특집논문 (Special Paper)
방송공학회논문지 제28권 제2호, 2023년 3월 (JBE Vol.28, No.2, March 2023)
https://doi.org/10.5909/JBE.2023.28.2.185
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

내용 기반의 정렬을 통한 HDR 동영상 생성 방법

정 혜 수[a)], 조 남 익[a)‡]

HDR Video Reconstruction via Content-based Alignment Network

Haesoo Chung[a)] and Nam Ik Cho[a)‡]

요 약

최근 인터넷을 통한 동영상 제공 서비스가 확대됨에 따라 높은 품질의 온라인 컨텐츠에 대한 수요가 급증하고 있다. 그런데 넓은 동적 범위 (dynamic range)를 표현할 수 있는 high dynamic range (HDR) 컨텐츠의 공급은 수요를 따라가지 못하고 있는 실정이다. 따라서 본 논문에서는 HDR 영상 제작의 한 방법으로서, 여러 노출값에서 촬영된 프레임들로 구성된 low dynamic range (LDR) 동영상을 이용해 HDR 영상을 생성하는 방법을 제안한다. 우선, 프레임들 사이에 움직임이 존재하기 때문에 정렬 과정을 통해 이웃 프레임들을 중심 프레임에 맞추어 정렬한다. 이때 내용 (content) 기반의 정렬을 하여 정확도를 높이고, 원래 크기의 입력을 그대로 이용하는 모듈을 함께 사용하여 세부 정보도 잘 살려준다. 그러고 나서 잘 정렬된 다중 프레임들을 합쳐서 하나의 HDR 프레임으로 만들어 준다. 실험을 통해 기존 방법들에 비해 우수한 성능을 보임을 확인하였다.

Abstract

As many different over-the-top (OTT) services become ubiquitous, demands for high-quality content are increasing. However, high dynamic range (HDR) contents, which can provide more realistic scenes, are still insufficient. In this regard, we propose a new HDR video reconstruction technique using multi-exposure low dynamic range (LDR) videos. First, we align a reference and its neighboring frames to compensate for motions between them. In the alignment stage, we perform content-based alignment to improve accuracy, and we also present a high-resolution (HR) module to enhance details. Then, we merge the aligned features to generate a final HDR frame. Experimental results demonstrate that our method outperforms existing methods.

Keyword : Image processing, Video, HDR

a) 서울대학교 전기·정보공학부 뉴미디어통신공동연구소(Department of ECE, INMC, Seoul National University)
‡ Corresponding Author : 조남익(Nam Ik Cho)
E-mail: nicho@snu.ac.kr
Tel: +82-2-880-8420
ORCID: https://orcid.org/0000-0001-5297-4649
· Manuscript January 20, 2023; Revised February 19, 2023; Accepted February 19, 2023.

Ⅰ. 서 론

최근 들어 온라인 동영상 제공 서비스의 확대와 함께 고품질의 동영상 컨텐츠에 대한 수요가 증가하고 있다. 디스플레이 기술의 발전으로 높은 화질과 동적 범위 (dynamic range)를 표현할 수 있는 기기는 많아졌지만, 그에 상응하는 High Dynamic Range (HDR) 컨텐츠들은 현저히 부족한 상황이다. HDR 동영상 생산을 위해 직접 특수 장비를 이용하여 취득하는 방법도 있지만, 촬영 장비의 가격이 비싸고 구하기 어렵기 때문에 Low Dynamic Range (LDR) 장비를 기반으로 한 HDR 동영상 제작 기술에 대한 연구도 필요하다. HDR 동영상 제작 기술은 다중 노출 영상을 이용한 HDR 영상 제작 기술과 유사하게, 노출 값을 달리하여 촬영한 여러 프레임들을 이용하여 한 프레임씩 생성하는 것이 일반적이다. 또는 일반적인 동영상의 밝기를 바꾸어서 다중 노출 동영상을 만들고, 이를 이용해 HDR 동영상을 제작할 수도 있다.

정지 영상에 대한 HDR 이미징 방법은 오랜 시간 활발히 연구되어 왔다[1,2,3,4,5,6,7,8]. 이 방법들은 주로 움직임이 있는 입력 영상들을 정렬하는 과정에 집중하였다. 딥 러닝의 발전과 더불어 방법 [4]는 처음으로 옵티컬 플로우 (optical flow)를 예측하여 입력 영상들을 정렬하는 딥 러닝 기반의 방법을 제안하였다. 방법 [5, 6]은 피쳐 (feature) 단계에서 영상들의 정보가 합쳐지게 해서 네트워크 내부적으로 움직임이 정렬되게 하였다. 방법 [7, 8]은 각각 디포머블 컨볼루션 (deformable convolution)과 어텐션 (attention)을 이용하여 움직임을 억제하였다.

반면 HDR 동영상 생성 방법에 대한 연구는 상대적으로 더디게 진행되었는데, 정지 영상에 대한 접근과 마찬가지로 노출 값을 변화시켜 가면서 촬영한 LDR 동영상을 통해 다양한 영역에 대한 충분한 정보를 취득할 수 있게 하려는 시도가 많았다. 그런데 입력으로 여러 장의 프레임을 사용할 때, 프레임 간의 밝기가 다를 뿐만 아니라 물체나 카메라의 움직임도 존재하기 때문에 이를 고려할 필요가 있다. 방법 [9]는 패치 기반의 최적화를, 방법 [10]은 옵티컬 플로우를 이용하여 정렬을 진행하였으나 보정이 정확하지 않고 시간이 많이 소요된다는 단점이 있었다. 최근에 제안된 방법 [11]은 옵티컬 플로우와 디포머블 컨볼루션을 이용한 네트워크를 학습하여 준수한 성능을 보였지만, 옵티컬 플로우에서 발생하는 오류가 누적되는 문제를 해결하지 못하였다.

본 논문에서는 밝기가 다른 여러 장의 프레임들을 효과

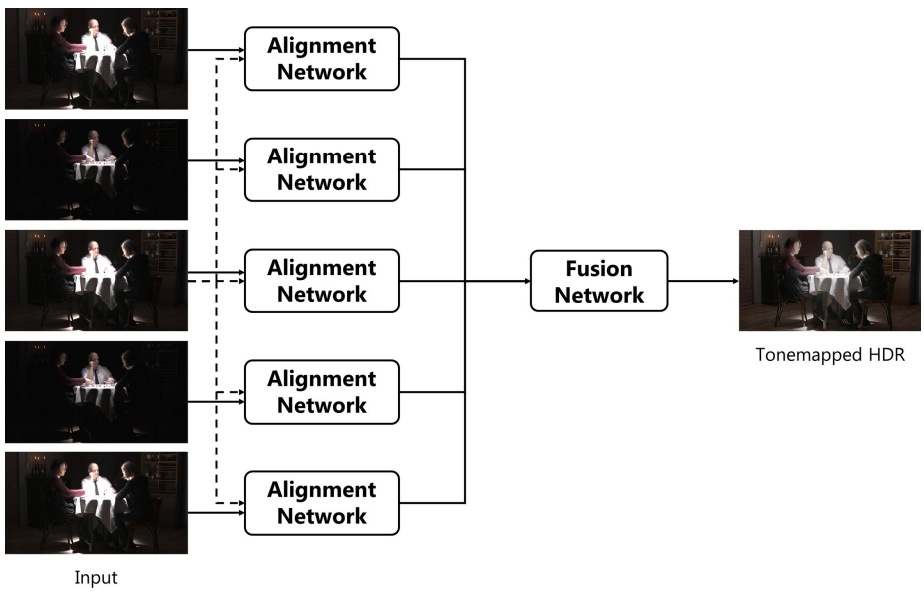

그림 1. 제안 방법의 전체 구조
Fig. 1. Overall architecture of the proposed algorithm

적으로 이용하기 위하여, 이웃 프레임들을 중심 프레임의 움직임에 맞추어 정렬한 뒤 정렬된 피쳐들을 하나로 합쳐 HDR 프레임을 만들어 낸다. 이때 그림 1에서 보이는 것과 같이 HDR 프레임 한 장을 만들기 위해 중심 프레임 한 장과 이웃 프레임 네 장을 입력으로 사용하고, 정렬 네트워크를 이용해 이웃 프레임의 움직임을 중심 프레임의 움직임에 맞추어 정렬해 준다. 정렬 네트워크는 크게 두 가지 모듈로 구성되는데, 첫번째 모듈은 다운 샘플링 (down-sampling)된 입력 프레임들을 어텐션을 이용해 내용 기반의 정렬을 수행하는 Low Resolution (LR) 모듈이고, 두번째 모듈은 두 장의 입력 프레임을 그대로 이용하여 LR 모듈에서 미처 생성하지 못한 세부 정보들을 만들어 내는 High Resolution (HR) 모듈이다. LR 모듈에서는 RGB 색상 값이 아닌 내용에 기반하여 추출한 키 (key)와 쿼리 (query)로 매칭을 진행함으로써 신뢰도 높은 정렬을 할 수 있다. 정렬 네트워크에서 움직임이 맞추어진 피쳐들을 만들어 내면, 합성 네트워크는 이 피쳐들을 합쳐 하나의 HDR 프레임을 생성한다.

본 논문의 구성은 다음과 같다. Ⅱ절에서 내용 기반의 정렬 방법을 이용한 HDR 동영상 생성 방법에 대해 설명하고, Ⅲ절에서는 제안 방법을 이용한 실험 결과를 제시한다. 마지막으로 Ⅳ절에서 결론을 지으며 마무리한다.

Ⅱ. 제안 방법

제안하는 방법은 연속된 다섯 장의 LDR 프레임 $\{I_{t-2}, I_{t-1}, I_t, I_{t+1}, I_{t+2}\}$을 이용하여 HDR 프레임 H_t을 생성하는 것을 목표로 한다. 다섯 장의 입력 프레임들은 교대로 다른 노출 값을 가진다. 그림 1에서 볼 수 있듯이 각각의

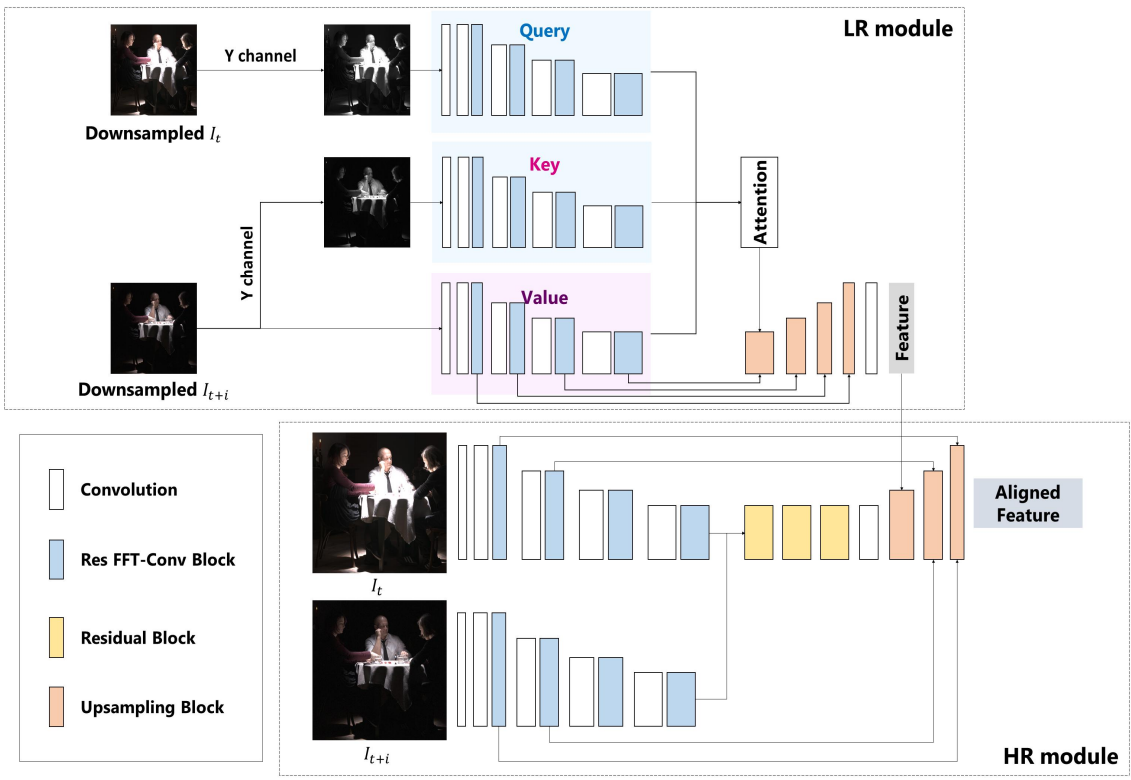

그림 2. 정렬 네트워크의 구조
Fig. 2. An architecture of the proposed alignment network

입력 프레임들은 중심 프레임과 함께 정렬 네트워크에 들어간다. 정렬 네트워크가 내용에 기반하여 이웃 프레임을 중심 프레임의 움직임에 맞추어 정렬해 주면 합성 네트워크는 잘 정렬된 피쳐들의 정보를 합쳐 최종 HDR 프레임을 생성한다. 정렬 네트워크는 그림 2와 같이 크게 두 가지 모듈로 구성된다. LR 모듈은 다운 샘플링된 중심 프레임 I_t와 이웃 프레임 I_{t+i} 간의 내용 기반 매칭을 통한 명시적인 정렬을 담당하고, HR 모듈은 인코더-디코더 구조를 이용해 고주파 세부 정보를 살려준다.

LR 모듈부터 자세히 살펴보면, 먼저 LR 모듈의 입력으로 원래의 입력 영상을 $1/4$ 크기로 다운 샘플링하여 넣어준다. 다운 샘플링된 영상을 이용하면 모듈 내에서 어텐션 계산을 할 때 메모리 사용량을 크게 줄여 줄 수 있다. 본격적으로 정렬을 위해 중심 프레임과 이웃 프레임 간에 유사한 부분을 찾을 때, 단순한 색상 정보를 이용한 매칭이 아닌 내용 기반의 매칭을 하기 위해 RGB 영상을 YCbCr로 변환한 뒤 색상 정보는 제외한 Y 채널 정보만 이용하여 키와 쿼리를 추출한다. 이렇게 하면 실제로는 상관관계가 낮지만 색상만 비슷한 영역끼리 매칭되는 것을 방지할 수 있다. 키와 쿼리는 각 인코더의 마지막 피쳐가 합성곱을 통해 임베딩 (embedding)이 된 피쳐이며, 두 인코더는 가중치를 공유한다. 한편 밸류 (value)는 이웃 프레임을 인코더에 통과시켜 얻은 피쳐를 사용하는데, 이때 사용하는 인코더는 키와 쿼리를 추출하는 데 사용한 것과 같은 구조를 가진다. 어텐션 연산 시에서는 주변 내용을 반영하기 위하여 픽셀 단위가 아닌 3×3 크기의 패치 단위로 진행한다. 즉, 하나의 쿼리 패치와 다른 모든 키 패치 간의 코사인 유사도를 계산한 후 가장 높은 값을 갖는 한 키 패치를 골라 사용한다. i번째 위치의 query 패치를 $\boldsymbol{q}_i$라고 하고 j번째 위치의 키 패치를 $\boldsymbol{k}_j$라고 하면 두 패치 간의 상관관계 $c_{i,j}$는 다음과 같이 계산할 수 있다.

$$c_{i,j}=\left\langle \frac{\boldsymbol{q}_i}{\|\boldsymbol{q}_i\|}, \frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|} \right\rangle. \tag{1}$$

j축에 대해 $c_{i,j}$ 값이 가장 높은 위치의 밸류 패치에 얻어진 유사도 값, 즉 신뢰도를 곱해주면 밸류 피쳐가 재정렬된다. 이때 단순히 어텐션 맵을 계산해 밸류 피쳐에 곱해 주는 것이 아니라 신뢰도가 높은 한 패치만 사용함으로써 블러 현상을 예방할 수 있다. 이렇게 내용 기반의 매칭을 통해 얻은 어텐션 값에 따라 이웃 프레임으로부터 얻은 밸류 피쳐를 적절히 가져온다. 즉, 계산된 어텐션 값이 높으면 해당 패치가 쿼리 패치와 유사하다는 의미이므로, 유사한 패치를 가져옴으로써 피쳐 수준에서 정렬이 수행된다. 이렇게 얻어진 피쳐는 디코더를 통과하여 정렬 네트워크의 입력 영상 크기의 피쳐로 복원된다. 디코더는 잔차 블록 (residual block)[12]과 합성곱 레이어들로 구성된다.

HR 모듈은 입력 영상의 크기를 줄이지 않고 그대로 이용해서, LR 모듈에서 만들어 내기 어려운 세부 정보들을 만들어낸다. HR 모듈은 중심 프레임과 이웃 프레임으로부터 각각 딥 피쳐를 추출한 뒤 잔차 블록을 통해 합쳐 준다. 이 과정에서 아티팩트를 야기할 수 있는 부분의 피쳐는 억제되고 정렬을 방해하지 않는 부분의 피쳐는 강조된다. 각 피쳐 추출을 위한 인코더는 간단한 합성곱 레이어와 주파수 도메인에서 연산을 수행하는 Res FFT-Conv Residual Block[13]으로 이루어져 있다. 주파수 도메인에서의 연산은 전체 영역을 볼 수 있기 때문에 좁은 영역만을 보는 합성곱 연산의 단점을 보완해 주는 효과가 있다. 이후 디코딩 과정에서 업 샘플링 블록 에서 LR 모듈의 출력을 받아 합쳐 주면서 복원을 진행한다. 해당 블록에서는 HR 모듈의 피쳐와 LR 모듈의 피쳐의 차원을 맞춰준 뒤, 두 피쳐를 더해주고 잔차 블록을 통과시켜준다.

정렬 과정이 끝나면 각 정렬 네트워크의 출력 피쳐는 합성 네트워크를 통과하여 하나의 HDR 영상으로 만들어 진다. 합성 네트워크는 합성곱 레이어와 Res FFT-Conv Residual Block으로 구성된다. 정렬된 피쳐들을 채널 축을 따라 이어 붙여준 뒤, 합성곱 레이어와 다섯 개의 Res FFT-Conv Residual Block을 통과시킨다. 마지막으로 3 채널로 줄여주고 시그모이드 함수를 통과시키면 최종 HDR 프레임이 만들어진다.

네트워크 학습을 위한 손실 함수로는 L_1 손실 함수와 VGG 손실 함수, 주파수 손실 함수를 사용하는데, 톤맵핑 (tonemapping)한 HDR 영상에 대해 계산한다. 이전 방법들

과 마찬가지로 HDR 영상 H를 톤맵하는 톤맵핑 함수로는 다음과 같이 정의되는 $\mu-law$를 사용한다.

$$T(H)=\frac{\log(1+\mu H)}{\log(1+\mu)} \tag{2}$$

$\mu-law$는 미분 가능하여 네트워크 학습에 이용하기 적합하다. 여기서 압축 정도를 결정하는 μ값은 5000으로 설정한다. VGG 손실 함수는 미리 학습된 VGG-19 네트워크의 피쳐 간의 거리를 계산하는 함수이다. 주파수 손실 함수는 각 HDR 프레임에 fast Fourier transform (FFT)을 적용한 뒤 주파수 영역에서 L_1 거리를 측정하는 함수로 다음과 같이 정의된다.

$$L_{freq}=\| F(T(\hat{H}))-F(T(H)) \|_1 \tag{3}$$

$T(\hat{H})$는 네트워크의 출력 HDR 영상을 톤맵핑한 영상이고, $T(H)$는 참값 영상을 톤맵핑한 영상이다. 각 손실 함수는 $1:0.001:0.1$의 가중치로 사용한다. 학습에는 $\beta_1=0.9$, $\beta_2=0.999$로 설정한 AdamW optimizer를 사용하고, learning rate는 $1e-4$로 설정한다. 학습 시 패치 크기는 128×128, 배치 크기는 8로 설정한다.

Ⅲ. 실험 결과 및 분석

제안하는 방법은 네트워크 학습 데이터로는 Vimeo-90K 데이터셋[14]을 사용하였고, 테스트 데이터로는 HDRVideo 데이터셋[9]과 DeepHDRVideo 데이터셋[11]을 이용하였다. HDRVideo 데이터셋은 참값 (ground truth, 이하 GT)을 갖고 있지 않기 때문에 정성적 평가에만 사용하였다. 학습 데이터와 테스트 데이터 모두 그림 3과 같이 노출값이 번갈아 가면서 바뀌게 구성되어 있고 프레임 간 움직임이 존재한다. 학습 데이터는 밝기가 일정한 일반적인 LDR 동영상 데이터를 이용해 합성하여 구성하였다. 선형화를 거친 LDR 영상을 X라고 할 때, 감마 보정 (gamma correction)을 이용하여 합성한 영상 Y는 다음과 같이 나타낼 수 있다.

$$Y=clip((Xt)^{1/\gamma}) \tag{4}$$

이때 t는 노출 시간을 의미하며 연속된 프레임에 대해 1과 4 또는 8을 번갈아 가며 적용한다. γ 값은 2.2를 사용하였다. X에 대응되는 HDR 영상 Z는 다음과 같이 생성한다.

$$Z=X^{\gamma}/t \tag{5}$$

비교를 위해 HDR 동영상 생성 방법 [10]의 결과를 포함하였다. 또한 다중 노출 영상을 이용한 HDR 이미징 방법 [5, 8]들도 매 프레임마다 HDR 영상 합성을 진행하면 HDR 동영상을 만들어 낼 수 있기 때문에, 해당 방법들을 적용한 결과와도 비교하였다. 공정한 비교를 위해 비교 방법들을 본 실험에서 사용한 것과 같은 데이터로 다시 학습하였다.

먼저 DeepHDRVideo 데이터셋에 대한 결과 영상을 그림 4에 나타내었다. 상단에 위치한 그림이 제안 방법을 이용하여 얻은 결과 영상이고, 하단에 위치한 그림들이 다른 방법들과 비교한 결과를 확대하여 나타낸 것이다. 다른 방법들의 결과에서는 프레임 간 움직임이 발생하는 부분에서 색

그림 3. 테스트 데이터 예시
Fig. 3. An example of the test data

그림 4. 정성적 결과 비교
Fig. 4. Qualitative comparison of results

상 왜곡이 나타나거나 주변 프레임의 움직임이 남아 있는 것을 확인할 수 있다. 반면 제안하는 방법은 프레임 간의 움직임으로 인한 왜곡을 최소화하고 깨끗한 HDR 프레임을 만들어 낸다. 그림 5는 HDRVideo 데이터셋에 대한 결과를 나타낸다. 해당 데이터셋의 입력 동영상에는 빠른 속도의 움직임이 존재함에도 불구하고, 제안하는 방법은 다른 방법들에 비해 움직이는 물체 주변에서 아티팩트 없이 선명한 결과를 만들어 내는 것을 확인할 수 있다.

그림 5. 정성적 결과 비교
Fig. 5. Qualitative comparison of results

또한 DeepHDRVideo 데이터셋에 대한 정량적 결과 비교를 위해 식 (2)를 이용해 HDR 영상을 톤맵핑한 후 측정한 PSNR과 SSIM을 평가 지표로 사용하였다. 표 1은 DeepHDRVideo 데이터셋의 전체 영상에 대한 평균값을 나타낸다. 제안하는 방법이 다른 방법들에 비해 모든 지표에서 높은 성능을 보이는 것을 확인할 수 있다.

표 1. 정량적 결과
Table 1. Quantitative results

	PSNR	SSIM
Prabhakar[5]	43.20	0.9646
Yan[6]	43.71	0.9682
Kalantari[8]	41.23	0.9553
제안 방법	44.67	0.9709

Ⅳ. 결 론

본 논문에서는 내용 기반의 정렬을 수행하는 LR 모듈과 세부 정보를 잘 만들어 내는 HR 모듈로 구성된 정렬 네트워크와 간단한 합성 네트워크를 이용한 HDR 동영상 생성 구조를 제안하였다. 제안 방법은 주파수 영역에서 연산을 수행함으로써 추출된 피쳐를 통하여 전체 영역을 볼 수 있기 때문에 좁은 영역만을 보는 합성곱 연산의 단점을 보완

해 주는 효과가 있다. 실험을 통해 제안하는 방법이 세부 정보들이 잘 표현되는 HDR 동영상을 만들어 낼 수 있음을 확인하였다.

참 고 문 헌 (References)

[1] T. Grosch, “Fast and robust high dynamic range image generation with camera and object movement,” Vision, Modeling and Visualization, RWTH Aachen, pp.277-284, 2006.

[2] S. Raman and S. Chaudhuri, “Reconstruction of high contrast images for dynamic scenes,” The Visual Computer, Vol.27, No.12, pp.1099 – 1114, 2011.
doi: https://doi.org/10.1007/s00371-011-0653-0

[3] Y. S. Heo, K. M. Lee, S. U. Lee, Y. Moon, and J. Cha, “Ghost-free high dynamic range imaging,” Asian Conference on Computer Vision, Berlin, Heidelberg, pp.486 – 500, 2010.

[4] N. K. Kalantari and R. Ramamoorthi, “Deep high dynamic range imaging of dynamic scenes,” ACM Trans. Graph., Vol.36, No.4, pp.144 – 1, 2017.
doi: http://dx.doi.org/10.1145/3072959.3073609

[5] K. R. Prabhakar, G. Senthil, S. Agrawal, R. V. Babu, and R. K. S. S. Gorthi, “Labeled from unlabeled: Exploiting unlabeled data for few-shot deep hdr deghosting,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.4875 – 4885, 2021.

[6] Y. Niu, J. Wu, W. Liu, W. Guo, and R. WH Lau, “Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions,” IEEE Transactionson Image Processing, pp.3885 – 3896, 2021.
doi: https://doi.org/10.1109/TIP.2021.3064433

[7] Z. Pu, P. Guo, M. S. Asif, and Z. Ma, “Robust high dynamic range (hdr) imaging with complex motion and parallax,” Proceedings of the Asian Conference on Computer Vision, 2020.

[8] Q. Yan, D. Gong, Q. Shi, A. V. D. Hengel, C. Shen, I. Reid, and Y. Zhang, “Attention-guided network for ghost-free high dynamic range imaging,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.1751-1760, 2019.

[9] N. K. Kalantari, E. Shechtman, C. Barnes, S. Darabi, D. B. Goldman, and P. Sen, “Patch-based high dynamic range video,” ACM Trans. Graph., Vol.32, No.6, pp.202-1, 2013.
doi: https://doi.org/10.1145/2508363.2508402

[10] N. K. Kalantari and R. Ramamoorthi, “Deep HDR video from sequences with alternating exposures,” Computer Graphics Forum, Vol.38, No.2, pp.193-205, 2019.
doi: https://doi.org/10.1111/cgf.13630

[11] G. Chen, C. Chen, S. Guo, Z. Liang, K. Y. K. Wong, L. Zhang, “HDR video reconstruction: A coarse-to-fine network and a real-world benchmark dataset,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.2502-2511, 2021.

[12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.770-778, 2016.

[13] X. Mao, Y. Liu, W. Shen, Q. Li, and Y. Wang, “Deep residual fourier transformation for single image deblurring,” arXiv preprint arXiv:2111.11745, 2021.

[14] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, “Video enhancement with task-oriented flow,” International Journal of Computer Vision, Vol.127, No.8, pp.1106-1125, 2019.
doi: https://doi.org/10.1007/s11263-018-01144-2

저 자 소 개

정 혜 수

- 서울대학교 전기정보공학부 박사과정
- ORCID : https://orcid.org/0000-0003-3804-9443
- 주관심분야 : 영상처리, 컴퓨터 비전, 딥러닝

저 자 소 개

조 남 익

- 서울대학교 전기정보공학부 교수
- ORCID : https://orcid.org/0000-0001-5297-4649
- 주관심분야 : 디지털 신호처리, 영상처리, 컴퓨터 비전