Modern High Dynamic Range Imaging at the Time of Deep Learning

Multiple Exposures Reconstruction

Francesco Banterle and Alessandro Artusi

Introduction

- HDR reconstruction from multiple-exposures:
  - If we don’t place the camera on a stable tripod the camera moves!
  - If we have wind or people, there will be movement!
  - All this means, we will have artifacts!

Introduction

- HDR reconstruction from multiple-exposures:
  - If we don’t place the camera on a stable tripod the camera moves!
  - If we have wind or people, there will be movement!
  - All this means, we will have artifacts!

Introduction: Camera Movement

- What if we capture a stack of exposure images free-hand without a tripod?

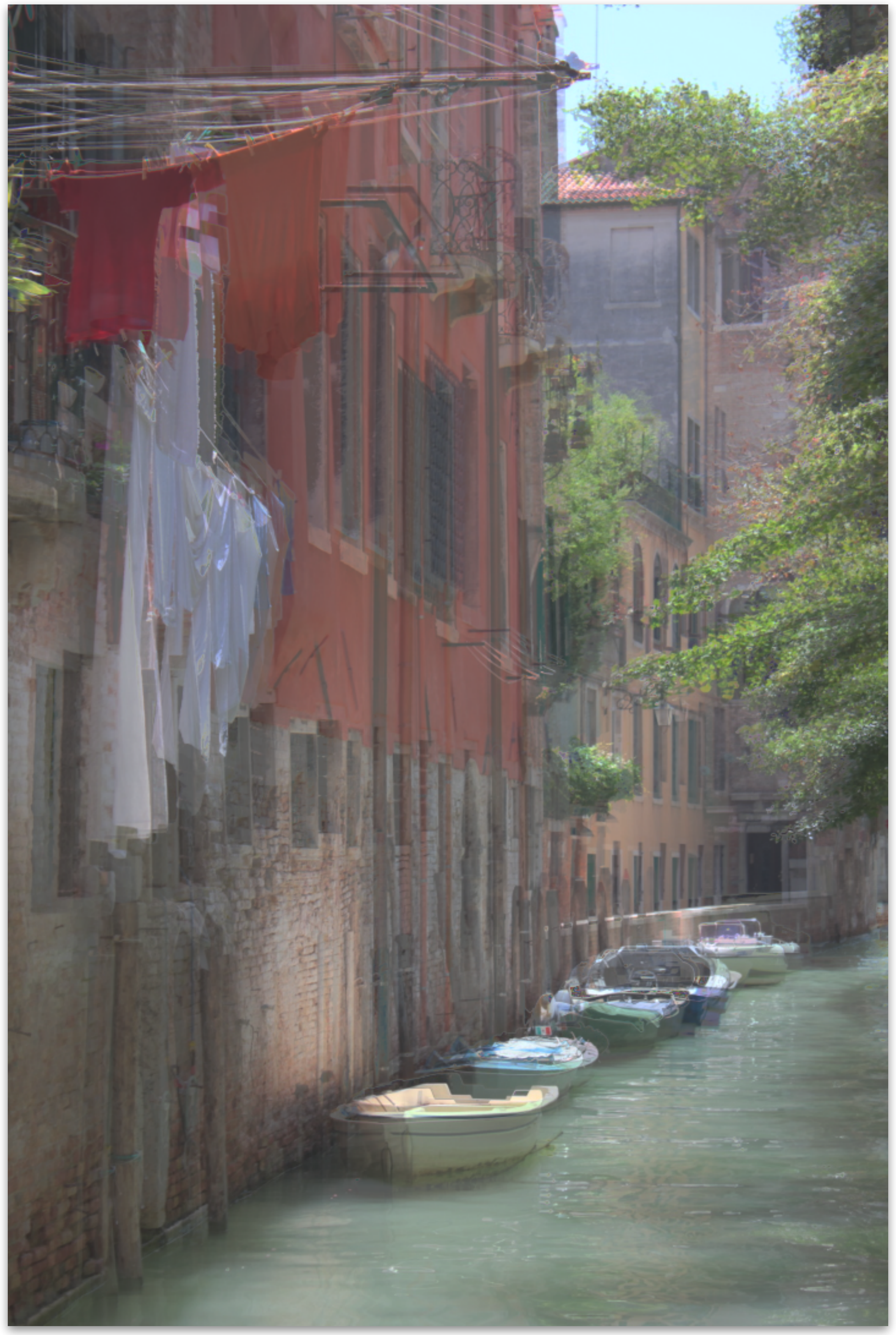
-2-stop

0-stop

+2-stop

Introduction: Camera Movement

Introduction: Camera Movement



Merged Stack and Tone Mapped

Introduction: Camera Movement

Introduction: Camera Movement

Merged Stack and Tone Mapped

Introduction: Camera Movement

Introduction: Camera Movement

- Typically, if we have **ONLY** camera movement, we can manage the merge:
  - We have only a single global movement.
- There are several robust algorithm to deal with such situations:
  - Greg Ward's MTB method.
  - Tomaszewska and Mantiuk's Homography algorithm.
  - Gallo's Multiple Homographies.

Introduction: Dynamic Scene

- What if we capture a stack of exposure images on a tripod in a dynamic scene?

-2-stop

0-stop

+2-stop

Introduction: Dynamic Scene

Introduction: Dynamic Scene

Merged Stack and Tone Mapped

Introduction: Dynamic Scene

Introduction: Dynamic Scene

Introduction: Dynamic Scene

Introduction: Dynamic Scene

Merged Stack and Tone Mapped

Introduction: Dynamic Scene

Introduction: Dynamic Scene

Introduction: Camera Movement

- Typically, if when the moving people/objects are small they can be fixed easily.
- There are several robust algorithm to deal with such situations:
  - Masks: Pece and Katuz 2010
  - Grandaos et al. 2013
  - PatchMatch-based: Sen et al./Hu et al. 2014

Datasets

Capturing Data: Kalantari’s Data

-2-stop

0-stop

+2-stop

Dynamic Stack

Video Courtesy of Jan Fröhlich - Stuttgart HDR Video Dataset

Capturing Data: Kalantari's Data

-2-stop

0-stop

+2-stop

Dynamic Stack

Static Stack

Video Courtesy of Jan Fröhlich - Stuttgart HDR Video Dataset

Capturing Data: Kalantari's Data

-2-stop

0-stop

+2-stop

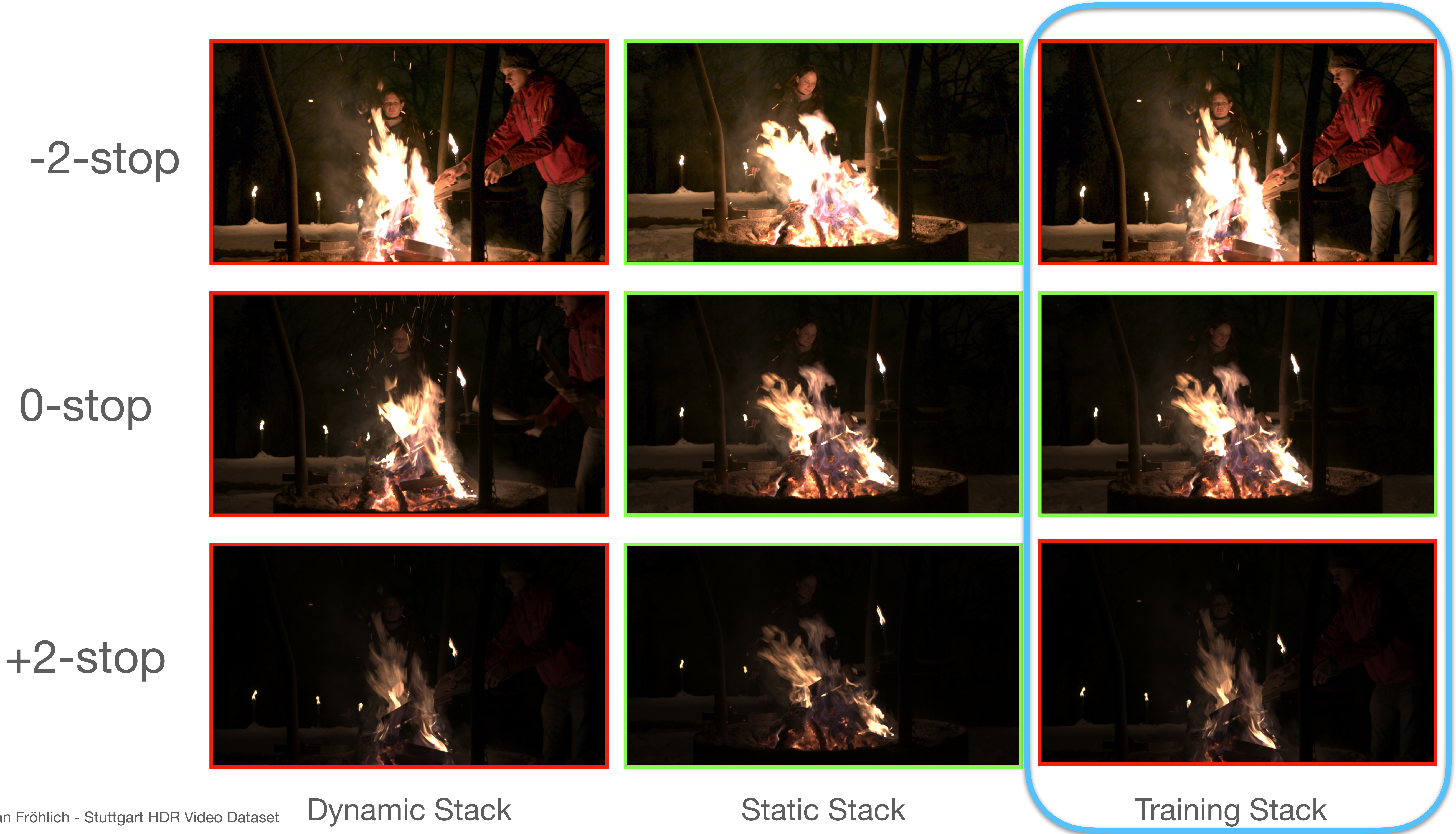
Dynamic Stack

Static Stack

Training Stack

Video Courtesy of Jan Fröhlich - Stuttgart HDR Video Dataset

Capturing Data: Kalantari's Data



Video Courtesy of Jan Fröhlich - Stuttgart HDR Video Dataset

Images

- For each SDR image I_i, we know:
  - The CRF, $f(\cdot)$; i.e, we know its inverse $g(\cdot) = f^{-1}(\cdot)$;
  - The exposure time $t_i = \dfrac{\text{ISO}_i \cdot t'_i}{K \cdot A_i^2}$
    - t'_i: Shutter speed.
    - A_i: Aperture value.
    - ISO_i: ISO value.
    - $K \in [30.6, 13.4]$: a constant depending on the camera.

Images

- Typically, we work with "calibrated" SDR image $\boldsymbol{H_i}$:

$$H_i = \frac{g(I_i)}{t_i}$$

- In many works, the CRF is assumed to be $f(x) = x^{\frac{1}{2.2}}$.
- Therefore, we have:

$$H_i = \frac{I_i^{2.2}}{t_i}$$

Images: Patches and Augmentations

- All methods are trained on patches of different size: 40×40, 256×256, 512×512.
- Patches may be create with or without overlap.
- We have different augmentations:
  - Rotation, Flips, etc.
  - Swapping color channels [Kalantari et al. 2017]

Preprocessing

- The problem can be “simplified” by using classic approach for a first alignment:
  - **Homography alignment** introduced by Wu et al. 2018;
  - **Optical flow alignment** introduced by Kalantari et al. 2017.
- This initial alignment reduces blur.
- Typically, it matches the background well:
  - Local mismatches are left.

HDR Image Datasets

Dataset Name	#Images	#Resolution	Calibrated	Website
Kalantari Dataset	74	1.5MPix	Uncalibrated	https://cseweb.ucsd.edu/~viscomp/projects/SIG17HDR/
Tursun Dataset	17	0.6Mpix	Uncalibrated	https://user.ceng.metu.edu.tr/~akyuz/files/eg2016/index.html

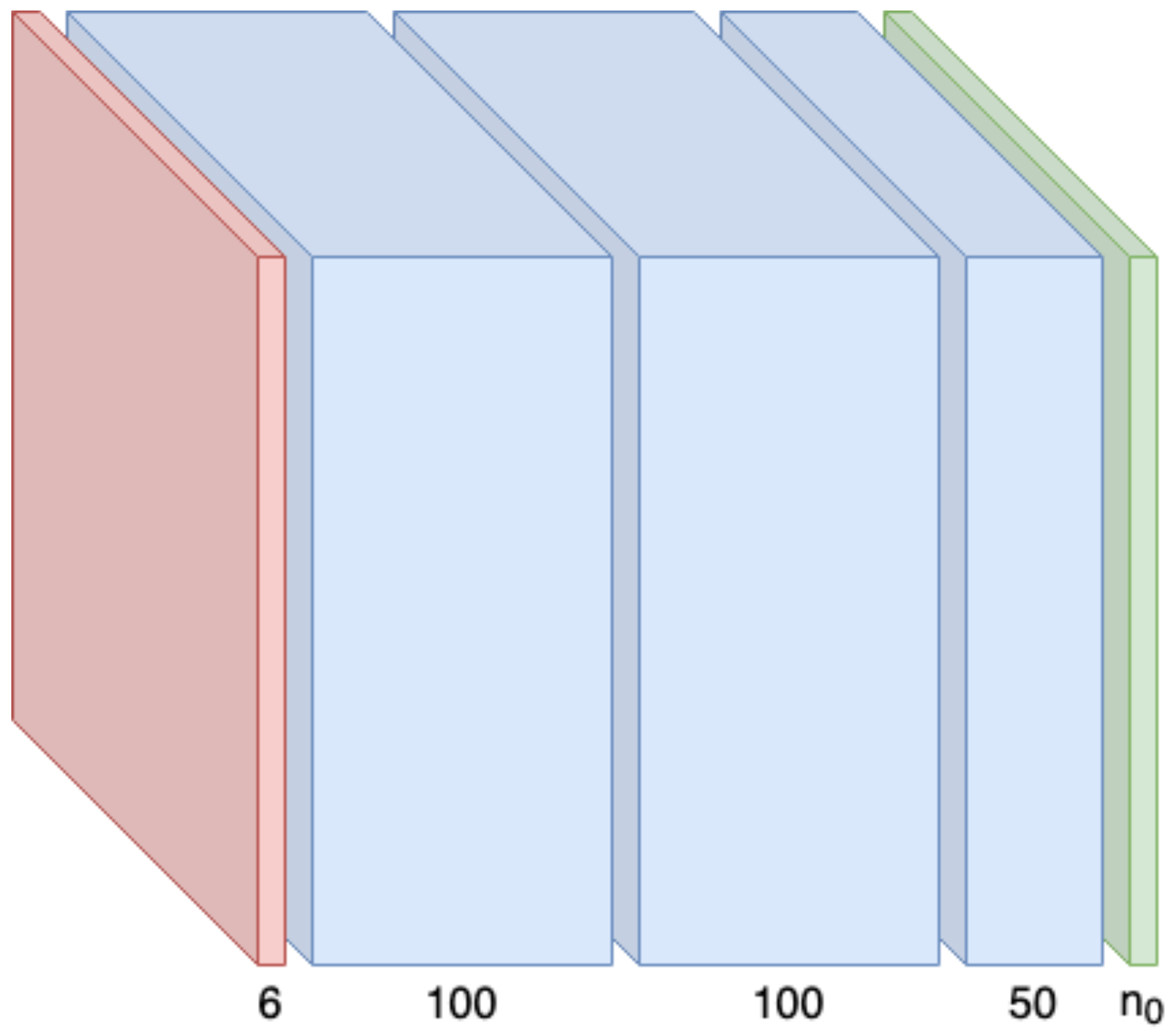
HDR Video Datasets

Dataset Name	#Videos	#Resolution	Length	FPS	Color Space	Format	Website
Stuttgart HDR Dataset	33	1920×1080	13s-100s	24/25	REC709	Floating Point	https://www.hdm-stuttgart.de/vmlab/projects/
UBC HDR Video Dataset	10	2048×1080	7s-10s	30	REC709	Floating Point	http://dml.ece.ubc.ca/data/DML-HDR/
LIVE HDR Video Quality Assessment Database	31 (310 at different bit-rates)	0.32Mpix	3s-10s	50/60	BT2020	HDR10	https://live.ece.utexas.edu/research/LIVEHDR/LIVEHDR_index.html
MPI HDR Video Dataset	2	0.3Mpix	24s-34s	24	REC709	Floating Point	https://resources.mpi-inf.mpg.de/hdr/video/
EBU HDR Video Dataset	10	3996×2160	10s-31s	50	BT2100	HLG	https://tech.ebu.ch/testsequences

End2End Architectures

Kalantari et al. 2017

- Kalantari et al. 2017 proposed a simple solution:
  - Optical Flow for the main alignment between exposures;
  - An end2end (a FCN) with ReLU in all layers except a sigmoid for the last layer:
    - Convolution varies in kernel size from large to small:
      - 7×7, 5×5, 3×3, and 1×1



Kalantari et al. 2017

- Kalantari et al. 2017 noted that the simple solution have some issues:
  - It is difficult to train; we need a huge dataset!
  - It does not fix alignment artifacts.
- The solution is to use the network to:
  - Compute Weights.
  - Refine images.

Kalantari et al. 2017

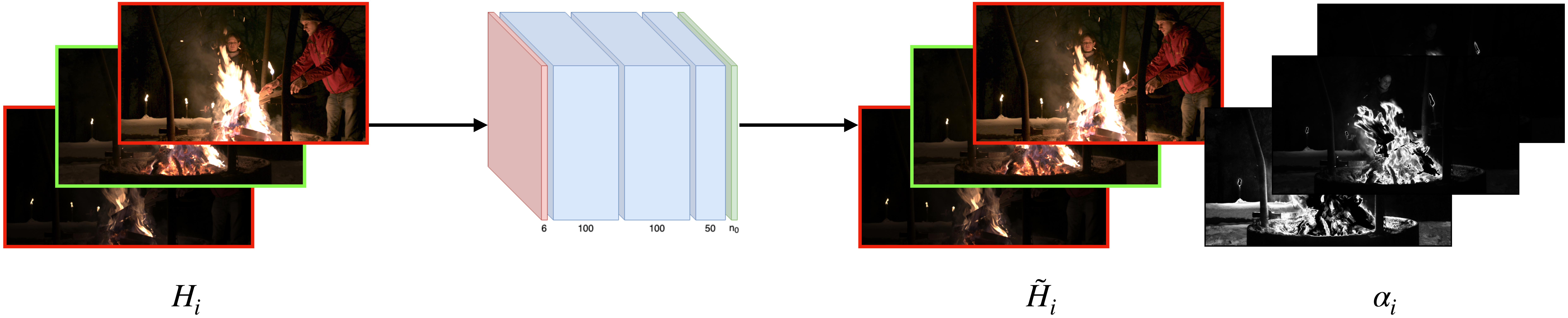
- Weight Estimator:
  - The shown architecture is used to compute the per-pixel weights, $\boldsymbol{\alpha}$, to obtain the estimated HDR image $\hat{H}$:

$$\hat{H} = \frac{\sum_i \alpha_i \cdot H_i}{\sum_i \alpha_i}$$

- Refined Images:
  - The network also refines the alignment obtaining new improved images $\tilde{H}_i$:

$$\hat{H} = \frac{\sum_i \alpha_i \cdot \tilde{H}_i}{\sum_i \alpha_i}$$

Kalantari et al. 2017



Video Courtesy of Jan Fröhlich - Stuttgart HDR Video Dataset

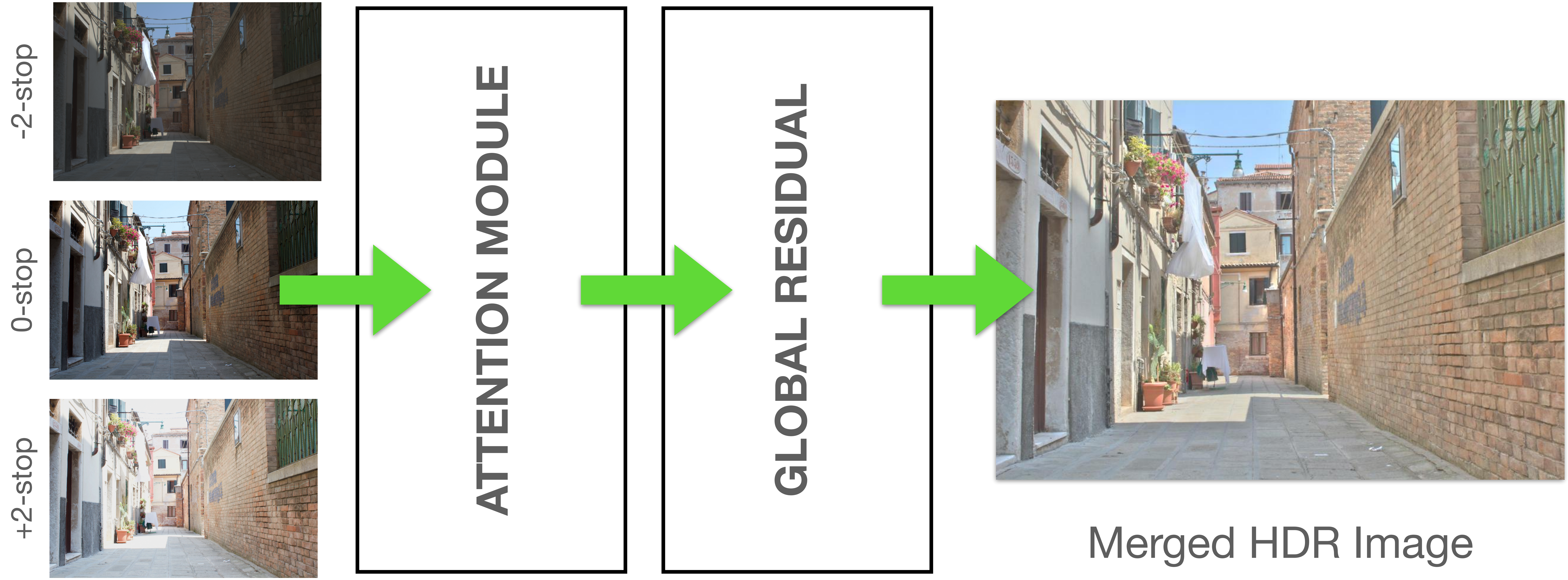
Encoder-Decoder - Wu et al. 2018

Video Courtesy of Jan Fröhlich - Stuttgart HDR Video Dataset

Attention HDR - Yan et al. 2019

- Yan et al. 2019 introduces two blocks:
  - Attention Module:
    - The attention is computed on low level features.
    - The attention is applied to features of images that are not the reference.
  - Residual Dense Blocks [Zhang et al. 2018] with dilated convolutions to have a larger receptive field.

Attention HDR - Yan et al. 2019

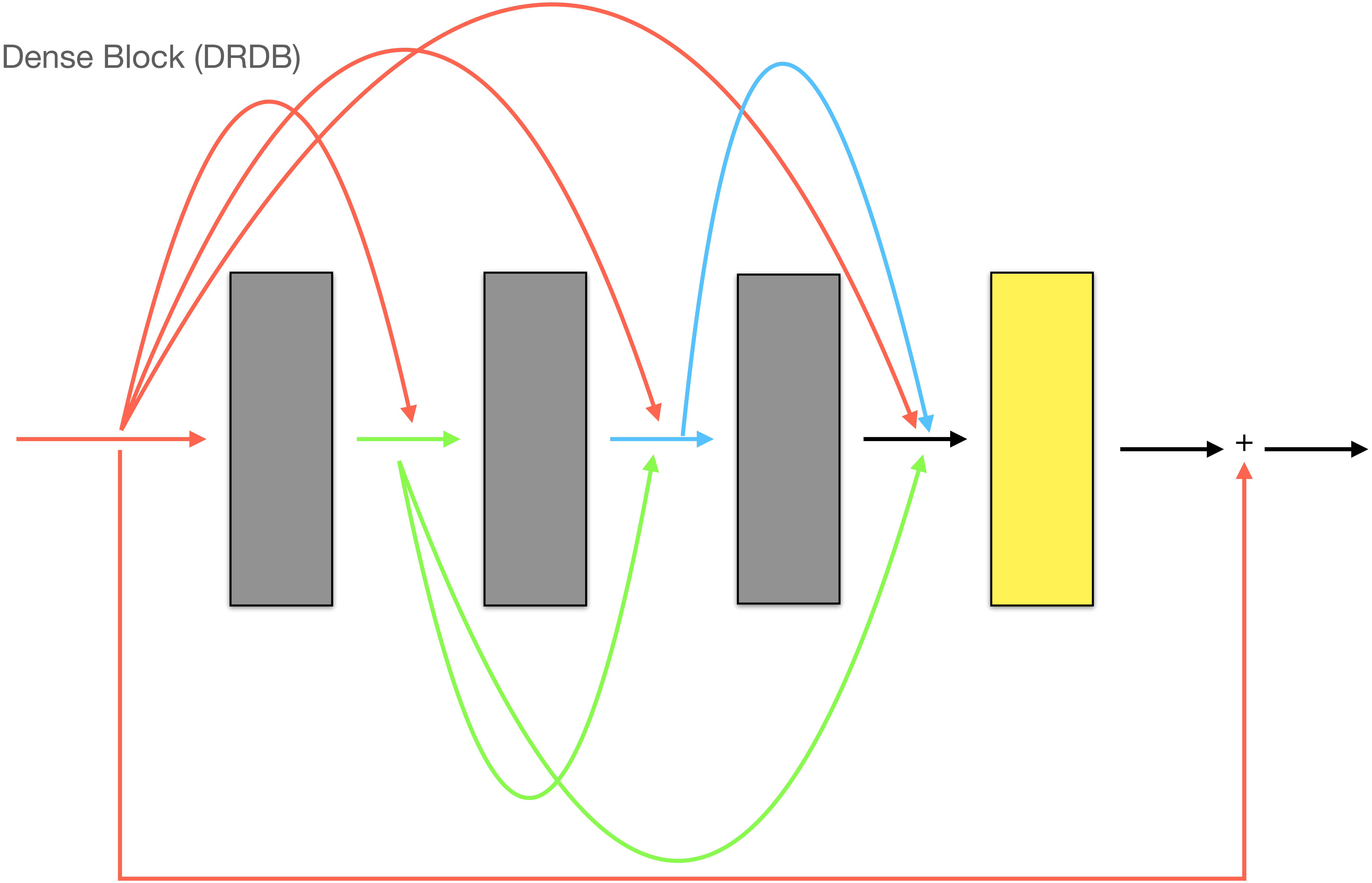


Merged HDR Image

Attention HDR - Yan et al. 2019

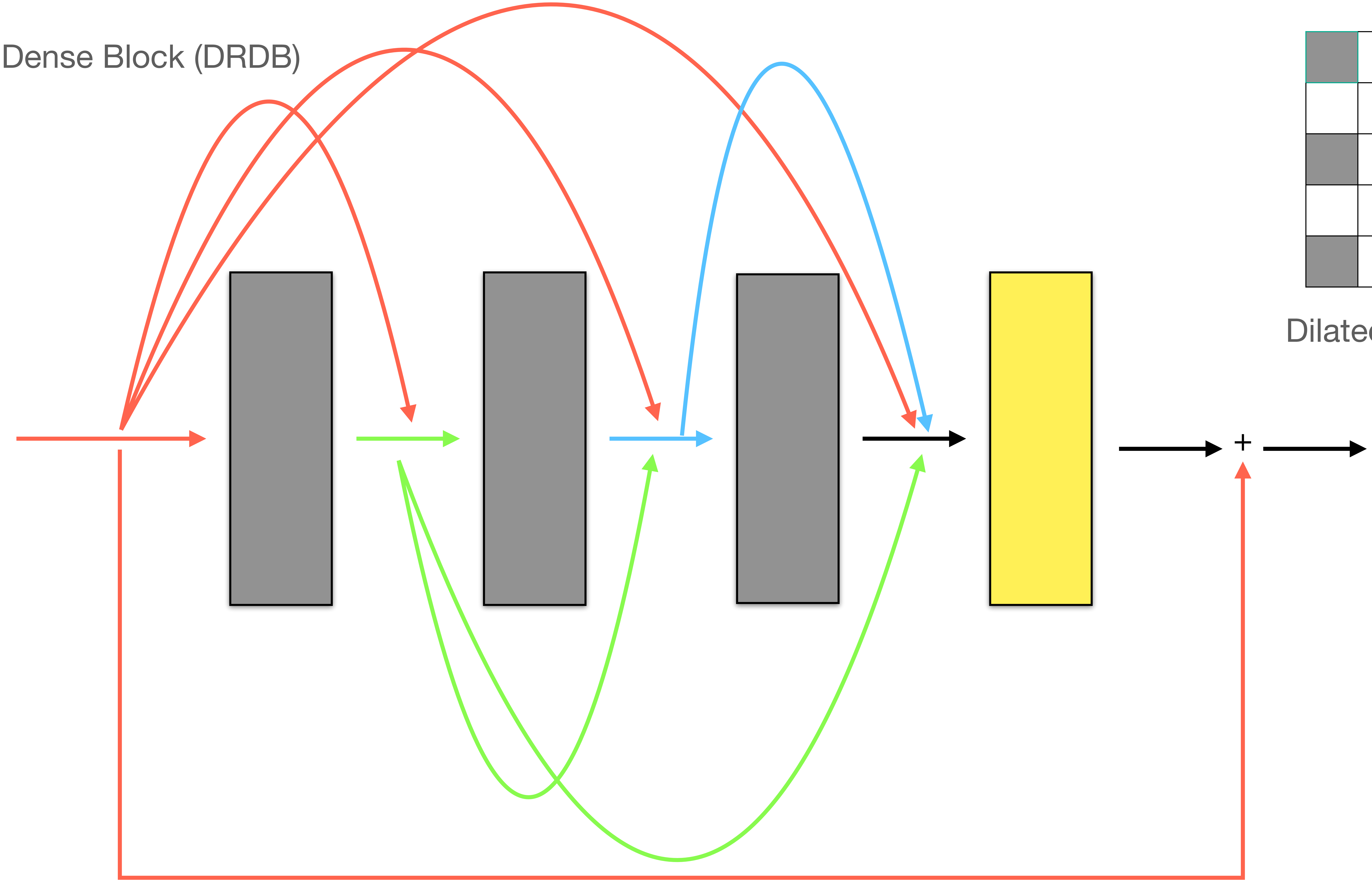
Attention HDR - Yan et al. 2019

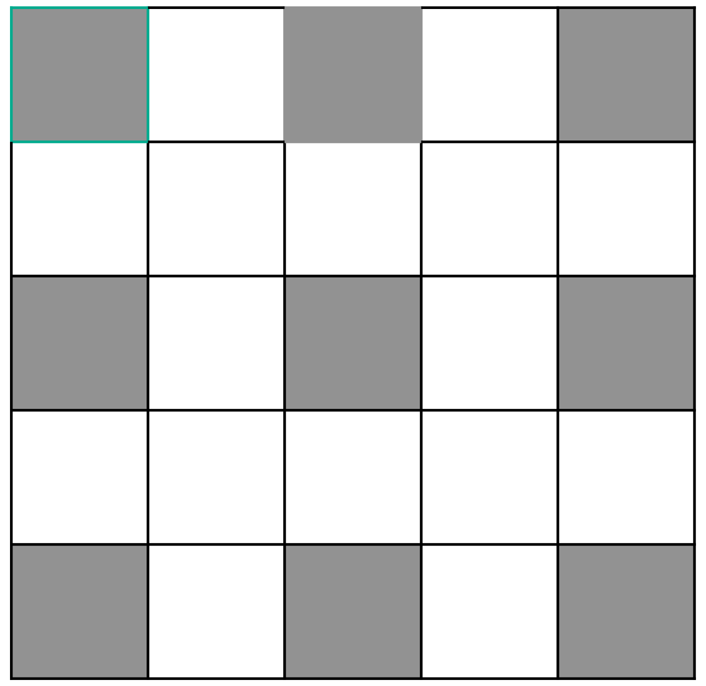
Dilated Residual Dense Block (DRDB)



Attention HDR - Yan et al. 2019

Dilated Residual Dense Block (DRDB)

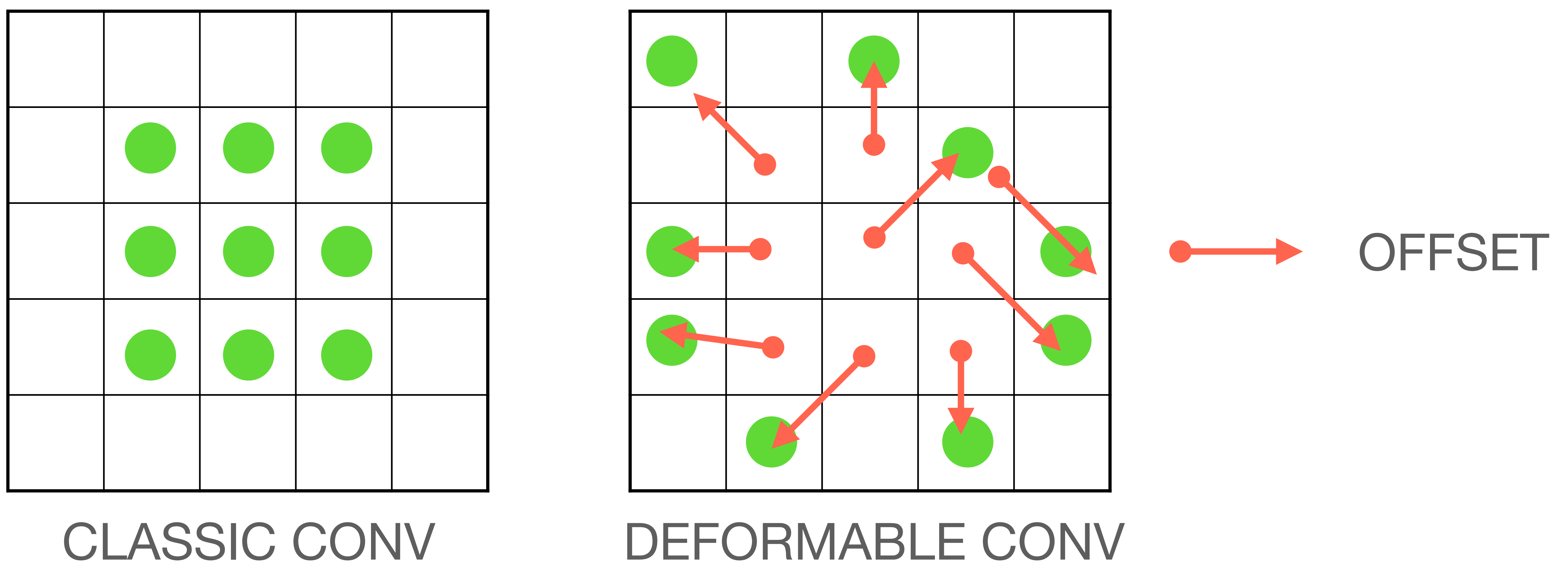




Dilated Convolutions

ADNet - Liu et al. 2021

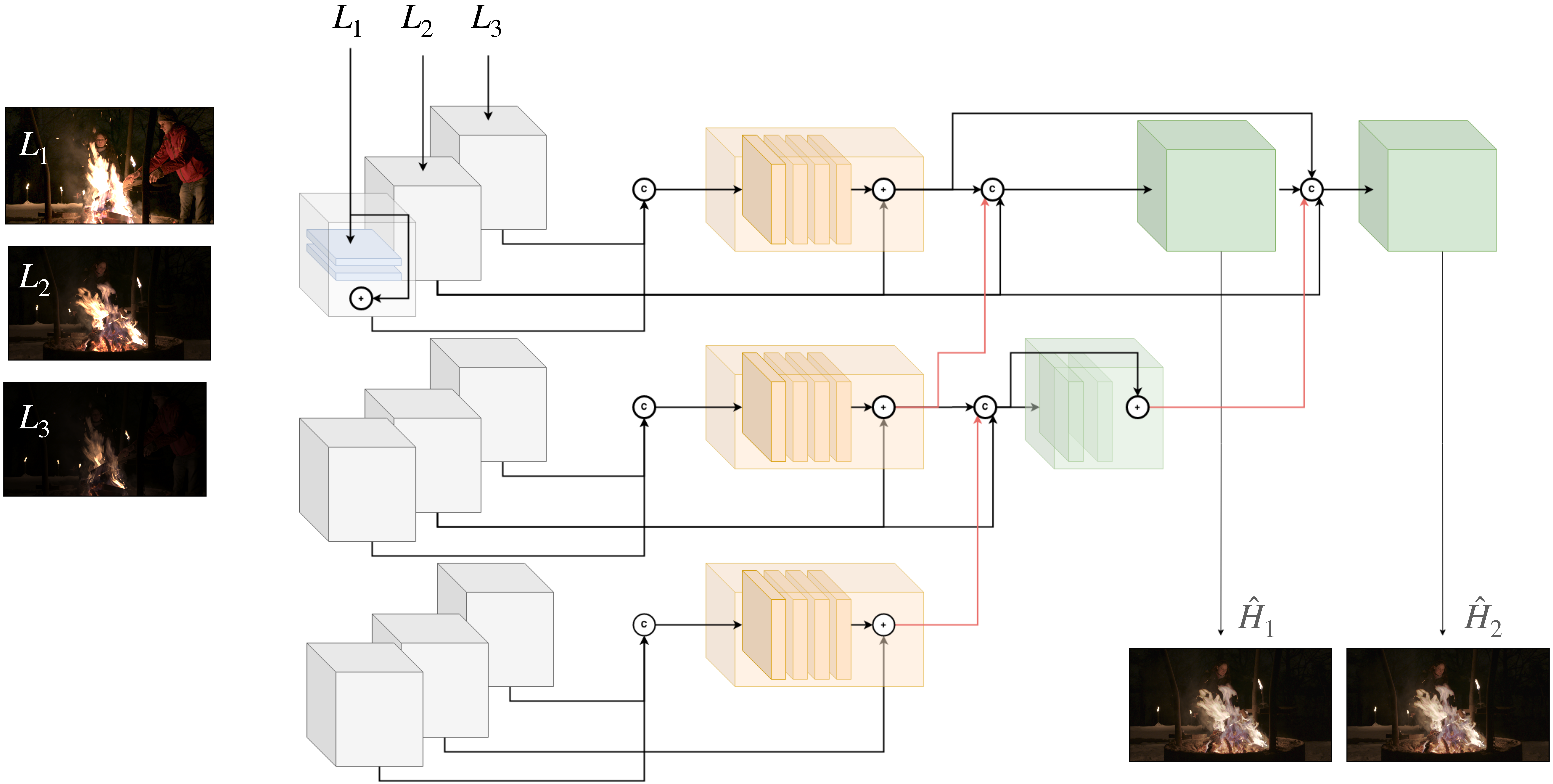
- Liu et al. 2021, similarly to Pu et al. 2020, proposed for NTIRE 2021 a network based on two main blocks:
    - Attention computed using the reference, similar to Yan et al. 2019.
    - Pyramid, Cascade and Deformable (PCD) module by Wang et al. 2019:
        - PCD is applied at the feature level of the gamma-corrected images.
        - This module uses deformable convolutions



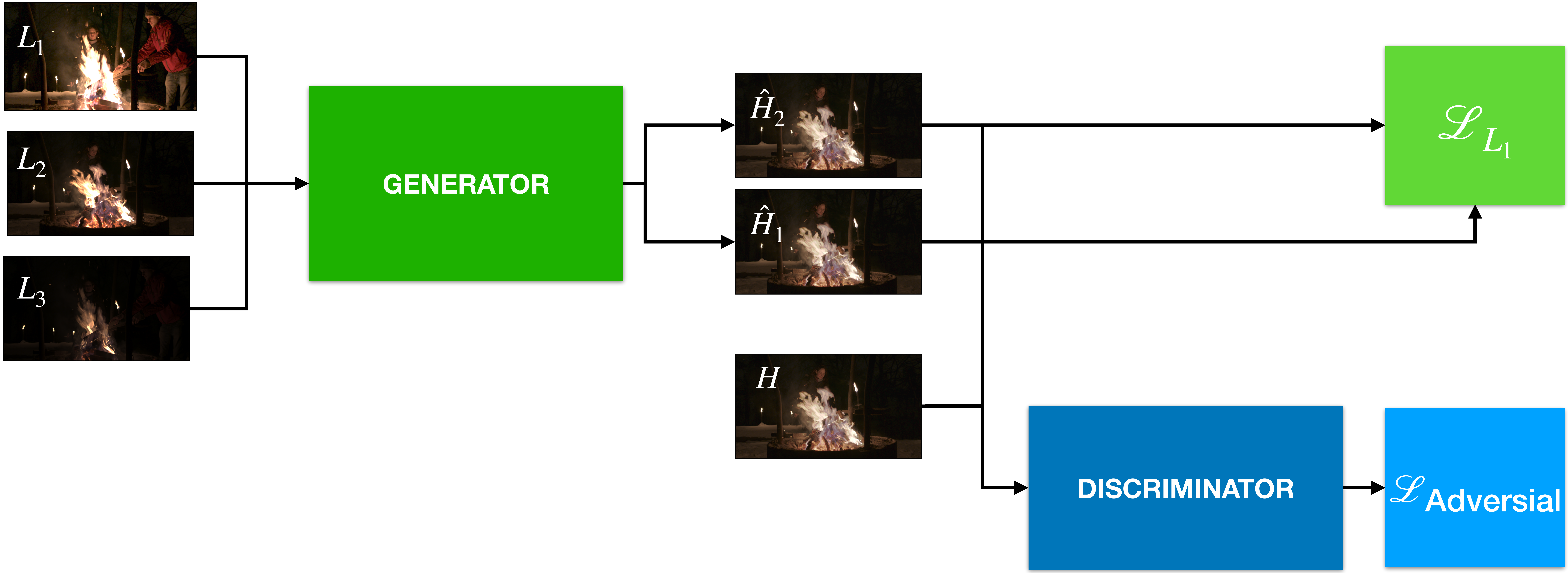
ADNet - PCD - Liu et al. 2021

GAN Architectures

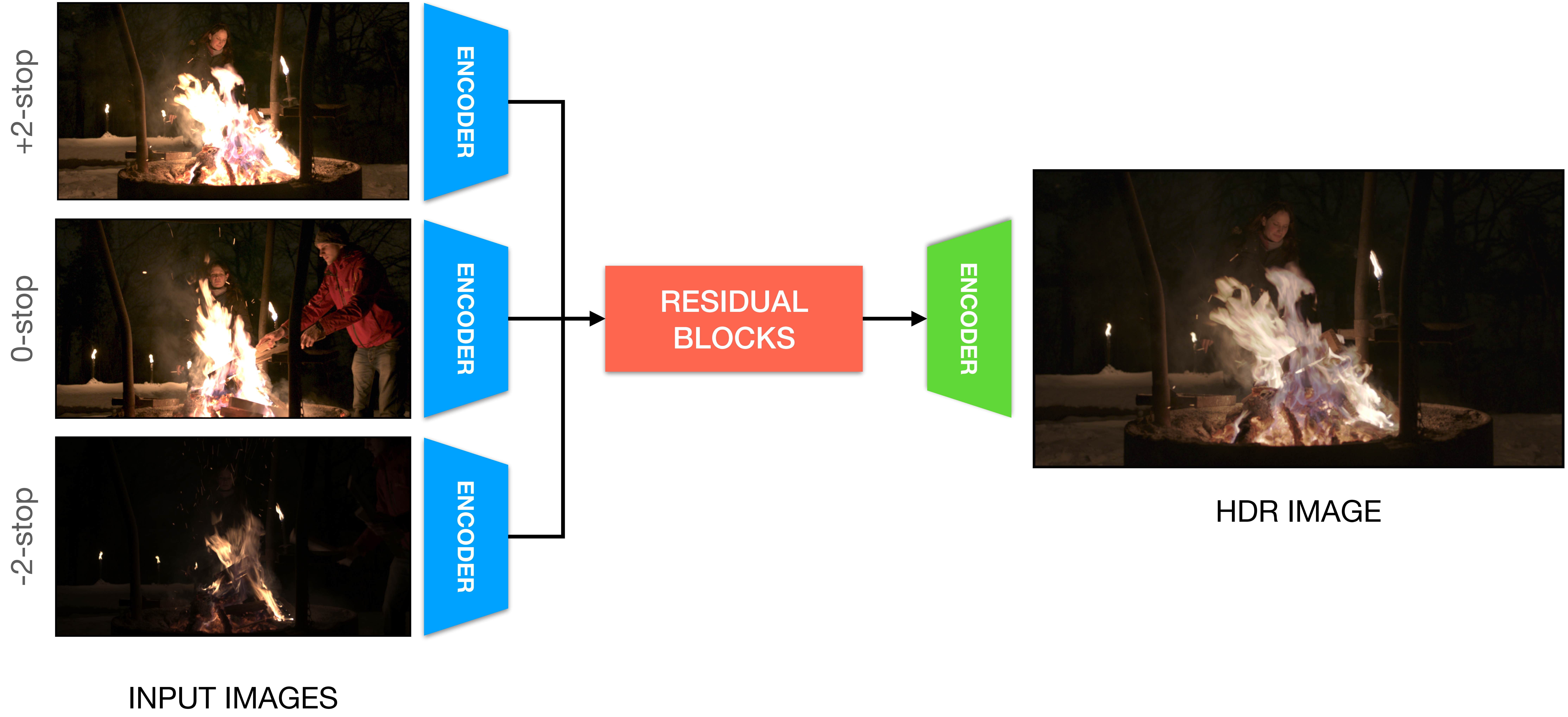
HDRGAN - Niu et al. 2021: Generator



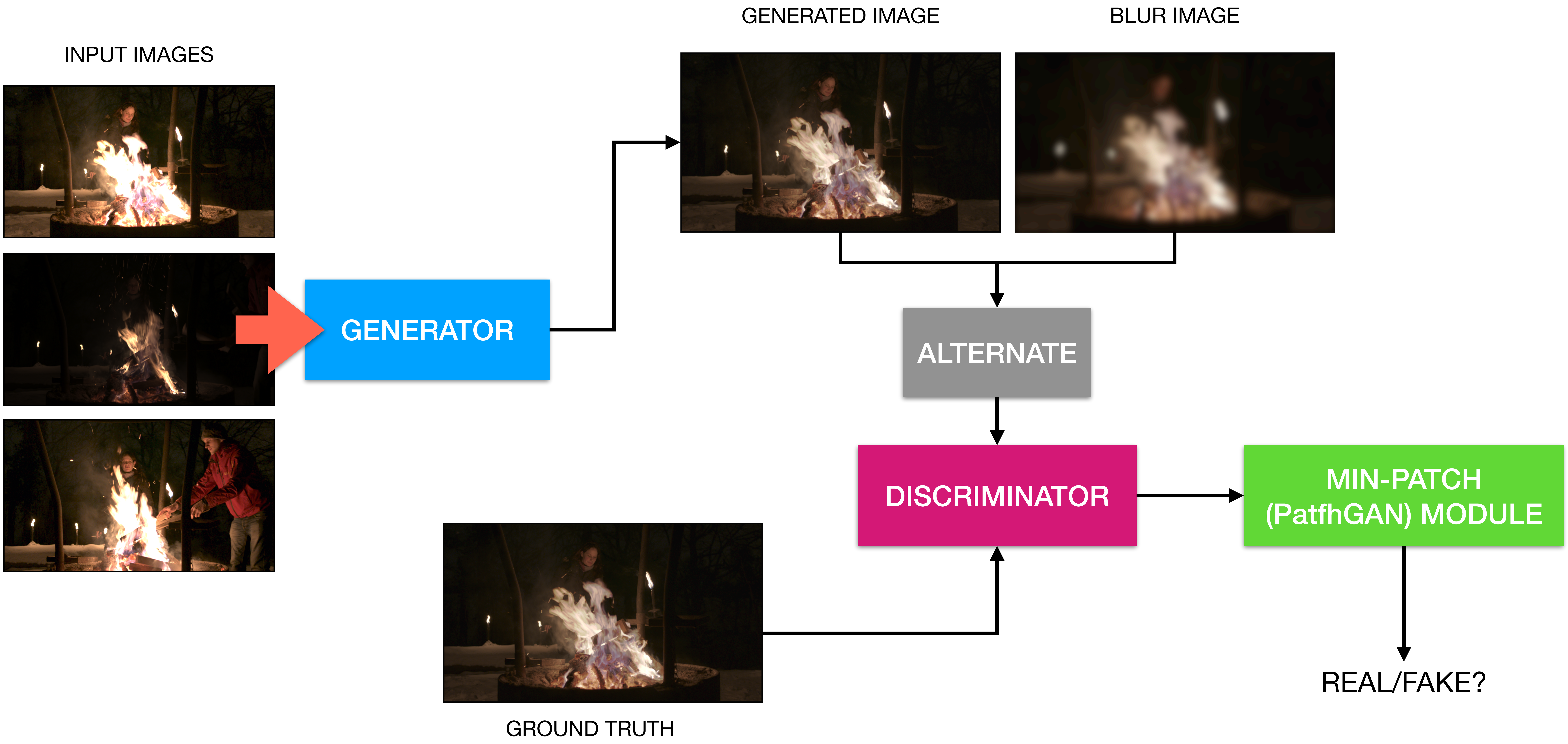
HDRGAN - Niu et al. 2021: Training



UPHDR-GAN - Li et al. 2022: Generator



UPHDR-GAN - Li et al. 2022: Training



Loss Functions

Loss Function in the μ-Law Domain

- Kalantari et al. 2017 introduced a L2 loss function in a tone-mapped domain:

$$\mathscr{L}_{\text{rec}}(\hat{I}, I) = \|\tau(I) - \tau(\hat{I})\|_2$$

where $\boldsymbol{\tau}(\cdot)$ is a differentiable tone mapping function based on the $\boldsymbol{\mu}$-law:

$$\tau(I) = \frac{\log(1 + \mu I)}{\log(1 + \mu)} \qquad \mu = 5000$$

- Note that there are variants of $\mathscr{L}_{\text{rec}}$ where we have L1 instead of L2.
- This loss function is **ubiquitous** in most HDR works for reconstruction and inverse tone mapping.

GAN Loss

- Our goal is:

$$\arg\min_{G}\max_{D} \mathcal{L}(G, D)$$

- Typically a GAN loss is defined as:

$$\mathcal{L}(G, D) = \alpha_1 \mathcal{L}_{\text{GAN}}(G, D) + \alpha_2 \mathcal{L}_{\text{rec}}(G)$$

where:

- $\mathcal{L}_{\text{GAN}}(G, D)$ is the adversial loss.
- $\mathcal{L}_{\text{rec}}(G)$ is the content/reconstruction loss.
- α_1 and α_2 are weights for balancing the two losses.

GAN Loss: HDRGAN

- Niu et al. 2021 has a GAN scheme with a content/reconstruction loss:

$$\mathscr{L}_{\text{rec}} = \min_G \left(\|\tau(\hat{H}_1) - \hat{H}\|_1 + \|\tau(\hat{H}_2) - \hat{H}\|_1 \right)$$

- And a GAN loss based on the sphere generative adverbial loss [Park and Kwon 2019], where the Discriminator output an $\boldsymbol{n}$-dimensional vector $\mathbf{q}$ which is projected on $\mathbf{p} \in \mathbb{S}^n$:

$$\mathscr{L}_{\text{GAN}} = \min_G \max_D \sum_r \mathbb{E}_{\mathbf{z}}[d_s^r(\mathbf{N}, D(\mathbf{z}))] - \sum_r \mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3} d_s^r(\mathbf{N}, D(G(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3))]$$

where $d_s(\mathbf{p}, \mathbf{p}')$ is the distance on the hypersphere, and $\mathbf{N} = [0,\ldots,0,1] \in \mathbb{R}^n$.

GAN Loss: UPHDR-GAN

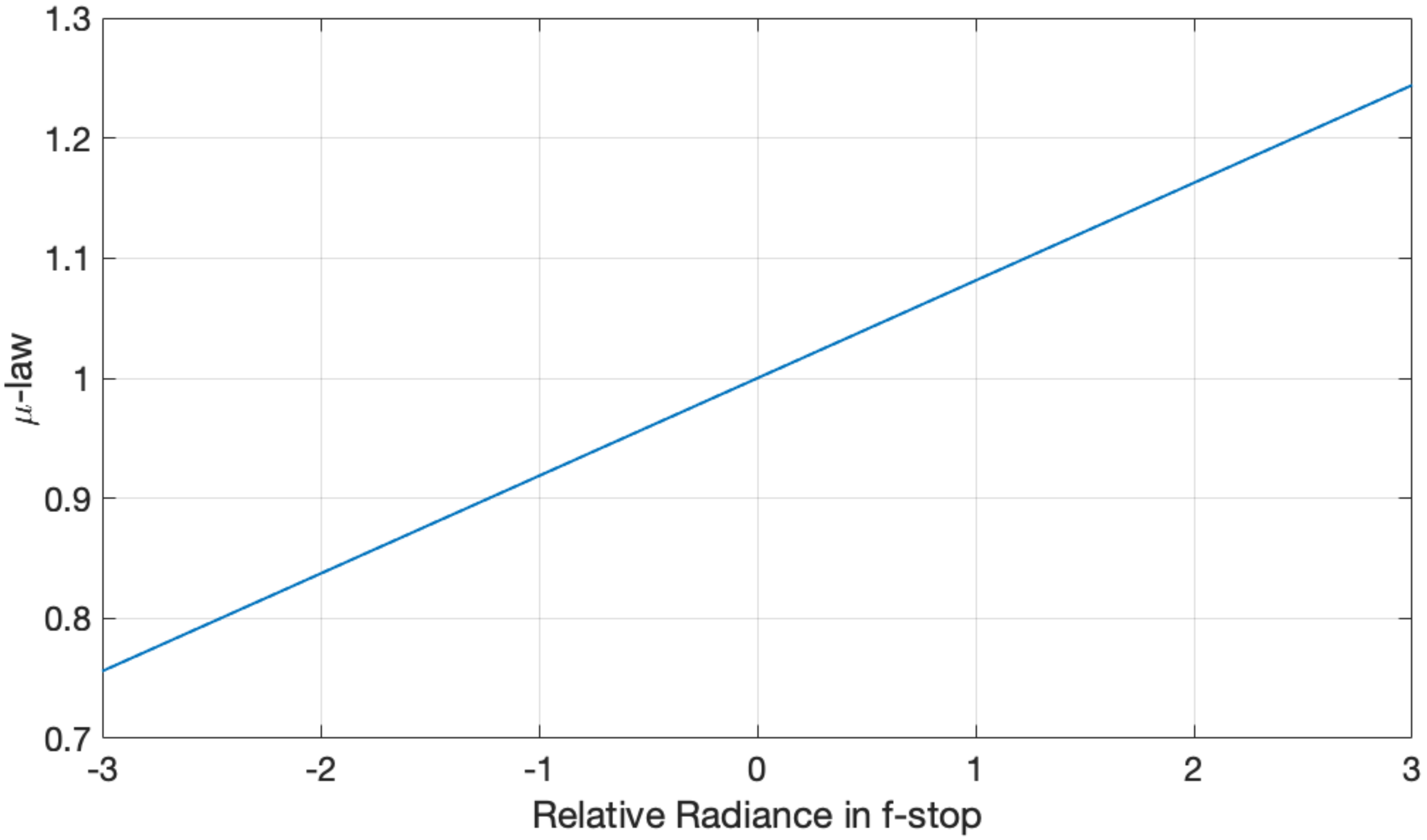
- Li et al. 2022 has a GAN scheme with a content/reconstruction loss:

$$\mathscr{L}_{\text{rec}} = \mathbb{E}_{x \sim p_{\text{data}}(x)}\left[\left\| VGG(G(x)) - VGG(x_2) \right\|_1\right]$$

- The GAN loss is defined as:

$$\mathscr{L}_{\text{GAN}} = \mathbb{E}_{y \sim p_{\text{data}}(y)}[\log D(y)] + \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log 1 - D(G(y))] + \mathbb{E}_{b \sim p_{\text{data}}(b)}[\log(1 - D(b))]$$

Loss Function in the μ-Law Domain



HDR Videos

HDR Videos: Temporally Varying Exposure Time

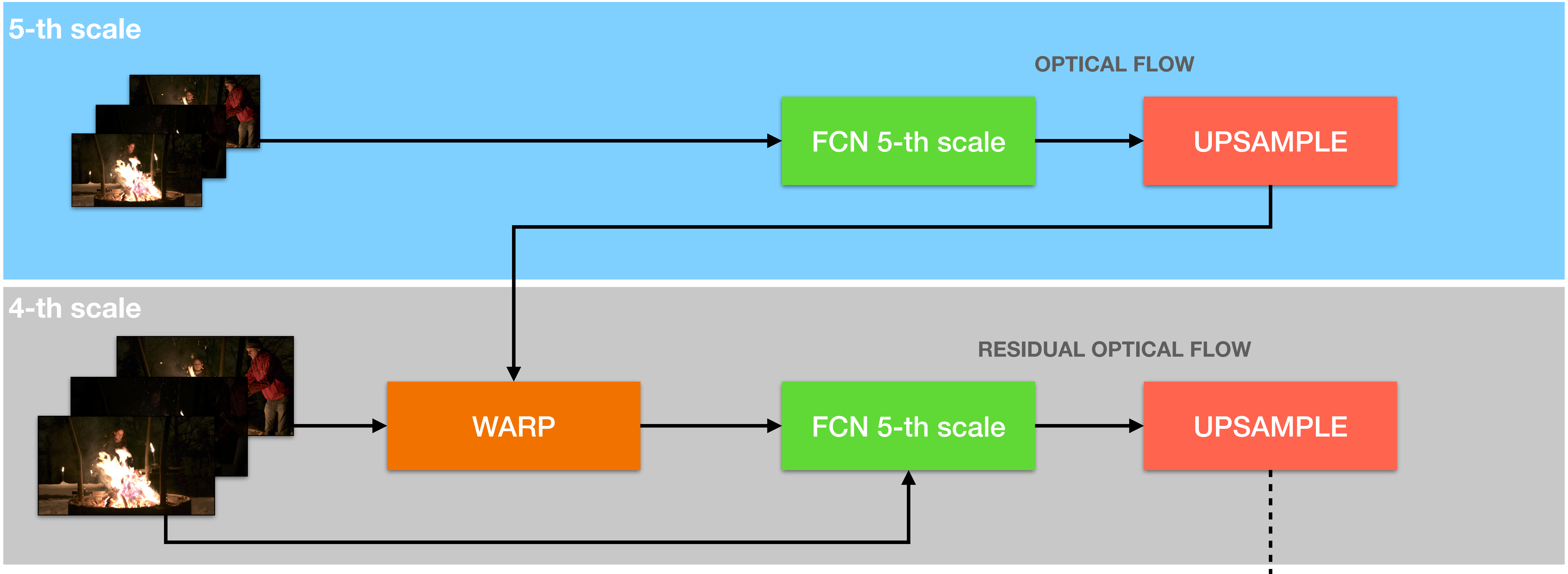
Stream

t_0 t_1 t_2

Video Courtesy of Jan Fröhlich - Stuttgart HDR Video Dataset

Video Strategies: Kalantari and Ramamoorthi 2019

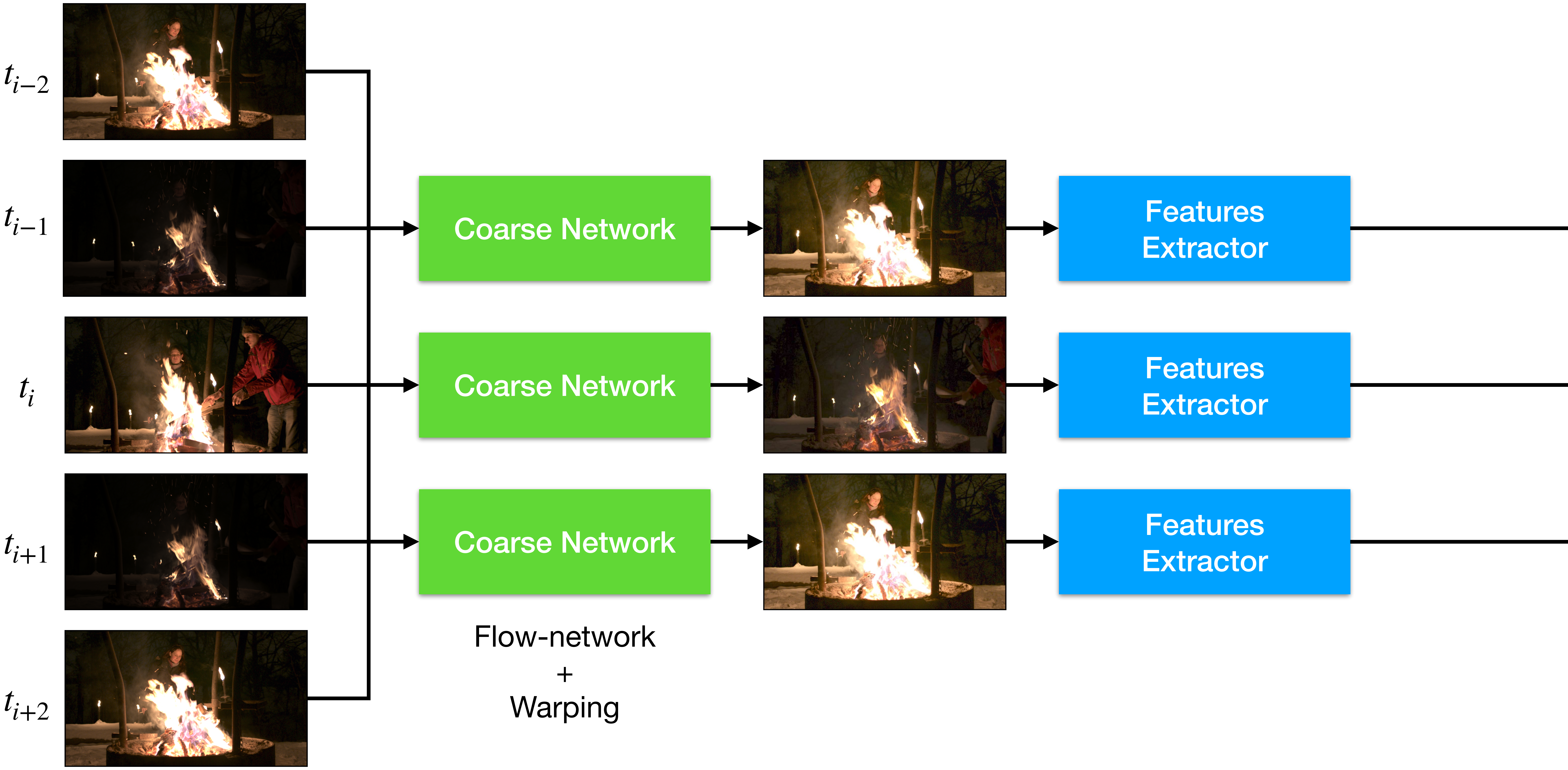
- A 5-scale pyramid for computing a multi-scale optical flow using a CNN for each scale a simple FCN:



Video Strategies: Kalantari and Ramamoorthi 2019

- Similar to the previous work by Kalantari et al. 2017, there is a merger (encoder-decoder).
- To enforce temporal coherency and reduce artifacts the merger uses neighbors frames at previous and next time.

Video Strategies: Chen et al. 2021



Video Strategies: Chen et al. 2021

Evaluation

Metrics

- Many works uses:
  - Linear domain PSNR and SSIM.
  - μ-law or Reinhard et al. 2002's TMO PSNR or SSIM
- These approaches have many issues:
  - Linear domain PSNR and SSIM are prone to outliers.
  - μ-law and Reinhard et al. 2002's TMO are empirical approaches that do not model the Human Visual System.
    - They may introduce distortions.

Metrics

- PSNR and SSIM should be computed using the PU21:
  - PU21 encodes absolute HDR linear value into approximately perceptually uniform (PU) values.
- HDR-VDP 2.2, and HDR-VDP 3.0.6.
- Deghosting artifacts: Tursun et al. 2016.
- Note that many HDR reference images and output images are **uncalibrated**:
  - If we do not have calibration data:
    - Display-referred values.

Limitations

Limitations

- The CRF needs to be known (a partial limitation);
- Most methods are limited to merge ONLY three images:
  - There is no method addressing an arbitrary number of images or more than threes.
- The difference in f-stop has to be fixed:
  - There is no method that can merge an image at -5-stop, 0-stop, and +1-stop.

Other Problems in Reconstruction

Other Reconstruction Problems

- We have other problems for HDR reconstruction with partial real information that can be solved using deep learning:
  - Assorted pixels/rows [Choi et al. 2017, Çogolan et al. 2020, Suda et al. 2020, Xu et al. 2021, Vien et al. 2022].
  - HDR from deep optics/masks [Alghamdi et al. 2019, Metzler et al. 2020]
  - HDR reconstruction using an event camera [Wang et al. 2019, Shaw et al. 2022, Messikommer et al. 2022].
  - HDR reconstruction for quanta sensors [Gnanasambandam et al. 2020, Gao et al. 2022].

Questions?