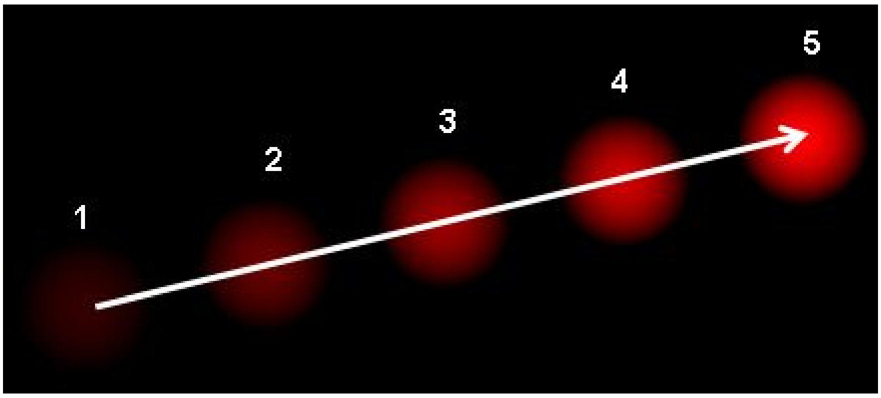
Deep High Dynamic Range Imaging with Large Foreground Motions

相关知识

> This paper proposes the first **non-flow-based** deep framework for high dynamic range (HDR) imaging of dynamic scenes with large-scale foreground motions. …In stark
> contrast to flow-based methods, we formulate HDR imaging as an image translation problem **without optical flows**.

Optical flows

光流是图像对象在两个连续帧之间由对象或相机的运动力矩引起的视在运动的模式。 它是2D向量场，其中每个向量都是位移向量，表示点从第一帧到第二帧的运动。



Optical flow works on several assumptions:

1. The pixel intensities of an object do not change between consecutive frames.
2. Neighbouring pixels have similar motion.

Consider a pixel $I(x,y,t)$ in first frame (Check a new dimension, time, is added here. Earlier we were working with images only, so no need of time). It moves by distance (dx, dy) in next frame taken after dt time. So since those pixels are the same and intensity does not change, we can say,

$$I(x,y,t) = I(x+dx, y+dy, t+dt)$$

Then take taylor series approximation of right-hand side, remove common terms and divide by dt to get the following equation:

$$f_x u + f_y v + f_t = 0$$

where:

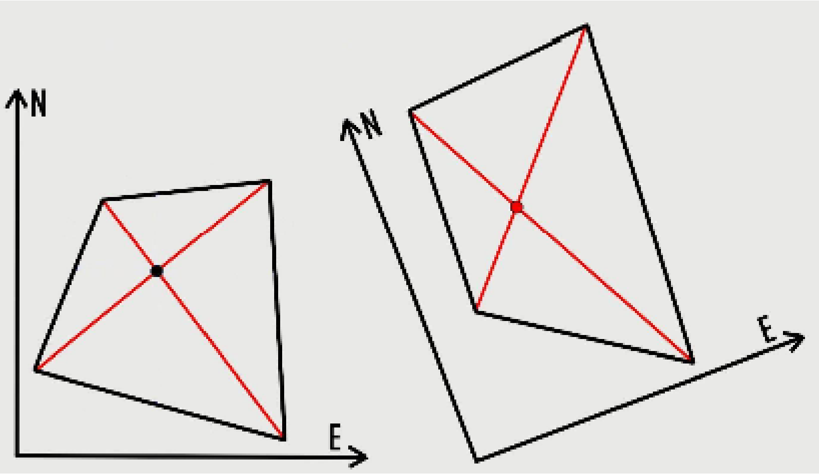
$$f_x = \frac{\partial f}{\partial x} \; ; \; f_y = \frac{\partial f}{\partial x}$$
$$u = \frac{dx}{dt} \; ; \; v = \frac{dy}{dt}$$

为了解出u,v,引出了Lucas-Kanade method：

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_i f_{x_i}{}^2 & \sum_i f_{x_i} f_{y_i} \\ \sum_i f_{x_i} f_{y_i} & \sum_i f_{y_i}{}^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i f_{x_i} f_{t_i} \\ -\sum_i f_{y_i} f_{t_i} \end{bmatrix}$$

> While the latter can be resolved to a large extent by **homography transformation** [26], foreground motions, on the other hand, will make the composition nontrivial.

homography transformation



> In [14], they first **used optical flow to align input LDR images**, followed by feeding the aligned LDRs into a convolutional neural network (CNN) to produce the final HDR image.

how to used optical flow to align input LDR images

> First, unlike [14], our network is trained **end-to-end** without optical flow alignment

What is an end-to-end network?

> 端到端指的是输入是原始数据，输出是最后的结果，原来输入端不是直接的原始数据，而是在原始数据中提取的特征，这一点在图像问题上尤为突出，因为图像像素数太多，数据维度高，会产生维度灾难，所以原来一个思路是手工提取图像的一些关键特征，这实际就是就一个降维的过程。
> 那么问题来了，特征怎么提？
> 特征提取的好坏异常关键，甚至比学习算法还重要，举个例子，对一系列人的数据分类，分类结果是性别，如果你提取的特征是头发的颜色，无论分类算法如何，分类效果都不会好，如果你提取的特征是头发的长短，这个特征就会好很多，但是还是会有错误，如果你提取了一个超强特征，比如染色体的数据，那你的分类基本就不会错了。
> 这就意味着，特征需要足够的经验去设计，这在数据量越来越大的情况下也越来越困难。
> 于是就出现了端到端网络，特征可以自己去学习，所以特征提取这一步也就融入到算法当中，不需要人来干预了。
>
> 作者：张旭
> 链接： https://www.zhihu.com/question/51435499/answer/129379006
> 来源：知乎

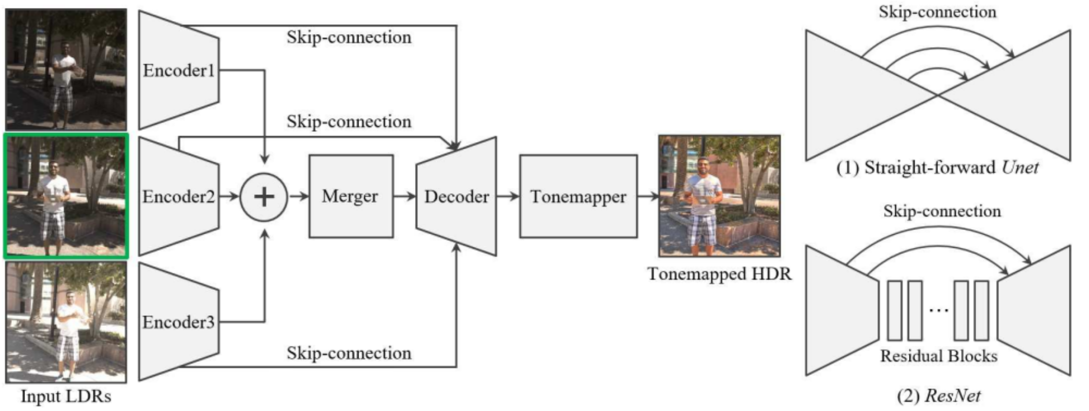
所以在本文中，end-to-end就是直接输入原始LDR images，不进行alignment。

image registration algorithm

图像配准是使用某种方法，基于某种评估标准，将一副或多副图片（局部）最优映射到目标图片上的方法。是基于某评估标准，将一副或多副图片（局部）最优映射到目标图片上的方法。通常情况下，它将一副图片（源图像，Moving Image）的坐标映射到另一幅图像（目标图像，Fixed Image）上，得到配准后的图像对（Moved Image）。

What is the dense correspondence between image pixels?

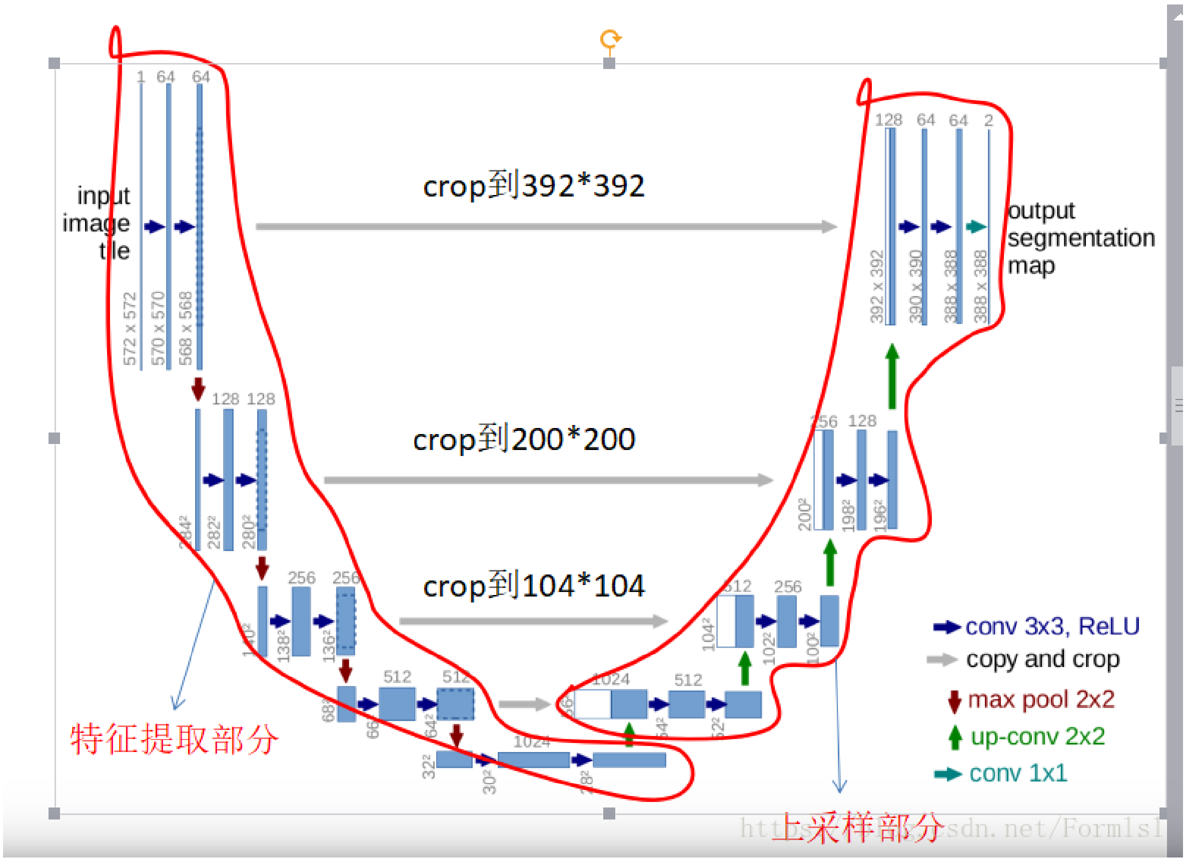
网络结构和训练方法



(a) Network Architecture (b) Structure

Fig. 2. Our framework is composed of three components: encoder, merger and decoder. Different exposure inputs are passed to different encoders, and concatenated before going through the merger and the decoder. We experimented with two structures, *Unet* and *ResNet*. We use skip-connections between the mirrored layers. The output HDR of the decoder is tonemapped before it can be displayed.

Unet



Unet主要用于图像分割，由FCN变种而来，先进行特征提取，再进行上采样。

https://blog.csdn.net/hduxiejun/article/details/71107285?depth_1-utm_source=distribute.pc_relevant.none-task&utm_source=distribute.pc_relevant.none-task

https://blog.csdn.net/weixin_43842265/article/details/96423588?depth_1-utm_source=distribute.pc_relevant.none-task&utm_source=distribute.pc_relevant.none-task

> Given a stack of LDR images, if they are not in RAW format, we first linearize the images using the estimated inverse of Camera Response Function (CRF) [6], which is often referred to as radiometric calibration. We then apply gamma correction to produce the input to our system. 不是很理解为什么进行这种预处理

We denote the set of input LDRs by I = {I1, I2, I3}, sorted by their exposure biases. We first map them to H = {H1, H2, H3} in the HDR domain. We use simple gamma encoding for this mapping ：

$$H_i = \frac{I_i^\gamma}{t_i}, \gamma > 1$$

We then concatenate I and H channel-wise into a 6-channel input and feed it directly to the network.

$$\hat{H} = f(\mathcal{I}, \mathcal{H})$$

tonemapping:

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

Loss Function:

$$\mathcal{L}_{\text{Unet}} = \|\mathcal{T}(\hat{H}) - \mathcal{T}(H)\|_2$$

具体细节

不是很懂skip-connection的作用体现在哪里，文章中应该只提到了两次，没说真正的作用

不足

One particular case where homography may not produce perfect alignment is the existence of parallax effects in saturated regions. The final HDR output may be blurry .

We also observe challenges of recovering massive saturated regions with minimal number of input LDRs.

HdrNet

相关知识

双边滤波器(bilateral filter)

双边滤波器（Bilateral filter）是一种可以保边去噪的滤波器。可以滤除图像数据中的噪声，且还会保留住图像的边缘、纹理等（因噪声是高频信号，边缘、纹理也是高频信息，高斯滤波会在滤除噪声的同时使得边缘模糊）。

双边滤波器的模板系数矩阵由高斯模板矩阵点乘（元素级相乘）值域系数获得。可以与其相比较的两个滤波器是：高斯低通滤波器和α-截尾均值滤波器（去掉百分率为α的最小值和最大之后剩下像素的均值作为滤波器）。

双边滤波器中，输出像素的值依赖于邻域像素的值的加权组合，

$$g(i,j) = \frac{\sum_{k,l} f(k,l)w(i,j,k,l)}{\sum_{k,l} w(i,j,k,l)}$$

权重系数w(i,j,k,l)取决于定义域核

$$d(i,j,k,l) = \exp\left(-\frac{(i-k)^2+(j-l)^2}{2\sigma_d^2}\right)$$

和值域核

$$r(i,j,k,l) = \exp\left(-\frac{\|f(i,j)-f(k,l)\|^2}{2\sigma_r^2}\right)$$

的乘积

$$w(i,j,k,l) = \exp\left(-\frac{(i-k)^2+(j-l)^2}{2\sigma_d^2} - \frac{\|f(i,j)-f(k,l)\|^2}{2\sigma_r^2}\right)$$

同时考虑了空间域与值域的差别，而Gaussian Filter和α均值滤波分别只考虑了空间域和值域差别。

原理：在平坦区域，像素差值较小，对应值域权重r接近于1，此时空域权重d起主要作用，相当于直接对此区域进行高斯模糊，在边缘区域，像素差值较大，值域系数下降，导致此处核函数下降（因w=r*d），当前像素受到的影响就越小，从而保持了边缘的细节信息。

思想：抑制与中心像素值差异较大的像素（即使你们空域相距较近）。

计算方法：对每一个邻域像素点，计算出其对应的空域系数和值域系数，相乘得到总的系数，然后进行加权求和。
原文链接： https://blog.csdn.net/MoFMan/article/details/77482794

双边滤波快速算法

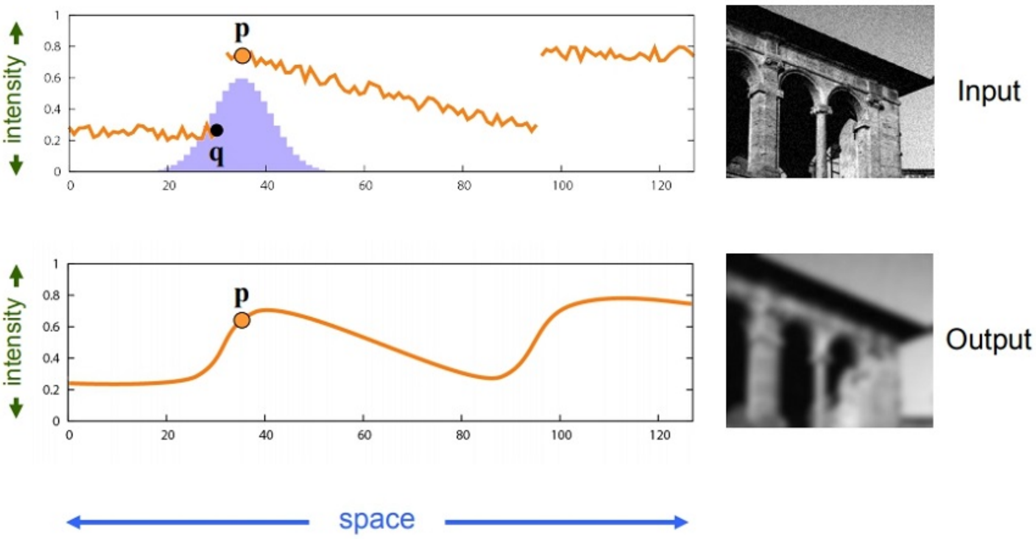
不理解，还得学习数字信号处理

A Fast Approximation of the Bilateral Filter using a Signal Processing Approach

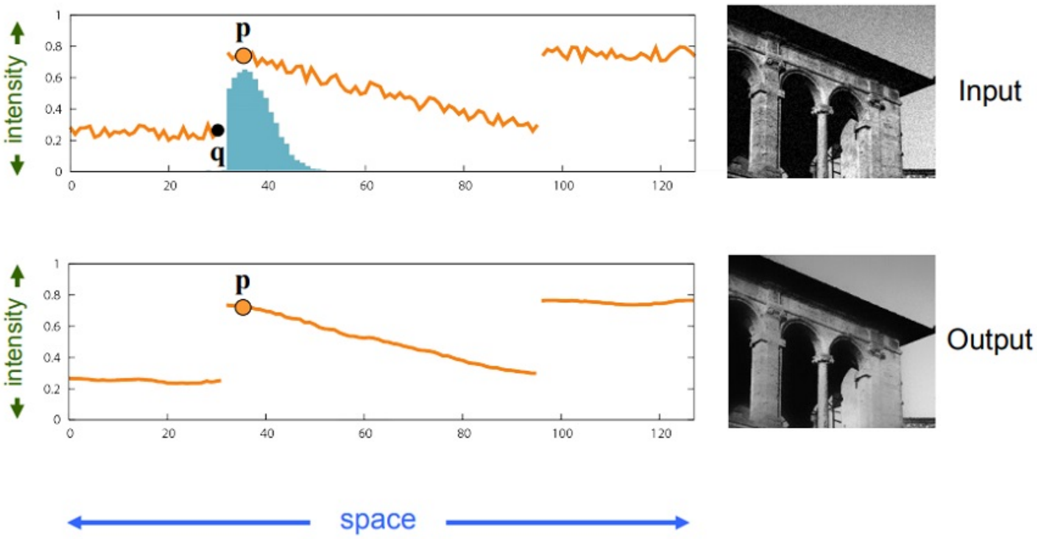
$$\textbf{linear: 3D convolution} \qquad (w^{\mathrm{b}}\, i^{\mathrm{b}}, w^{\mathrm{b}}) = g_{\sigma_{\mathrm{s}},\sigma_{\mathrm{r}}} \otimes (wi, w)$$

$$\textbf{nonlinear: slicing+division} \qquad I^{\mathrm{b}}_{\mathbf{p}} = \frac{w^{\mathrm{b}}(\mathbf{p}, I_{\mathbf{p}})\; i^{\mathrm{b}}(\mathbf{p}, I_{\mathbf{p}})}{w^{\mathrm{b}}(\mathbf{p}, I_{\mathbf{p}})}$$

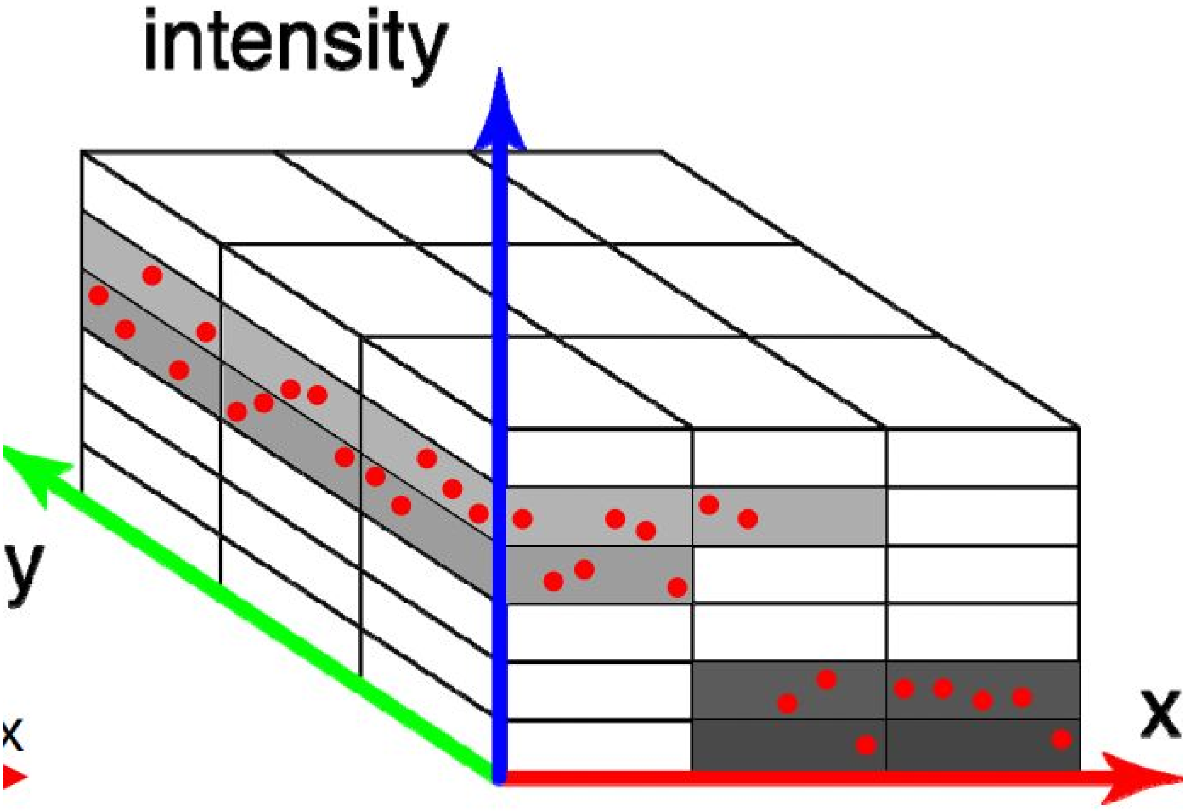
$$gb(I)_{\mathbf{p}} = \sum_{\mathbf{q}} G_{\sigma}(\|\mathbf{p}-\mathbf{q}\|)\; I_{\mathbf{q}}$$



$$bf(I)_{\mathbf{p}} = \frac{1}{W_{\mathbf{p}}} \sum_{\mathbf{q}} G_{\sigma_{\mathrm{s}}}(\|\mathbf{p}-\mathbf{q}\|)\; G_{\sigma_{\mathrm{r}}}(|I_{\mathbf{p}}-I_{\mathbf{q}}|)\; I_{\mathbf{q}}$$



bilateral grid



Bilateral Grid

https://groups.csail.mit.edu/graphics/bilagrid/bilagrid_web.pdf

双边网格的滤波主要分以下步骤:

- 创建双边网格, 将整个网格初始化为 `0` , 通过在像素值域(range)和空间域(space)进行采样, 每个值域内的亮度值可由加权平均获得, 将采样点取整后放入对应的坐标上形成二维的双边网格, 对于图像来说就是一个3D双边网格, 这样任何处理图片的操作都可以处理这个Grid
- 处理双边网格, 通过高斯核对填充后的双边网格进行卷积操作生成了更加平滑的双边网格, 但是由于采样对于图像来说就是低分辨率的图像
- 通过将平滑后的双边网格和输入信号进行slice操作(上采样)就得到输出信号, 其中Slice操作就是选择一个参考图(一般由输入信号生成), 对参考图任意一个像素进行空间域和像素值域进行采样, 然后使用三线性插值的方法实现未知范围的亮度值的计算得到高分辨率的输出

http://simtalk.cn/2018/06/05/HDRNET/

BGU

https://people.csail.mit.edu/hasinoff/pubs/ChenEtAl16-bgu.pdf

JBF:
$$J_p = \frac{1}{k_p} \sum_{q \in \Omega} I_q \, f(||p-q||) \, g(||\tilde{I}_p - \tilde{I}_q||). \qquad (2)$$

上采样是选取一个参考图，对其任意一个空间的像素进行空域和值域的采样，找到其在网格上的位置，这里不进行取整而是采用三线性插值的方法，实现未知范围内亮度值的计算，这个过程被称为slicing。

BGU核心思想是:

1. 任何滤镜效果，在局部小区域内都可以看做是一个线性变换
2. 利用 bilateral grid 可以从一个低分辨率的图上 slice 得到高分辨率的结果
3. upsample 针对的是变换系数，而不是直接针对像素。这样对细节方面损失降低到最小

具体实现的步骤如下：

1. 对原图 downsample 得到一个小图
2. 在小图上应用滤镜
3. 在小图上划分网格（bilateral graid），拟合每一个网格中的线性变换
4. 线性变换的系数在网格间做平滑
5. 利用这个网格，根据原始大图在这个网格上做 slicing，得到高分辨率的线性变换系数，进一步得到高分辨率的结果

这里有两点比较重要，一是利用 bilaeral grid 做 slicing 来 upsample，二是用线性变换的系数做中间媒介，而不是直接 upsample 小图，这样得到的结果更为自然，大图的细节损失很小。

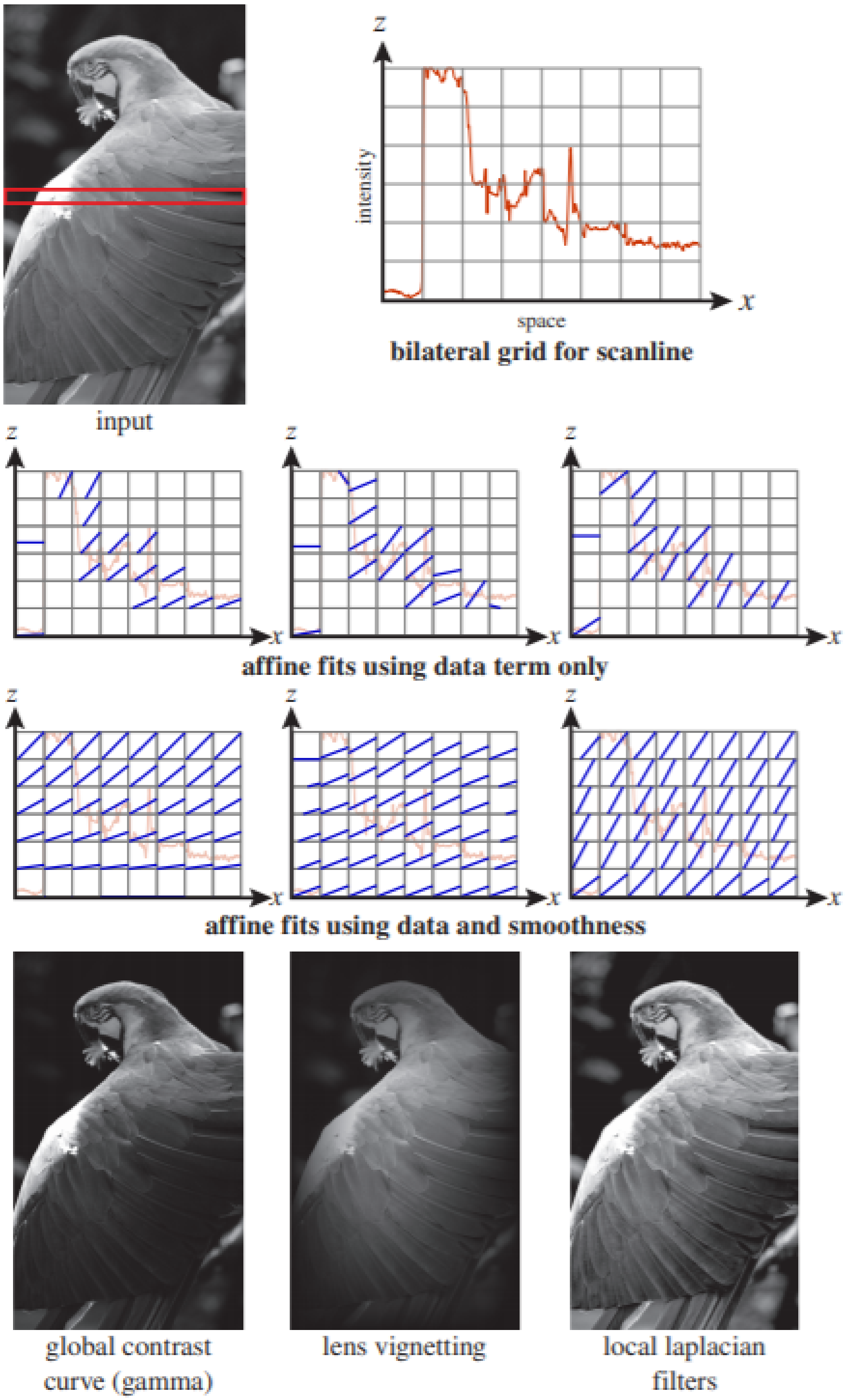

z
intensity
space
x
bilateral grid for scanline
input
affine fits using data term only
affine fits using data and smoothness
global contrast curve (gamma)
lens vignetting
local laplacian filters

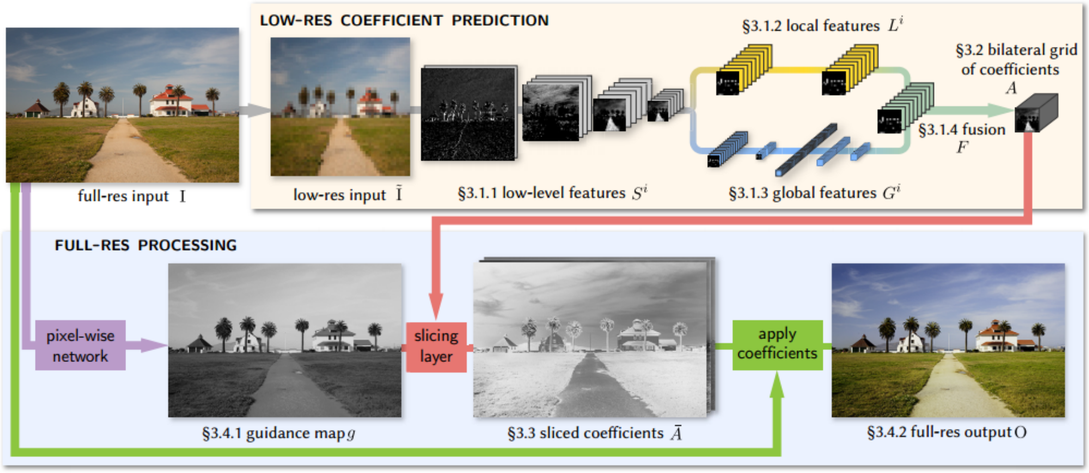
Figure 4: *Our algorithm modeling three different operators. Top: input image and bilateral grid for highlighted scanline. Second row: local affine models fit to the input/output pair with a data term only. Note that grid cells with no data are empty. Third row: affine models fit everywhere using data and smoothness. A global curve (bottom left) results in affine models that vary with intensity (z), but not spatially. A vignette can be expressed as an affine model (a scaling) that varies with position but not intensity. A more complex effect (bottom right) produces affine models that varies with both.*

网络结构和思路

Strategy

1. 将大部分的预测在低分辨的双边网格下进行，每个像素包括的x，y维和第三维可以表示颜色功能，可以用来在对3D双边网格做slicing操作的时候考虑到输入的颜色。
2. 学习输入到输出的变换过程，而不是直接学习输出，因此整个结构学习的是一个仿射变换。
3. 虽然大部分的操作是在低分辨下进行的，但是损失函数最终建立在原来的分辨率上，从而使得低分辨下的操作去优化原分辨下的图像。

HDRNet


LOW-RES COEFFICIENT PREDICTION
§3.1.2 local features L^i
§3.2 bilateral grid of coefficients A
§3.1.4 fusion F
full-res input I
low-res input $\tilde{I}$
§3.1.1 low-level features S^i
§3.1.3 global features G^i
FULL-RES PROCESSING
pixel-wise network
slicing layer
apply coefficients
§3.4.1 guidance map g
§3.3 sliced coefficients $\bar{A}$
§3.4.2 full-res output O

流程解释：

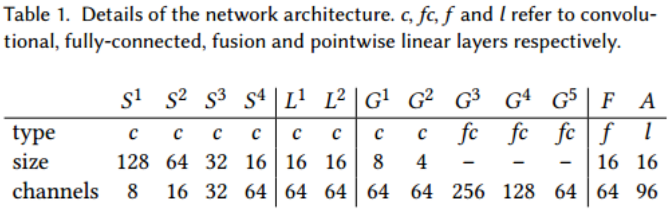


Table 1. Details of the network architecture. *c*, *fc*, *f* and *l* refer to convolutional, fully-connected, fusion and pointwise linear layers respectively.

	S^1	S^2	S^3	S^4	L^1	L^2	G^1	G^2	G^3	G^4	G^5	F	A
type	c	c	c	c	c	c	c	c	fc	fc	fc	f	l
size	128	64	32	16	16	16	8	4	–	–	–	16	16
channels	8	16	32	64	64	64	64	64	256	128	64	64	96

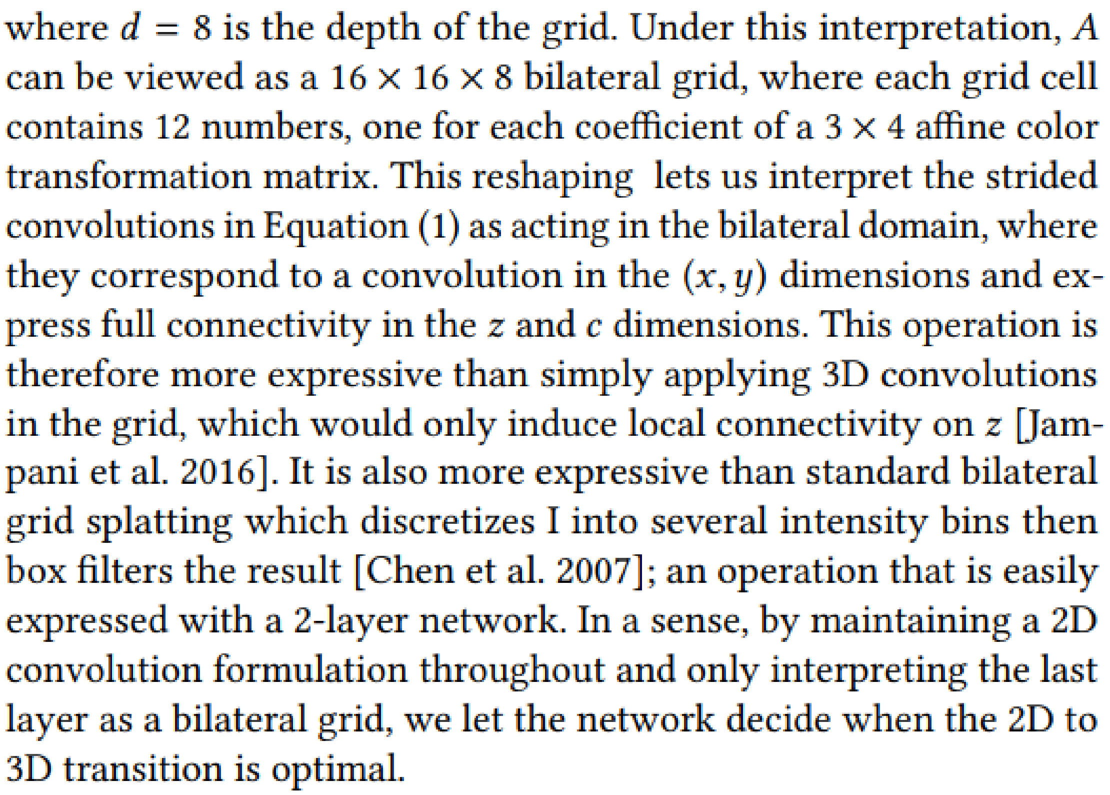
首先full-res input I downsampling 到low-res input $\hat{I}$,然后用strided convolutional layers提取low-level features,这里选择$n^S = 4$,然后输入到两条线:local features和global features。对于local features，使用两层stride = 1的卷积层提取；对于global features, 通过两次 stride=2的卷积层，再使用3层全连接层。然后**融合特征**，再通过一个1*1卷积核进行升维(why?)。再将特征展开作为双边网格（还没细看）。使用可训练的slicing layer进行上采样。最后实现全分辨的最终输出。

这篇文章回溯的论文有点多，比如最早的bilateral filter,到快速算法的实现,再到增加一维形成bilateral grid进行加速计算，然后又提出BGU实现滤镜加速，最后引出神经网络和BGU的结合，训练比如线性映射模型的系数，特征提取的参数等等，形成一个实时的图像增强。论文中还有很多细节没有看，还需要深究。

update 2020.3.25

1*1卷积核进行升维应该是为了转换成双边网格时，使得网格线性变换矩阵大一些。

$$A_{dc+z}[x, y] \leftrightarrow A_c[x, y, z]$$

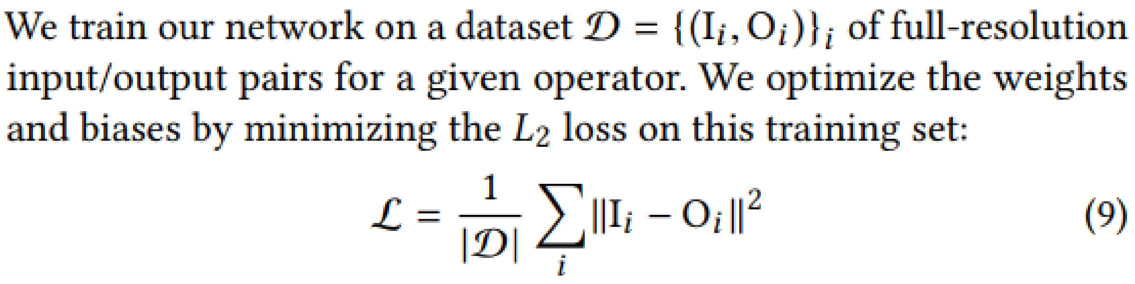


where $d = 8$ is the depth of the grid. Under this interpretation, A can be viewed as a $16 \times 16 \times 8$ bilateral grid, where each grid cell contains 12 numbers, one for each coefficient of a 3×4 affine color transformation matrix. This reshaping lets us interpret the strided convolutions in Equation (1) as acting in the bilateral domain, where they correspond to a convolution in the (x, y) dimensions and express full connectivity in the z and c dimensions. This operation is therefore more expressive than simply applying 3D convolutions in the grid, which would only induce local connectivity on z [Jampani et al. 2016]. It is also more expressive than standard bilateral grid splatting which discretizes I into several intensity bins then box filters the result [Chen et al. 2007]; an operation that is easily expressed with a 2-layer network. In a sense, by maintaining a 2D convolution formulation throughout and only interpreting the last layer as a bilateral grid, we let the network decide when the 2D to 3D transition is optimal.

使用slicing上采样：学习一个引导图，配合三线性插值法

生成full-res的方法实在不是很理解

训练：



We train our network on a dataset $\mathcal{D} = \{(\mathrm{I}_i, \mathrm{O}_i)\}_i$ of full-resolution input/output pairs for a given operator. We optimize the weights and biases by minimizing the L_2 loss on this training set:

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_i ||\mathrm{I}_i - \mathrm{O}_i||^2 \tag{9}$$

优点:

1. 神经网络提取出一系列特征，包含低级特征、局部特征、全局特征。对于图像处理来说，非常完整。考虑到人类摄影师修图的过程，也无非考虑这几方面的特征，因此这个方法适用范围非常广。
2. 神经网络强大的学习能力，可以学习非常复杂的变换；局部线性变换的假设，一定程度上防止了过拟合。这个方法鲁棒性很好。
3. 相比 BGU 在像素上做线性变换，这里是在神经网络提取的特征上做变换，结果更稳定，并且能 handle 极为复杂、非线性极强的变换形式。
4. 利用 bilateral grid 做 upsample，可以得到较高的质量，从而在前期可以放心做 downsample，一方面减少计算量加快计算速度，一方面也防止过拟合到一些局部细节上。

不足:

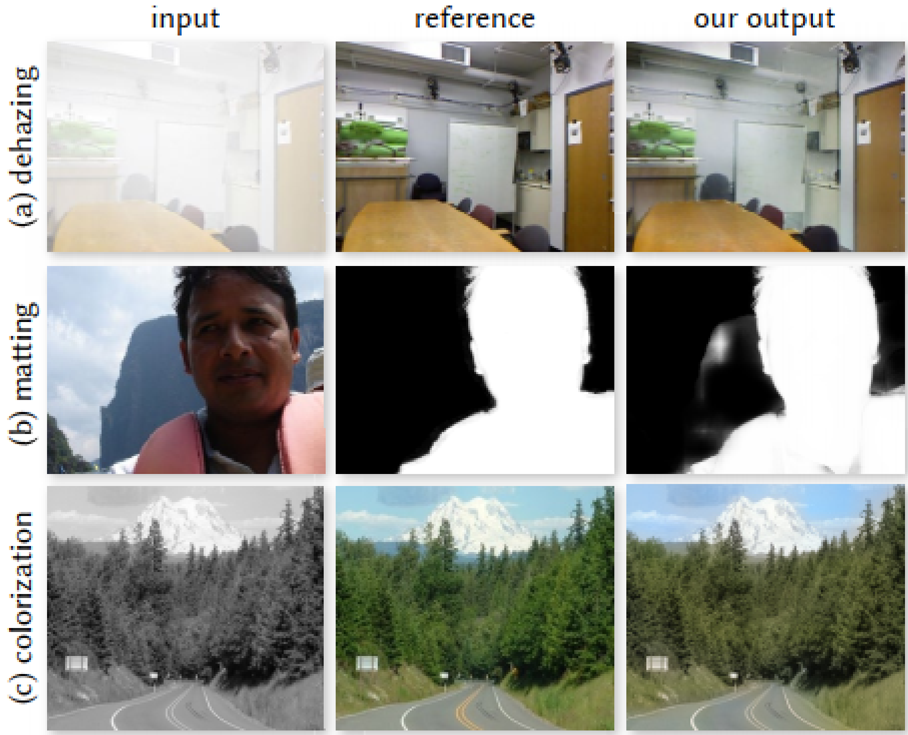


Fig. 12. Our algorithm fails when the image operator strongly violates our modeling assumptions. (a) Haze reduces local contrast, which limits the usefulness of our guidance map. It also destroys image details that cannot be recovered with our affine model (e.g., on the whiteboard). (b) Matting has successfully been modeled by locally affine models on 3 × 3 neighborhoods [Levin et al. 2008]. However, this affine relationship breaks down at larger scales (like a grid cell in our model) where the matte no longer follows tonal or color variations and is mostly binary. This limits the usefulness of our bilateral grid. (c) For colorization, the learned guidance map is at best a nonlinear remapping of the grayscale input. Our model can thus only learn a local color per discrete intensity level, at a spatial resolution dictated by the grid's resolution. Our output is plagued with coarse variations of colors that are muted due to our L_2 loss (see the road line, and the tree/sky boundary).

High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs

相关知识