本质图像论文记录

CVPR2018

1. SfSNet: Learning Shape, Reflectance and Illuminance of Faces `in the Wild'

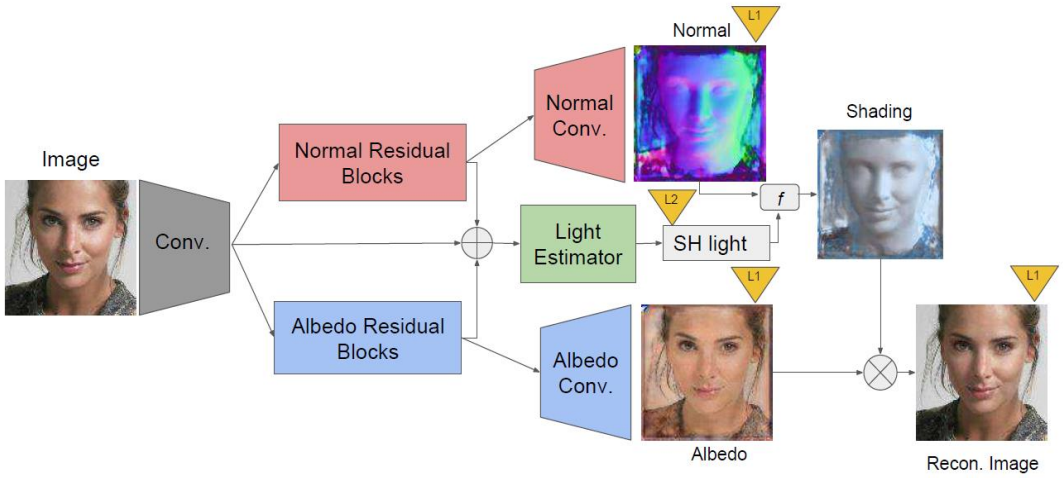
之前相关人脸本质图像分解的工作都是在合成数据集中完成的， 但到真实的人脸，不同分布使得泛化效果很差，这篇论文的特色是提出了一种新的训练范式（SFS-supervision），从真实无标签的真实人脸数据中学习形状，反射以及光照，并且还提出了一种更强大的网络模型（SFS-Net）。

SFS-supervision 分为以下三步：

a) 先使用 3DMM 中合成的数据集训练 SFS-Net；
b) 然后用训练好的网络对真实的人脸数据集生成伪标签；
c) 最后共同训练合成数据集以及带有伪标签的真实数据集。

直接对真实图像使用重建损失进行反向传播会使分解过程中各个组件发生崩溃而产生平凡解，这里的伪标签是很大程度上缓解这种情况的产生。

SFS-Net 网络结构如下：



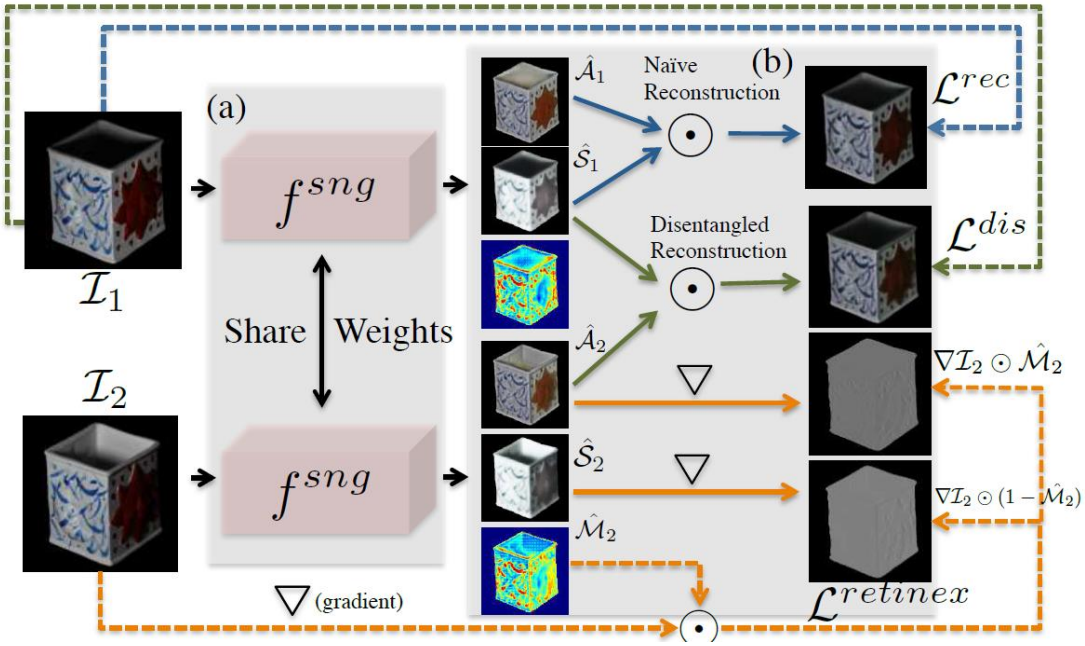
网络并没有采用传统的 U-Net 结构，作者指出了 U-Net 的缺点，由于高频特征可以直接通过远跳跃连接从编码器传到解码器，所以对于胡须以及皱纹这种高频特征是不知道它来自于 Normal 还是 Albedo 的，潜在空间所具有的信息性弱，所以作者提出了 SFS-Net，通过一个共享 Conv，然后分两路通过 Normal Residual Blocks 和 Albedo Residual Blocks（残差块可以进行低频与高频的共同学习），得到 Normal features 和 Albedo features，最后 Normal features 和 Albedo features 分别通过各自的 Conv 得到形状图以及反射图，生成光照信息则是将 image features，Normal features 和 Albedo features 三者进行 concat，然后通过一个 Light Estimator 得到 SH light，最后形状图和光照信息联合通过一个函数得到光照图，光照图和反射图相乘重建出原图。网络有四个 LOSS，除了 SH light 是 L2 loss，Normal，Albedo 以及 Recon 都是 L1 loss。网络更多细节参考论文附录以及代码。

ECCV2018

1. Single Image Intrinsic Decomposition without a Single Intrinsic Image

本质图像分解按照图片的数量可以分为 single-image based 和 multi-image based，基于单张图片的方法的缺点在于缺少标签，而基于多张图片的算法虽然不需要标签，但由于需要多张图像，这在现实情况下很难应用。

本文提出了一种全新的思路，通过多张图片进行无 GT 训练，但在测试过程中使用单张图片进行预测，还可以联合带标签的数据进一步提升分解效果，实验表明当使用 50%的标签图像时就可以达到 SOTA。网络结构图如下：



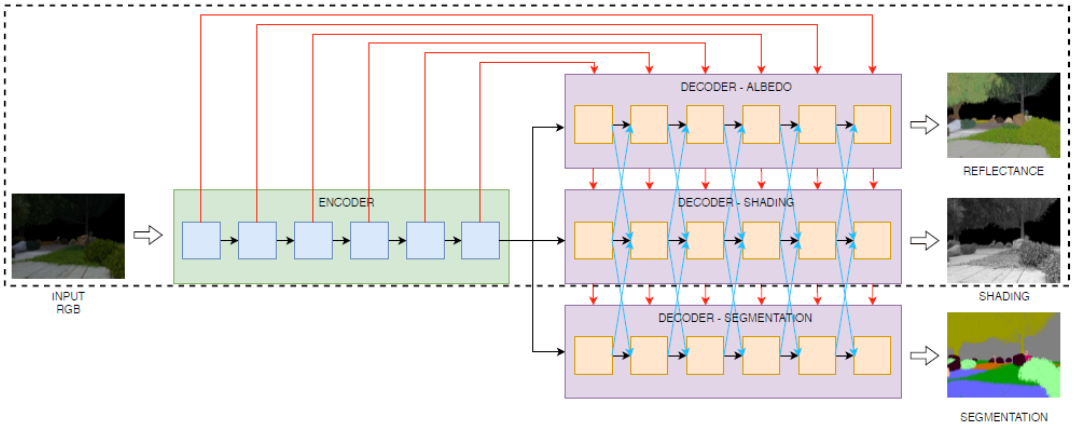
网络分析：将不同光照条件的同一物体的两个图像通过一个共享参数的孪生网络（用的常规 U-Net），得到两个图像的反射图，光照图以及软分配掩模，首先I_1通过孪生网络得到反射图A_1以及光照图S_1，反射图A_1和光照图S_1可以相乘重建原图I_1，I_2通过孪生网络同样得到得到反射图A_2以及光照图S_2， 由于反射图是光照不变性的，所以A_2和S_1同样可以重建原图I_1，两个重建可以与I_1构成两个**重建 L1_loss**，软分配掩模是预测图片在每点像素值的梯度是属于反射图的概率，所以 I_2的梯度与软分配掩模 M_2相乘代表的是反射图 A_2的梯度，与预测的 A_2的梯度进行**梯度 L1_loss**，反射图和光照图的概率和为 1，通过 1 减去软分配掩模 M_2的概率即使光照图的梯度概率，剩下的操作与反射图一样。另外，反射图可能会出现全白像素的图像，这种图像也是光照不变性的， 为了防止这种退化情况的产生，作者加了一个额外的 **embedding loss** $\boldsymbol{L_1^{ebd}}$用来正则化，让两个反射图尽可能一样，同时随机采样要保持两个反射图有差异。

2. Joint Learning of Intrinsic Images and Semantic Segmentation

这篇论文是本质图像分解与语义分割的结合，本质图像去除了光照的影响，会促进语义分割的准确度，而语义分割的标签给图像分块，使得图像具有像素块的颜色信息，边界导向的信息，同质反射值相同信息等等，所以语义分割应该也会促进本质图像分解的性能。

由于缺少既有本质图像以及语义分割的数据集，所以作者自己做了一个，场景级下的自然环境合成数据集，含有本质图像 GT 以及语义分割 GT。另外，作者提出了一个新的层级 CNN 架构用来联合训练本质图像分解以及语义分割，最后用实验分析了两种问题联合训练所带来的好处。

CNN 层级架构如下：



网络结构与 U-Net 有些不同，一个是多一个生成语义分割图的 decoder，另外光照图的 decoder 和反射图以及语义分割的 decoder 进行了互相层间级联 concat，这种网络有一点比较有特色，通过这样级联以及共用一个编码器，可以是本质图像分解以及语义分割互相影响，相互监督并促进性能提升。作者做了多个实验验证了两个任务的确有促进作用，联合训练效果更佳。

损失函数如下图，反射图和光照图使用的 MSE 以及带尺度的 MSE，语义分割图用的则是交叉熵损失函数，p_x^L代表给定像素 x 属于类别 L 的概率。

$$\mathcal{L}_{MSE}(J,\hat{J}) = \frac{1}{n}\sum_{\boldsymbol{x},c} ||\hat{J} - J||_2^2$$

$$\mathcal{L}_{SMSE}(J,\hat{J}) = \mathcal{L}_{MSE}(\alpha J,\hat{J})$$

$$\mathcal{L}_{CL}(J,\hat{J}) = \gamma_{SMSE}\,\mathcal{L}_{SMSE}(J,\hat{J}) + \gamma_{MSE}\,\mathcal{L}_{MSE}(J,\hat{J})$$

$$\mathcal{L}_{IL}(R,\hat{R},S,\hat{S}) = \gamma_R\,\mathcal{L}_{CL}(R,\hat{R}) + \gamma_S\,\mathcal{L}_{CL}(S,\hat{S})$$

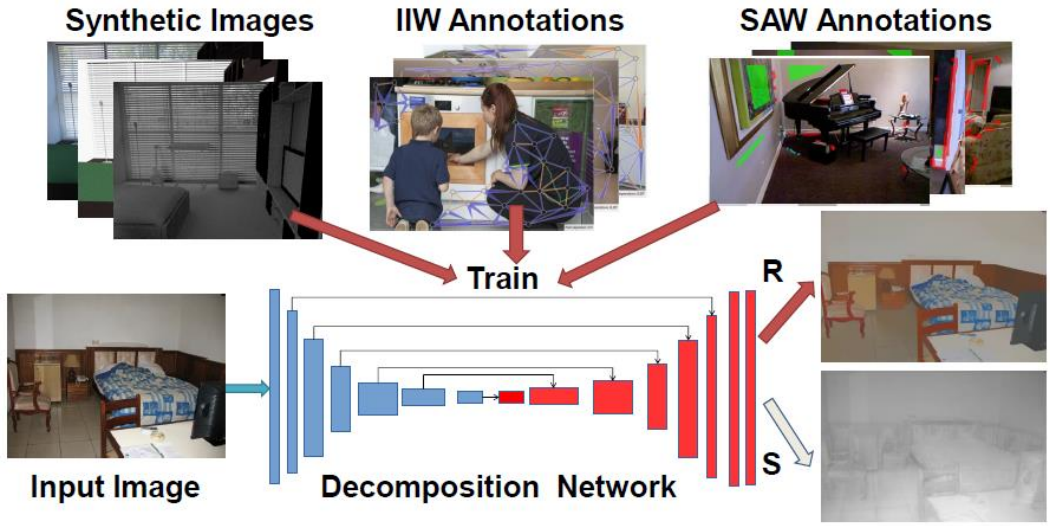
$$\mathcal{L}_{CE} = -\frac{1}{n}\sum_{\boldsymbol{x}}\sum_{L\in O_{\boldsymbol{x}}} log(p_{\boldsymbol{x}}^L)\,,$$

$$\mathcal{L}_{JL}(I,R,\hat{R},S,\hat{S}) = \gamma_{CE}\,\mathcal{L}_{CE} + \gamma_{IL}\,\mathcal{L}_{IL}(R,\hat{R},S,\hat{S}).$$

项目主页：https://ivi.fnwi.uva.nl/cv/intrinseg （数据集和模型还未公开）

3. CGIntrinsics: Better Intrinsic Image Decomposition through Physically-Based Rendering

这篇论文分析现有本质图像数据集存在的一些问题，如合成数据集受限于单个物体（shapeNet），不真实的光照（CG Animation），缺少细节以及低信噪比（SUNCG），而真实本质图像数据集是稀疏的（IIW 和 SAW），并且难以收集富集标签，作者在这篇论文中提出了一个**高质量，高信噪比，真实的，仔细渲染的合成数据集 CGI**（基于 SUNCG，拥有大于 20000 张图片并带有 GT）。另外，作者用半监督学习方式来联合训练带标签的 CGI 以及无标签的 IIW 和 SAW，最后在 IIW 以及 SAW 两种数据集下达到了 SOTA。使用的网络还是基本的 U-Net，如下图。



损失函数如下图，详细公式可以参考原论文。

$$\mathcal{L} = \mathcal{L}_{\text{CGI}} + \lambda_{\text{IIW}}\mathcal{L}_{\text{IIW}} + \lambda_{\text{SAW}}\mathcal{L}_{\text{SAW}}.$$

$$\mathcal{L}_{\text{CGI}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{ord}}\mathcal{L}_{\text{ord}} + \lambda_{\text{rec}}\mathcal{L}_{\text{reconstruct}}$$

$$\mathcal{L}_{\text{IIW}} = \lambda_{\text{ord}}\mathcal{L}_{\text{ord}} + \lambda_{\text{rs}}\mathcal{L}_{\text{rsmooth}} + \lambda_{\text{ss}}\mathcal{L}_{\text{ssmooth}} + \mathcal{L}_{\text{reconstruct}}$$

$$\mathcal{L}_{\text{SAW}} = \lambda_{\text{S/NS}}\mathcal{L}_{\text{S/NS}} + \lambda_{\text{rs}}\mathcal{L}_{\text{rsmooth}} + \lambda_{\text{ss}}\mathcal{L}_{\text{ssmooth}} + \mathcal{L}_{\text{reconstruct}}$$

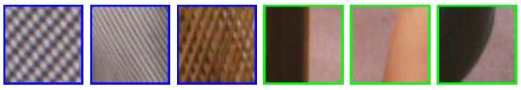
相关资源：项目主页和代码

ICCV2017

1. A Joint Intrinsic-Extrinsic Prior Model for Retinex

这篇文章的主要特点在于首次将形状先验带入到 Retinex 中，提出了 Local Variation deviation（LVD）的概念，通过将 LVD 作为形状先验来保护结构的完整性。然后联合纹理先验，光照先验以及重建损失构成最终的优化函数，求解最优解，与之前的 retinex 方法相比，达到了 SOTA。（论文中 S 代表观测图像，I 代表 Illumination，R 代表 Reflectance）

LVD 可以分成两个部分来看，第一个部分是 LV，即局部变化，代表的是梯度特征，然后第二部分是 D，即偏差，指的是梯度的偏差。LVD 可以看作是对局域梯度进行一种规范化，去除均值的影响，得到梯度变化的方差相关性，纹理是趋向于弱相关性的，而结构是趋向于强相关性的，LVD 则正好对纹理和结构有非常强的鉴别能力。LVD 的公式如下所示，这里的没有使用减去均值的绝对偏差，而采用的是相对偏差，相对偏差更能放大相关性差异。从下图中人物中纹理（蓝色方框）与结构（绿色方框）的在绝对偏差和相对偏差中数值可以看出，纹理和结构确实在 LVD 中有明显差别，而且使用相对偏差能够放大差异。



	A	B	C	D	E	F
$\overline{\mathcal{D}}$	0.073	0.079	0.043	0.013	0.007	0.013
$\overline{\mathcal{R}}$	3.672	3.971	2.168	0.572	0.336	0.637

$$\mathcal{D}_{x/y} = \left| \nabla_{x/y} I - \frac{1}{|\Omega|} \sum_{\Omega} \nabla_{x/y} I \right| \quad \text{rewritten} \rightarrow \quad \mathcal{R}_{x/y} = \left| \frac{\nabla_{x/y} I}{\frac{1}{|\Omega|} \sum_{\Omega} \nabla_{x/y} I + \epsilon} \right|$$

另外，文章还给出了纹理先验以及光照先验，纹理先验是保持反射图间断连续，则纹理先验可以形成如下公式：

$$E_t(R) = \|\nabla_x R\|_1 + \|\nabla_y R\|_1$$

作者对$S = I \bullet R$进行反转，变换成了$(1 - S) = 1 - I \bullet R = (1 - R) \bullet I + (1 - I)$，通过让$H = 1 - S$，$J = 1 - R$，$T = I$以及$a = 1$，原始变换成了$H = J \bullet T + a(1 - T)$，最后公式类似雾霾图像的形成模型，H 代表有雾霾的观测图像，J 是需要还原出的图像，T 是媒介传播，a 是全球大气光，作者引用了何凯明等人提出的去雾算法的黑通道先验，黑通道先验公式如下所示，更具体地推导可以看原始论文

$$T = 1 - \min_{\Omega} \left(\min_{c \in \{r,g,b\}} \frac{H^c}{a} \right)$$

黑通道先验是说在绝大多数非天空的局部区域内，某一些像素至少一个颜色通道具有很低的值，这是何凯明等人基于 5000 多张自然图像的统计得到的定理。作者根据公式推导出了亮通道先验，公式如下：

$$I = 1 - \min_{\Omega}\left(\min_{c\in\{r,g,b\}}(1-S)^{c}\right) = \max_{\Omega}\left(\max_{c\in\{r,g,b\}}S^{c}\right)$$

然后令B = $max_{\Omega}(max_{c}S^{c})$，最后使用 L2 距离损失最小化估计光照和亮通道先验。

$$E_l\left(I\right) = \left\|I - B\right\|_2^2$$

联合优化的最后公式如下：

$$E\left(I,R\right) = \left\|I\cdot R - S\right\|_2^2 + \alpha E_s\left(I\right) + \beta E_t\left(R\right) + \lambda E_l\left(I\right)$$

然后由于$E_s(I)$以及$E_t(I)$都是 L1 范数，非凸，所以作者这两个先验进行了改进，变成了 L2 范数，公式如下：

$$\begin{cases} E_s\left(I\right) = u_x\left\|\nabla_x I\right\|_2^2 + u_y\left\|\nabla_y I\right\|_2^2 \\ E_t\left(R\right) = v_x\left\|\nabla_x R\right\|_2^2 + v_y\left\|\nabla_y R\right\|_2^2 \end{cases},$$

$$\text{where}\begin{cases} u_{x/y} = \left(\left|\frac{1}{\Omega}\sum_{\Omega}\nabla_{x/y}I\right|\left|\nabla_{x/y}I\right| + \epsilon\right)^{-1} \\ v_{x/y} = \left(\left|\nabla_{x/y}R\right| + \epsilon\right)^{-1} \end{cases}$$

项目主页：https://caibolun.github.io/JieP/
代码: https://github.com/caibolun/JieP