doi: 10.12068/j.issn.1005-3026.2020.08.001

一种基于深度网络的视图重建方法

张之敏, 乔建忠, 林树宽, 王品贺
(东北大学 计算机科学与工程学院, 辽宁 沈阳 110169)

摘 要: 为了解决在仅有单目视图的环境下实现立体匹配的问题,在现有视图重构网络模型 Deep3D 的基础上,提出了基于加权局部对比归一化约束的全卷积重构模型. 该模型采用改进的全卷积神经网络架构作为模型的特征提取模块,以期减少训练参数,降低训练时间,增加模型的非线性. 为了进一步提高重构精度,设计了新的基于加权局部对比归一化的约束条件,并采用结构相似性成本(SSIM)与 $L1$ 成本相结合的损失优化函数对模型进行优化. 在 KITTI 2015 数据集上展开实验,并与 Deep3D 模型及其后续的改进方法进行比较. 实验结果表明,在只使用左视图作为训练数据的情况下,生成的右视图在 SSIM 和峰值信噪比两个指标上有很大提升,能够满足立体匹配方法中右视图的精度要求.

关 键 词: 视图重构;卷积神经网路;立体匹配;全卷积网络;加权局部对比归一化
中图分类号: TP 18 **文献标志码**: A **文章编号**: 1005-3026(2020)08-1065-05

A View Reconstruction Method Based on Deep Network

ZHANG Zhi-min, *QIAO Jian-zhong*, *LIN Shu-kuan*, *WANG Pin-he*
(School of Computer Science & Engineering, Northeastern University, Shenyang 110169, China. Corresponding author: QIAO Jian-zhong, E-mail: qiaojz_neu@163.com)

Abstract: To deal with stereo matching in the environment of only a single view, a full convolution reconstruction model with weighted local contrast normalization constraint is proposed on the basis of the existing view reconstruction network model Deep3D. This model adopts the improved full convolutional neural network architecture as the feature extraction module of the model to reduce the training parameters and training time, and to increase the nonlinearity of the model. In order to further improve the accuracy of reconstruction, a new constraint condition based on weighted local comparison normalization is designed, and a loss optimization function combining structural similarity (SSIM) cost and $L1$ cost is used to optimize the model. Experiments were carried out on the KITTI 2015 dataset, and compared with the Deep3D model and subsequent improvements. The experimental results show that the generated right view has a great improvement in SSIM and peak signal to noise ratio when only the left view is used as the training data, which can meet the accuracy requirements of the right view in the stereo matching method.

Key words: view reconstruction; convolutional neural network; stereo matching; fully convolutional network; weighted local contrast normalization

立体匹配是三维场景重建的主要方法,在虚拟现实、无人驾驶汽车等领域得到广泛应用[1]. 然而,双目相机在设置时存在标定误差和同步问题,因此单目相机在实际应用场景中更受青睐. 理论上,用单张图片构建三维场景的思路是一个不适定、几何意义模糊的问题,因为传统的单目深度估计方法很难在单个图像中直接获取几何三维线索[2]. 与几何正确性的立体匹配方法相比,当前最先进的基于深度学习的单目视觉方法存在一定的问题:①该方法几乎完全依赖于高级语义信息,直接利用输入图像和与其对应的真实深度图之间的关系构建预测模型;由于对其需要逼近的函数没有任何先验知识,因此学习语义信息是很困难的[3];②即使有效地学习,该方法也需要大量昂

收稿日期: 2020-01-18
基金项目: 国家自然科学基金青年基金资助项目(61902261).
作者简介: 张之敏(1988-),男,山东德州人,东北大学博士研究生; 乔建忠(1954-),男,辽宁葫芦岛人,东北大学教授,博士生导师.

贵且高质量的相匹配的真实深度图作为监督标签[4]. 为了解决这些问题,Luo 等[5]首次证明了单目深度估计问题可以被重新表述为两个子问题,即视图合成问题和立体匹配问题,并取得了精确的估计结果. 然而,该方法中的立体匹配效率,很大程度上取决于视图合成过程的精度,因此,本文将视图合成作为文章的研究重点. 当前有许多文献对图像重构方法做了研究,Flynn 等[6]首次通过卷积深度网络从其他视图中获取像素来重构不可见视图. Zhou 等[7]利用卷积神经网络学习相同实例的不同视角的相关性来预测外观流. Xie 等[8]提出基于 Deep3D 网络从左视图中产生相匹配的右视图来解决当前 2D 电影转变成 3D 电影的问题. Luo 等[5]在 Deep3D 网络基础上通过改变网络结构提出了视图合成模型.

然而,上述方法对于视图合成的效率和精度都有一定的局限性,影响立体匹配方法的预测结果. 为了能够提高重构视图的效率,本文基于当前的视图重建方法,提出一种新的基于改进全卷积神经网络的视图合成模型. 并且设计了一种加权局部对比归一化(LCN)约束条件,并采用结构相似性成本(SSIM)与 $L1$ 成本相结合的损失优化函数对模型进行优化,使网络对于遮挡更具鲁棒性且不受亮度和低纹理影响.

1 视图重建方法

这一部分描述了本文视图重建方法及网络模型和损失函数. 本文在现有网络模型的基础上进行了创新,引入新的加权局部对比归一化损失来实现视图重建.

1.1 视图重建方法分析

本文视图合成方法是在 Deep3D[8]方法的基础上改进的,图 1 为视图重建过程.

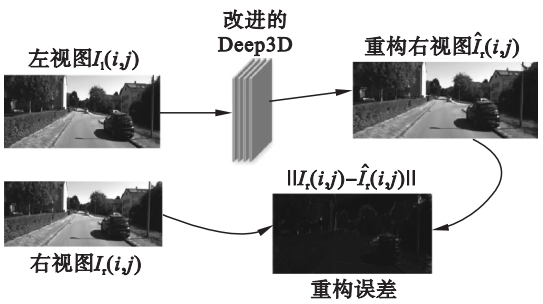


图 1 视图重建方法

Fig. 1 View reconstruction method

从图 1 中可以看出左视图 I_l 被改进的 Deep3D 网络模型处理后生成重构后的右视图 $\hat{I}_r$,然后通过重构的右视图与原始右视图之间的成本函数来优化网络模型. 从图 2 的网络架构中可以看出视图重建方法的详细过程:从不同的中间层对特征图进行向上采样,以获得相同的分辨率,以便将低层次的特性合并到最终的使用中;然后对这些特征进行汇总,进一步生成不同视差值的概率视差图 D;最后将该概率视差图 D 和左视图 I_l 提供给选择层,经过选择层处理后,输出重构后的右视图. 该过程通常用公式描述为

$$\hat{I}_r(i,j) = I_l((i,j) + D_{i,j}). \tag{1}$$

式中,(i,j) 分别指的是左视图 I_l 或视差图 D 的行和列. 该过程是网络模型的前馈过程,神经网络是个迭代优化的模型,因此需要后反馈过程来优化权重值;然而,式(1)对于概率视差图 D 是不可导的,因此不能用通常的梯度下降法来训练深度神经网络. 为了使式(1)可微,对选择模块采用概率求和方式,该网络通过预测在每个像素位置的一个可能的视差值 d 的概率分布 $D_{i,j}^d$,获得一个约束公式:$\sum_d D_{i,j}^d = 1$. 这样可以重新得到一个可微的右视图公式:

$$\hat{I}_r = \sum_d I_l\left((i,j) + d\right)D_{i,j}^d. \tag{2}$$

1.2 网络模型

图 2 为视图重建网络模型. 为了更好地提取图片特征,该网络编码部分的基准网络采用 VGG19 网络架构;然而在 VGG 网络中,最后的三个全连接层占用了大量参数,导致训练占用内存过大,训练速度较慢. 受文献[9]的启发,模型采用三层卷积网络代替全连接层,整个编码结构采用全卷积神经网络作为编码网络,从而节省了训练时间和内存成本. 除此之外,根据文献[10]实验结果得出的相关结论:最大池化层的引入没有增加模型的性能,反而使该操作引入了非线性过程,因此将该模型中 VGG19 网络的池化层用步长为 2 的卷积层和 RELU 激活层代替. 本文仍然采用反卷积双线性插值法作为上采样层,即当上采样系数为 S 时,反卷积层的核为 $2S \times 2S$,步长为 S,边缘填充为 $S/2$. 核权重 W 初始化:

$$W_{ij} = \left(1 - \left|\frac{i}{S-C}\right|\right)\left(1 - \left|\frac{j}{S-C}\right|\right), \tag{3}$$

其中,

$$C = \left(\frac{2S - 1 - (S \bmod 2)}{2S}\right). \tag{4}$$

为了能够将最终的预测与低层级特征的信息相结合,本文在每个卷积模块后增加一个跳跃层,然后对该分支进行批量归一化加上 3×3 的

卷积层，接下来初始化成双线性上采样的反卷积层. 其中反卷积的参数由该层的深度决定. 每一层得到的特征图都会放大成与原始图像相同的大小. 实验结果表明，改进后的模型在性能和精度上与 Deep3D 模型和文献[5]模型相比都有所提高.

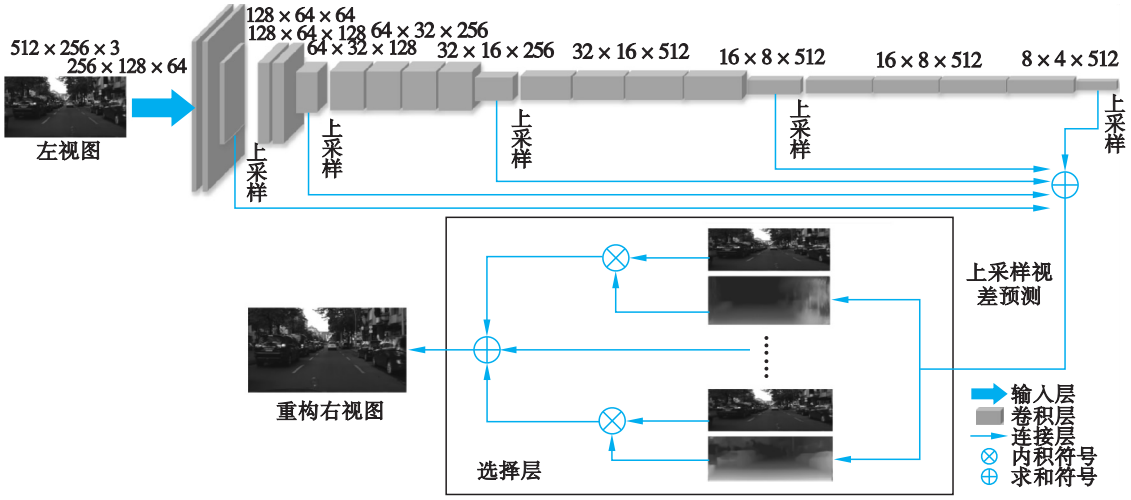


图 2　视图合成网络模型

Fig. 2　View synthesis network model

1.3　损失函数

作为一种图像重构的深度学习模型，通常的损失函数采用原始右视图与重构右视图之间每个像素的光度误差作为损失函数 $L(\theta)$：

$$L(\theta) = \sum_{i,j} \| I_r(i,j) - \hat{I}_r(i,j) \| . \quad (5)$$

然而，根据文献[11]可知，光度损失对于图像重建问题并不是一个很好的选择. 因此本文采用局部对比归一化(LCN)方法，该方法不仅能够消除强度和时差的依赖，而且可以为被遮挡区域提供更好的残差，并且对于左右视图的亮度变化具有不变性. 对于每一个像素，以该像素为中心的邻域为 9 ×9 的小像素块中的均值 μ 和标准差 σ 来归一化当前像素亮度：

$$I_{\mathrm{LCN}} = \frac{I - \mu}{\sigma + \eta}. \quad (6)$$

式中：I 为待归一化视图；η 为一个小常数. 但是 LCN 方法在标准差接近 0 的弱纹理区域表现不好. 为了解决该问题，本文采用局部标准差加权的方法来计算损失函数：

$$L_{\mathrm{LCN}} = \sum_{ij} \| \sigma_{ij}(I^r_{\mathrm{LCN}} - \hat{I}^r_{\mathrm{LCN}}) \| . \quad (7)$$

式中：I^r_{LCN} 和 $\hat{I}^r_{\mathrm{LCN}}$ 分别为原始和重构右视图归一化；σ_{ij} 为局部像素标准差.

受文献[12]的启发，采用 $L1$ 成本和结构相似性(SSIM)成本的组合作为模型的光度图像重建损失函数.

$$L = \frac{1}{N}\sum_{ij} \alpha \frac{1 - \mathrm{SSIM}(I_r(i,j) - \hat{I}_r(i,j))}{2} + (1-\alpha) \| \sigma_{ij}(I^r_{\mathrm{LCN}} - \hat{I}^r_{\mathrm{LCN}}) \| . \quad (8)$$

式中 $\alpha = 0.85$.

2　实验过程及结果分析

2.1　实验过程

为了更好地训练模型，在实验平台上(E5 - 2620v4 处理器，32GB RAM 和两个 11GB RAM NVIDIA GTX 1080TI GPU)用 pytorch 框架实现了网络架构. 使用流行的 KITTI 2015 数据集来训练和评估本文方法. 选择 VGG19 作为基线网络，并使用 ImageNet 预训练模型初始化其权值，所有其他权值通过标准差为 0.01 的正态分布进行初始化. 为了适应 KITTI 数据集，网络模型的输入分辨率设置成 512 ×256. 为了获取精确的视差结果，本文设置 193 通道概率图，表示从 0 到 192 可能的视差范围，作为最终的特征. 模型设置批处理大小为 8，训练轮数为 50 次，并经过 1.8×10^5 的迭代训练. 为了更好地训练网络，模型设置了动态变化的学习率：初始学习速率设置为 0.002，在前 30 个训练轮次中保持不变，然后，每 10 个训练轮次将其减少 2 倍，直到最后. 算法 1 是该方法的实现过程. 图 3 描述了该算法在训练过程中训练损失和迭代次数之间的关系.

算法 1：一种基于深度网络的视图重建方法

输入：左视图数据集 $X_l = \{x_l^1, x_l^2, \cdots, x_l^n\}$，右视图数据集 $X_r = \{x_r^1, x_r^2, \cdots, x_r^n\}$；网络参数 θ；网络模型 $Y_r = f(\theta, X_l)$.

①用预训练模型和正态分布方法初始化网络参数 θ，并设置网络超参数，包括学习率和迭代次数；

②用改进的 Deep3D 模型按式(8)来计算损失函数 $L(\theta)$;

③用批量梯度下降法优化参数 θ;

④$\hat{\theta} = \arg\min_{\theta} L(\theta)$.

输出:$\hat{\theta}$.

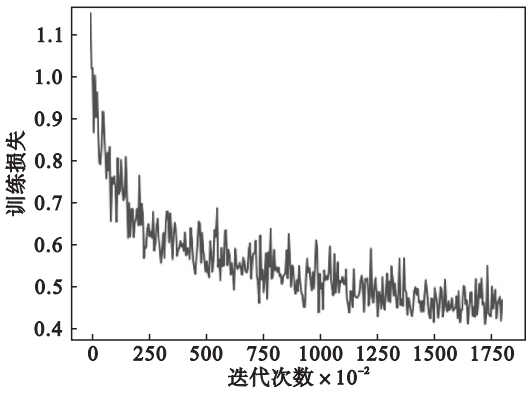


图 3 模型训练过程

Fig. 3 Model training process

2.2 实验结果分析

2.2.1 评价模型

为了定量地对比改进模型与现有模型之间的性能,本文建立了重构视图的评价公式,计算了重构视图与原始视图的峰值信噪比(PSNR)和结构相似性指数(SSIM):

$$\mathrm{MSE} = \frac{1}{N}\sum_{ij} \| I_{\mathrm{r}}(i,j) - \hat{I}_{\mathrm{r}}(i,j) \|^2, \quad (9)$$

$$\mathrm{PSNR} = 10 \times \lg\left(\frac{(2^n-1)^2}{\mathrm{MSE}}\right). \quad (10)$$

式中:N 是图像的像素个数;i,j 指图像中像素的横纵坐标;n 是每个像素的比特数,通常是 8.

$$\mathrm{SSIM}(I,\hat{I}) = \frac{2\mu_I\mu_{\hat{I}} + C_1}{\mu_I^2\mu_{\hat{I}}^2 + C_1}\frac{2\sigma_{I\hat{I}} + C_2}{\sigma_I^2\sigma_{\hat{I}}^2 + C_2}. \quad (11)$$

式中:μ_I,$\mu_{\hat{I}}$分别是原始视图 I 和重构视图$\hat{I}_{\mathrm{r}}$ 的均值;σ_I^2,$\sigma_{\hat{I}}^2$分别是原始视图I和重构视图$\hat{I}_{\mathrm{r}}$的方差;$\sigma_{I\hat{I}}$原始视图 I 和重构视图 $\hat{I}_{\mathrm{r}}$ 的协方差;$C_1 = k_1L_1^2$,$C_2 = k_2L_2^2$ 是保持稳定的常数,L_1,L_2 是像素值的动态范围,$k_1 = 0.01$,$k_2 = 0.03$.

2.2.2 结果分析

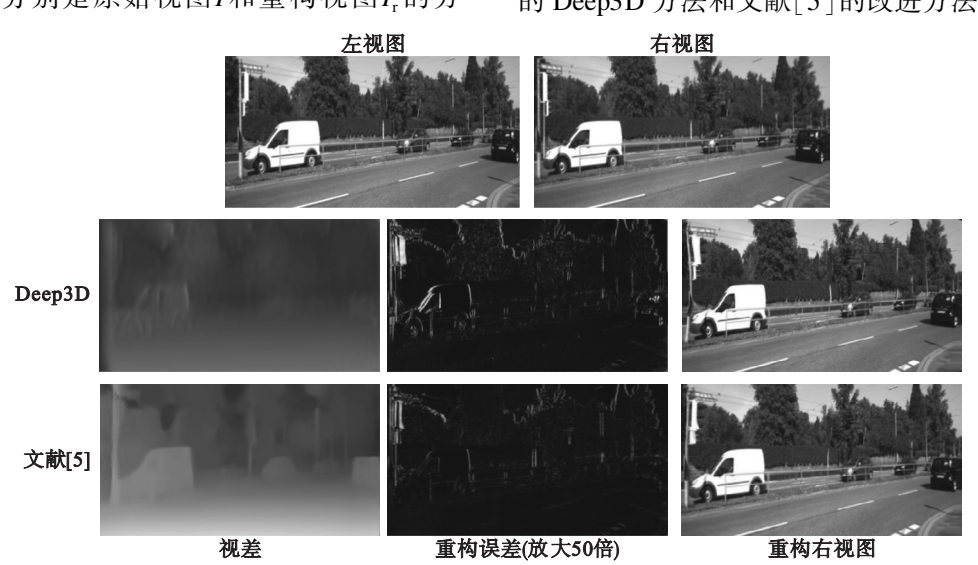
表 1 给出了本文视图重建方法与当前的视图合成方法在 PSNR 和 SSIM 两个指标上的量化比较结果. 本文从数据集中获取了 200 张原始图像作为测试集,该数值是取这 200 张图像评价标准的平均值. 表中为通过各个模型转换出来的 200 张图片与其真正的右图像对比的各个参数的平均值.

表 1 各模型的评估分析

Table 1 Evaluation analysis of models

模型	SSIM	PNSR
Deep3D	0.783	20.209
文献[5]	0.810	22.310
Deep3D – VGG19	0.815	22.510
本文最终模型	0.831	25.065

根据表 1 的数据可以得知,原始的 Deep3D 方法采用 VGG16 网络结构和 $L1$ 损失函数来优化深度模型,其 PSNR 和 SSIM 指标是最低的;文献[5]对原始 Deep3D 进行了改进,增加了可能的视差范围,并修改了部分网络结构;本文的 Deep3D – VGG19 模型较原始 Deep3D 模型和文献[5]在 PSNR 和 SSIM 两个指标上都有所提高. 表 1 中的最终模型与文献[5]模型相比,在 PSNR 标准上提高了 2.755 dB,在 SSIM 标准上提高了 0.21. 图 4 在视差图、重构错误和重构右视图上定性地展示了不同方法的视图重建效果. 为了更直观地展示对比结果,将重建误差增加了 50 倍. 从图中的视差图上可以看到,本文方法能够更加清晰地展示出物体之间的边缘,重建误差优于原始的 Deep3D 方法和文献[5]的改进方法.



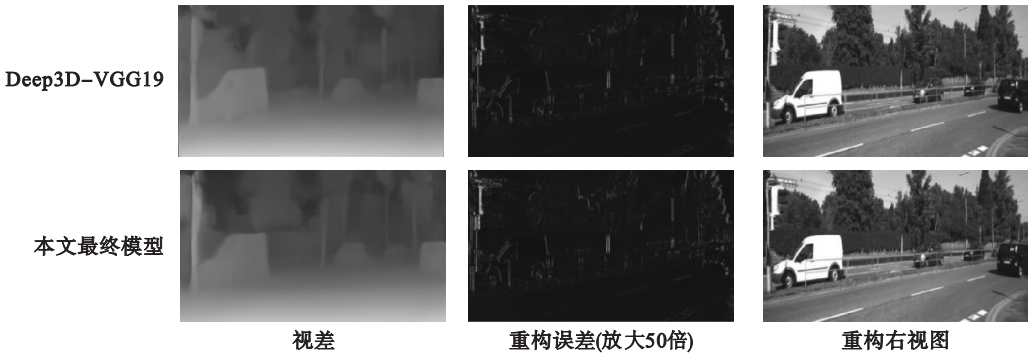


图 4　不同重建方法的视差、重建误差和重建的右视图

Fig. 4　Parallax, reconstruction error and reconstructed right view with different reconstruction methods

3　结　　语

提出了一种基于神经网络的视图重建方法，能够在只有单目视图的情况下重构出右视图，实现立体匹配算法. 本文改进当前的 Deep3D 网络模型和相关损失函数，并在公开数据集 KITTI 2015 上验证了本文方法，取得了较为理想的效果. 然而，虽然本文方法能够在一定程度上满足立体匹配方法中对于右视图的要求，但要进一步提高立体匹配的精度，需要获取更加精确的右视图. 今后将尝试结合传统的视图重构方法和约束条件，进一步改进深度学习网络，以获取更加精确的右视图.

参考文献：

[1] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C/OL]// Advances in Neural Information Processing Systems 27 (NIPS2014). Kuching, Malaysia, 2014 [2019 - 12 - 15]. https://www.researchgate.net/publication/262974294_Depth_Map_Prediction_from_a_Single_Image_using_a_Multi-Scale_Deep_Network.

[2] Saxena A, Sun M, Ng A Y. Learning 3-D scene structure from a single still image[C]//2007 IEEE 11th International Conference on Computer Vision. New York: IEEE, 2007: 1 - 8.

[3] Garg R, Kumar V, Carneiro G, et al. Unsupervised CNN for single view depth estimation: geometry to the rescue[C]// European Conference on Computer Vision. Amsterdam, 2016: 740 - 756.

[4] Godard C, Mac Aodha O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 270 - 279.

[5] Luo Y, Ren J, Lin M, et al. Single view stereo matching [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 155 - 163.

[6] Flynn J, Neulander I, Philbin J, et al. DeepStereo: learning to predict new views from the world's imagery [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 5515 - 5524.

[7] Zhou T, Tulsiani S, Sun W, et al. View synthesis by appearance flow [C]//European Conference on Computer Vision. Amsterdam, 2016: 286 - 301.

[8] Xie J, Girshick R, Farhadi A. Deep3D: fully automatic 2D-to-3D video conversion with deep convolutional neural networks [C]//European Conference on Computer Vision. Amsterdam, 2016: 842 - 857.

[9] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2015: 3431 - 3440.

[10] Springenberg J T, Dosovitskiy A, Brox T, et al. Striving for simplicity: the all convolutional net[C/OL]//International Conference on Learning Representations. San Diego, 2015 [2019 - 12 - 15]. http://arxiv.org/pdf/1412.6806.pdf.

[11] Zhang Y, Khamis S, Rhemann C, et al. Activestereonet: end-to-end self-supervised learning for active stereo systems [C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, 2018: 784 - 801.

[12] Zhao H, Gallo O, Frosio I, et al. Loss functions for image restoration with neural networks[J]. *IEEE Transactions on Computational Imaging*, 2016, 3(1): 47 - 57.