Modern High Dynamic Range Imaging at the Time of Deep Learning

Inverse Tone Mapping

Francesco Banterle and Alessandro Artusi

Introduction

- Acquisition is tedious:
  - Images alignment.
  - Ghosts removal.
- What can we do without bracketing or modified/expensive hardware?

Introduction

- Acquisition is tedious:
  - Images alignment.
  - Ghosts removal.
- What can we do without bracketing or modified/expensive hardware?

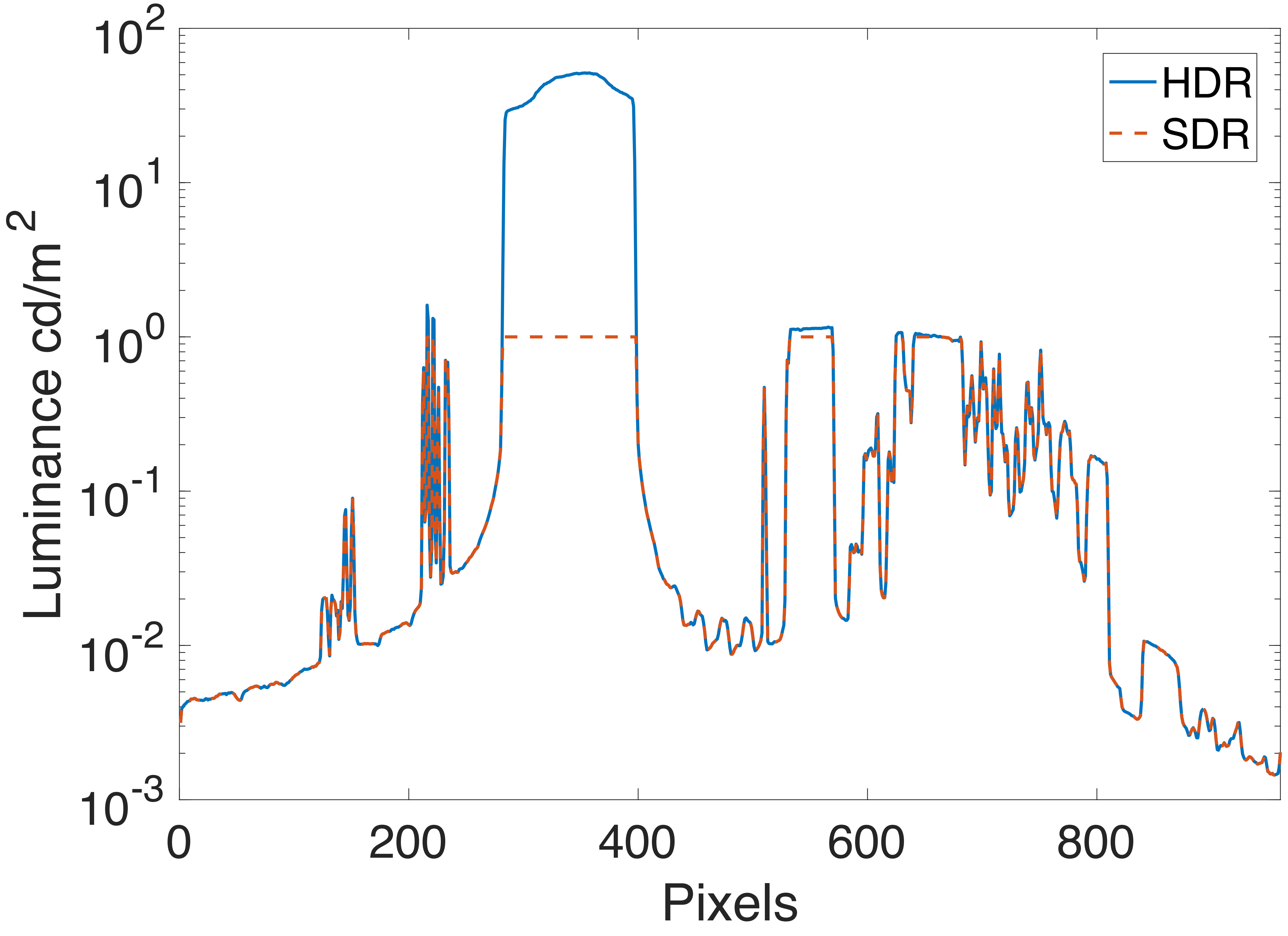
Introduction

- Acquisition is tedious:
  - Images alignment.
  - Ghosts removal.

- What can we do without bracketing or modified/expensive hardware?

The Problem

Image

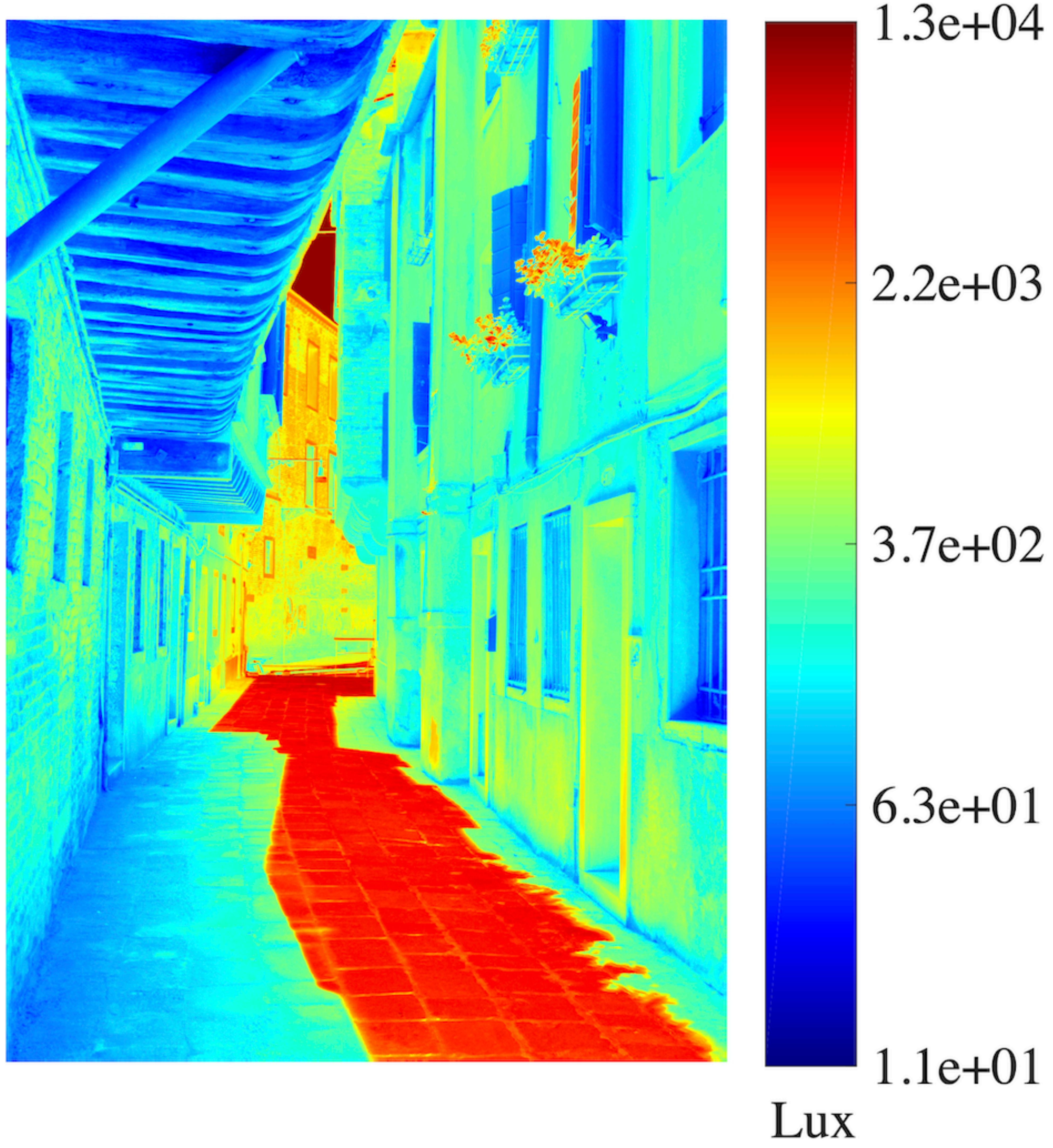


Histogram of the red dotted line

The Problem

SDR Image

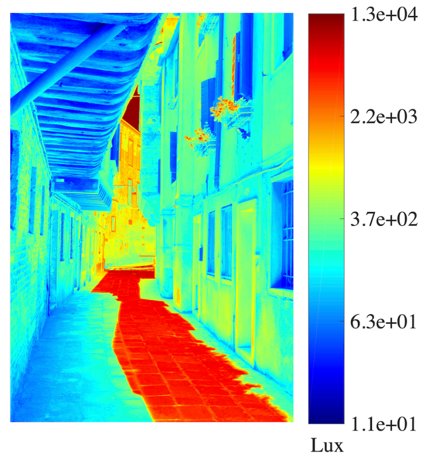
ITMO



HDR Image

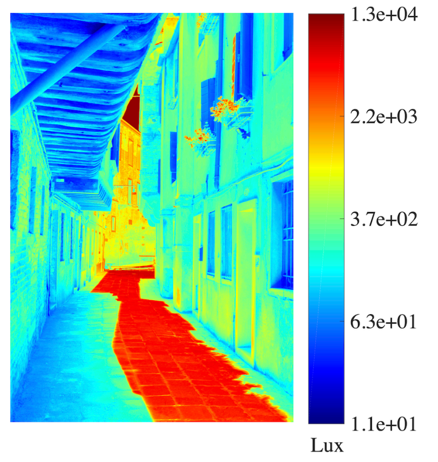
The Full Pipeline

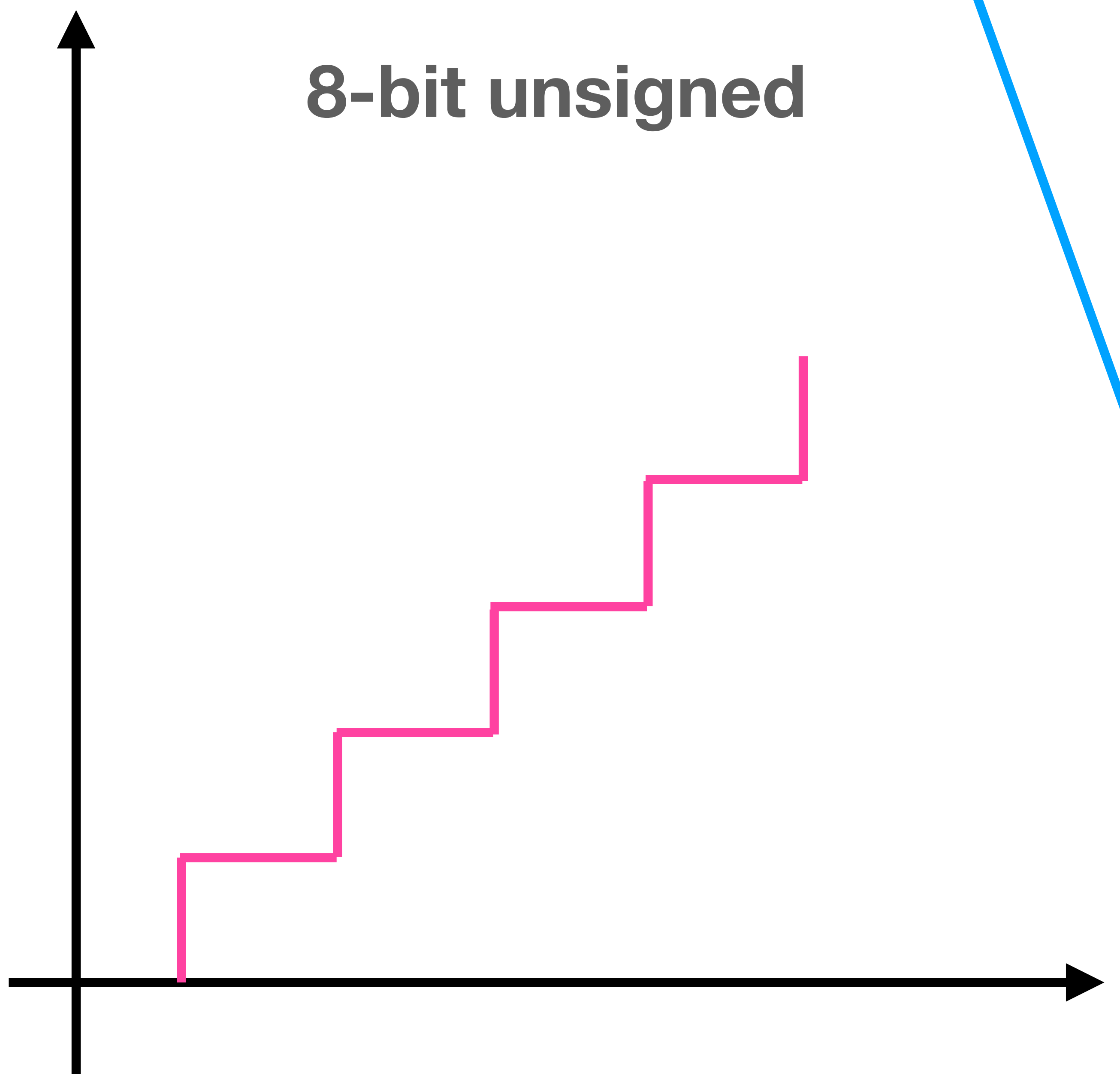
Dequantization → Linearization → Hallucination →



The Full Pipeline

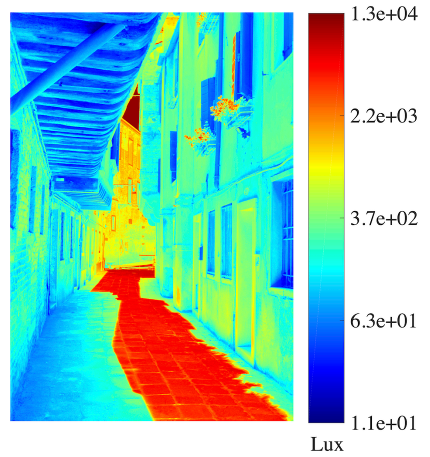
Dequantization → Linearization → Hallucination

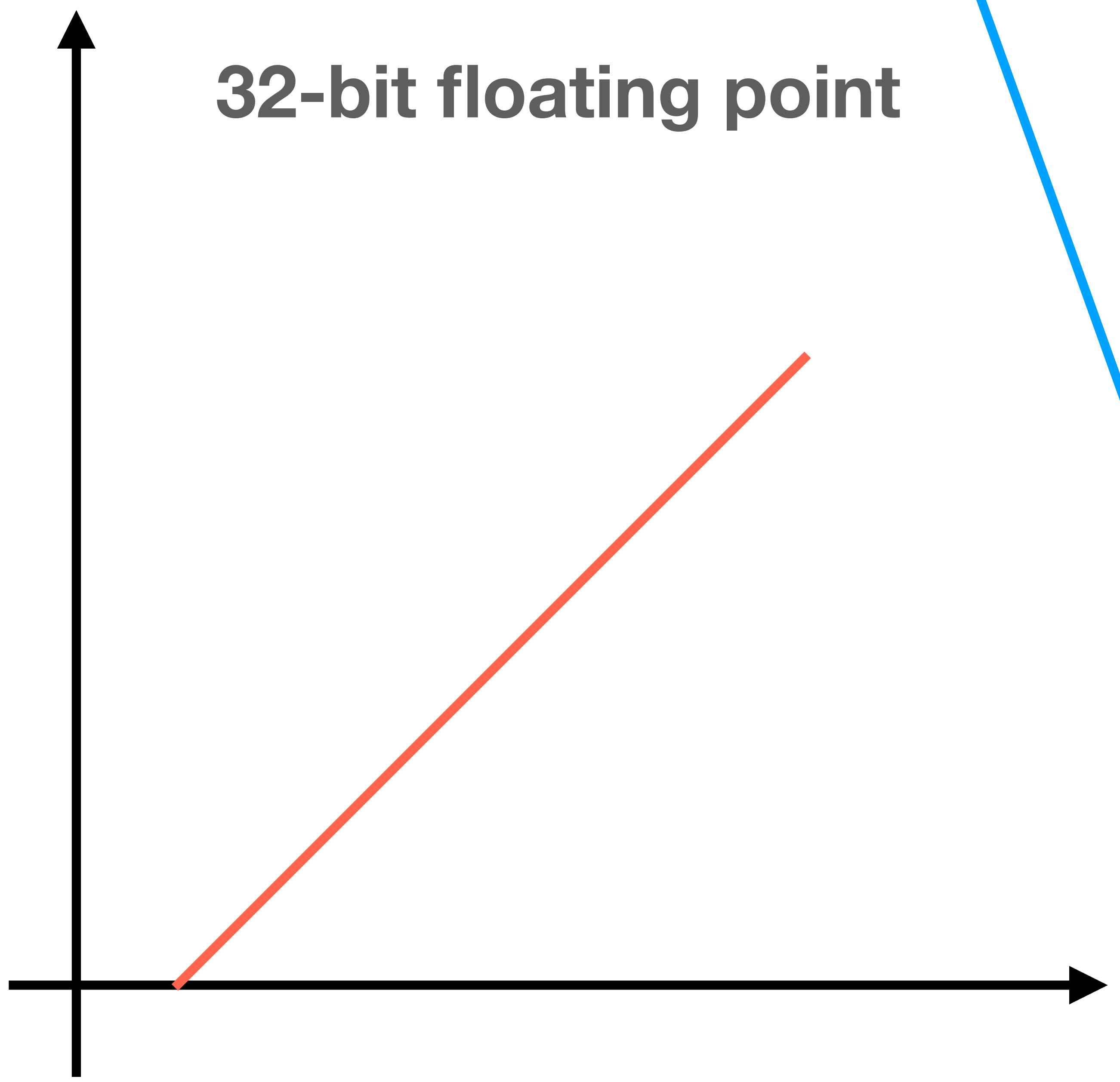




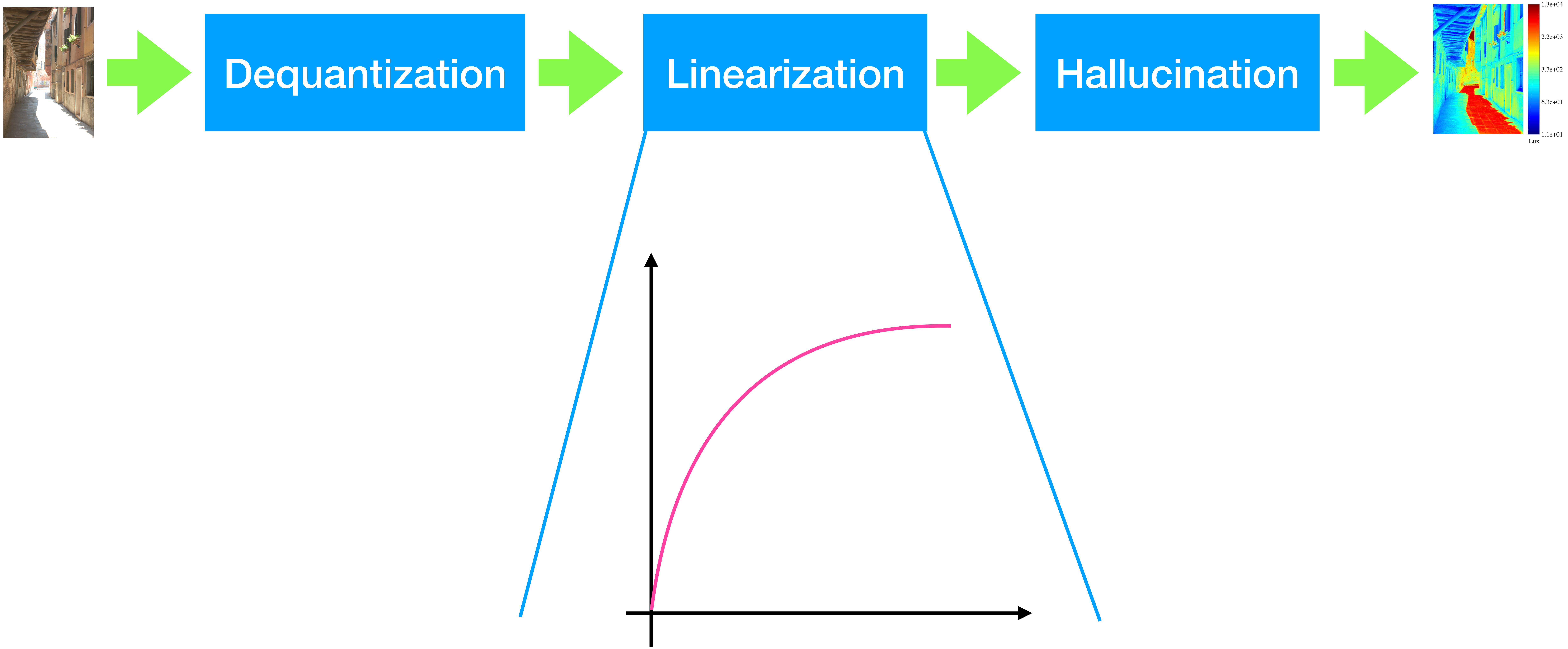
The Full Pipeline

Dequantization → Linearization → Hallucination →



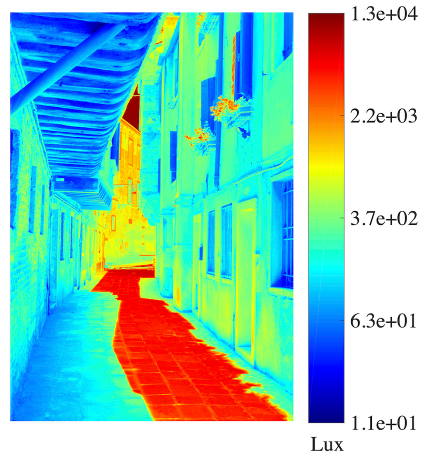


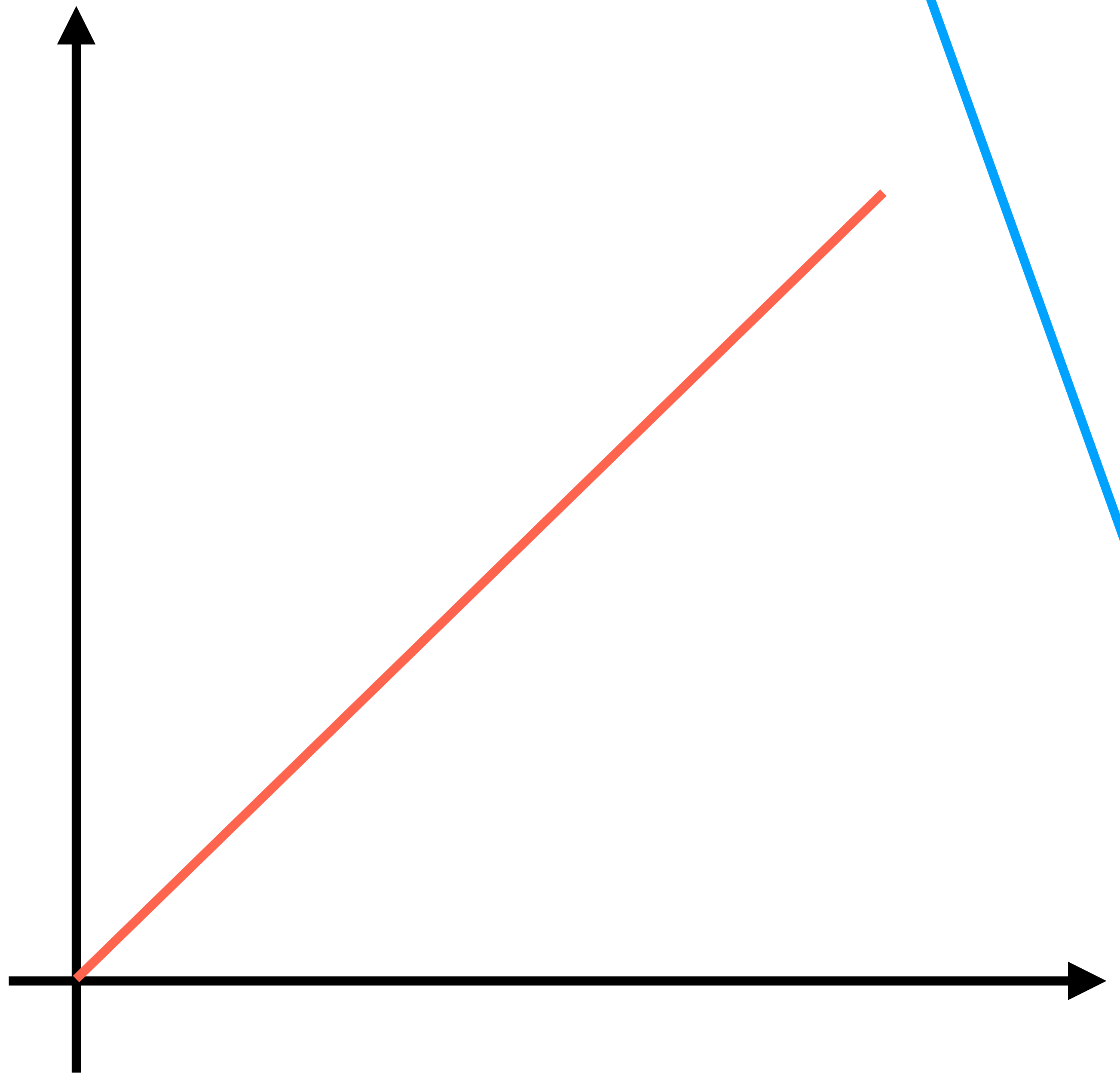
The Full Pipeline



The Full Pipeline

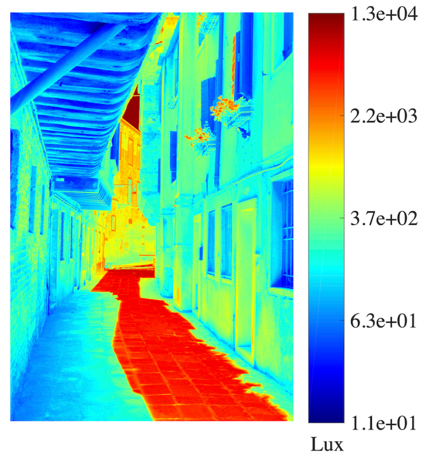
Dequantization → Linearization → Hallucination →





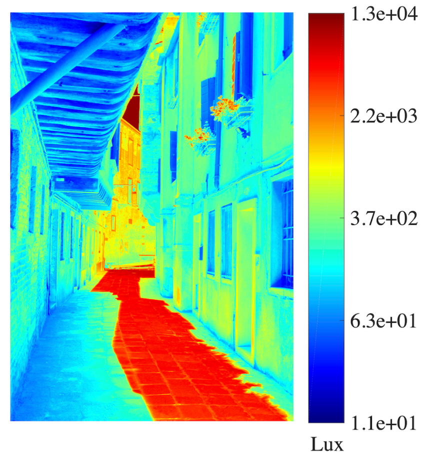
The Full Pipeline

Dequantization → Linearization → Hallucination



The Full Pipeline

Dequantization → Linearization → Hallucination →



The Linearization Dilemma

The Linarization Dilemma

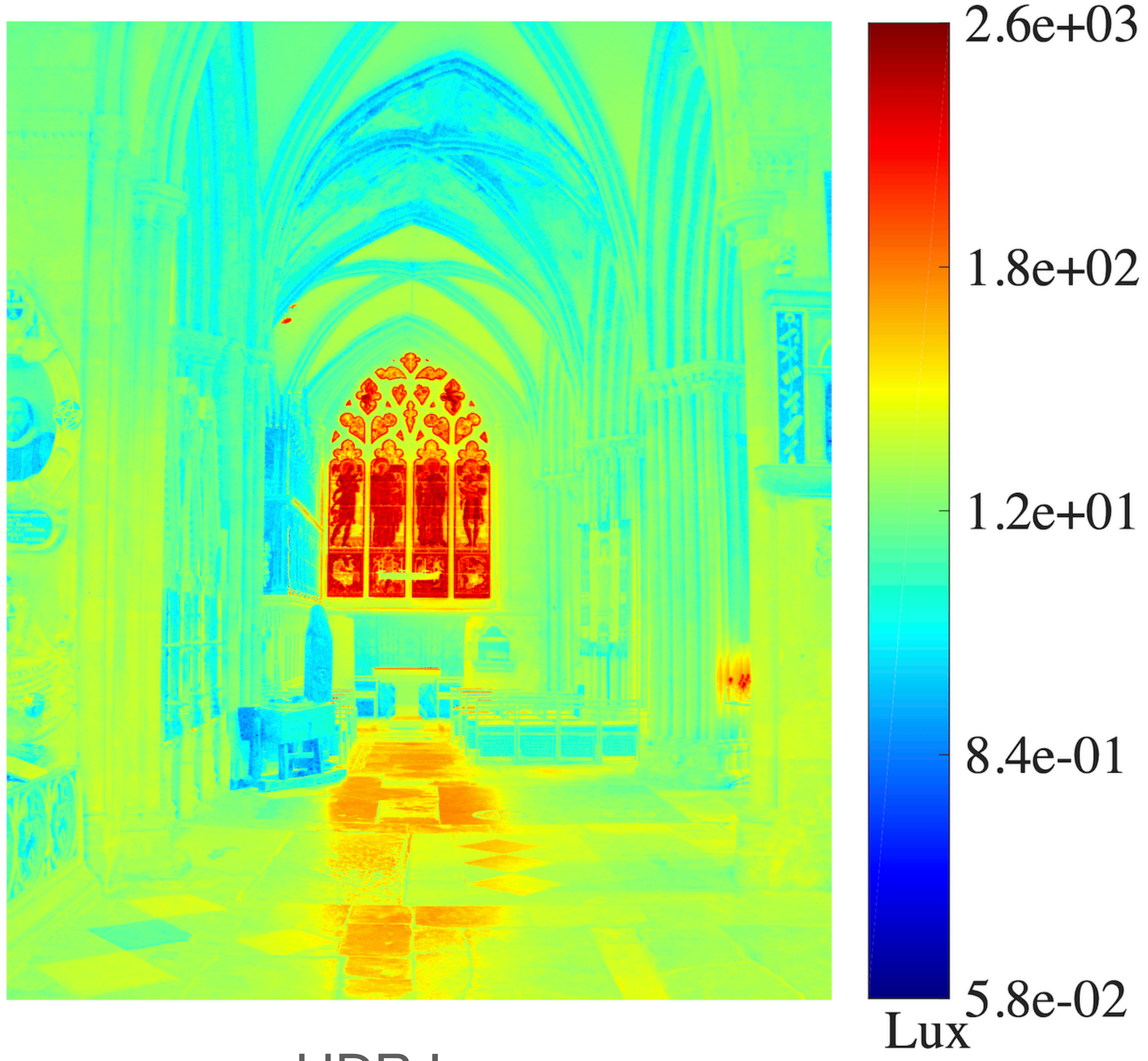
- One of the first step to decide is how we linearize the input SDR image.
- Many methods uses a standard $\gamma = 2$ or $\gamma = 2.2$:
  - Eilertsen et al. 2017, Marnerides et al. 2018, etc.
- Note that many modern cameras encode images using common CRF such as sRGB, PQ, and HLG.

Architectures

Architectures

- Here, we have **two possibilities** to solve the problem:
    - Approach 1: Given an input image, we generate directly a HDR image

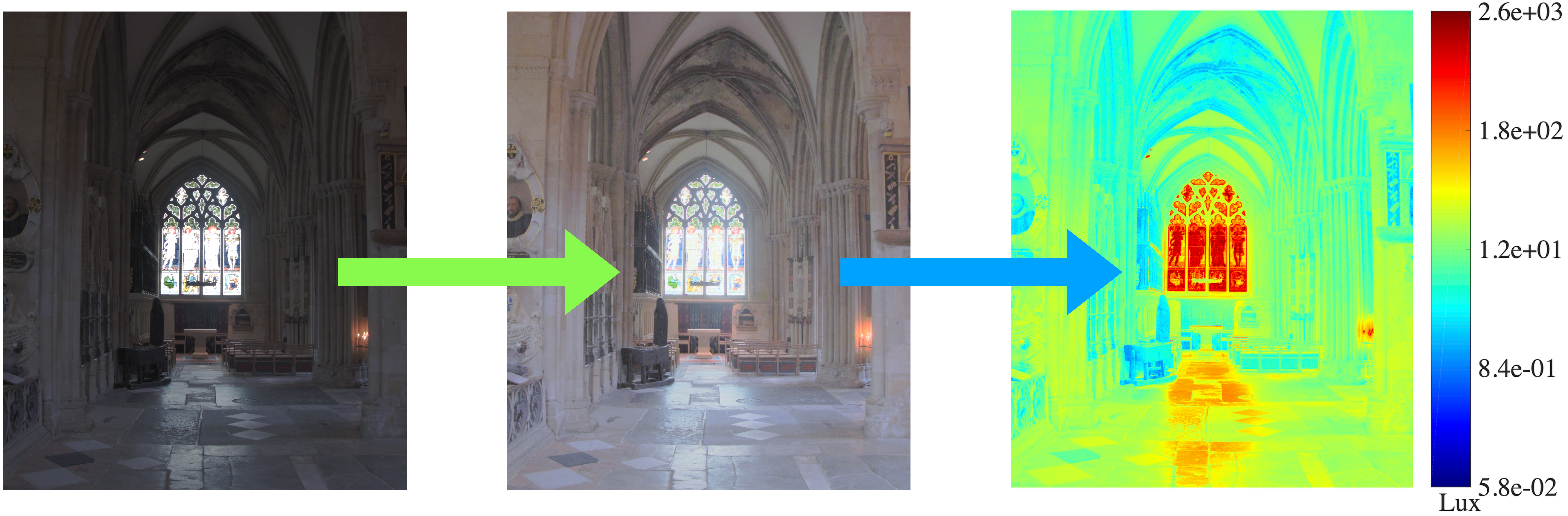
SDR Image



HDR Image

Architectures

- This approach may also compute a tone mapped version of the radiance map to recover. If the tone mapper is invertible, we can obtain a radiance map.



SDR Image   Tone Mapped HDR Image   HDR Image

Architectures

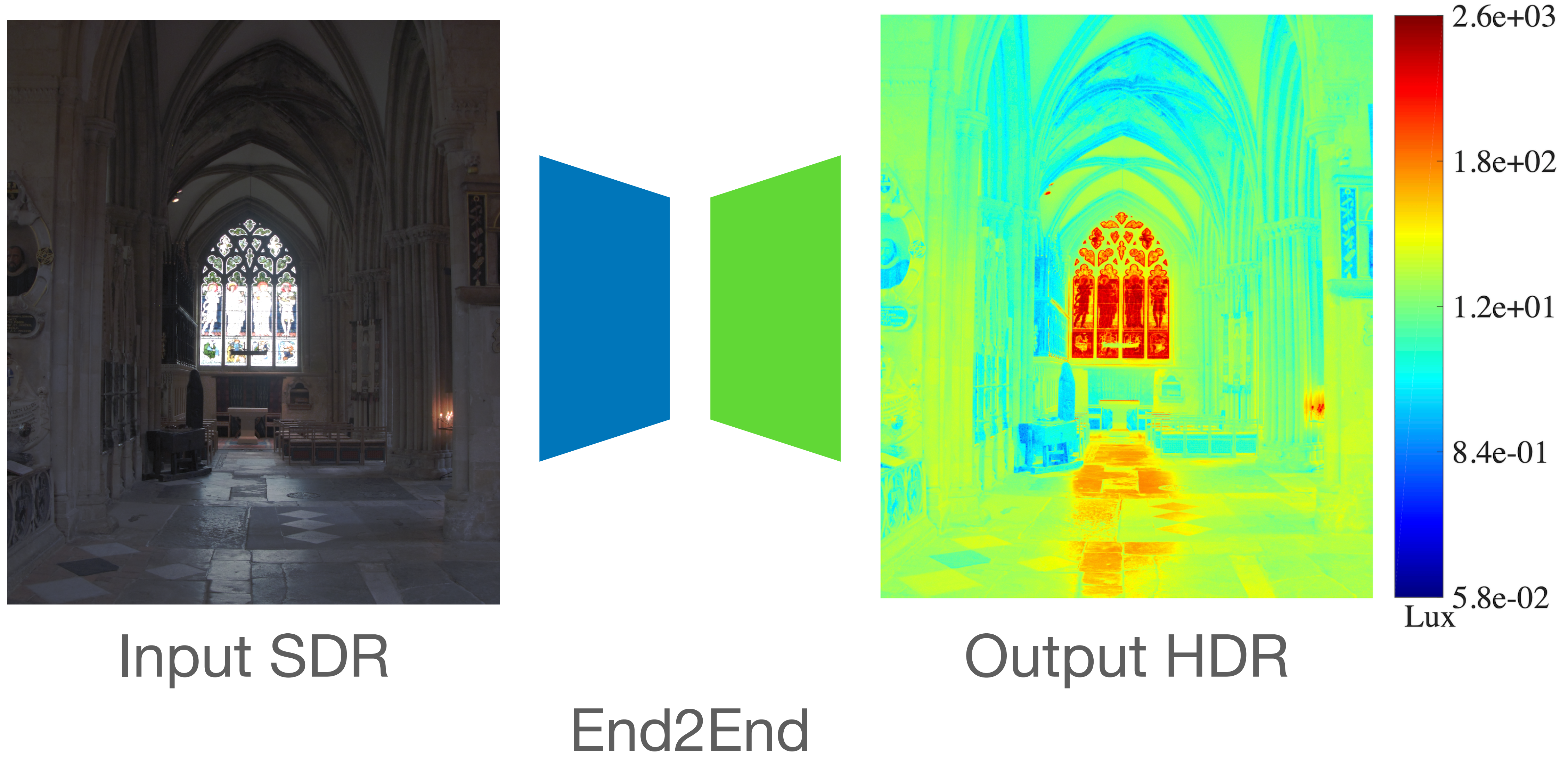
- Another possibility is:
    - Approach 2: Given an input SDR image, we generate a stack of n SDR images at different exposure times.

SDR Image

HDR Image

Which Architecture?

- The bread and butter of most iTMO are:
  - FCN.
  - U-Net [Eilertsen et al 2017].
  - Residual Blocks [Kim et al. 2019].
- They are simple models that generally works.

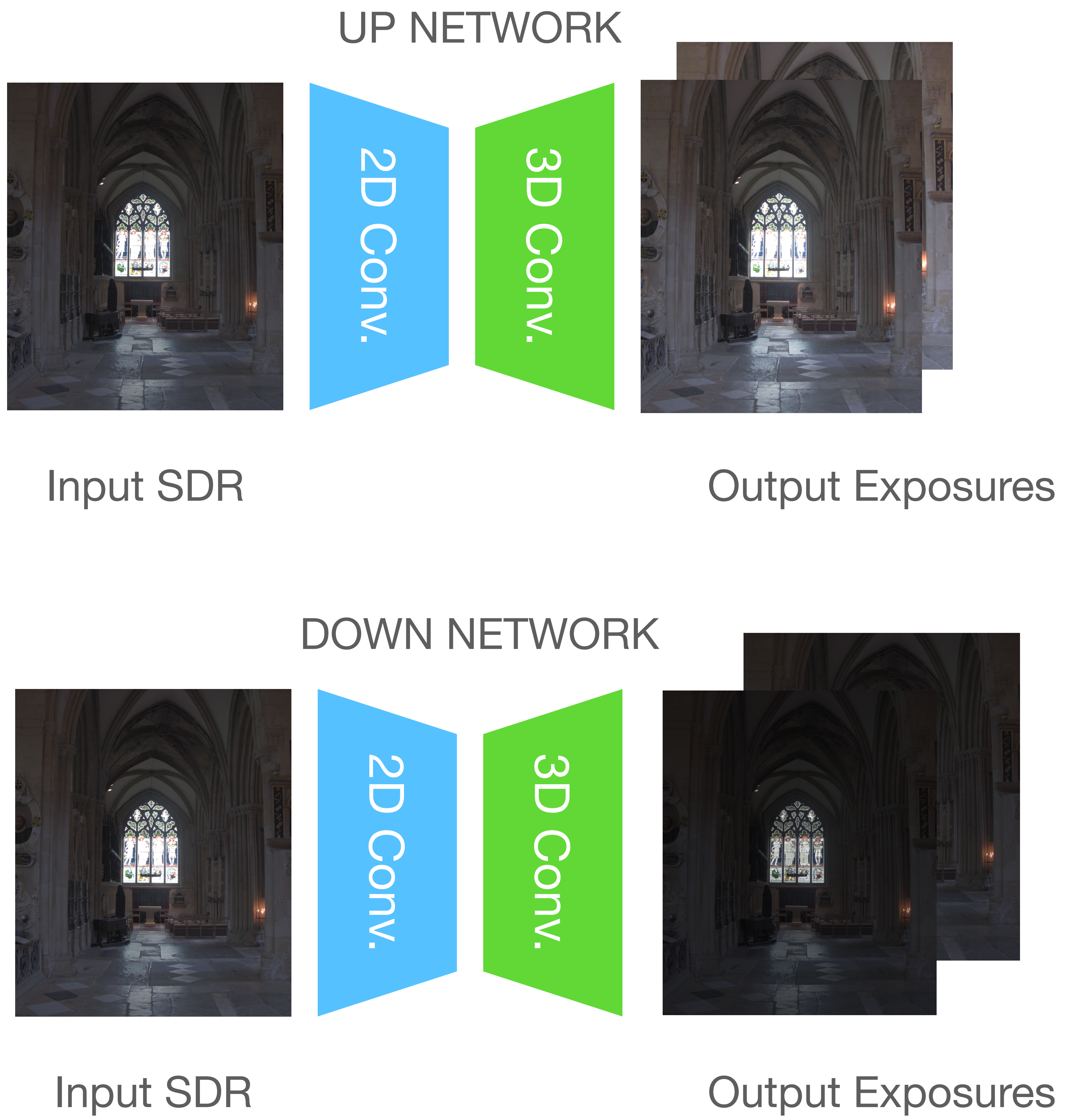


Which Architecture?

- Activation function:
  - LeakyReLU/GeLU in the encoder part.
  - ReLU in the decoder part.
  - The last layer:
    - Sigmoid: tone mapped results or single exposures.

Which Architecture?

- Endo et al. 2017 employs a classic U-Net with a twist:
  - Encoder has 2D convolutions.
  - Decoders has 3D convolutions:
    - Generate in a single network all exposures.
    - Limitations: the number of exposures are limited.



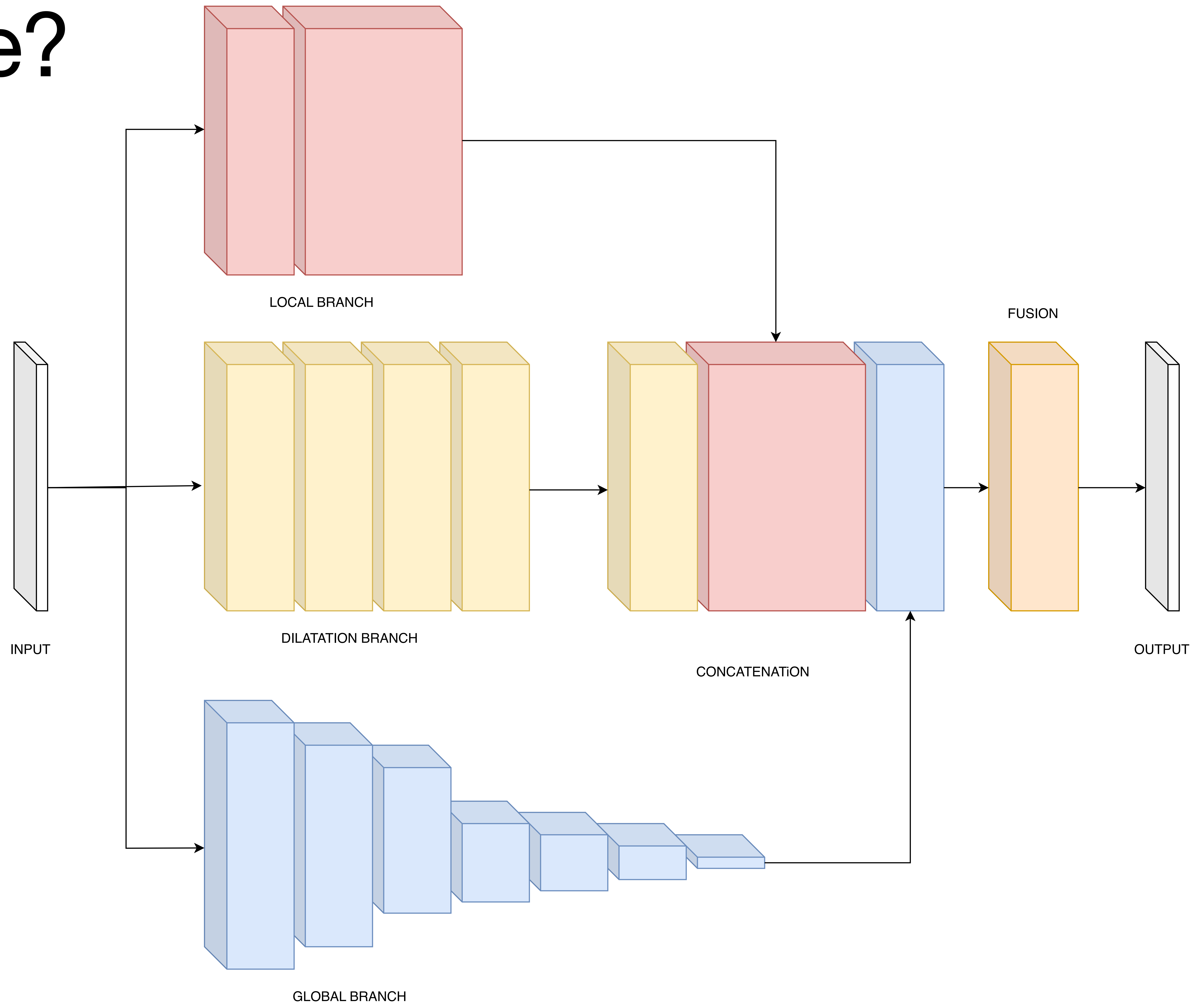
Which Architecture?

- Marnerides et al. 2018 proposed a multi-branch architecture to overcome U-Net limits (i.e., blocking artifacts):

    - Local features;

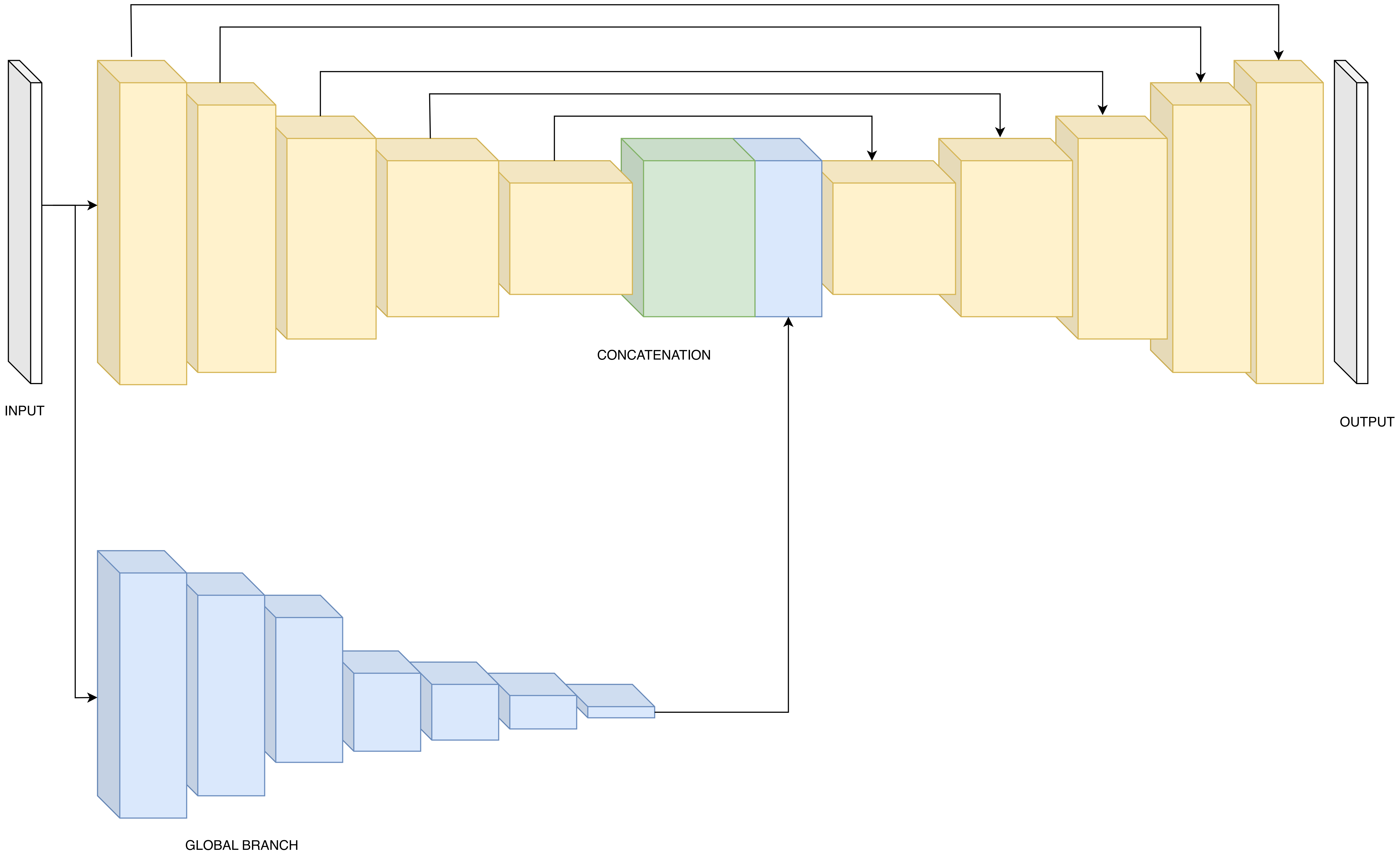
    - Medium features;

    - Global features.



Which Architecture?

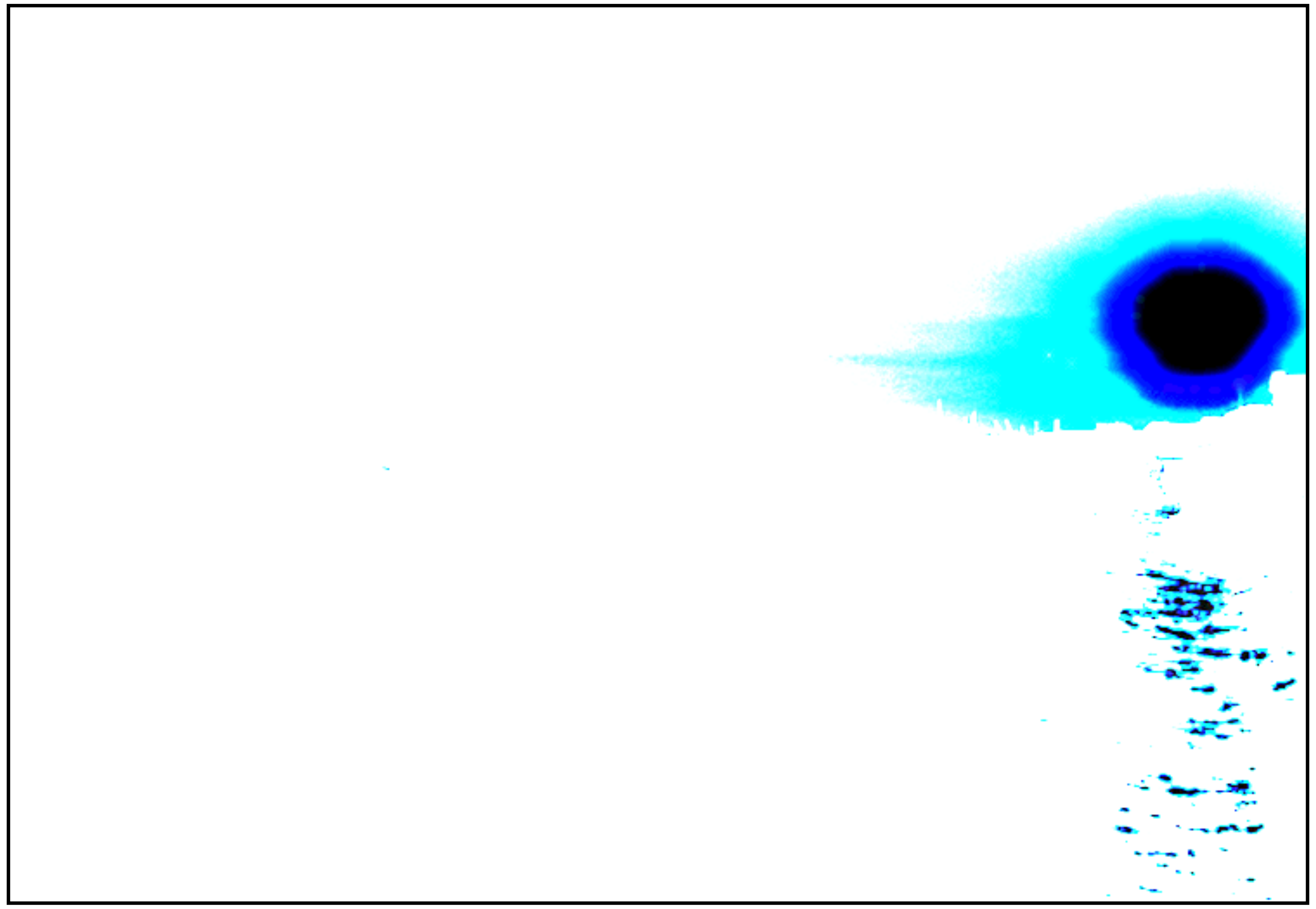
- Kinoshita and Kiya 2019 paired the global branch with U-Net to solve similar issues of Marnerides et al. 2018
- This network is trained on tone mapped images.



Which Architecture? Feature Masking

- Santos et al. 2020 introduces masking:
    - We can see inverse tone mapping as an inpainting problem, where our mask is defined using over-exposed pixels.

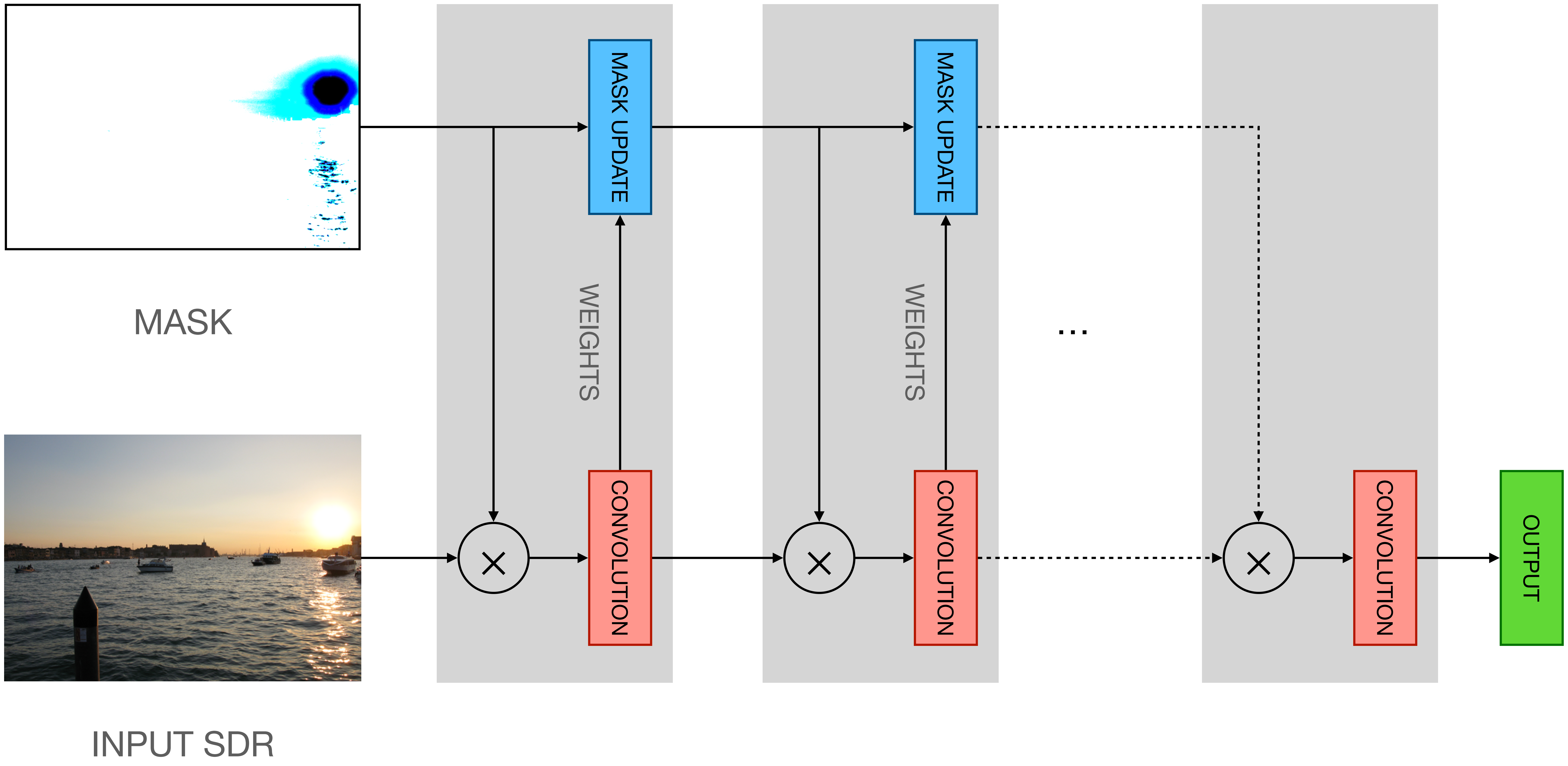
INPUT SDR



MASK

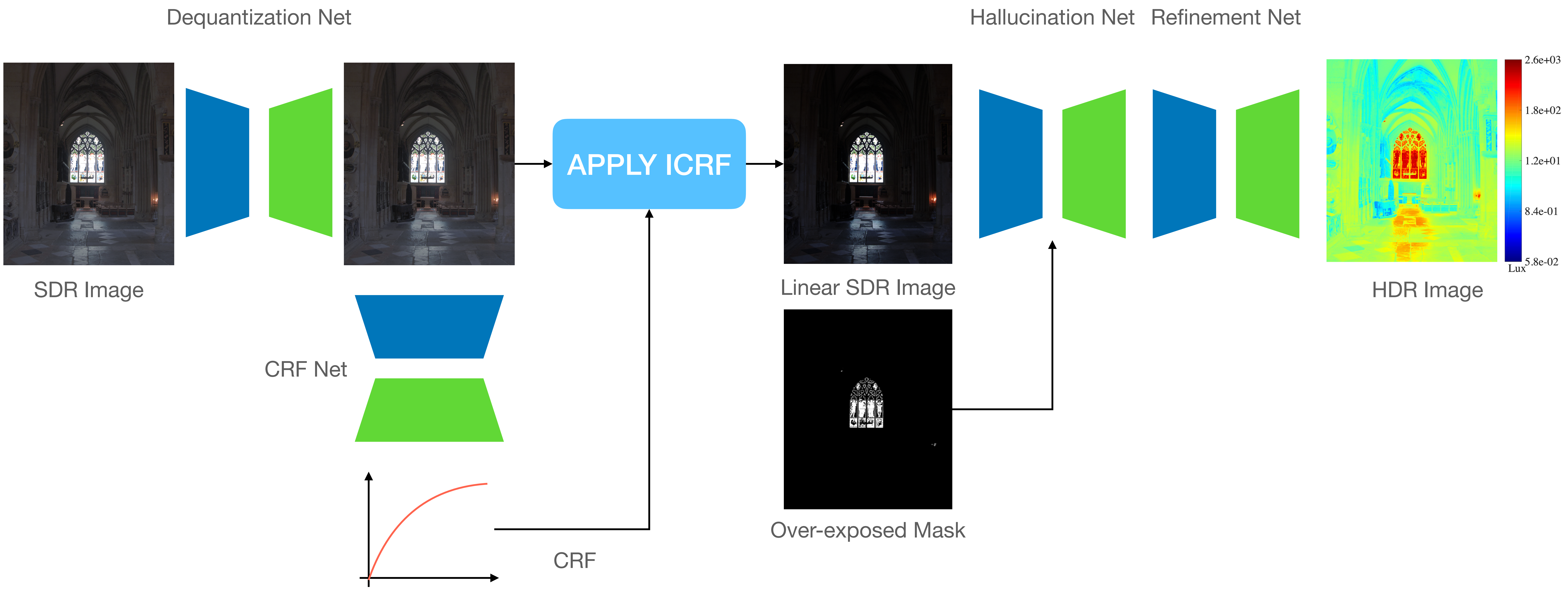
Which Architecture? Feature Masking

- Santos et al. 2020 apply the mask at each convolution step:



Which Architecture? Feature Masking

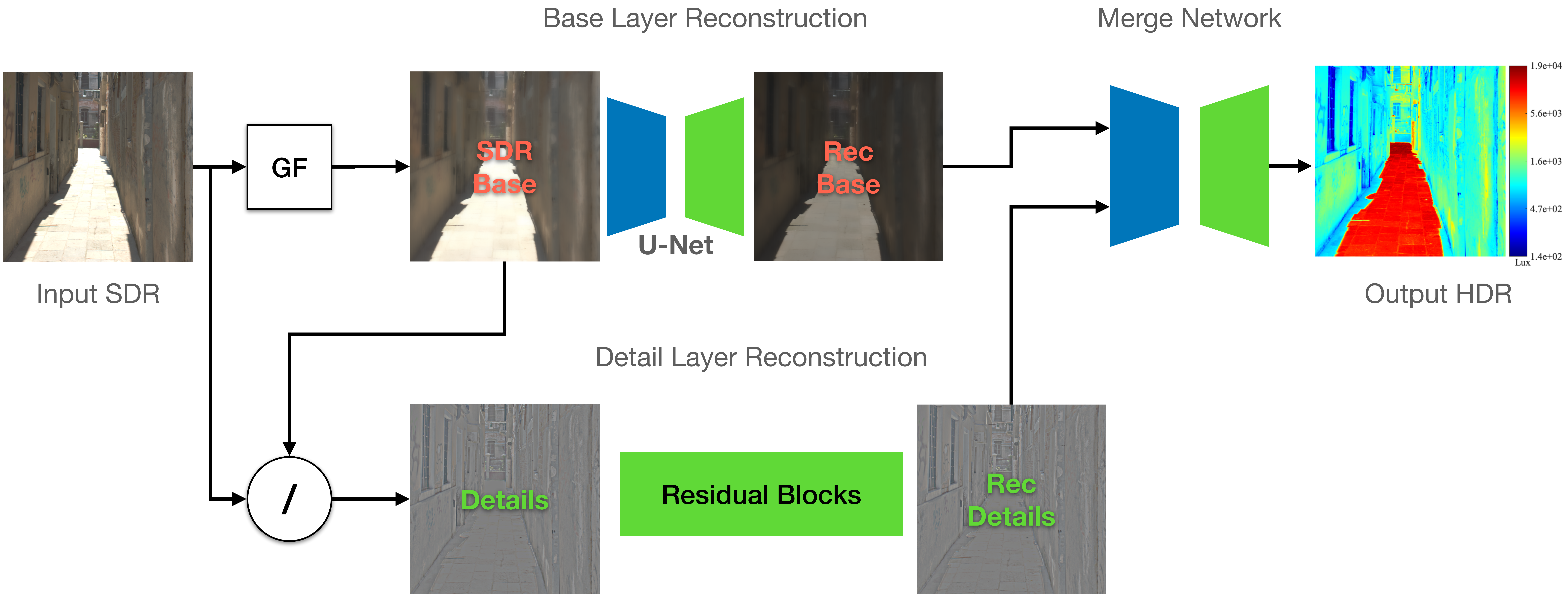
- Liu et al. 2020 has a network that recovers the inverse camera pipeline:



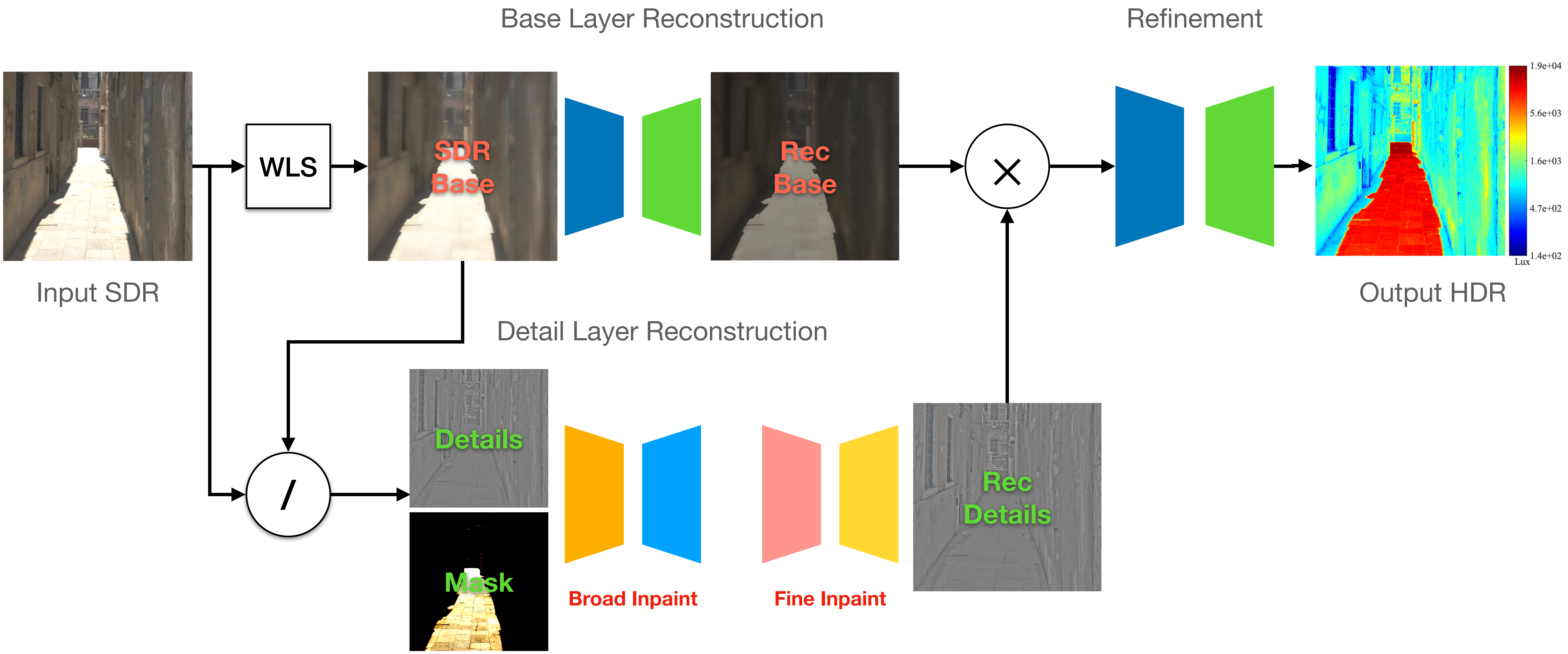
Which Architecture? Frequencies Separation

- Adopted a classic end2end encoding paired with a GAN, so nothing special right now… The novelty:
  - A network for each frequency:
    - Base image or I_b: is the output of filtering the input image, I, filtered using an edge-aware filter:
      - Bilateral Filter, Guided Filter, WLS, etc.
    - Detail image or I_d: is an image encoding the high-frequency details, and it is computed as: $I_d = I / I_b$.
  - A similar work with more refinement networks was proposed by Zhang and Aydın 2021 using WLS instead of the bilateral filter.

Which Architecture? Frequencies Separation - Wang et al. 2019



Which Architecture? Frequencies Separation - Zhang and Aydın 2021



Datasets

HDR Image Datasets

- Proper HDR images/videos (≥ 18-stop) are scarce on the Internet.
- There are few datasets of real HDR images.
- These datasets are typically uncalibrated:
  - This means that luminance values are relative; i.e., they do not have absolute values in cd/m^2.
  - Colors may not match the real colors.
- They are stored in different formats without the use of a standard. Typically, using the Radiance (.hdr) or OpenEXR (.exr) format files.

HDR Image Datasets

Dataset Name	#Images	#Resolution	Calibrated	Website
HDR Survey	108	5MPix	Scene-referred	http://markfairchild.org/HDR.html
HDR Eye	47	2MPix (full-HD)	Display-referred	
Stanford HDR Dataset	88	0.32Mpix	Scene-referred	https://qualinet.github.io/databases/image/high_dynamic_range_imaging_dataset_of_natural_scenes/
Laval HDR Indoor	2100	2MPix (2:1 ratio)	Relative values	http://indoor.hdrdb.com/
Laval HDR Outdoor	205	2Mpix (2:1 ratio)	Relative values	http://outdoor.hdrdb.com/
Akyuz HDR Images	10	5MPix	Relative values	https://user.ceng.metu.edu.tr/~akyuz/hdrdisp_eval/hdrdisp_project.html
Debevec HDR Images	21	0.3-2Mpix	Relative values	https://www.pauldebevec.com/
MPI HDR Images	7	3MPix	Scene-referred	https://resources.mpi-inf.mpg.de/hdr/gallery.html
Classic HDR Images	10	<1Mpix	Relative values	https://www.cs.huji.ac.il/w~danix/hdr/results.html
Funt HDR Dataset	105	3Mpix	Scene-referred	https://www2.cs.sfu.ca/~colour/data/funt_hdr/

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

HDR Content Datasets

- Are these tables complete?
  - No, they are not.
- There are more datasets, but it can happen they may be not be available for some time. For example:
  - LiU HDR Video Dataset: high-quality dataset that is not currently available on the web.
  - MPEG HDR Video Dataset: not freely available.
  - …

Augmentation Strategies

- Classic flips and rotations;
- Cropping from high-resolution images;
- Channel swapping [Kalantari et al. 2017]:
  - RGB channels are randomly swapped;

Creating Images for Training

- The training dataset:
    - <Input SDR, Output HDR>
- How do we compute the input?

$$Z = f(E \cdot \delta t)$$

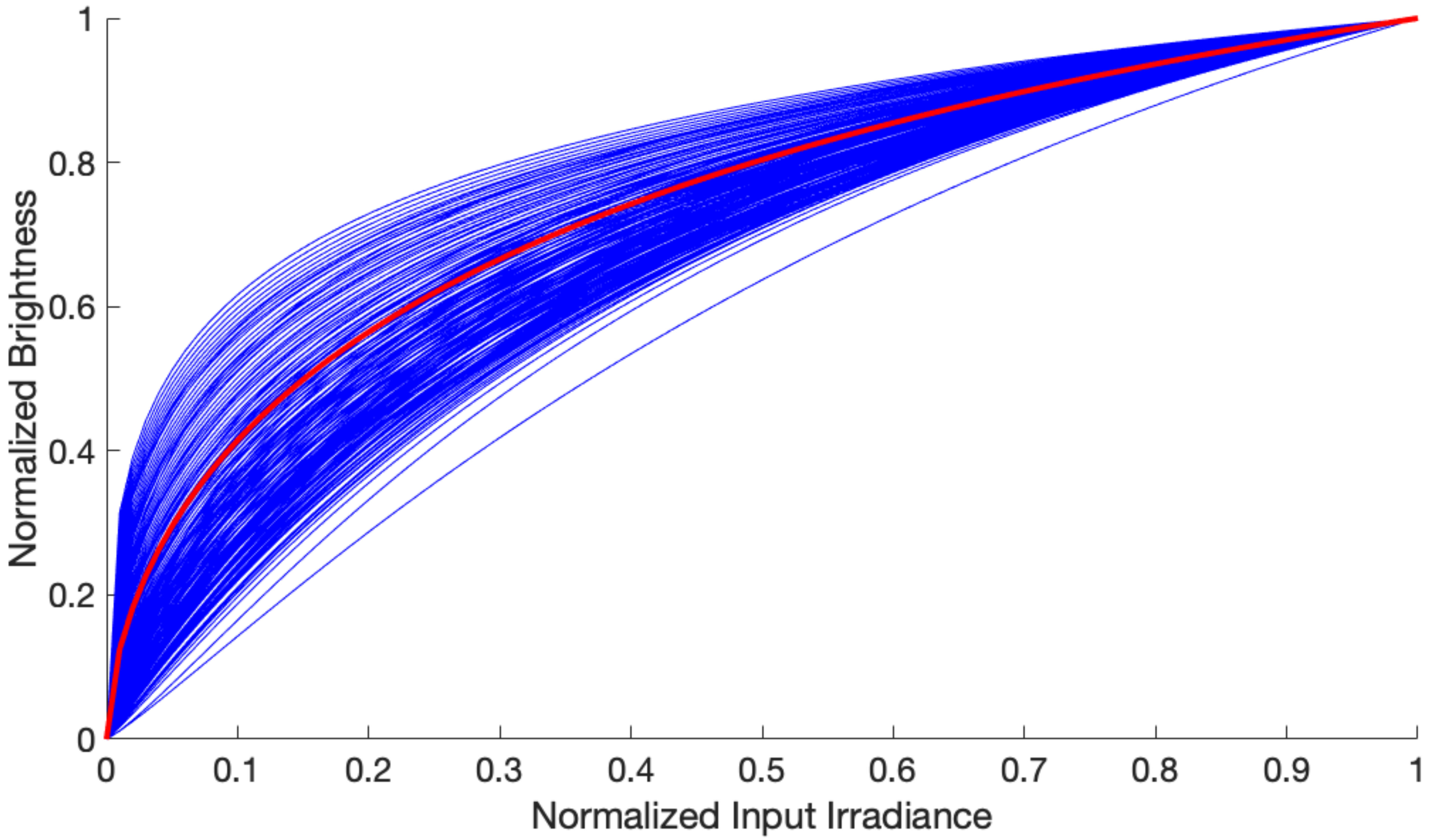
- δt is the virtual exposure value.
- $f(x)$ is the camera response function where the simplest to be used is:

$$f(x) = x^{\frac{1}{2.2}}$$

Creating Images for Training

- Many methods employs a random function from Grossberg and Nayar 2003 dataset of CRFs:
    - Eilertsen et al. 2017 showed that meaningful CRF can be modeled as:

$$f(x) = (1+\sigma) \cdot \frac{x^n}{x^n + \sigma} \qquad n \sim \mathcal{N}(0.9, 0.1) \quad \sigma \sim \mathcal{N}(0.6, 0.1)$$



Creating Images for Training

- δt is an important value to be picked up:
  - Its range is $[1/I_{\min}, 1/I_{\max}]$
- Automatic exposure:
  - $\delta t = \dfrac{1}{4I_{\text{mean}}}$
  - We pick the δt that maximizes the well-exposed pixels in the range $[0.05, 0.95]$:
    - We do not want too dark images.
    - We do not want too bright images.

Creating Images for Training

- We may perform a random augmentation:

$$\delta t \sim [1/I_{\min}, 1/I_{\max}]$$

- In this case, we need to skip extremely bright and dark images:
    - These are difficult cases.
    - We need a minimum of well-exposed pixels in order to draw something of meaningful from our methods:
        - 50-75% of well-exposed pixels:
            - Half/Quarter of the image totally white or totally black.

Selecting Patches

- Eilertsen et al. 2017:
  - For each HDR, 10 patches are selected at 320×320 using random cropping.
    - Lee et al. uses random crops at 256×256
- Endo et al. 2017:
  - Images are downsampled at 512×512.
- Marnerides et al. 2018:
  - Random crop with Gaussian distribution (center image) at 384×384.
- Santos et al. 2020:
  - Selection of patches with texture; i.e., mean gradient of the detail layer over 0.85 (bilateral separation).

Training

The Loss Function

- Eilertsen et al. 2017:
  - MSE in the log domain.
  - We have a loss function for the luminance and the reflectance component:
    - Equal weight in the paper for both losses.
- Marnerides et al. 2018:
  - L1 + Cosine Loss (for colors in under-exposed areas):

$$\mathscr{L}_{\cos}(\hat{I}, I) = 1 - \frac{1}{N} \sum_{i,j} \frac{\hat{I}(i,j) \cdot I(i,j)}{\|\hat{I}(i,j)\|_2 \cdot \|I(i,j)\|_2} ,$$

where I is the reference image and $\hat{I}$ is the results of the network.

The Loss Function

- Lee et al. 2018 employs as content loss L_1 and classic GAN loss:

$$\mathcal{L}_{\text{GAN}}(D) = \frac{1}{2}\mathbb{E}_{x,y}[(D(y,x)-1)^2] + \frac{1}{2}\mathbb{E}_{x,z}[(D(G(y,z),x))^2]$$

$$\mathcal{L}_{\text{GAN}}(G) = \mathbb{E}_{x,z}[(D(G(y,z),x)-1)^2]$$

$$\mathcal{L}_{L_1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1]$$

- Wang et al. 2019, Santos et al. 2020, Liu et al. 2020 uses a perceptual loss (VGG network) together with L_1:

$$\mathcal{L}_P(I,\hat{I}) = \|\psi(I) - \psi(\hat{I})\|_2$$

- Liu et al. 2020 has a complex loss where the main contribution is the reconstruction loss (L_1) TV loss and a CRF loss (MSE)

HDR Videos

What's about video?

- There are many papers treating videos:
  - In many cases, these works on a single frame:
    - There is no temporal coherence mechanisms in place:
      - **Not working on multiple frames at the same time;**
      - **No temporal loss;**

What's about video?

- Why are these considered videos methods?
  - They use HDR10/HDR10+ video datasets with wide gamut (e.g., RECO2020 or REC2100 color space).
  - They output directly PQ/HLG values.
  - They work on YUV input values.

What's about video? Video Stabilization

- Eilertsen et al. 2019 showed how to make imaging method temporal coherent: colorization, inverse tone mapping etc.
- The key is the introduction of a new loss:

$$\mathcal{L}(I, \hat{I}) = \mathcal{L}_{\text{rec}}(I, \hat{I}) \cdot (1 - \alpha) + \alpha \mathcal{L}_{\text{reg}}(I, \hat{I})$$

  where $\alpha \in [0.85, 0.95]$.

- Given that it is difficult to have good video dataset, the idea is to approximate a "video movement" by a small Euclidian Transformation T, which can be: a translation, a rotation, and a scaling.

What's about video?

- If our network is $f(\cdot)$ and its input I_{in} we can define the regularization as:

$$\mathcal{L}_{\text{reg}}(I, \hat{I}) = \mathcal{L}_{\text{reg}}(I, f(I_{in})) = \left\| \boxed{\Big(f(T(I_{in})) - T(I)\Big)} - \boxed{\Big(f(I_{in}) - I\Big)} \right\|_2$$

The difference between ground-truth and the network results after T; i.e., the "next frame"

The difference between ground-truth and the network results.

- $T(\cdot)$ is a random transformation:
  - Translation $[-2,2]^2$ pixels;
  - Rotation $\pm 1°$;
  - Scaling $[0.97, 1.03]$;

Evaluation

Evaluation

- Main metrics recommended for evaluations are [Hanji et al. 2022]:
  - If we have a reference:
    - HDR-VDP 2.2, HDR-VDP 3.0.6, PU-VSI, and PU21-PSNR.
  - If we do not have a reference:
    - PU21-PIQE.
- To focus evaluation on the generated content, we should remove influence of the CRF. A possibility is to estimate the CRF using the reference (if available).

Future Directions

The Status

- Currently, 2-3 new methods appears every month on arXiv!
- Many works just get old or new datasets and they train the latest architecture on them:
  - Diffusion networks;
  - Transformers;
  - etc.

Promising Approaches

- The main limitations of doing HDR and especially inverse tone mapping is that datasets are very small:
  - There are a small amount of images achieving 20-stops.
  - The few datasets may disappear due to maintenance!

Promising Approaches

- On the other hand there are large datasets available online of SDR image that could be used to copy well-exposed data in over-exposed areas:
    - Banterle et al. 2021: unsupervised generation of HDR videos from SDR videos.
    - Wang et al. 2022: unsupervised generation of HDR images from SDR images.

Dark SDR Images

- Very dark images without over-exposed areas do not process the image:
    - There is no recover of the dark areas.
- When we have large over-exposed (e.g., 25% of over-exposed pixels) areas is a challenging case.

Dark SDR Images: Well-exposed Example

SDR Input 4% over-exposed pixels

Dark SDR Images: Well-exposed Example

Ground Truth -1-stop

Eilertsen et al. 2017 -1-stop

Dark SDR Images: Well-exposed Example

Ground Truth -1-stop

Santos et al. 2020 -1-stop

Dark SDR Images: Over-exposed Example

SDR Input 25% over-exposed pixels

Dark SDR Images: Over-exposed Example

Ground Truth -1-stop

Eilertsen et al. 2017 -1-stop

Dark SDR Images: Over-exposed Example

Ground Truth -1-stop

Santos et al. 2020 -1-stop

Questions?