GL-GAN보다 더 근사적인 값을 사용할 수 있다.

Ⅳ. 손실 함수

1. 영어-영어 텍스트 스타일 전이 학습 손실 함수

영어-영어 텍스트 스타일 전이 학습의 손실값을 계산하는 방법은 TSB[1]와 동일한 방식을 채택하며, 종합적인 손실은 아래와 같이 계산된다.

$$L = L_D + L_S + \lambda_3 L_T + \lambda_4 L_R + \lambda_5 L_{rec} + \lambda_6 L_{cyc}$$

[수식 4-1] 영어-영어 스타일 전이 학습 합산 손실

(1) 판별자 기반의 적대적 손실(L_D), (2) 스타일 손실(L_S), (3) 글꼴 손실(L_T), (4) 콘텐츠 손실(L_R), (5) 재건 손실(L_{rec})과 순환 재건 손실(L_{cyc})의 합산으로 계산된다.

(1) 판별자 기반의 적대적 손실(L_D, adversarial loss)은 생성자가 현실적인 이미지를 생성하도록 유도하기 위해 사용된다. PatchGAN[24] 기반의 판별자에 의해 계산되며, 그 방법은 아래와 같다.

$$L_D = \mathbb{E}_{x \sim P_{data}(x)}[\log(D(x|y)) + \mathbb{E}_{x \sim p_z(z)}[\log(1 - D(G(z|y)))]$$

[수식 4-2] 판별자 기반의 적대적 손실

(2) 스타일 손실(L_S)은 재건 이미지가 스타일 이미지와 유사한 텍스트 스타일을 표현할 수 있도록 돕기 위해 사용된다. 재건 이미지와 스타일 이미지 간의 전체적인 구조적 특징의 유사성을 측정하는 지각 손실(L_{per}, perceptual loss)과 질감

유사성을 측정하는 질감 손실(L_{tex}, texture loss)의 합산 결과로 계산된다. 지각 손실과 질감 손실은 사전 학습된 VGG16을 사용해 계산되며, 계산 방법은 아래와 같다.

$$L_S = \lambda_1 L_{per} + \lambda_2 L_{tex}$$

[수식 4-3] 스타일 손실

$$L_{per} = \mathbb{E}[\sum_i \frac{1}{M_i} \| \phi_i(I_s) - \phi_i(O_{c1}) \|_1]$$

[수식 4-4] 지각 손실

$$L_{tex} = \mathbb{E}_i[\| G_i^{\phi}(I_s) - G_i^{\phi}(O_{c1}) \|_1]$$

[수식 4-5] 질감 손실

(3) 글꼴 손실(L_T)은 재건 이미지가 스타일 이미지와 동일한 글꼴의 텍스트를 생성할 수 있도록 유도하기 위해 사용된다. 이를 위해 글꼴 분류기의 마지막 분류 레이어를 제거하여 재건 이미지와 스타일 이미지의 특징값을 추출한 후, L1 손실을 계산한다.

$$L_T = \mathbb{E}_i[\| \psi(I_s) - \psi(O_{c1}) \|_1]$$

[수식 4-5] 글꼴 손실

(4) 콘텐츠 손실(L_R)은 재건 이미지와 스타일 적용 이미지가 영어 텍스트 형태를 유지할 수 있도록 촉진하기 위해 사용된다. 콘텐츠 손실을 계산하기 위해 텍스트 인식기가 사용된다. 텍스트 인식기가 재건 이미지와 스타일 적용 이미지의 영어 인식 결과 ($c1'$, $c2'$)를 제공하면, 이를 원본 텍스트와 신규 텍스트인 ($c1$,

$c2$)와 비교하여 크로스 엔트로피를 계산한 뒤 평균을 취한다. 계산 식은 아래와 같이 계산되며, N은 단어의 길이, C는 클래스의 수, y는 클래스를 원-핫 인코딩 한 값, p는 예측한 확률을 나타낸다.

$$CE = -\frac{1}{N}\Sigma_{i=1}^{N}\Sigma_{c=1}^{C} y_{i,c}\log(p_{i,c})$$

[수식 4-6] 크로스 엔트로피

(5) 재건 손실(L_{rec}, reconstruction loss)과 순환 재건 손실(L_{cyc}, cyclic reconstruction loss)은 생성된 이미지가 스타일 이미지와 유사하도록 유도하기 위해 사용되는 손실이다. 재건 손실은 재건 이미지와 스타일 이미지의 간의 L1 손실값을 계산하여 사용된다. 더 효과적인 스타일 전이 학습을 위해 재건 이미지를 생성기에 입력하여 스타일 전이를 실행하고, 이로부터 순환 재건 이미지를 생성한다. 순환 재건 손실은 순환 재건 이미지와 재건 이미지 간의 L1 손실값을 계산하여 사용된다.

2. 한국어-영어 텍스트 스타일 전이 학습 손실 함수

한국어-영어 텍스트 스타일 전이 학습은 영어-영어 텍스트 스타일 전이 학습과는 달리 이중 언어 인식기, 이중 언어 글꼴 분류기, 그리고 동적인 전역적 및 지역적 이중 수준 최적화 모듈을 사용하여 평가된다. 따라서 손실 함수 일부가 영어-영어 스타일 전이 학습과 다르게 구성되며, 아래와 같이 계산된다.

$$L = L_D + L_{BS} + \lambda_3 L_{BT} + \lambda_4 L_{BR} + \lambda_5 L_{rec} + \lambda_6 L_{cyc}$$

[수식 4-7] 한국어-영어 스타일 전이 학습 합산 손실 함수

영어-영어 스타일 전이 학습 모델의 손실 계산과 다른 부분에 집중하여 기술하였다.

1) 이중 언어 스타일 손실(L_{BS})

이중 언어 스타일 손실(L_{BS})은 동적인 전역적 및 지역적 이중 수준 최적화 모듈에 의해 결정된 최적화 방법에 따라 손실 함수가 달라진다. 이 손실은 전역 최적화를 진행할 때는 영어-영어 텍스트 스타일 전이 학습과 동일하게 계산되고, 지역 최적화를 진행하는 경우는 다음과 같이 계산된다.

$$L_{BS} = \lambda_1 L_{per-local} + \lambda_2 L_{tex-local}$$

[수식 4-8] 이중 언어 스타일 손실

$$L_{per-local} = (\mathbb{E}[\sum_i \frac{1}{M_i} \| \phi_i(I_s) - \phi_i(O_{c1}) \|_1] + \mathbb{E}[\sum_i \frac{1}{m_i} \| \phi_i(I_s) - \phi_i(O_{c1} \odot M_{c1}) \|_1])/2$$

[수식 4-9] 지역 지각 손실

$$L_{tex-local} = (\mathbb{E}_i[\| G_i^{\phi}(I_s) - G_i^{\phi}(O_{c1}) \|_1] + \mathbb{E}_i[\| G_i^{\phi}(I_s) - G_i^{\phi}(O_{c1} \odot M_{c1}) \|_1])/2$$

[수식 4-10] 지역 질감 손실

지역 최적화를 진행하면 이중 언어 스타일 손실값은 두 개의 이미지에 대한 스타일 손실값의 평균으로 계산된다. 지역 최적화를 할 때 재건 이미지에 저품질이 아닌 지역에 마스크가 적용된 결과를 얻을 수 있다. 마스크가 적용되지 않은 이미지와 마스크가 적용된 이미지 두 장을 사용하여 각각 스타일 손실값을 계산한 뒤 평균을 취한다. 평균값을 사용하는 이유는 마스크가 적용된 이미지만을 사

용해 스타일 손실값을 계산하게 되면 이미지의 일부분 스타일이 전체 이미지의 스타일로 인식되어 스타일 전이가 제대로 이루어지지 않을 수 있기 때문이다. 따라서 전체의 이미지 스타일은 유지하면서 추가로 이미지의 저품질 지역의 스타일을 중점적으로 고려하도록 설계되었다.

2) 이중 언어 글꼴 손실(L_{BT})

영어-영어 텍스트 스타일 전이 학습에서는 스타일 이미지와 재건 이미지를 사용해 글꼴 손실을 계산하였다. 영어-영어 텍스트 스타일 전이 학습은 재건 이미지와 스타일 적용 이미지가 모두 영어로 이루어진 동일 언어 간 스타일 전이 학습을 수행하기 때문에, 재건 이미지만으로도 충분히 동일한 글꼴을 만들어 낼 수 있었다. 그러나 한국어-영어 텍스트 스타일 전이 학습에서는 재건 이미지는 한국어로, 스타일 적용 이미지는 영어로 이루어진 이중 언어 간 텍스트 스타일 전이 학습을 수행한다. 이에 따라 재건 이미지만을 이용한 손실값 계산은 한국어에 대한 글꼴 손실만을 측정하게 된다. 이는 영어의 글꼴 생성에 대해 충분한 학습이 이루어지지 않아 스타일 이미지와 동일한 글꼴의 영어를 생성하지 못할 수 있음을 의미한다. 그러므로 스타일 적용 이미지도 글꼴 손실 계산에 포함하여 모델이 영어의 글꼴 생성에 대해서도 효과적으로 학습할 수 있도록 변경하였다. 변경된 이중 언어 글꼴 손실 계산 수식은 아래와 같다.

$$L_{BT} = (\mathbb{E}_i[\,\|\,\psi(I_s) - \psi(O_{c1})\,\|_1] + \mathbb{E}_i[\,\|\,\psi(I_s) - \psi(O_{c2})\,\|_1])/2$$

[수식 4-10] 지역 질감 손실

영어-영어 텍스트 스타일 전이 학습 방법과 동일하게 이중 언어 글꼴 분류기의 마지막 분류 레이어를 제거하여 이중 언어 글꼴 손실 계산에 사용하였다. 스

타일 이미지와 재건 이미지의 글꼴 특징값 차이에 대한 L1 손실값과 스타일 이미지와 스타일 적용 이미지의 글꼴 특징값 차이에 대한 L1 손실값의 평균값을 채택하였다.

3) 이중 언어 콘텐츠 손실(L_{BR})

영어-영어 텍스트 스타일 전이 학습에서는 재건 이미지와 스타일 적용 이미지가 모두 영어로 구성되어 있어서 영어 인식기만을 이용해 얻은 인식 결과를 기반으로 콘텐츠 손실을 계산하였다. 그러나 한국어-영어 텍스트 스타일 전이 학습에서는 재건 이미지는 한국어로 작성되고, 스타일 적용 이미지는 영어로 작성되었기 때문에 한국어 인식기와 영어 인식기를 모두 포함한 이중 언어 인식기를 사용해야 한다. 이중 언어 콘텐츠 손실은 이중 언어 인식기를 사용해 얻은 재건 이미지와 스타일 적용 이미지의 인식 결과를 기반으로 계산되며, 계산 방법은 영어-영어 텍스트 스타일 전이 학습과 동일하다.

V. 실험 및 결과

1. 데이터 세트

[그림 5-1] IMGUR5K 데이터 세트 예시

영어-영어 텍스트 스타일 전이 학습에서는 IMGUR5K[31] 데이터 세트를 활용하였다. IMGUR5K는 손 글씨로 작성된 영어를 포함한 이미지와 해당 이미지 내의 단어들을 감싸고 있는 바운딩 박스의 좌표와 크기, 그리고 바운딩 박스 안에 있는 텍스트 정보를 포함하고 있다. 주어진 바운딩 박스의 크기에 맞게 이미지를

31) IMGUR5K-Handwriting-Dataset, Github repository, https://github.com/facebookresearch/IMGUR5K-Handwriting-Dataset

잘라내고, 잘라낸 이미지 안에 있는 텍스트로 라벨링 하여 전처리 작업을 수행하였다. 이 과정을 통해 총 228,264장의 (이미지 - 영어 텍스트) 쌍을 얻을 수 있었고, 8:1:1의 비율로 나누어 학습 데이터 182,092장, 검증 데이터 22,453장, 테스트 데이터 23,819장으로 구성하여 학습에 활용하였다.

[그림 5-2] AI 허브 '야외 실제 촬영 한글 이미지' 데이터 세트 예시

한국어-영어 스타일 전이 학습에서는 AI 허브[27]의 '야외 실제 촬영 한글 이미지' 데이터 세트를 활용하였다. 이 데이터 세트에는 실내외에서 촬영된 다양한 한글을 포함한 이미지들로 구성되어 있으며, 각 이미지 내의 단어들을 감싸고 있는 바운딩 박스의 좌표와 크기, 그리고 바운딩 박스 안에 있는 텍스트 정보가 주어진다. 이 중에서도 가로형 간판 데이터 151,042장을 선별하여 사용하였다. IMGUR5K 데이터 세트와 동일한 방법을 사용해 전처리하였다. 이를 통해 총

202,113장의 (이미지 – 한국어 텍스트) 쌍을 얻었고, 이를 8:1:1의 비율로 나누어 학습 데이터 161,691장, 검증 데이터 20,211장, 테스트 데이터 20,211장으로 구성하여 학습에 활용하였다.

2. 실험 세부 사항

장면 텍스트 스타일 전이 모델은 TSB[1)]를 기반으로 하되, 마스크를 제외한 형태로 설계되었다. TSB에서의 마스크는 스타일 전이 모델이 생성한 이미지에서 추출한 전경을 나타내며, 올바른 형태의 글자를 생성했는지 평가하는 데 사용된다. 전경을 사용하지 않더라도, 스타일 전이 모델이 생성한 이미지 자체만 사용해서 학습해도 올바른 형태의 글자를 생성할 수 있다고 판단했다.

생성기는 StyleGAN[2)]을 사용하며, 점진적 성장(progressive growing)과 노이즈 입력 기술은 채택하지 않았다. 점진적 성장은 학습을 진행하면서 생성자와 판별자의 레이어 개수를 점진적으로 확장하여 이미지의 해상도를 향상시키는 기술이다. 이러한 기술은 이미지의 품질을 향상시키지만, 모델의 크기와 계산 리소스가 증가하는 단점이 있다. 점진적 성장 기술은 PGGAN[32)]을 기반으로 하며, 인물 얼굴을 생성할 때 피부 질감, 눈, 입술, 머리카락 등의 세부적인 부분을 표현할 수 있게 해준다. 그러나 텍스트 생성은 얼굴 생성에 비해 세부 사항을 정교하게 표현할 필요성이 상대적으로 낮으므로, 이 기술을 사용하지 않고 모델의 복잡성을 줄이는 것이 더 적절하다고 판단하였다. 노이즈 입력 기술 또한 동일한 이유로 사용되지 않았다.

사전 학습된 네트워크인 글꼴 분류기, 텍스트 인식기, 이중 언어 글꼴 분류기, 이중 언어 인식기, 그리고 스타일 추출기는 평가에만 사용하기 위해 매개변수를 고정하여 업데이트하지 않고 그대로 사용하였다.

---

32) Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196.

구체적인 시스템 환경과 하이퍼 파라미터는 [표 5-1]과 [표 5-2]에서 확인할 수 있다.

[표 5-1] 시스템 환경

Computing Environment	Workstation
Operating System	Ubuntu 20.04.6
GPU	NVIDIA GeForce RTX 3080 Ti X4
Memory	12GB X4
Framework	PyTorch

[표 5-2] 하이퍼 파라미터

Parameters	first-stage training	second-stage training
Image size	64x196	
Font name	Verily Serif Mono	Verily Serif Mono(English), Malgun Gothic(Korean)
Batch size	16	
Optimizer	AdamW	
Learning rate	1e-3	1e-4
λ_1	1.0	
λ_2	500.0	
λ_3	1.0	
λ_4	1.0	
λ_5	10.0	
λ_6	1.0	

3. 평가 지표

한국어-영어 스타일 전이 시스템의 성능 평가를 위해 5가지 지표를 활용하였다. 글자 생성 능력을 평가하기 위해서는 텍스트 인식 정확도를 측정하였다. 스타일 전이 능력을 판단하기 위해서는 이미지 생성에서 보편적으로 사용되는 지표를

도입하여 분석하였다. 픽셀 수준의 차이에 중점을 두고 계산되는 평균 제곱 오차(MSE; mean Square Error), 최대 신호 대 잡음 비율(PSNR; peak Signal-to-Noise Ratio), 구조적 유사성 지수(SSIM; Structural Similarity Index Map)를 선택하였고, 더불어 시각적 품질을 반영하는 프레셰 인셉션 거리(FID; Fréchet Inception Distance)[33]를 선택하였다. 이러한 다양한 지표를 통해 스타일 이미지와 생성된 이미지 간의 차이를 정량적으로 평가하였다. MSE, FID는 낮을수록, 인식 정확도, PSNR, SSIM은 높을수록 성능이 우수하다고 평가된다.

(1) 텍스트 인식 정확도는 사전 학습된 텍스트 인식기를 사용하여 스타일 전이 모델이 생성한 이미지에 대한 광학 문자 인식(OCR)을 수행하고, 이를 생성하고자 했던 실제 텍스트와의 일치 여부를 비교하였다.

$$Acc = \frac{1}{\#test}(\sum_{i}\mathbb{I}(R(O_{c_i}) == c_i)$$

[수식 5-1] 텍스트 인식 정확도

(2) MSE는 원본 이미지와 생성된 이미지의 간의 각 픽셀 차이를 제곱한 후 평균한 값으로, 값이 작을수록 유사하다고 판단된다.

$$MSE = \frac{1}{n}\Sigma_{i=1}^{n}(y_i - \hat{y_i})^2$$

[수식 5-2] MSE

(3) PSNR은 이미지 품질을 평가하는 지표 중 하나로, 원본 이미지와 생성된 이미지 사이의 차이를 신호와 잡음 간의 비율로 표현하여 이미지 사이의 품질 손

33) Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30.

실을 측정하는 데 사용된다. 값이 클수록 두 이미지 간의 차이가 적다는 것을 의미하며, 이미지의 품질이 높다고 판단된다.

$$PSNR = 10 \times log_{10}\left(\frac{H \times W \times MAX_I^2}{\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}[R(i,j) - G(i,j)]}\right)$$

[수식 5-3] PSNR

(4) SSIM은 원본 이미지와 생성된 이미지 간의 유사성을 측정하여 평가한다. 이미지의 유사성은 구조, 밝기, 대비 등의 요소를 고려하여 측정하며, 값이 1에 가까울수록 두 이미지의 유사성이 높다고 여겨진다.

$$SSIM = \frac{(2\mu_R\mu_G + C_1)(2\sigma_{RG} + C_2)}{(\mu_R^2 + \mu_G^2 + C_1)(\sigma_R^2 + \sigma_G^2 + C_2)}$$

[수식 5-4] SSIM

(5) FID는 원본 이미지와 생성된 이미지의 간의 차이를 측정하여 평가하는 데 활용된다. 값이 작을수록 두 이미지 간의 차이가 작다는 것을 나타낸다.

$$FID = \| \mu_R - \mu_G \|^2 + Tr(\Sigma_R + \Sigma_G - 2(\Sigma_R\Sigma_G)^{1/2})$$

[수식 5-5] FID

4. 모델 비교

1) 텍스트 스타일 전이 모델 비교

제안하는 시스템의 우수성을 기존 모델들과의 비교를 통해 입증하고자 하였다. [표 5-3]에는 기존의 텍스트 스타일 전이 모델인 TSB[1)]와 SRNet[23)]의 한국어-영어 텍스트 스타일 전이 성능을 정리하였다.

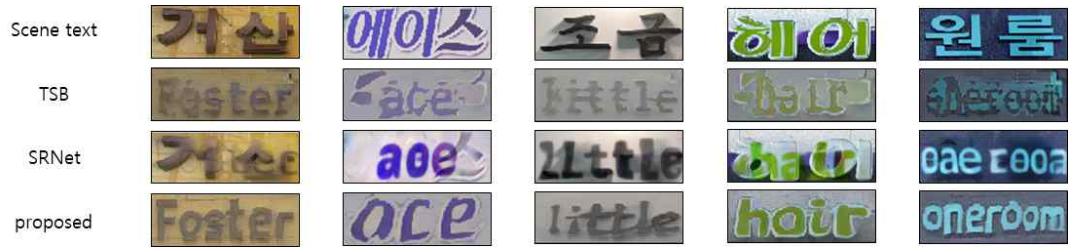
[표 5-3] 텍스트 스타일 전이 모델 성능 비교

Method	MSE↓	PSNR↑	SSIM↑	FID↓	Acc↑
TSB	5587.2628	27.9123	0.2413	162.4936	0.0929
SRNet	**4445.2828**	28.1414	**0.4951**	67.0277	0.1406
proposed	5990.2674	**28.1642**	0.2715	**65.7774**	**0.7408**

- TSB는 StyleGAN 기반의 영어-영어 텍스트 스타일 전이 모델이며, 원-샷 학습 환경에서 단일 장면 텍스트를 사용해 스타일 전이가 가능하다.
- SRNet : SRNet은 영어-영어 텍스트 스타일 전이 모델이며, 텍스트 변환 모듈(Text convention module), 배경 인페인팅 모듈(Background inpainting module), 합성 모듈(fusion module)로 구성된다. 여러 가지 글꼴, 색상, 각도를 가진 텍스트를 다양한 배경 데이터에 합성하여 학습 데이터로 사용한다.

대부분의 성능 지표에서 제안하는 시스템이 우수한 점수를 기록하였으나, 일부 지표에서는 SRNet이 더 높은 점수를 기록하였다. SRNet은 MSE와 SSIM에서 높은 성능을 보였으나, 영어 인식 정확도에서 저조한 결과를 나타냈다. 이는 SRNet의 글자 생성 능력이 부족하다는 것을 시사한다. SRNet은 합성 데이터를 학습에 사용하기 때문에 배경 분리가 어려운 장면 텍스트에서는 스타일 전이에

한계가 있었다. 또한, TSB는 모든 평가 지표에서 낮은 점수를 기록하였다. TSB는 학습 데이터 자체를 정답으로 사용하여 약한 자기 지도 학습을 수행하는 방식으로 학습되었는데, 이 과정에서 영어가 포함되지 않은 학습 데이터로 인해 알파벳에 대한 적절한 학습이 이루어지지 않았기 때문이다.



[그림 5-3] 텍스트 스타일 전이 모델 성능 비교 예시

[그림 5-3]에서 각 모델의 텍스트 스타일 전이 결과를 시각적으로 확인할 수 있다. 첫 번째 행은 다양한 배경과 글꼴을 가진 한국어 장면 텍스트를 나타내고 있으며, 두 번째 행부터는 TSB, SRNet, 제안하는 시스템의 스타일 전이 결과를 순서대로 보여준다. TSB는 글자를 읽기 어려운 수준으로 생성하였고, 일부 경우에서는 배경 색상과 글자의 색상을 제대로 표현하지 못했다. SRNet은 장면 텍스트와 번역된 영어가 겹쳐있는 듯한 이미지를 생성하였는데, 이는 합성 데이터를 사용해 학습했기 때문에 발생한 것으로, MSE와 SSIM에서 좋은 성능을 보인 원인으로 분석된다. 반면, 제안하는 시스템은 장면 텍스트와 유사한 스타일로 영어 글자를 제대로 생성하여 가장 우수한 결과를 나타냈다.

2) TSB 비교

제안하는 시스템은 TSB를 기반으로 한 모델이므로, TSB와의 기능 차이를 단계별로 비교하여 이중 언어 간 스타일 전이에 효과적인 모델임을 입증하고자 하였다.

[표 5-4] TSB 기능 확장을 통한 성능 비교

Method	Training time(hrs)	MSE↓	PSNR↑	SSIM↑	FID↓	Acc↑
TSB	68.375	**5587.2628**	27.9123	0.2413	162.4936	0.0929
+ Bilingual recognizer	**34.88**	5992.6329	27.9440	0.2179	157.2062	0.7365
+ Bilingual typeface classifier	90.08	6040.9558	27.9522	0.2155	155.6558	0.7154
+ Two-stage training	61.14(34.52+26.62)	6075.5438	28.1009	0.2561	75.4160	0.7383
+ Dynamic global and local bilevel optimization(**proposed**)	85.75(34.52+51.23)	5990.2674	**28.1642**	**0.2715**	**65.7774**	**0.7408**

[표 5-4]는 TSB에 단계별로 추가된 기능에 대한 한국어-영어 스타일 전이 결과를 나타낸다. 이중 언어 인식기, 이중 언어 글꼴 분류기, 2단계 학습 그리고 동적인 전역적 및 지역적 이중 수준 최적화를 차례로 적용한 결과이다. 2단계 학습을 적용하면 학습 시간이 34.52시간 추가되었는데, 이는 영어-영어 텍스트 스타일 전이 학습을 1차로 수행한 시간이다. 결과적으로, 모든 기능을 적용한 마지막 단계에서 스타일 전이 성능이 가장 우수하게 측정되었으며, 영어 인식 정확도 역시 최고 수준을 기록했다. 학습 시간 측면에서는 마지막 단계에서 TSB와 비교하여 약 1.25배의 증가를 보였지만, FID 점수는 2배 이상 향상되었다. MSE는 TSB가 가장 좋은 결과를 보였지만, 다른 모든 지표에서는 좋지 않은 결과를 보여 우수한 모델로 평가하기에는 어려워 보인다.

[그림 5-4] TSB 기능 확장을 통한 성능 비교 예시

TSB의 기능 확장에 따른 실험의 결과는 [그림 5-4]에서 시각적으로 확인할 수 있다. 첫 번째 행은 한국어 장면 텍스트를 나타내고, 두 번째 행부터 마지막 행까지는 TSB에 추가한 기능에 따른 한국어-영어 스타일 전이 결과를 나타낸다. TSB는 글자를 읽을 수 있는 수준으로 생성하지 못했다. 이중 언어 인식기를 추가하였을 때 글자의 형태가 갖춰졌고, 이중 언어 글꼴 분류기를 추가하였을 때 글자의 형태가 뚜렷해졌다. 2단계 학습을 진행하면서 글자의 두께, 각도, 색상이 더욱 명확하게 표현되었지만, 이미지의 일부분에 아티팩트가 발생했다. 동적인 전역적 및 지역적 이중 수준 최적화까지 적용한 결과, 아티팩트 없이 장면 텍스트와 가장 유사한 스타일을 가진 이미지가 생성되었다.

3) 데이터 세트 생성 모델 비교

Xie, Yangchen, et al.[10]은 TSB를 기반으로 한 구조에 '통합된 어텐션 모듈(Integrated Attention module)'을 도입하여 장면 텍스트를 사용한 이중 언어 간(영어-카자흐어, 중국어-한국어) 스타일 전이를 시도하였다. Xie, Yangchen, et al.이 제안하는 모델과 한국어-영어 텍스트 스타일 전이 성능을 비교하려 했으나, 코드가 공개되지 않았다. 따라서 제안하는 시스템으로 중국어-한국어 텍스트 스

타일 전이 학습을 수행한 후 그 결과를 비교하였다. 학습에는 BCTR[34]의 중국어 장면 텍스트 152,282장을 8:1:1의 비율로 나누어 학습 데이터 127,734장, 검증 데이터 15,766장, 테스트 데이터 15,782장으로 구성하여 학습에 사용하였으며, 자세한 내용은 6장에서 다루었다.

[표 5-5] 텍스트 인식 모델 성능 비교

Training data	Korean(Acc)	Korean(ED)
SRNet*	-	-
TUNIT*	17.539	0.444
DGFont*	39.237	0.625
Xie, Yangchen, et al.*	**64.837**	**0.806**
proposed	60.8524	0.786

- TUNIT[35] : TUNIT은 동물 이미지의 스타일과 콘텐츠를 분리하고, 콘텐츠의 구조적 정보를 유지하면서 다양한 스타일을 적용하는 비지도 학습 방식의 이미지-이미지 번역 모델이다.
- DGFont[36] : DGFont는 특징 변형 스킵 연결(FDSC:feature deformation skip connection)을 도입하여 텍스트를 다른 글꼴로 변형하는 비지도 학습 방식의 글꼴 생성 모델이다.

Xie, Yangchen, et al.은 한국어 인식 모델을 위한 한국어 데이터 세트 생성을

34) Yu, H., Chen, J., Li, B., Ma, J., Guan, M., Xu, X., ... & Xue, X. (2021). Benchmarking chinese text recognition: Datasets, baselines, and an empirical study. arXiv preprint arXiv:2112.15093.

35) Baek, K., Choi, Y., Uh, Y., Yoo, J., & Shim, H. (2021). Rethinking the truly unsupervised image-to-image translation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 14154-14163).

36) Xie, Y., Chen, X., Sun, L., & Lu, Y. (2021). Dg-font: Deformable generative networks for unsupervised font generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5130-5140).

목표로 중국어-한국어 텍스트 스타일 전이를 수행하였다. 따라서 다른 텍스트 스타일 전이 모델들이 생성한 데이터 세트에 대한 비교 실험을 진행하였다. 그 실험 방법은 다음과 같다. 각 모델의 중국어-한국어 텍스트 스타일 전이를 통해 생성한 한국어 데이터 세트를 사용하여 텍스트 인식 모델인 TRBA[11]를 학습한다. 그 후, 학습된 한국어 인식 모델을 ICDAR2019-MLT[37]의 크롭 된 한국어 이미지 4,060장을 평가 데이터 세트로 사용하여 한국어 인식 정확도와 한국어 ED(edit distance)를 측정하여 평가하였다. 그 결과는 [표 5-5]에 정리되었으며, * 표시된 부분들은 Xie, Yangchen, et al.의 논문에서 발췌한 정보이므로 정확한 비교는 어려우나 참고용으로 작성되었다.

한국어 인식 정확도와 한국어 ED 측면에서 Xie, Yangchen, et al.가 가장 우수한 성능을 보였으며, 그다음 미세한 차이로 제안하는 시스템이 우수하였다. 이러한 결과는 Xie, Yangchen, et al.과 본 연구의 목표 차이에 의해 비롯된 것으로 판단된다. 본 연구는 주로 스타일 전이에 초점을 맞추었지만, Xie, Yangchen, et al.은 한국어 인식 정확도를 높이기 위해 글자 생성에 주력하였다. 그럼에도 불구하고, 제안하는 모델이 Xie, Yangchen, et al.의 모델과 미세한 차이를 기록한 것을 고려하면 우수한 모델임을 입증할 수 있다.

37) Nayef, N., Patel, Y., Busta, M., Chowdhury, P. N., Karatzas, D., Khlif, W., ... & Ogier, J. M. (2019, September). ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition—RRC-MLT-2019. In 2019 International conference on document analysis and recognition (ICDAR) (pp. 1582-1587). IEEE.

[그림 5-5] 중국어-한국어 텍스트 스타일 전이 모델 성능 비교 예시

Xie, Yangchen, et al.이 제공한 TUNIT, DG-Font와의 중국어-한국어 텍스트 스타일 전이 결과에 제안하는 시스템의 결과를 추가하여 비교한 결과는 [그림 5-5]에서 시각적으로 확인할 수 있다. 첫 번째 행은 중국어 장면 텍스트를 나타내고, 두 번째 행부터 마지막 행까지는 TUNIT, DG-Font, Xie, Yangchen, et al. 그리고 제안하는 시스템의 중국어-한국어 텍스트 스타일 전이 결과를 나타낸다. *로 표시된 모델의 결과는 Xie, Yangchen, et al.의 논문에서 제공한 이미지를 인용하였다. TUNIT의 경우 글자의 형태가 왜곡되었고, DG-Font는 색상이나 글꼴을 잘 표현하지 못했다. Xie, Yangchen, et al.은 스타일과 글자의 형태를 잘 표현했지만, 글자의 일부 뒷부분에 약간의 번짐이 나타났다. 제안하는 시스템은 스타일을 잘 표현하면서 정교한 글자를 생성하였다.

4) 최적화 모델 비교

본 연구에서 제안하는 동적 최적화의 성능을 평가를 위해 Zhang, Yucun, et al.[38], GL-GAN[13]과 비교하였다. Zhang, Yucun, et al.과 GL-GAN의 최적화 방법론을 장면 텍스트 스타일 전이 모델에 적용한 결과를 본 연구에서 제안하는 동적인 전역적 및 지역적 이중 수준 최적화와 비교하여 [표 5-6]에 정리하였다.

[표 5-6] 최적화 모델 성능 비교

Method	Training time(hrs)	MSE↓	PSNR↑	SSIM↑	FID↓	Acc↑
Zhang, Yucun, et al.	369.87	**5987.7665**	28.0932	0.2570	77.6269	0.7517
GL-GAN	164.35	6000.8105	28.0906	0.2565	76.0199	**0.7558**
proposed	**51.23**	5990.2674	**28.1642**	**0.2715**	**65.7774**	0.7408

- Zhang, Yucun, et al. : Zhang, Yucun, et al.은 이미지 모션 디블러링(image motion deblurring)의 성능 향상을 위해 이중 판별자(dual discriminator)를 사용한 최적화를 제안하였다. 이중 판별자는 생성된 이미지의 전체를 최적화하는 전역 판별자(Global discriminator)와 생성된 이미지에서 무작위로 크롭한 5개의 패치를 최적화하는 지역 판별자(Local discriminator)를 사용한다.
- GL-GAN : GL-GAN는 이미지 생성 모델의 품질 향상을 위해 적응적인 전역적 및 지역적 이중 수준 최적화(Adaptive global and local bilevel optimization)를 제안하였다. 전역적 최적화는 전체 이미지를 사용해 모델을 최적화하고, 지역적 최적화는 이미지의 저품질 지역만으로 모델을 최적화한다. 이미지의 품질에 따라서 두 가지의 최적화 방법 중 하나가 적응적으로 선택된다.

38) Zhang, Y., Li, T., Li, Q., Fu, X., & Kong, T. (2023). Image motion deblurring via attention generative adversarial network. Computers & Graphics, 111, 122-132.

대부분의 성능 지표에서 제안하는 최적화 방법론이 가장 우수한 성능을 기록하였다. Zhang, Yucun, et al.은 MSE 점수에서는 가장 우수한 성능을 보였으나, 학습 시간에 가장 많은 시간이 소요되었다. 이는 Zhang, Yucun, et al.이 두 개의 판별자를 사용하며, 각각 독립적인 파라미터를 가진 데서 오는 연산량 증가로 인해 더 많은 학습 시간을 필요로 했기 때문이다. GL-GAN는 Zhang, Yucun, et al.과 학습 시간을 제외하고는 유사한 성능을 보였다. 제안하는 방법론은 GL-GAN을 기반으로 하되, 동적인 임곗값 설정을 도입하여 학습 시간을 3분의 1로 단축하였으며, PSNR, SSIM, FID 점수 모두 향상시켰다. 영어 인식 정확도는 GL-GAN이 더 높았으나, 이는 미미한 차이에 불과하였다.

[그림 5-6] 최적화 모델 성능 비교 예시

최적화 모델 비교 결과는 [그림 5-6]에서 시각적으로 확인할 수 있다. 첫 번째 행은 한국어 장면 텍스트를 나타내고, 두 번째 행부터 마지막 행까지는 Zhang, Yucun, et al., GL-GAN, 제안하는 방법론의 스타일 전이 결과를 나타낸다. Zhang, Yucun, et al.과 GL-GAN은 일부 아티팩트가 여전히 존재하는 반면, 제안하는 방법론은 아티팩트가 없는 이미지를 생성한 것을 보아 제안하는 방법론이 아티팩트 제거 성능에서 가장 우수하였다.

5) GL-GAN 비교

[표 5-7] GL-GAN과 다양한 언어 데이터 세트를 활용한 성능 비교

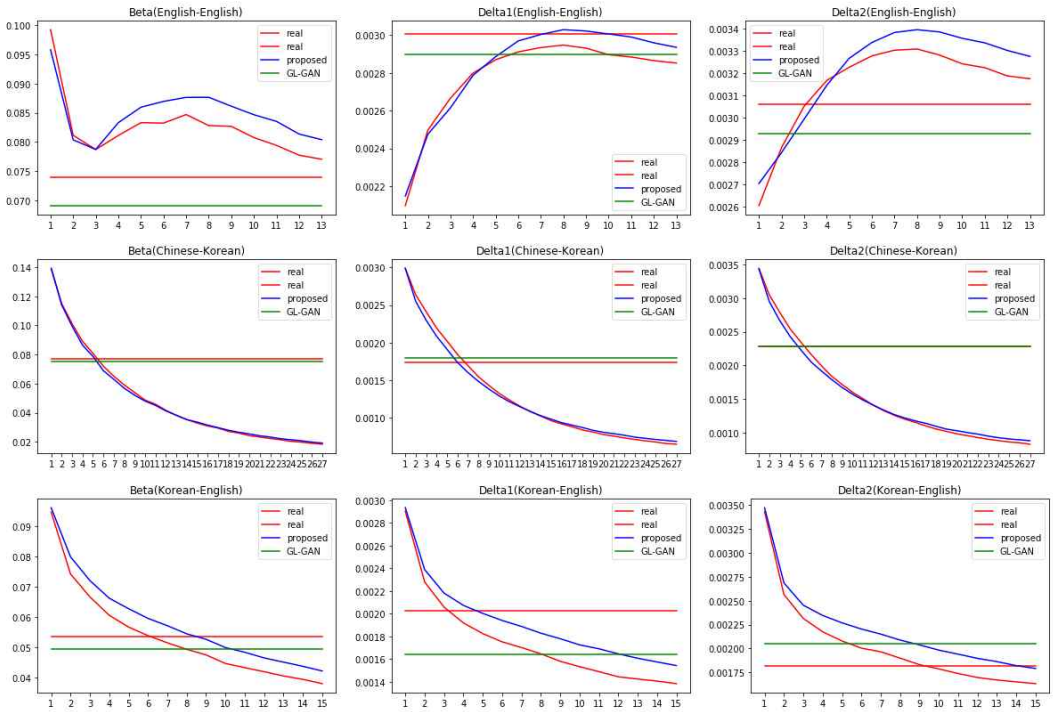
Language	Method	Training time(hrs)	MSE↓	PSNR↑	SSIM↑	FID↓	Acc↑
English-English	GL-GAN	69.54	4356.7419	**28.5769**	0.3708	74.9804	0.6798
	proposed	**36.50**	**4232.0857**	28.5767	**0.3755**	**73.9807**	**0.6895**
Chinese-Korean	GL-GAN	47.55	**3665.8668**	28.0645	**0.4478**	85.7335	0.8000
	proposed	**20.4475**	3669.1174	**28.0658**	0.4449	**84.9907**	**0.8148**
Korean-English	GL-GAN	164.35	6000.8105	28.0906	0.2565	76.01996	**0.7558**
	proposed	**51.23**	**5990.2674**	**28.1642**	**0.2715**	**65.7774**	0.7408

제안하는 동적 최적화 방법론의 일반화 성능과 범용성을 평가하기 위해 추가로 다양한 언어 데이터 세트를 사용해 GL-GAN과 성능을 비교하였다. 영어 장면 텍스트 데이터 세트를 사용한 영어-영어 텍스트 스타일 전이 학습과 중국어 장면 텍스트 데이터 세트를 사용한 중국어-한국어 텍스트 스타일 전이 학습 결과를 [표 5-7]에 정리하였다. GL-GAN의 방법론은 최적화를 적용하지 않은 기본 모델을 수렴할 때까지 학습해야 하지만, 제안하는 방법론은 기본 모델을 단 1 에포크만 학습해도 최적화할 수 있어 학습 시간이 적게 소요되었다. 또한, 대부분 지표에서 우수한 성능을 보였으며, 일부 지표에서 좋지 않은 성능을 보였으나 그 차이는 미세하였다.

[표 5-8] 임곗값 차이 비교

Language	Method	Beta	Delta 1	Delta 2
English-English	GL-GAN	0.004863	0.00011	0.000133
	proposed	**0.003009**	**0.000065**	**0.000078**
Chinese-Korean	GL-GAN	0.001583	0.000057	**0.000002**
	proposed	**0.001108**	**0.000044**	0.00005
Korean-English	GL-GAN	**0.003949**	0.000390	0.000232
	proposed	0.004929	**0.000164**	**0.000171**

이러한 결과의 원인은 임곗값 설정 방법의 차이에서 비롯된 것으로 분석된다. GL-GAN은 임곗값을 설정할 때, 최적화를 적용하지 않은 기본 모델이 생성한 이미지들의 특성 행렬을 사용하여 산출한다. 그러나 기본 모델과 최적화를 적용한 모델이 생성하는 이미지에는 차이가 존재하기 때문에, 이 임곗값이 실제 최적화 품질의 기준으로 사용되기에는 정확하지 않다. 반면, 제안하는 최적화 방법은 가장 유사한 성능을 보이는 이전 에포크에서 생성된 이미지들의 특성 행렬을 사용하여 임곗값을 산출하므로, GL-GAN보다 더 근사적인 값을 임곗값으로 사용할 수 있다. [표 5-8]은 다양한 언어 데이터 세트를 사용한 텍스트 스타일 전이 학습에서 GL-GAN의 방법으로 설정한 임곗값 및 제안하는 방법으로 설정한 임곗값과 실제 임곗값의 차이를 나타낸다. 제안하는 방법으로 설정한 임곗값이 대부분은 실제 임곗값과 근사한 값을 보였다.



[그림 5-7] 임곗값 그래프

[그림 5-7]은 텍스트 스타일 전이 학습에서 각 에포크마다의 실제 임곗값과 GL-GAN 및 제안하는 방법으로 계산된 임곗값을 시각화한 것이다. 그래프의 x축은 에포크, y축은 임곗값을 나타내며, 영어-영어, 중국어-한국어, 한국어-영어 텍스트 스타일 전이 학습에서 각 에포크에서 실제로 계산된 임곗값과의 차이를 확인할 수 있다. 이 그래프들을 통해 제안하는 임곗값 설정 방법이 GL-GAN보다 좁은 차이를 보여 더 근사한 값을 제공함을 알 수 있다.

5. 절제 연구

1) 손실 함수 절제 연구

[표 5-9] 손실 함수의 영향 비교

Loss/Method	Language scale	MSE↓	PSNR↑	SSIM↑	FID↓	Acc↑
L_D	-	**5809.2398**	28.1065	0.2322	56.4816	0.0754
$L_D + \underline{L_R}$	S	6611.9091	28.0720	0.2203	109.5734	0.4902
$L_D + \underline{L_{BR}}$	M	9022.1511	27.8350	0.0057	233.1462	**0.9157**
$L_D + \underline{L_{BR}} + L_{rec}$	M	6142.2742	28.1022	0.2907	114.0821	0.8769
$L_D + \underline{L_{BR}} + L_{rec} + L_{cyc}$	M	6311.9718	28.1573	**0.2816**	91.1293	0.8478
$L_D + \underline{L_{BR}} + L_{rec} + L_{cyc} + L_S + \underline{L_T}$	M, S	6011.6899	28.1593	0.2673	70.7157	0.7223
$L_D + \underline{L_{BR}} + L_{rec} + L_{cyc} + L_S + \underline{L_{BT}}$	M, M	5990.2674	**28.1642**	0.2715	**65.7774**	0.7408

[표 5-9]는 한국어-영어 텍스트 스타일 전이 학습에서 다양한 손실 함수 및 언어 네트워크의 규모를 평가하는 절제 연구 결과를 나타낸다. 언어 네트워크는 텍스트 인식기 및 글꼴 분류기를 말하며, 이 네트워크가 단일 언어를 지원할 때는 “S”로 표시되고, 이중 언어를 지원할 때는 “M”으로 표시된다. L_D는 현실적인 이미지를 생성하여 스타일 전이 성능에 대해서는 점수가 높았지만, 올바른 글자를 생성하는 능력이 부족하여 영어 인식 정확도는 찾게 측정되었다. L_D+L_R는 단일 언어 인식기와 이중 언어 인식기 간의 차이를 보였는데, 한국어와 영어를 모두 인식하는 이중 언어 인식기를 사용할 때 올바른 글자 생성에 더 나은 성능을 나타내어 인식 정확도가 상승하였다. 그러나 생성된 이미지의 스타일이 원본 스타일과 유사하지 않아 스타일 성능에 대해서는 낮게 측정되었다. 스타일 일관성을 높여주는 L_{rec}와 L_{cyc}를 추가하여 스타일 전이 성능을 향상시킬 수 있었다. 마지막으로 직접적으로 스타일에 관여하는 손실값인 L_S와L_T를 추가하여 가장 좋은 FID 점수를 얻을 수 있었다. FID는 사람이 볼 때와 가장 유사한 시각적 지표라고 할

수 있어 텍스트 스타일 전이 학습에서 가장 중요한 지표라 생각된다. L_T은 단일 언어 글꼴 분류기와 이중 언어 글꼴 분류기 간의 성능 차이가 있었는데, 한국어와 영어의 글꼴을 모두 분류할 수 있는 이중 언어 글꼴 분류기를 사용할 때 한국어와 영어의 글꼴 차이를 좁힐 수 있어 더 좋은 FID 점수를 보여주었다.

2) 최적화 절제 연구

[표 5-10] 최적화 적용 손실 성능 비교

Method	Training time(hrs)	MSE↓	PSNR↑	SSIM↑	FID↓	Acc↑
Base	**26.62**	6075.5438	28.1009	0.2561	75.4160	0.7383
Base + quality optimization	65.51	6078.2795	28.1350	0.2587	69.7651	0.7316
Base + style optimization	51.23	**5990.2674**	**28.1642**	**0.2715**	**65.7774**	**0.7408**
Base + quality & style optimization	55.25	6034.4158	28.1430	0.2685	69.5927	0.7320

동적인 전역적 및 지역적 이중 수준 최적화를 적용한 손실값에 대한 실험 결과는 [표 5-10]에서 확인할 수 있다. 이미지 생성 품질에 영향을 미치는 손실값 L_D를 사용해 최적화한 경우를 '품질 최적화(quality optimization)'로 정의하고, 이미지의 구조적 특성과 질감에 영향을 미치는 손실값 L_{per}와 L_{tex}를 사용하여 최적화한 경우를 '스타일 최적화(style optimization)'라고 정의했다. 스타일 최적화, 품질 및 스타일 최적화, 품질 최적화 순서로 텍스트 스타일 전이 성능이 좋았다. 이는 텍스트 스타일 전이에서 이미지 스타일 전이 성능이 이미지 품질보다 더 많은 영향을 미치기 때문이다.

[그림 5-8] 최적화 적용 손실 성능 비교 예시

[그림 5-8]은 동적인 전역적 및 지역적 이중 수준 최적화를 적용한 손실값에 대한 한국어-영어 텍스트 스타일 전이 실험 결과를 시각적으로 확인할 수 있다. 첫 번째 행은 한국어 장면 텍스트를 나타내며, 두 번째 행부터 마지막 행까지는 품질 최적화, 스타일 최적화, 품질 및 스타일 최적화의 한국어-영어 텍스트 스타일 전이 결과를 나타낸다. 기본 모델에서 생성된 이미지 일부에 발생된 아티팩트를 최적화를 통해 제거하고자 하였다. 품질 및 스타일 최적화와 품질 최적화를 적용하면 일부 아티팩트가 미세하게 남아있었으나, 스타일 최적화를 적용하면 대부분의 아티팩트를 제거할 수 있어 이미지의 품질이 가장 우수하였다.

[표 5-11] 최적화 방법 성능 비교

Method	Training time(hrs)	MSE↓	PSNR↑	SSIM↑	FID↓	Acc↑
Base	**26.62**	6075.5438	28.1009	0.2561	75.4160	0.7383
+ local optimization	98.53 (26.62+71.91)	6090.1529	28.1295	0.2537	77.0773	0.7175
+ global optimization	111.06 (26.62+84.44)	6045.0326	28.1174	0.2613	75.3979	**0.7575**
+ dynamic optimization (**proposed**)	51.23	**5990.2674**	**28.1642**	**0.2715**	**65.7774**	0.7408

제안하는 동적인 전역적 및 지역적 이중 수준 최적화의 성능을 입증하기 위해서 전역 최적화와 지역 최적화에 대해서 비교 분석과 함께, 동적인 방법을 정적

인 방법과 비교 분석한 결과를 [표 5-11]에 기록하였다. 기본 모델은 최적화를 수행하지 않은 모델로, 2단계 학습까지 진행한 모델이다. 두 번째 모델은 이미지의 저품질 지역만을 사용한 지역 최적화를 적용하였다. 세 번째 모델은 전역 최적화 및 지역 최적화를 적응적으로 선택하였다. 마지막으로 네 번째 모델은 제안하는 동적인 전역적 및 지역적 이중 수준 최적화를 적용한 결과이다.

제안하는 최적화 모델이 MSE, PSNR, SSIM, FID 점수에서 가장 우수한 성능을 보였지만 미세한 차이로 영어 인식 정확도가 낮게 측정되었다. 또한, 기본 모델보다 더 많은 학습 시간이 소요되었다. 이는 제안하는 시스템의 목적은 단순히 이미지를 생성이 아니라 스타일 전이이므로, 이미지의 품질 상승보다는 더 복잡성이 높은 스타일 전이 성능 향상에 우선순위를 두기 때문에 최적화 과정에서 더 많은 시간이 소요된 것으로 보인다.

[그림 5-9] 최적화 방법 성능 비교 예시

[그림 5-9]는 전역 최적화와 지역 최적화에 대한 비교 분석과 함께, 동적인 방법을 정적인 방법과 비교 분석한 결과를 시각적으로 확인할 수 있다. 첫 번째 행은 한국어 장면 텍스트를 보여주고, 두 번째 행부터 마지막 행까지는 기본 모델에 적용한 최적화 방법에 따른 한국어-영어 텍스트 스타일 전이 결과를 보여준다. 기본 모델이 생성한 이미지의 일부분에 아티팩트가 발생하였다. 두 번째 모델은 기본 모델에서 발생한 아티팩트를 제거하는 데 효과적이었지만, 또 다른 아티

팩트를 생성하거나 이미지의 전체적인 색감이 연해지는 문제가 있다. 이는 이미지 전체를 고려하지 않고, 일부 지역만을 최적화하였기 때문이다. 네 번째 모델은 전역 최적화 및 지역 최적화를 적응적으로 선택하여 전체 이미지와 저품질 지역을 모두 고려할 수 있어 더 좋은 성능을 보였지만, 아티팩트 제거에는 효과적이지 못했다. 네 번째 모델은 동적 최적화를 적용한 모델의 결과를 나타낸다. 모델의 학습 동향에 맞게 최적화 방법을 조절하여 다른 모델에 비해 효과적으로 아티팩트를 제거할 수 있었다.

3) 네트워크 절제 연구

[표 5-12] 네트워크 적용 여부에 따른 성능 비교

Baseline	+ Bilingual network	Two-stage training	Dynamic global and local bilevel optimization	FID↓	Acc↑
✓				162.4936	0.0929
✓	✓			155.6558	0.7154
✓	✓	✓		75.4160	0.7383
✓	✓		✓	159.37	0.7058
✓	✓	✓	✓	**65.7774**	**0.7408**

제안하는 시스템의 여러 기능에 대해 효과를 분석하고 정리하기 위해 [표 5-12]에 작성하였다. 이중 언어 네트워크, 2단계 학습, 그리고 동적인 전역적 및 지역적 이중 수준 최적화를 차례로 적용한 결과를 보여준다. 텍스트 스타일 전이 성능에서 가장 중요하다고 생각되는 FID 점수와 텍스트 인식 정확도는 점차 향상되었다. 결론적으로, 모든 기능을 적용한 결과가 가장 우수한 스타일 전이 결과를 보였다.

6. 실패 사례

[그림 5-10] 실패 사례 예시

[그림 5-10]은 한국어-영어 장면 텍스트 스타일 전이에서 발생한 실패 사례를 제시한다. 왼쪽은 여러 스타일을 동시에 가진 장면 텍스트의 스타일 전이 결과를 나타낸다. 제안하는 시스템은 스타일 인코더를 사용하여 스타일을 추출하는데, 이 인코더는 한 번에 하나의 스타일만 추출할 수 있어 다중 스타일을 가진 장면 텍스트의 경우 스타일 전이가 불가능하였다. 오른쪽은 텍스트 일부가 변형되어 다른 글꼴이나 기울기를 가진 경우를 보여준다. 제안하는 시스템은 글꼴 분류기를 활용하여 글꼴을 추출하는데, 한 번에 하나의 글꼴만 추출할 수 있어 다중 글꼴을 동시에 처리할 수 없어 스타일 전이가 불가능했다.

Ⅵ. 보충 자료

1. 네트워크 세부 사항

제안하는 시스템의 구조는 TSB[1)]를 기반으로 설계되었으며, 학습에 사용된 코드는 Olga Kozlova[28)]의 공개된 코드를 기반으로 작성되었다. 시스템은 총 9개의 네트워크로 구성되며, 스타일 인코더, 콘텐츠 인코더, StyleGAN[2)] 기반의 생성자, 텍스트 인식기, 글꼴 분류기, 스타일 추출기, PatchGAN[24)] 기반의 판별자, 동적 최적화 모듈로 이루어진다. 네트워크의 각 구성 요소를 설명하기 위해 다음과 같은 표기법을 사용한다.

- 합성곱 계층(Conv), 풀링 계층(Pool), 잔여 합성곱 블록(ResBlock), 최대 풀링 계층(MaxPool), 평균 풀링 계층(AvgPool), 완전 연결 계층(FC), 스트라이드(s), 업샘플링 요소(up), 커널 크기(k), 채널(c)

[표 6-1] 스타일 인코더 구조

Layers	Configurations	Output
Input	RGB Image(I_s)	64 x 192
conv1	c : 64, k : 7 x 7	62 x 96
MaxPool1	s :2, k : 3 x 3	16 x 48
ResBlock1	[c : 64, k : 3 x 3 c : 64, k : 3 x 3] x 2	16 x 48
ResBlock2	[c : 128, k : 3 x 3 c : 128, k : 3 x 3] x 2	8 x 24
ResBlock3	[c : 256, k : 3 x 3 c : 256, k : 3 x 3] x 2	4 x 12
ResBlock4	[c : 512, k : 3 x 3 c : 512, k : 3 x 3] x 2	2 x 6
AvgPool1	c : 512	1 x 1

[표 6-2] 콘텐츠 인코더 구조

Layers	Configurations	Output
Input	RGB Image(I_c)	64 x 192
conv1	c : 64, k : 7 x 7	62 x 96
MaxPool1	s :2, k : 3 x 3	16 x 48
ResBlock1	[c : 64, k : 3 x 3 c : 64, k : 3 x 3] x 2	16 x 48
ResBlock2	[c : 128, k : 3 x 3 c : 128, k : 3 x 3] x 2	8 x 24
ResBlock3	[c : 256, k : 3 x 3 c : 256, k : 3 x 3] x 2	4 x 12
ResBlock4	[c : 512, k : 3 x 3 c : 512, k : 3 x 3] x 2	2 x 6

[표 6-1]과 [표 6-2]는 각각 스타일 인코더와 콘텐츠 인코더의 구조를 제시한다. 이 인코더들은 ResNet18의 구조를 기반으로 하며, 마지막 레이어를 제외하고 동일한 구조를 사용한다. 스타일 인코더의 마지막 계층은 풀링을 통해 512차원의 특징 벡터로 표현된다.

[표 6-3] 스타일 매핑 네트워크 구조

Layers	Configurations	Output
Input	Style Vector	1 x 512
PixelNorm		1 x 512
FC1	c :512, k : 1 x 1	1 x 512
FC2	c :512, k : 1 x 1	1 x 512

[표 6-3]은 생성자 내부에 포함된 스타일 매핑 네트워크의 구조를 제시한다. 이 네트워크는 스타일 인코더에서 추출된 특징 벡터를 입력으로 사용하며, 레이어별 스타일 구성 요소를 출력한다. 여기서 PixelNorm은 정규화 계층을 나타낸다.

$$PixelNorm = \frac{e_s}{\sqrt{\frac{1}{512}\sum_{i=1}^{512} e_{s,i}^2 + \epsilon}}$$

[수식 6-1] 픽셀 정규화

[표 6-4] 생성자 구조

Layers	Configurations	Output
Input	Content Feature Matrix	2 x 6
Block1	[StyleConv, $w_{s,1}$, up = 2 StyleConv, $w_{s,2}$, up = 1]	4 x 12
Block2	[StyleConv, $w_{s,3}$, up = 2 StyleConv, $w_{s,4}$, up = 1]	8 x 24
Block3	[StyleConv, $w_{s,5}$, up = 2 StyleConv, $w_{s,6}$, up = 1]	16 x 48
Block4	[StyleConv, $w_{s,7}$, up = 2 StyleConv, $w_{s,8}$, up = 1]	32 x 96
Block5	[StyleConv, $w_{s,9}$, up = 2 StyleConv, $w_{s,10}$, up = 1 RGBConv, $w_{s,11}$, up = 1]	64 x 192

[표 6-4]는 StyleGAN 기반의 생성자 구조를 제시한다. 이 생성자는 점진적 성장(progressive growing)은 사용하지 않았다. 생성자의 입력은 512 x 2 x 6 차원의 콘텐츠 특성 행렬이다. 표에 표시된 각 블록에는 합성곱 스타일 레이어(StyleConv), RGB 레이어(RGB-Conv)가 포함된다. 이러한 레이어들은 $w_{s,i}$로 표시된 해당 스타일 벡터의 가중치를 조절하는 데 사용된다. 생성자의 출력은 마지막 RGB 레이어에서 생성되며, 64 x 192 차원의 이미지와 동일한 차원으로 생성된다.

[표 6-5]는 PatchGAN기반의 판별자 구조를 제시한다. 판별자의 입력은 3 x 64 x 192 차원이며, 진위를 판별한 점수를 특성 행렬로 출력한다.

[표 6-5] 판별자 구조

Layers	Configurations	Output
Input	RGB Image(O_c)	64 x 192
conv1	c : 64, s : 2, k : 4 x 4	32 x 96
conv2	c : 128, s : 2, k : 4 x 4	16 x 48
conv3	c : 256, s : 2, k : 4 x 4	8 x 24
conv4	c : 512, k : 4 x 4	7 x 23
conv5	c : 1, k : 4 x 4	6 x 22

텍스트 인식기(영어 인식기, 한국어 인식기)와 글꼴 분류기(영어 글꼴 분류기, 이중 언어 글꼴 분류기)에 대해서는 표준적인 오프-더-셸프(Off-The-Shelf) 구조를 사용하였다. 영어 인식기는 Olga Kozlova[28]가 제공한 모델을 사용하였으며, 한국어 인식기는 AI 허브[27]의 '야외 실제 촬영 한글 이미지' 데이터 세트를 사용하여 직접 학습하였다. 이때, 학습 모델은 텍스트 인식 분야에서 최고 성능을 달성한 Baek, Jeonghun, et al.[11]의 TPS-STN-ResNet-BiLSTM를 사용하였다. 영어 글꼴 분류기는 Olga Kozlova가 제공하는 VGG19 기반의 분류 모델을 사용하였으며, 이중 언어 글꼴 분류기는 VGG16 구조를 사용하여 학습하였다. 동적 최적화 모듈의 경우 GL-GAN[13]의 최적화 방법론에 동적인 임곗값 설정 방법을 도입하였다.

2. 데이터 세트 세부 사항

1) 합성 데이터 세트

[그림 6-1] 합성 데이터 세트 예시

이중 언어 글꼴 분류기를 학습할 때 사용된 합성 데이터 세트는 [그림 6-1]에 제시된 바와 같다. 이 데이터 세트는 Gupta, Ankush, et al.[29]에서 제시한 방법을 따라 직접 생성하였다. 구체적으로는 SynthText[39]의 원본 코드 저장소에서 배경 이미지를 내려받아, 'Google Font[30]'에서 제공하는 37개의 글꼴을 사용해 한국어와 영어 텍스트를 합성하여 생성하였다. 사용된 글꼴은 한국어와 영어를 동

39) SynthText, Github repository, https://github.com/ankush-me/SynthText

시에 지원하는 글꼴로, 그 목록은 [표 6-6]에서 확인할 수 있다.

[표 6-6] 글꼴 목록

Font list
BagelFatOne-Regular
BlackAndWhitePicture-Regular
BlackHanSans-Regular
CuteFont-Regular
Diphylleia-Regular
DoHyeon-Regular
Dokdo-Regular
Dongle-Regular
EastSeaDokdo-Regular
Gaegu-Regular
GamjaFlower-Regular
GasoekOne-Regular
GothicA1-Regular
GowunBatang-Regular
GowunDodum-Regular
GrandifloraOne-Regular
Gugi-Regular
Hahmlet-Regular
HiMelody-Regular
IBMPlexSansKR-Regular
Jua-Regular
KirangHaerang-Regular
MoiraiOne-Regular
NanumBrushScript-Regular
NanumGothic-Regular
NanumGothicCoding-Regular
NanumMyeongjo-Regular
NanumPenScript-Regular
NotoSansKR-Regular
NotoSerifKR-Regular
Orbit-Regular
PoorStory-Regular
SingleDay-Regular
SongMyung-Regular
Stylish-Regular
Sunflower-Medium
YeonSung-Regular

2) BCTR 데이터 세트

[그림 6-2] BCTR 데이터 세트 예시

Xie, Yangchen, et al.[10]이 제안하는 이중 언어 간 스타일 전이 모델과 비교를 위해, 본 연구에서 중국어-한국어 텍스트 스타일 전이 학습을 수행하였다. 학습에 사용된 데이터 셋은 중국어 장면 텍스트 데이터 세트인 BCTR[34]이며, 예시는 [그림 6-2]에서 확인할 수 있다. 이 데이터 세트는 RCTW[40], ReCTS[41], LSVT[42], ArT[43], CTW[44]에서 공개적으로 사용할 수 있는 중국어 장면 텍스트를 수집하여

40) Shi, B., Yao, C., Liao, M., Yang, M., Xu, P., Cui, L., ... & Bai, X. (2017, November). Icdar2017 competition on reading chinese text in the wild (rctw-17). In 2017 14th iapr international conference on document analysis and recognition (ICDAR) (Vol. 1, pp. 1429-1434). IEEE.

41) Zhang, R., Zhou, Y., Jiang, Q., Song, Q., Li, N., Zhou, K., ... & Jawahar, C. V. (2019, September). Icdar 2019 robust reading challenge on reading chinese text on signboard. In 2019 international conference on document analysis and recognition (ICDAR) (pp. 1577-1581). IEEE.

42) Sun, Y., Ni, Z., Chng, C. K., Liu, Y., Luo, C., Ng, C. C., ... & Jin, L. (2019, September). ICDAR 2019 competition on large-scale street view text with partial labeling-RRC-LSVT. In 2019 International Conference on Document Analysis and Recognition (ICDAR) (pp. 1557-1562). IEEE.

43) Chng, C. K., Liu, Y., Sun, Y., Ng, C. C., Luo, C., Ni, Z., ... & Jin, L. (2019, September). Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In 2019 International Conference on Document Analysis and Recognition (ICDAR) (pp. 1571-1576). IEEE.

44) Yuan, T. L., Zhu, Z., Xu, K., Li, C. J., Mu, T. J., & Hu, S. M. (2019). A large chinese text dataset in the wild. Journal of Computer Science and Technology, 34, 509-521.

무작위로 섞은 뒤, 8:1:1의 비율로 나누어 학습, 검증, 테스트 데이터 세트로 구성되었다.

3) ICDAR2019-MLT 데이터 세트

[그림 6-3] ICDAR2019-MLT 데이터 세트 예시

중국어-한국어 텍스트 스타일 전이를 통해 생성된 한국어 데이터 세트를 사용해 한국어 인식 모델을 학습하였다. 이후 한국어 인식률을 평가하기 위해 ICDAR2019-MLT[37]데이터 세트가 사용되었다. 이 데이터 세트는 'ICDAR 2019 Robust Reading Challenge on Multi-lingual scene text detection and recognition'에서 제공되었으며, 아랍어, 영어, 프랑스어 등 10개 언어로 이루어진 (장면 텍스트 - 텍스트) 쌍으로 구성된다. 이 중 한국어 장면 텍스트를 사용했으며, 그 예시는 [그림 6-3]에서 확인할 수 있다.

3. 장면 텍스트 스타일 전이 결과

이 연구에서는 텍스트 스타일 전이 결과의 정성적 평가를 제시한다. 다양한 언어에 대한 단어 수준의 스타일 전이 결과는 [그림 6-4], [그림 6-5], [그림 6-6], [그림 6-7]

에서 확인할 수 있다.

[그림 6-4] 한국어-영어 텍스트 스타일 전이 결과

[그림 6-5] 영어-영어 텍스트 스타일 전이 결과

[그림 6-6] 중국어-한국어 텍스트 스타일 전이 결과

[그림 6-7] 한국어-한국어 텍스트 스타일 전이 결과

Ⅷ. 결론 및 향후 연구

본 연구에서는 외형 차이가 큰 한국어-영어 간 장면 텍스트 스타일 전이 시스템을 제안하였다. 이 시스템은 StyleGAN 기반의 단일 언어 간 텍스트 스타일 전이 모델인 TSB를 이중 언어 간 텍스트 스타일 전이 모델로 확장하였다.

TSB는 입력 이미지를 정답 이미지로 활용하여 약한 자기 지도 학습을 수행함으로써, 스타일 전이 결과에 대한 정답 이미지가 없어도 스타일 전이를 수행할 수 있는 학습 방법을 제안하였다. 그러나 이러한 방식은 한국어-영어 간 텍스트 스타일 전이에는 적용하기 어려웠다. 이는 한국어와 영어의 임베딩 공간의 거리 차가 컸기 때문이다. 이를 해결하기 위해, 첫 번째 단계에서는 영어-영어 텍스트 스타일 전이를 수행하고, 두 번째 단계에서는 한국어-영어 텍스트 스타일 전이 학습을 수행하여 이중 언어 간 텍스트 스타일 전이를 가능하게 하였다. 이러한 접근은 장면 텍스트처럼 스타일 전이 결과에 대한 정답 이미지를 구하기 어려운 상황에서도 이중 언어 간 텍스트 스타일 전이를 가능하게 하는 데 의의가 있다.

영어에 특화된 TSB를 이중 언어 간 스타일 전이로 확장하기 위해 한국어 인식기와 한국어 및 영어 글꼴 분류기를 도입하였다. 한국어 인식기는 텍스트 인식 분야에서 최고 성능을 달성한 TRBA 모델의 구조를 기반으로 학습하여, 생성된 이미지의 텍스트를 인식하고 올바른 형태의 글자를 생성했는지 평가하는 데 사용하였다. 한국어 및 영어 글꼴 분류기는 VGG16 구조를 기반으로, 다양한 배경 이미지에 한국어 및 영어 텍스트를 합성하여 생성한 데이터를 사용해 학습하였다. 이 글꼴 분류기는 생성된 이미지가 입력된 이미지와 동일한 글꼴을 표현하는지 평가하는 데 사용하였다. 한국어에 대한 평가 네트워크를 도입함으로써 이중 언어 간 텍스트 스타일 전이에서 더 좋은 성능을 달성할 수 있었다.

자음과 모음을 조합하는 한국어와 알파벳을 나열하는 영어는 구조적인 특성에서 차이가 있다. 이러한 글자의 외형 차이에 의해 이중 언어 간 텍스트 스타일

전이에서 발생한 아티팩트를 제거하기 위해, 기존의 최적화 방법론인 GL-GAN에 동적인 임곗값 설정 방식을 도입하였다. 생성된 이미지의 품질에 따라 전역 최적화 또는 지역 최적화 중 하나를 선택하여 수행하였다. 전역 최적화는 이미지 전체를 사용하여 최적화하고, 지역 최적화는 저품질 지역만을 사용해 최적화하였다. 최적화 학습에서는 임곗값이 사용되며, 이 임곗값은 최적화 학습의 방법을 결정하고 지역 최적화 시 저품질 지역을 판단하는 데 사용된다. 스타일 전이 학습에서 에포크마다 최적화를 적용하는 동시에 다음 에포크에서 사용할 임곗값을 계산함으로써 학습 방향을 동적으로 조절할 수 있었다. 기존의 GL-GAN과 비교하여 학습 시간을 단축시키고 효과적으로 아티팩트를 제거함으로써 이미지의 품질을 향상시킬 수 있었다는 점에서 큰 의미가 있다.

제안한 시스템의 한국어-영어 텍스트 스타일 전이 성능은 기존의 스타일 전이 모델인 TSB와 SRNet과의 비교를 통해 입증하였다. TSB는 학습 데이터에 포함되지 않은 언어에 대한 스타일 전이가 어려워 이중 언어 간 스타일 전이에서 저조한 성능을 보였다. 또한, SRNet은 합성 데이터를 사용하여 학습하기 때문에, 실제 촬영된 이미지인 장면 텍스트와 같이 배경 분리가 어려운 이미지의 경우 스타일 전이에 한계가 있었다. 이어서, TSB와의 기능 차이를 단계별로 비교하여 추가 적용한 네트워크들의 유효성을 입증하였다. TSB에 차례로 이중 언어 인식기, 이중 언어 글꼴 분류기, 2단계 학습, 그리고 동적인 전역적 및 지역적 이중 수준 최적화를 적용하였다. 이 모든 기능을 적용한 경우, 가장 우수한 텍스트 스타일 전이 성능을 달성하였다.

그러나 장면 텍스트가 다중 스타일을 가지는 경우, 여러 스타일 전이가 불가능하였다. 이는 스타일 인코더가 한 가지의 스타일만 추출할 수 있고, 글꼴 분류기가 하나의 글꼴만 추출할 수 있기 때문이다. 따라서 다중 스타일 전이가 가능하도록 추가적인 연구가 필요하다. 또한, 생성된 이미지의 색상이 전체적으로 탁한 경우가 많아, 색상 품질을 선명하게 향상시키는 연구도 필요하다.

참 고 문 헌

[1] Krishnan, P., Kovvuri, R., Pang, G., Vassilev, B., & Hassner, T. (2023). Textstylebrush: transfer of text aesthetics from a single example. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[2] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4401-4410).

[3] Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. Advances in neural information processing systems, 30.

[4] Li, C., Taniguchi, Y., Lu, M., & Konomi, S. I. (2021). Few-shot font style transfer between different languages. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 433-442).

[5] Pan, W., Zhu, A., Zhou, X., Iwana, B. K., & Li, S. (2023). Few shot font generation via transferring similarity guided global style and quantization local style. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 19506-19516).

[6] Wang, C., Zhou, M., Ge, T., Jiang, Y., Bao, H., & Xu, W. (2023). Cf-font: Content fusion for few-shot font generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition (pp. 1858-1867).

[7] Vinyals, O., Blundell, C., Lillicrap, T., & Wierstra, D. (2016). Matching networks for one shot learning. Advances in neural information processing systems, 29.

[8] Azadi, S., Fisher, M., Kim, V. G., Wang, Z., Shechtman, E., & Darrell, T. (2018). Multi-content gan for few-shot font style transfer. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7564-7573).

[9] Kong, Y., Luo, C., Ma, W., Zhu, Q., Zhu, S., Yuan, N., & Jin, L. (2022). Look closer to supervise better: One-shot font generation via component-based discriminator. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13482-13491).

[10] Xie, Y., Chen, X., Zhan, H., Shivakumara, P., Yin, B., Liu, C., & Lu, Y. (2024). Weakly supervised scene text generation for low-resource languages. Expert Systems with Applications, 237, 121622.

[11] Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., ... & Lee, H. (2019). What is wrong with scene text recognition model comparisons? dataset and model analysis. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4715-4723).

[12] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[13] Liu, Y., Fan, H., Yuan, X., & Xiang, J. (2022). GL-GAN: Adaptive global and local bilevel optimization for generative adversarial network. Pattern Recognition, 123, 108375.

[14] Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2414-2423).

[15] Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14 (pp. 694-711). Springer International Publishing.

[16] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2223-2232).

[17] Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8789-8797).

[18] Liu, M. Y., Breuel, T., & Kautz, J. (2017). Unsupervised image-to-image translation networks. Advances in neural information processing systems, 30.

[19] Huang, X., 8iu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In Proceedings of the

European conference on computer vision (ECCV) (pp. 172-189).

[20] Liu, M. Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., & Kautz, J. (2019). Few-shot unsupervised image-to-image translation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10551-10560).

[21] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8110-8119).

[22] Park, S., Chun, S., Cha, J., Lee, B., & Shim, H. (2021). Multiple heads are better than one: Few-shot font generation with multiple localized experts. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 13900-13909).

[23] Wu, L., Zhang, C., Liu, J., Han, J., Liu, J., Ding, E., & Bai, X. (2019, October). Editing text in the wild. In Proceedings of the 27th ACM international conference on multimedia (pp. 1500-1508).

[24] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-1134).

[25] Zhao, J., & Zhang, H. (2022). Thin-plate spline motion model for image animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3657-3666).

[26] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE

conference on computer vision and pattern recognition (pp. 770-778).

[27] AIHub, Available at: https://www.aihub.or.kr/

[28] deep-text-edit, Github repository, https://github.com/grenlayk/deep-text-edit

[29] Gupta, A., Vedaldi, A., & Zisserman, A. (2016). Synthetic data for text localisation in natural images. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2315-2324).

[30] Google Fonts, Available at : https://fonts.google.com/

[31] IMGUR5K-Handwriting-Dataset, Github repository, https://github.com/facebookresearch/IMGUR5K-Handwriting-Dataset

[32] Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196.

[33] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30.

[34] Yu, H., Chen, J., Li, B., Ma, J., Guan, M., Xu, X., ... & Xue, X. (2021). Benchmarking chinese text recognition: Datasets, baselines, and an empirical study. arXiv preprint arXiv:2112.15093.

[35] Baek, K., Choi, Y., Uh, Y., Yoo, J., & Shim, H. (2021). Rethinking the truly unsupervised image-to-image translation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp.

14154-14163).

[36] Xie, Y., Chen, X., Sun, L., & Lu, Y. (2021). Dg-font: Deformable generative networks for unsupervised font generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5130-5140).

[37] Nayef, N., Patel, Y., Busta, M., Chowdhury, P. N., Karatzas, D., Khlif, W., ... & Ogier, J. M. (2019, September). ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition—RRC-MLT-2019. In 2019 International conference on document analysis and recognition (ICDAR) (pp. 1582-1587). IEEE.

[38] Zhang, Y., Li, T., Li, Q., Fu, X., & Kong, T. (2023). Image motion deblurring via attention generative adversarial network. Computers & Graphics, 111, 122-132.

[39] SynthText, Github repository, https://github.com/ankush-me/SynthText

[40] Shi, B., Yao, C., Liao, M., Yang, M., Xu, P., Cui, L., ... & Bai, X. (2017, November). Icdar2017 competition on reading chinese text in the wild (rctw-17). In 2017 14th iapr international conference on document analysis and recognition (ICDAR) (Vol. 1, pp. 1429-1434). IEEE.

[41] Zhang, R., Zhou, Y., Jiang, Q., Song, Q., Li, N., Zhou, K., ... & Jawahar, C. V. (2019, September). Icdar 2019 robust reading challenge on reading chinese text on signboard. In 2019 international conference on document analysis and recognition (ICDAR) (pp. 1577-1581). IEEE.

[42] Sun, Y., Ni, Z., Chng, C. K., Liu, Y., Luo, C., Ng, C. C., ... & Jin, L. (2019, September). ICDAR 2019 competition on large-scale street view text with partial labeling-RRC-LSVT. In 2019 International Conference on Document Analysis and Recognition (ICDAR) (pp. 1557-1562). IEEE.

[43] Chng, C. K., Liu, Y., Sun, Y., Ng, C. C., Luo, C., Ni, Z., ... & Jin, L. (2019, September). Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In 2019 International Conference on Document Analysis and Recognition (ICDAR) (pp. 1571-1576). IEEE.

[44] Yuan, T. L., Zhu, Z., Xu, K., Li, C. J., Mu, T. J., & Hu, S. M. (2019). A large chinese text dataset in the wild. Journal of Computer Science and Technology, 34, 509-521.

ABSTRACT

Korean-English Scene Text Style Transfer System based on Dynamic Optimization

Ye Rim Kim
Department of Future Convergence Technology Engineering
Graduate School of
Sungshin University

This study proposes a scene text style transfer system for Korean-English pairs, where the languages have significant visual differences. The system extends the single-language text style transfer model, TSB (Text Style Brush) based on StyleGAN, to a bilingual text style transfer model. In the first phase, we conduct English-to-English text style transfer training, and in the second phase, we perform Korean-to-English text style transfer training. This two-step training process enables bilingual style transfer without the need for ground truth images. We utilize a Korean text recognizer, trained based on the TRBA model architecture, which has achieved state-of-the-art performance in the text recognition field, to evaluate whether the generated images during text style transfer correctly render the text. A font classifier for both Korean and English, trained using the VGG16 architecture, is employed to assess whether the generated images preserve the font style of the input images during the text style transfer process. Dynamic optimization is

conducted by introducing a dynamic threshold setting method to the existing GL-GAN optimization framework. This dynamic optimization helps to eliminate artifacts arising from the visual differences between Korean and English during the text style transfer and improves the overall quality. The performance of the proposed system is validated through comparisons with existing text style transfer and optimization studies, as well as through our own ablation studies.