레터논문 (Letter Paper)
방송공학회논문지 제29권 제1호, 2024년 1월 (JBE Vol.29, No.1, January 2024)
https://doi.org/10.5909/JBE.2024.29.1.105
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

# Squeeze and Excitation Block을 사용한 BasicVSR 모델 필터링 연구

우 승 택[a)], 김 동 휘[a)], 김 아 로[a)], 갈리 바니 프리얀카[a)], 박 상 효[a)‡]

# BasicVSR Model Filtering Study Using Squeeze and Excitation Block

Seung-taek Woo[a)], Dong-hwi Kim[a)], Aro Kim[a)], Vani Priyanka Gali[a)], and Sang-hyo Park[a)‡]

## 요 약

본 논문은 BasicVSR 모델을 개량하여 특징 맵을 필터링할 수 있도록 개선된 모델을 제안한다. 기존 모델은 인접한 프레임을 바탕으로 학습하지만, 학습에 악영향을 끼칠 수 있는 특징까지 학습에 이용하는 단점이 있다. 따라서 기존 모델에 필터링 모듈을 추가하여 모델이 학습에 적절한 특징 맵을 선택하도록 개선한다. 필터링의 성능을 효과적으로 검증하기 위해 2D 애니메이션 동영상 데이터 세트를 사용한다. 필터링이 추가된 BasicVSR 모델은 AVC-Train 데이터 세트로 학습하며, AVC-Test 데이터 세트로 테스트한다. 성능 측정 지표 결과 필터링이 동영상 초해상화 모델의 학습에 효과적임을 알 수 있다.

## Abstract

This paper introduces an improved model that filters feature maps by refining the BasicVSR model. While existing models learn from adjacent frames, they exhibit a drawback in incorporating features that can negatively impact the learning process. To address this, we enhance the existing model by integrating filtering modules, enabling the model to selectively choose suitable feature maps for effective learning. To assess the filtering performance, a dataset of 2D animation video is employed. The BasicVSR model with filtering is trained using the AVC-Train dataset and subsequently evaluated using the AVC-Test dataset. As a result of the performance metric, it proves that filtering is effective in learning the video super-resolution model.

Keywords : BasicVSR, Filtering, SEBlock, Video Super-Resolution

## Ⅰ. 서 론

본 논문은 BasicVSR[1] 동영상 초해상화 모델을 개선하기 위해 필터링 모듈을 추가한 연구에 대해 다룬다. 동영상 초해상화 딥러닝 모델로는 컨볼루션 신경망 기반 모델[1][2], 트랜스포머 기반 모델[3][4]과 적대적 신경망 기반 모델[5][6]이 있다. 앞선 모델들은 모두 시간적 차원을 고려하며, 현재 프레임과 인접한 프레임을 참조하여 현재 프레임의 초해상화

a) 경북대학교 컴퓨터학부(School of Computer Science and Engineering, Kyungpook National University)
‡ Corresponding Author : 박상효(Sang-hyo Park)
E-mail: s.park@knu.ac.kr
Tel: +82-53-950-6373
ORCID: https://orcid.org/0000-0002-7282-7686
※ 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00167169, 이동형 로봇 기반 실사 메타버스 실감형 비디오의 획득 및 처리 기술 개발)
· Manuscript December 11, 2023; Revised January 16, 2024; Accepted January 25, 2024.

를 수행한다. 그러나 이러한 방식은 빠르게 움직이는 객체나 장면 전환과 같은 상황에서 참조된 특징들이 의도한 목표와 다를 수 있다는 문제가 존재한다. 이는 인접한 프레임으로부터 얻어낼 수 있는 정보의 양을 제한하고, 학습에 악영향을 끼치는 정보가 될 가능성이 있다.

## Ⅱ. 제안 모델

BasicVSR은 현재프레임과 인접한 프레임을 딥러닝 모델[7]에 입력하여 flow를 얻고, 이를 정제된 특징 맵에 warping 하여 특징을 정렬한다. 정렬된 특징 맵은 현재 프레임과 함께 잔차 블록에 입력되어 정제 과정을 거치고, pixel shuffle 방식으로 업스케일링 된다. BasicVSR이 초해상화를 수행하는 수식은 아래와 같다.

$$s_i^{(f,b)} = S(x_i, x_{i\pm1}),\quad \overline{h}_i^{(f,b)} = W(h_{i\pm1}^{(f,b)}, s_i^{(f,b)}) \qquad (1)$$
$$h_i^{(f,b)} = R_{(f,b)}(x_i, \overline{h}_i^{(f,b)})$$

$$y_i = U(h_i^f, h_i^b) \qquad (2)$$

$x_i, x_{i\pm1}$은 각각 현재 프레임과 인접한 프레임을 의미한다. $s_i^{(f,b)}, \overline{h}_i^{(f,b)}, h_i^{(f,b)}$는각각 flow, 정렬된 특징 맵, 정제된 특징 맵을 의미한다. $U$는 업스케일링 모듈을 의미한다.

이때 장면 전환이 발생하는 경우, 측정된 flow는 정보가 존재하지 않거나, 정렬된 특징 맵에 악영향을 끼칠 여지가 있다. 이후 특징 맵이 정제되었을 때, 일부 특징 맵들은 학습에 악영향을 주는 특징 맵이 되어 모델의 학습이 적절히 수행되지 못할 가능성이 있다.

이를 해결하기 위해 기존 모델에 SEBlock[8]을 추가하여 기존 모델을 개선하고자 한다. SEBlock을 통해 학습에 악영향을 끼치는 특징 맵에 낮은 가중치를 부여한다면, 해당 특징 맵이 결과에 미치는 영향을 최소화하고, 학습에 유리한 특징 맵을 강조할 수 있다. 그림 1.a는 SEBlock의 구조이며, squeeze 연산은 특징 맵을 벡터로 만든다. excitation 연산은 벡터에 가중치를 적용하고, 이를 특징 맵과 곱하여 특징 맵의 각 채널에 가중치를 부여한다.

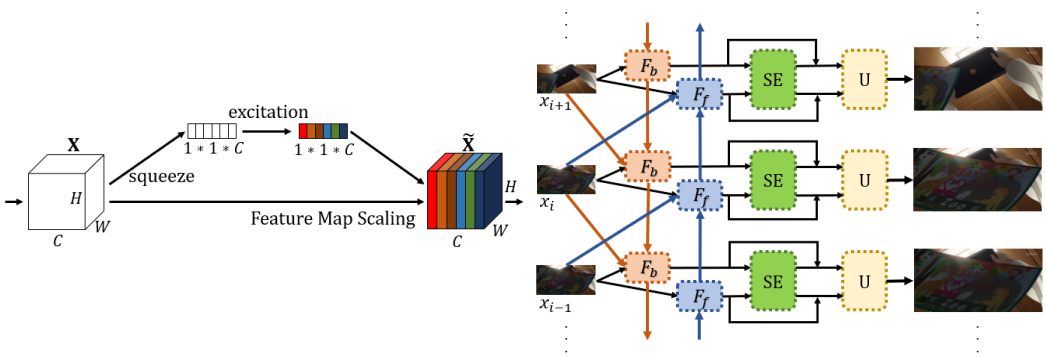
SEBlock은 학습에 유리한 특징과 시너지 효과가 좋은 특징들에 높은 가중치를 부여하도록 학습되고, 정제된 특징 맵을 필터링하며, 그림 1.b와 같은 구조가 된다. SEBlock은 위의 수식(2) 대신에 수식(3)을 수행한다.

$$h_{(i,w)}^{(f,b)} = SE(h_i^f, h_i^b)\ ,\quad y_i = U(h_i^{(f,b)} * h_{(i,w)}^{(f,b)}) \qquad (3)$$

$SE$는 SEBlock을 의미한다. $h_{(i,w)}^{(f,b)}$는 각 특징 맵의 가중치 값을 의미한다.

## Ⅲ. 데이터 세트 및 실험

일반적으로 비디오 초해상화 모델은 장면 전환을 고려하지 않지만, 해당 실험에서는 필터링 효과를 검증하기 위해 장면 전환이 존재하는 데이터 세트를 사용했다. 이때 자연의 동영상인 REDS[9], Vimeo-90K[10]에 비해 애니메이션 동




(a) (b)

그림 1. (a)는 SEBlock의 구조, (b)는 SEBlock이 적용된 BasicVSR(ours)
Fig. 1. (a) is Structure of SEBlock, (b) is the BasicVSR with SEBlock(ours)

영상은 에지가 많이 존재하고, 자막은 에지로 이루어져 있어 약간의 손상이 큰 화질 저하로 이어질 가능성이 있다. 따라서 차이점을 에지를 통해 두드러지게 파악할 수 있으므로 데이터 세트로 애니메이션 동영상인 AVC-Train[11]과 AVC-Test를 선택했다. 배치 사이즈는 4이고, 총 9개의 프레임을 입력했다. 다른 세팅은 기존 모델과 동일하다.

## Ⅳ. 실험 결과

모델의 정성적 성능을 평가하기 위해 기존 모델과 제안 모델의 초해상화 결과를 구하였고, 그 결과는 그림 2와 같다. 그림 2.a에서 제안 모델이 기존 모델에 비해 글자가 명확한데, 이는 제안 모델이 인접한 프레임의 배경에 해당하는 특징 맵을 필터링하기 때문에 글자가 덜 뭉개짐을 알 수 있다. 그림 2.b에서 제안 모델이 기존 모델보다 글자의 형태를 더 잘 유지하고, 노이즈가 덜한 모습을 볼 수 있다. 이는 필터링 모듈이 일반적인 상황에서도 학습에 유리한 특징에 높은 가중치를 줄 여지가 있음을 알 수 있다.

그림 3은 그림 2.a에 필터링을 적용하기 전과 후의 특징 맵 차이이다. 빨간색 점은 특징 맵이 낮은 가중치를 받았음을 의미한다. 그림 3에서 n번째 프레임의 특징 맵 차이가 $diff_n$ 일 때, $diff_{73}$이 $diff_{72}$에 비해 크거나 작은 경우가 다른 프레


그림 2. 장면 전환이 발생했을 때 기존 모델과 제안 모델의 초해상화 결과
Fig. 2. Video Super-Resolution results of existing and proposed models when scene change occur

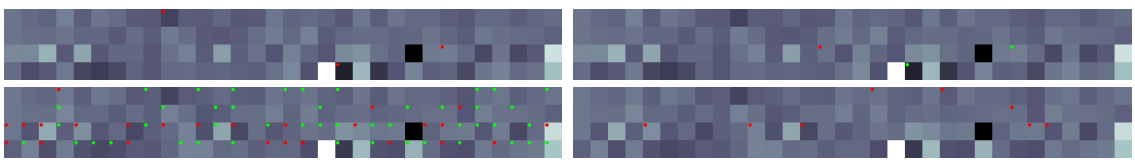


그림 3. 그림 2.a에서 SEBlock을 적용하기 전과 적용한 후의 특징 맵 차이. 검은색에 가까울수록 차이가 줄어듦. 좌측 상단부터 71번째, 72번째, 73번째, 74번째 프레임의 특징 맵 차이
Fig. 3. The difference between the feature map before and after applying SEBlock to Fig 2.a. The closer to black, the smaller feature map difference. Differences in feature maps for frames 71, 72, 73, and 74 from top left

표 1. 그림 2.a에서 장면 전환이 있는 프레임에서의 PSNR, SSIM 결과
Table 1. PSNR and SSIM results in frames when scene change occurs in Fig 2.a

| | BasicVSR | BasicVSR + SEBlock (ours) | Proposed Model Improvement | BasicVSR | BasicVSR + SEBlock (ours) | Proposed Model Improvement |
|---|---|---|---|---|---|---|
| Frame | 72nd | 72nd | 72nd | 73rd | 73rd | 73rd |
| PSNR | 36.9472 | 37.1496 (↑) | 0.2024(0.55%) | 36.9311 | 37.0393 (↑) | 0.1082(0.29%) |
| SSIM | 0.9693 | 0.9704 (↑) | 0.0011(0.11%) | 0.9806 | 0.9809 (↑) | 0.0003(0.03%) |

임에 비해 많은 것을 확인할 수 있다. 이를 통해 장면 전환이 발생했을 때, SEBlock이 특징 맵을 필터링하는 것을 알 수 있다. $mean(diff_{n-1}) = \alpha$, $mean(diff_n) = \beta$일 때, $diff_n$이 $diff_{n-1} + 2^* |\alpha - \beta|$보다 큰 부분은 초록색 점, 작은 부분은 빨간색 점으로 표시했다. 표 1은 그림 2.a에서 PSNR, SSIM을 측정한 결과이다. 표 2는 AVC-Test에서 PNSR, SSIM, LPIPS 그리고 VMAF를 측정한 수치이다. 제안 모델은 기존 모델에 비해 PNSR, SSIM, LPIPS 그리고 VMAF가 각각 0.0378(0.11%)dB, 0.0002(0.02%), 0.000383(0.63%), 0.160407(0.20%)만큼 향상되었다.

표 2. 기존 모델과 제안 모델의 PSNR, SSIM, LPIPS, VMAF 결과
Table 2. PSNR, SSIM, LPIPS and VMAF results of existing model, proposed model

| | AVC-Test | | |
|---|---|---|---|
| | BasicVSR | BasicVSR + SEBlock (ours) | Proposed Model Improvement |
| PSNR | 34.1057 | 34.1435 (↑) | 0.0378(0.11%) |
| SSIM | 0.9777 | 0.9779 (↑) | 0.0002(0.02%) |
| LPIPS | 0.061224 | 0.060841 (↓) | 0.000383(0.63%) |
| VMAF | 80.723477 | 80.883884 (↑) | 0.160407(0.20%) |

## V. 결 론

본 연구에서는 모델이 인접한 프레임을 참조하여 학습하는 중에 특징 맵을 필터링하여 학습에 도움이 될 수 있도록, 필터링 모듈을 기존 모델에 추가했다.

필터링의 유무에 따른 성능 차이를 효과적으로 검증하기 위해 2D 애니메이션 데이터를 사용했다. 기존 모델 대비 제안 모델에서 PSNR, SSIM, LPIPS 그리고 VMAF 결과가 각각 0.0378(0.11%)dB, 0.0002(0.02%), 0.000383(0.63%), 0.160407(0.20%)만큼 향상되었고, 이를 통해 특징 맵의 필터링이 모델의 학습에 도움을 준다는 것을 알 수 있었다.

## 참 고 문 헌 (References)


[1] K. C. K. Chan, X. Wang, K. Yu, C. Dong and C. C. Loy, “BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond,” 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 4945-4954, 2021.
doi: https://doi.org/10.1109/CVPR46437.2021.00491

[2] K. C. K. Chan, S. Zhou, X. Xu and C. C. Loy, “BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment,” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 5962-5971, 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.00588

[3] J. Liang, J. Cao, Y. Fan, K. Zhang, R. Ranjan, Y. Li, R. Timofte and L. V. Gool, “VRT: A Video Restoration Transformer,” arXiv preprint arXiv:2201.12288, 2022.
doi: https://doi.org/10.48550/arXiv.2201.12288

[4] J. Liang, Y. Fan, X. Xiang, R. Ranjan, E. Ilg, S. Green, J. Cao, K. Zhang, R. Timofte, L. V. Gool, “Recurrent Video Restoration Transformer with Guided Deformable Attention,” Advances in Neural Information Processing Systems, 35, 378-393, 2022.
doi: https://doi.org/10.48550/arXiv.2206.02146

[5] M. Chu, Y. Xie, J. Mayer, L. Leal-Taixé, and N. Thuerey. “Learning temporal coherence via self-supervision for GAN-based video generation,” ACM Trans. Graph., 39(4), 75-1, 2020.
doi: https://doi.org/10.1145/3386569.3392457

[6] A. Lucas, A. K. Katsaggelos, S. Lopez-Tapuia and R. Molina, “Generative Adversarial Networks and Perceptual Losses for Video Super-Resolution,” 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, pp. 51-55, 2018.
doi: https://doi.org/10.1109/TIP.2019.2895768

[7] A. Ranjan and M. J. Black, “Optical Flow Estimation Using a Spatial Pyramid Network,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 2720-2729, 2017.
doi: https://doi.org/10.1109/CVPR.2017.291

[8] J. Hu, L. Shen and G. Sun, “Squeeze-and-Excitation Networks,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 7132-7141, 2018.
doi: https://doi.org/10.1109/CVPR.2018.00745

[9] S. Nah, S. Baik, S. Hong, G. Moon, S. Son, R. Timofte, K. M. Lee, “NTIRE 2019 Challenge on Video Deblurring and Super-Resolution: Dataset and Study,” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, pp. 1996-2005, 2019.
doi: https://doi.org/10.1109/CVPRW.2019.00251

[10] T. Xue, B. Chen, J. Wu, D. Wei and W. T. Freeman, “Video enhancement with task-oriented flow,” International Journal of Computer Vision, 127, 1106-1125, 2019.
doi: https://doi.org/10.1007/s11263-018-01144-2

[11] Y. Wu. X. Wang, G. Li, Y. Shan, “AnimeSR: Learning Real-World Super-Resolution Models for Animation Videos,” Advances in Neural Information Processing Systems, 35, 11241-11252, 2022.
doi: https://doi.org/10.48550/arXiv.2206.07038