基于学习的图像超分辨率算法

林宙辰

微软亚洲研究院，北京 100190

1 引言

超分辨率（superresolution）算法是增强图像或视频分辨率的技术，它的目的是要使得输出的图像或视频的分辨率比任意一幅输入的图像或输入视频的任意一帧的分辨率都要高。这里的“提高分辨率”意味着已有内容更加清晰或者用户能看到原来没有的细节。在获取高质量的图像或视频比较困难或者代价比较昂贵的时候，使用超分辨率算法是很有必要的。比如在视频监控（video surveillance）中，人脸所占的区域往往只有几十个像素；在遥感（remote sensing）中，超高分辨率器材的价格会远远高于一般分辨率器材的价格；而且用户对提高分辨率的需求没有止境。

超分辨率技术自 Tsai 和 Huang [1] 1984 年提出以来算法甚多，按照其主要原理大致可分为四类[2–4]。第一类是基于插值的算法。这类算法先把低分辨率图像配准（register）到要计算的高分辨率图像的格点上，然后运用非均匀插值（non-uniform interpolation）技术把高分辨率图像每一像素的值插值出来，最后再反卷积以进一步提高清晰度。第二类是基于频率的算法。这类算法利用了傅立叶变换（Fourier transform）空域上的平移对应于频域上的相移的性质，从具有不同相位的低分辨率图像的频谱中估计出高分辨率图像的频谱，然后做傅立叶反变换重构出高分辨率图像。第三类算法是基于重构（reconstruction-based）的算法。这类算法先是根据低分辨率图像和高分辨率图像之间的配准关系，得出每个高分辨率像素对每个低分辨率像素灰度值的贡献，由此得到一个联系高分辨率像素构成的矢量和低分辨率像素构成的矢量的线性方程组，再通过求解该线性方程组获得高分辨率图像。第四类算法是近年来才涌现出来的新型算法，即基于学习的算法。相比之下，前三类算法只是把图像作为信号来处理，而基于学习的算法更注重对图像内容和结构的理解，它利用和问题及数据相关的先验知识来提供更强的约束，因此经常能得到更好的结果。

现有的基于学习的超分辨率算法已有不少，如果按照适用的图像来分，它可以分成通用算法和专用算法两种。通用算法指的是该算法可适用于各种类型、各种尺寸的图像或视频，比如[5–14]。而专用算法指的是该算法只适用于某种类型、某一尺寸的图像或视频，比如用于人脸幻构（face hallucination）的算法[15–25]。通用算法的特点是要把图像分块，先逐块处理，再联合处理以消除相邻块之间的不一致。而专用算法目前基本上只处理人脸图像或视频，这既是应用上的驱动，也是由人脸的特殊性决定的，因为人脸有非常强的结构，而这种结构又比较好表示，比如用特征脸（eigenface）[10, 21]、张量脸（tensorface）[23]等。另一方面，基于学习的超分辨率算法如果按照它的运行细节来分，则可以分成直接最大后验算法和间接最大后验算法两类，其中后者还可以再细分成全局算法和局部算法两类。以下我们就按后一种分类法简要介绍现有的基于学习的超分辨率算法的思想，然后探讨基于学习的超分辨率算法的极限，即它最多能“有效放大”图像多少倍。

2 现有的基于学习的超分辨率算法综述

2.1 间接最大后验算法

抽象地说，间接最大后验算法是把超分辨率问题表述成如下形式：

$$\mathbf{H} = \arg\max_{\mathbf{H}} P\left(\left\{\dot{\mathbf{L}}_i\right\}_{i=1}^{N} \middle| \dot{\mathbf{H}}\right) P\left(\dot{\mathbf{H}}\right), \quad (1)$$

其中$\mathbf{H}$是要求的高分辨率图像，$\dot{\mathbf{H}}$（$\dot{\mathbf{L}}_i$）是和高（低）分辨率图像有关的数量或特征（$\dot{\mathbf{H}}$（$\dot{\mathbf{L}}_i$）可以就是高（低）分辨率图像本身）。不同算法的差别在于似然$P\left(\left\{\dot{\mathbf{L}}_i\right\}_{i=1}^{N} \middle| \dot{\mathbf{H}}\right)$和先验概率$P\left(\dot{\mathbf{H}}\right)$的定义。

2.1.1 局部间接最大后验算法

局部间接最大后验算法先逐块估计高分辨率图像，然后再解决相邻重叠的高分辨率图像块之间的不一致性，从而得到最终的高分辨率图像。比较有代表性的算法是 Freeman 和 Pasztor [5] 1999 年提出来的 Markov 网络（Markov network）方法。这个算法也是最早的基于学习的超分辨率算法，它属于通用超分辨率算法。它把超分辨率算法表述成高分辨率图像高频成分的推断问题[1]：

$$\mathbf{H} = \bar{\mathbf{L}} + \hat{\mathbf{H}},$$

其中$\bar{\mathbf{L}}$是把低分辨率图像插值到高分辨率图像的尺寸所得到的高分辨率图像的低频成分，$\hat{\mathbf{H}}$是缺失的高频成分。$\hat{\mathbf{H}}$通过如下方式估计：

$$\hat{\mathbf{H}} = \arg\max_{\hat{\mathbf{H}}} P\left(\tilde{\mathbf{L}} \middle| \hat{\mathbf{H}}\right) P\left(\hat{\mathbf{H}}\right),$$

其中$\tilde{\mathbf{L}}$是$\bar{\mathbf{L}}$的中频成分。请注意这里$\hat{\mathbf{H}}$的估计方式和（1）式相近。在[5]中，$P\left(\tilde{\mathbf{L}} \middle| \hat{\mathbf{H}}\right)$和$P\left(\hat{\mathbf{H}}\right)$都分块定义：

$$P\left(\tilde{\mathbf{L}} \middle| \hat{\mathbf{H}}\right) = \prod_k P\left(\tilde{\mathbf{l}}_k \middle| \hat{\mathbf{h}}_k\right), \quad P\left(\hat{\mathbf{H}}\right) = \prod_{\substack{i, \\ \hat{\mathbf{h}}_j \in N(\hat{\mathbf{h}}_i)}} P\left(\hat{\mathbf{h}}_i \middle| \hat{\mathbf{h}}_j\right),$$

其中$\left\{\hat{\mathbf{h}}_k\right\}$和$\left\{\tilde{\mathbf{l}}_k\right\}$分别是$\hat{\mathbf{H}}$和$\tilde{\mathbf{L}}$里对应的图像块，$N(\hat{\mathbf{h}}_i)$是$\hat{\mathbf{h}}_i$相邻的图像块集合。因此，上述定义式可以用图 1 里的 Markov 网络表示图像块之间的关系。网络各顶点间的权重$P\left(\tilde{\mathbf{l}}_k \middle| \hat{\mathbf{h}}_k\right)$和$P\left(\hat{\mathbf{h}}_i \middle| \hat{\mathbf{h}}_j\right)$由从样本图像中估计出的高斯混合模型（mixture of Gaussians）表达。要求的高频成分$\hat{\mathbf{H}}$可用标准的信念传播（Belief Propagation）算法迭代求得。

[1] 为了本文符号上的统一，描述算法时采用的符号和原文不同。下同。

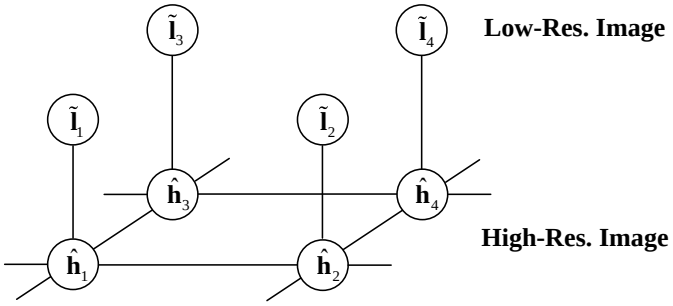

图 1. Freeman 和 Pasztor 采用的 Markov 网络模型，参考[5]重画。

Simon 和 Kanade [15]的人脸幻构（face hallucination）算法可以说是第二个基于学习的超分辨率算法，它是专门针对人脸图像的。在他们的论文里，似然按传统的方式定义：

$$P\left(\left\{\mathbf{L}_i\right\}_{i=1}^{N}\middle|\mathbf{H}\right)\sim\exp\left(-\lambda\left\|\mathbf{H}-\mathbf{PL}\right\|^2\right), \tag{2}$$

其中 $\mathbf{L}$ 是把低分辨率图像中所有像素拼在一起得到的矢量，$\mathbf{P}$ 是联系高低分辨率像素的矩阵。而先验 $P\left(\mathbf{H}\right)$ 按照输入的低分辨率图像的每一像素的特征在样本图像中的匹配情况进行定义，其特征通过图像的 Gaussian 和 Laplacian 金字塔（pyramid）计算。

Liu 等人[21]随后也提出了人脸幻构算法。他们的算法是全局和局部算法的混合。他们假定高分辨率人脸图像 $\mathbf{H}$ 能够被分解为全局的人脸结构图像 $\mathbf{H}^g$ 和局部的人脸特征图像 $\mathbf{H}^l$：

$$\mathbf{H}=\mathbf{H}^g+\mathbf{H}^l.$$

全局图像 $\mathbf{H}^g$ 可以建模成特征脸的线性组合：

$$\mathbf{H}^g=\mathbf{Fx}+\mathbf{f},$$

其中 $\mathbf{F}$ 是以特征脸为列构成的矩阵，$\mathbf{f}$ 是平均脸。因此，$\mathbf{H}^g$ 的计算就转化成寻求最佳的线性组合系数 $\mathbf{x}$：

$$\mathbf{x}=\arg\max_{\mathbf{x}}P(\mathbf{x})P\left(\mathbf{L}\middle|\mathbf{x}\right),$$

其中 $\mathbf{L}$ 是低分辨率人脸图像。 $P(\mathbf{x})$ 和 $P\left(\mathbf{L}\middle|\mathbf{x}\right)$ 分别定义成：

$$P(\mathbf{x})\sim\exp\left(-\mathbf{x}^T\boldsymbol{\Lambda}^{-1}\mathbf{x}\right),\quad P\left(\mathbf{L}\middle|\mathbf{x}\right)\sim\exp\left\{-\lambda\left\|\mathbf{P}(\mathbf{Fx}+\mathbf{f})-\mathbf{L}\right\|^2\right\},$$

其中 $\boldsymbol{\Lambda}$ 是特征脸的协方差矩阵，$\mathbf{P}$ 是联系高低分辨率像素的矩阵。至于局部特征图像 $\mathbf{H}^l$，作者把高、低分辨率图像分块并采用 Markov 随机场模型通过 $\mathbf{H}^g$ 推断 $\mathbf{H}^l$：

$$\mathbf{H}^l=\arg\max_{\mathbf{H}^l}P\left(\mathbf{H}^l\middle|\mathbf{H}^g\right).$$

继[5, 15, 21]之后，对超分辨率感兴趣的学者开始掀起了设计新的基于学习的算法的热潮，出现了很多变形。属于本类的算法就有若干种。[12]里似然的定义和（2）一样，但是先验 $P(\mathbf{H})$

的定义改为高分辨率图像每一个像素 p 的灰度值的分布是以 $m(p)$ 为均值的高斯分布，其中 $m(p)$ 是样本中和 p 的邻域（不含 p ）最匹配的图像块的中心的灰度值。[6]里也采用了 Freeman 和 Pasztor [5]的 Markov 网络模型，但是他们认为要得到好的超分辨率效果，和输入的低分辨率图像块相匹配的样本图像块的点扩散函数（point spread function, PSF）要和输入的低分辨率图像块的点扩散函数一样。因此，他们的算法和 Freeman 和 Pasztor [5]的基本一样，只是要增加 PSF 的估计和匹配这一步骤。Bishop 等人[7]把 Freeman 和 Pasztor [5]的算法简化后拓展到视频。但是视频比图像复杂些，因为如果仅考虑提高帧内的分辨率会导致帧间不连贯，产生闪烁效应。为了解决这个问题，他们又添加了帧间的平滑性约束。[17]里，Dedeoğlu 等人允许匹配得到的高分辨率图像块在运用到要求的高分辨率图像上时可以在改变均值后再用，因此他们需要多预测高分辨率图像块的均值。

2.1.2 全局间接最大后验算法

这类算法不是逐块估计高分辨率图像，而是假定高分辨率图像可以分解成一些“基”的组合，因此只需估计“基”间的组合系数就行了。“基”的存在就要求高分辨率图像有一致的结构，否则重构出来的图像将缺少细节或细节不合理，不能达到超分辨率的目的。所以目前为止所有全局间接最大后验算法都是针对人脸图像，因此都是专用算法。

Gunturk 等人[18]把高、低分辨率人脸图像都表示成相应尺寸的特征脸的线性组合。其超分辨率模型变成：

$$\mathbf{h} = \arg\max_{\mathbf{h}} P(\mathbf{h}) P\left(\{\mathbf{l}_i\}_{i=1}^{N} \middle| \mathbf{h} \right),$$

其中 $\mathbf{h}$ 和 $\mathbf{l}_i$ 分别是高、低分辨率人脸图像的组合系数。 $P(\mathbf{h})$ 和 $P\left(\{\mathbf{l}_i\}_{i=1}^{N} \middle| \mathbf{h} \right)$ 都假定是高斯：

$$P(\mathbf{h}) \sim \exp\left\{ -(\mathbf{h}-\boldsymbol{\mu}_h)^T \boldsymbol{\Lambda}^{-1} (\mathbf{h}-\boldsymbol{\mu}_h) \right\},$$

$$P\left(\{\mathbf{l}_i\}_{i=1}^{N} \middle| \mathbf{h} \right) \sim \exp\left\{ -\sum_{i=1}^{N} (\mathbf{l}_i - \mathbf{F}_l^T \mathbf{P}_i \mathbf{F}_h^T \mathbf{h} - \boldsymbol{\eta})^T \mathbf{Q}^{-1} (\mathbf{l}_i - \mathbf{F}_l^T \mathbf{P}_i \mathbf{F}_h^T \mathbf{h} - \boldsymbol{\eta}) \right\},$$

其中 $\mathbf{F}_h$ 和 $\mathbf{F}_l$ 分别是以高、低分辨率特征脸为列构成的矩阵， $\mathbf{P}_i$ 是联系第 i 幅低分辨率图像和要求的高分辨率图像的矩阵， $\boldsymbol{\mu}_h$ 和 $\boldsymbol{\eta}$ 为均值， $\boldsymbol{\Lambda}$ 和 $\mathbf{Q}$ 为协方差矩阵。 $\mathbf{F}_h$ 、 $\mathbf{F}_l$ 、 $\boldsymbol{\mu}_h$ 、 $\boldsymbol{\eta}$ 、 $\boldsymbol{\Lambda}$ 和 $\mathbf{Q}$ 都从样本中获得。Gunturk 等人的算法很有影响，后来有很多改进。比如，Li 和 Lin [20]进一步考虑了人脸的朝向。Capel 和 Zisserman [16]进一步把人脸细分成六个区域：左眼、右眼、鼻子、左颊、右颊和嘴，在每一区域上做超分辨率后再拼成一个完整的人脸（因此其缺点是区域的边界显得不连续）。

前面说过，Liu 等人[21]的人脸幻构算法是全局和局部算法的混合。其全局脸 $\mathbf{H}^g$ 也用特征脸计算，而局部脸 $\mathbf{H}^l$ 可看成是额外的细节（residue）。这种“global + residue”的思想后来也有很多人采用，比如 Li 和 Lin [19]、Liu 等人[22–25]（后者属于直接最大后验算法，见下文）等等。

2.2 直接最大后验算法

这类算法直接逐块推断高分辨率图像块，所以都是局部方法。因此这类算法无需再细分是局部还是全局算法。抽象地说，直接最大后验算法是把超分辨率问题表述成如下形式[2]：

$$\mathbf{H} = \arg\max_{\mathbf{H}} P\left(\dot{\mathbf{H}}\middle|\left\{\dot{\mathbf{L}}_i\right\}_{i=1}^{N}\right).$$

Sun 等人[13]采用了 primary sketch 先验来推断缺失的高频：

$$\mathbf{H} = \bar{\mathbf{L}} + \mathbf{H}^p,$$

其中 $\bar{\mathbf{L}}$ 是把低分辨率图像进行插值得到的图像，而 $\mathbf{H}^p$ 按下式进行估计：

$$\mathbf{H}^p = \arg\max_{\mathbf{H}^p} P\left(\mathbf{H}^p\middle|\bar{\mathbf{L}}\right).$$

他们假定：

$$P\left(\mathbf{H}^p\middle|\bar{\mathbf{L}}\right) \sim \prod_k P(C_k|\bar{\mathbf{L}}),$$

其中 C_k 是 $\bar{\mathbf{L}}$ 里检测到的边缘线（contour）并已经用许多可能的高分辨率图像块近似。为了解决相邻高分辨率图像块之间的不一致性，每条轮廓线上的可能的高分辨率图像块被看成形成一条一阶 Markov 链，这个思想和 Freeman 和 Pasztor [5]的 Markov 网络模型有点相似。最终的高分辨率图像块也是通过信念传播算法确定。

Hertzmann 等人[9]提出了一个非常简单的图像间风格迁移（style transfer）的算法。给定一幅样本图像对 A 和 A' 和测试图像 B，作者希望能把 B 变成和 A' 同样风格的图像 B'。该算法是按扫描线次序逐像素生成 B'。先构造好 A、A' 和 B 的高斯金字塔，在金字塔的每一层，从粗到细，B' 的每一个像素 Q' 提取出适当特征后与 A 和 A' 里的像素的特征进行匹配，A' 里最匹配的像素的值就赋给 B'。当 A 和 A' 分别是低、高分辨率图像时，以上算法就变成超分辨率算法。

Lin 等人[22]和 Liu 等人[23–25]按照 Liu 等人[21]的两步合成思想又提出了一些算法。这些算法都假定如果输入的低分辨率图像（块）是它在训练样本中最近邻的 k 个低分辨率图像（块）的线性组合，那么相应的高分辨率图像（块）也是它在训练样本中和 k 个低分辨率图像（块）对应的 k 个高分辨率图像（块）的线性组合，且组合系数相同。在处理额外的细节（residue）时，也采用这一假定。这种假定显然是借鉴了 LLE（locally linear embedding）[26]的思想。

另外，由于超分辨率是一个非线性过程，而神经网络（neural network）又擅长拟合非线性映射，因此也有一些学者尝试把神经网络应用于超分辨率。现有的基于神经网络的算法都属于直接最大后验算法。比如 Zhang 和 Pan [14]训练了将低分辨率细节映射到高分辨率细节的神经网络。估计出高分辨率细节后叠加到插值后的图像上就得到要求的高分辨率图像。他们还证明了在一定意义下他们的算法是最优的。Candocia 和 Principe [8]假定相似的低分辨率图像块应当对应相似的高分辨率块。因此他们把训练图像分块，把低分辨率块去掉均值后用自组织映射神经网络聚类，然后对每一类训练一个联想记忆（associative memory）神经网络。做超分辨率时，每个低分辨率块去掉均值后从样本里找到它最近邻的低分辨率块，然后用训练好的联想记忆神经网络映射成高分辨率块。把均值加回去再处理好相邻块的互相重叠部分，就得到高分辨率图像。Miravet 和 Rodr ́guez [11]使用了混合多层感知机（hybrid multilayer

[2] 符号的含义请见（1）。

perceptron）和概率神经网络（probabilistic neural network）对低分辨率图像进行非均匀插值。在配准低分辨率图像后，高分辨率像素的值要从它周围的低分辨率像素的值插值得到，但其权重不再按照传统的随离高分辨率像素距离的平方指数衰减（即使用高斯核）的方式计算，而是通过混合多层感知机把距离映射成权重。插值后再做一次图像反卷积以进一步提高图像清晰度。因此，这个算法也属于第一类超分辨率（基于插值的算法，参见引言）。Kursun 和 Favorov [10]受仿生学启发，仿照视觉皮层里的神经网络结构（SINBAD 细胞）构建神经网络。SINBAD 细胞之间有抑制和反馈作用。该神经网络需要大量的图像来训练。SINBAD 神经网络有很好的识别图像中高阶统计特性的能力，因此有良好的超分辨率性能。

2.3 基于学习的超分辨率算法的优缺点

现有的基于学习的超分辨率算法所使用的高/低分辨率样本图像对（或图像块对）都是固定放大倍数的，这就决定了训练得到的算法不能随便改变放大倍数。全局间接最大后验算法的局限更为严重，因为其高/低分辨率样本图像对的尺寸也是固定的，训练得到的算法甚至不能应用到不是指定尺寸的图像。而且基于学习的超分辨率算法的性能非常依赖于训练样本库，如果样本选择不好，结果就不会很好。不幸的是，现在还没有相关理论来指导样本的选择。另外，虽然有些论文报告了在人脸幻构中达到了很高的放大倍数，但那只是在理想情况下的实验结果，比如低分辨率人脸图像配准无误、噪声类型和水平都已知等，这些在实际情况下都是很困难的。

但是基于学习的超分辨率算法也有它的优点。首先，它一般需要较少的低分辨率图像就能得到较好的超分辨率图像，因为图像的先验信息已经包含在算法中了。许多算法甚至只需要一幅低分辨率图像，这是其他类型的超分辨率算法所做不到的。其次，虽然需要的低分辨率图像少，它能达到的放大倍数却比其他类型的超分辨率算法的高。第三，它不但能提高图像分辨率，如果把样本换成其他的，那么它还能做其他事情。比如，Freeman 和 Pasztor [5]和 Pickup 等人[12]的实验显示，样本适当改变后，它能输出带艺术风格的图像。这也是其他类型的超分辨率算法所做不到的。第四，很多基于学习的超分辨率算法，比如基于特征脸的人脸幻构算法，计算速度很快；相比之下其他类型的超分辨率算法，除了仅使用非均匀插值的算法外，基本都是迭代算法，计算速度相对较慢。

3 基于学习的超分辨率算法的性能极限

虽然研究人员在理想的实验环境下得到的基于学习的超分辨率算法能达到较高的放大倍数，在实际情况下，基于学习的超分辨率算法的性能并不远远优于其他类型的算法，能达到的放大倍数还是比较小。因此，有必要考察基于学习的超分辨率算法是否有一个极限。如果能估计出一个能达到的放大倍数的上界，那么人们就不需要再盲目、徒劳地去试更大的放大倍数，这样就能节省很多时间和空间的资源。Lin 等人的[27]是唯一一篇讨论这个问题的论文。对于基于重构的算法，其性能的界曾被 Lin 等人[28]和 Baker 等人[15]探讨过。由于没有特殊类型的图像（比如人脸和文字图像）的统计模型，下面我们只讨论基于学习的**通用**超分辨率算法的极限。我们后面的分析需要用到一般自然图像的如下统计特性：

1. 高分辨率图像的分布不集中在少数几种高分辨率图像附近，低分辨率图像也是如此。这个性质比较明显，因为一般自然图像不能被分类成少数几类图像。
2. 平滑的低分辨率图像的先验概率要比纹理丰富的低分辨率图像的高。这实际上就是图像出理中最常用的“平滑先验”（smoothness prior）。

需要了解细节的读者可参考[27]。

3.1 什么是基于学习的超分辨率算法的极限？

为了能够定量地分析，我们必须定义什么是“基于学习的超分辨率算法的极限”。这个极限不能单纯地由能计算的放大倍数决定，因为理论上只要计算机性能足够好，给定任何样本，它都能训练出一个超分辨率算法，而且这个算法对任何输入图像都能给一个输出图像，只是这个输出图像的质量并没有任何保证。[28]中已经讨论了如何定义基于重构的超分辨率算法的极限的问题，作者建议把能达到的“有效放大倍数”的上界定义为超分辨率算法的极限。所谓“有效放大倍数”指的是在该放大倍数下算得的高分辨率图像的确会比在较低的放大倍数下算得的高分辨率图像更清晰或细节更多。但是这个定义是同样无法量化计算的，因为“更清晰或细节更多”依赖于对图像内容的理解，它不能由一些低级视觉（low-level vision）的概念获得，比如它不能定义成图像的高频成分的能量，因为单纯加噪声也能增加图像的高频成分的能量而噪声实际上并不增加图像的细节。于是[28]中接着论证需要增加的细节应当是在高分辨率图像中真正存在的细节，最终得出结论，超分辨率算法的极限应当定义成能使得超分辨率算法输出的“高分辨率”图像和真正的高分辨率图像的差不超过某适当阈值的放大倍数的上界。虽然我们不可能知道真正的高分辨率图像是什么样的，但是一些数学技巧，比如先验误差估计，可以克服这个困难。

基于[28]的奠基性工作，我们也可以同样定义基于学习的超分辨率算法的极限。但是基于学习的超分辨率算法又有自身的特殊性，主要是它对不同输入图像的性能不同。如果输入图像比较接近训练图像，显然输出图像的质量就会高；否则质量有可能会差很多。这就促使我们考虑它的平均性能。如果我们把基于学习的超分辨率算法抽象成从低维空间往高维空间的映射函数 s，那么它的平均性能可以按照它的期望风险来定义：

$$\widetilde{g}_s(N,m)=\int_{\mathbf{h}}\left\|\mathbf{h}-s(\mathbf{Dh})\right\|^2 p_h(\mathbf{h})\mathrm{d}\mathbf{h}, \tag{3}$$

其中我们假定低分辨率图像的尺寸为 $N\times N$，放大倍数为 m（即高分辨率图像的尺寸为 $mN\times mN$），$\mathbf{D}$ 对应于把原高分辨率图像降为低分辨率图像的成像过程，先暂时假定为简单的下采样（downsampling），$p_h(\mathbf{h})$ 是高分辨率图像的分布密度函数。我们把损失函数选为平方误差是因为它对应着均方误差，而均方误差是衡量图像差异的最常用的度量。

这里要特别注意的是（3）式只是单个超分辨率算法 s 的期望风险，我们研究它的大小对探讨基于学习的超分辨率算法的极限并没有帮助，因为单独一个算法的极限就是它设定的放大倍数的大小（第 2.3 节里已经指出每个基于学习的超分辨率算法都不能随意改变放大倍数），而基于学习的超分辨率算法的极限要探究的是**所有**算法都不能达到的有效放大倍数。注意到我们可以换一种说法，基于学习的超分辨率算法的极限就是一个上界，放大倍数超过这一界后任何算法的期望风险都不可能小于某一合适阈值。因此，我们可以考虑（3）式定义的期望风险在**所有可能的**超分辨率算法 s 下的下界。如果这个下界大于某一合适阈值，那么任何算法的期望风险都不可能小于这一阈值。这就促使我们探讨（3）式定义的期望风险在所有可能的超分辨率算法 s 下的下界。

3.2 期望风险的下界

为了估计下界，我们需要对（3）做变量替换。首先找一个矩阵 $\mathbf{Q}$ 使得 $\begin{pmatrix}\mathbf{D}\\ \mathbf{Q}\end{pmatrix}$ 非奇异且 $\mathbf{QU}=\mathbf{0}$，其中 $\mathbf{U}$ 是把低分辨率图像插值到高分辨率图像尺寸的上采样（upsampling）矩阵，它具有性质 $\mathbf{DU}=\mathbf{I}$。这个矩阵 $\mathbf{Q}$ 是不唯一的，但它在我们的最终结果中不出现，只是起辅助

推导的作用。接下来定义$\mathbf{M}=\begin{pmatrix}\mathbf{R} & \mathbf{V}\end{pmatrix}$为$\begin{pmatrix}\mathbf{D}\\ \mathbf{Q}\end{pmatrix}$的逆。从$\begin{pmatrix}\mathbf{D}\\ \mathbf{Q}\end{pmatrix}\begin{pmatrix}\mathbf{R} & \mathbf{V}\end{pmatrix}=\mathbf{I}$，我们知道$\mathbf{R}=\mathbf{U}$。

下面我们做变量替换$\mathbf{h}=\mathbf{M}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}$，则（3）变成

$$\begin{aligned}\tilde{g}_s(N,m)&=\int_{\mathbf{x},\mathbf{y}}\left\|\begin{pmatrix}\mathbf{U} & \mathbf{V}\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}-s(\mathbf{x})\right\|^2 p_{x,y}\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\\&=\int_{\mathbf{x}}p_x(\mathbf{x})V_s(\mathbf{x})\mathrm{d}\mathbf{x},\end{aligned}\tag{4}$$

其中

$$p_{x,y}\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right)=\left|\det(\mathbf{M})\right|p_h\left(\mathbf{M}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right),$$

$$V_s(\mathbf{x})=\int_{\mathbf{y}}\left\|\mathbf{V}\mathbf{y}-\phi_s(\mathbf{x})\right\|^2\tilde{p}_y(\mathbf{y}\mid\mathbf{x})\mathrm{d}\mathbf{y},$$

$p_x(\mathbf{x})$是 **x** 的边际分布，$\tilde{p}_y(\mathbf{y}\mid\mathbf{x})$是 **y** 的条件分布，$\phi_s(\mathbf{x})=s(\mathbf{x})-\mathbf{U}\mathbf{x}$是从低分辨率图像 **x** 恢复的高频成分（因此我们称$\phi_s(\mathbf{x})$为高频函数）。

注意到$V_s(\mathbf{x})$对所有的超分辨率算法s有一个下界，这个下界由如下最优的高频函数达到：

$$\phi_{opt}(\mathbf{x};\tilde{p}_y)=\mathbf{V}\int_{\mathbf{y}}\mathbf{y}\tilde{p}_y(\mathbf{y}\mid\mathbf{x})\mathrm{d}\mathbf{y}\,.$$

它意味着最优的高频分量应当是给定低分辨率图像 **x** 时所有可能的高频分量的期望（注意到$\mathbf{V}\mathbf{y}=\mathbf{h}-\mathbf{U}\mathbf{x}$是高分辨率图像 **h** 的高频成分）。相应地，“最优”的超分辨率算法是

$$s_{opt}(\mathbf{x})=\phi_{opt}(\mathbf{x})+\mathbf{U}\mathbf{x}\,,$$

但是这个“最优”的超分辨率算法只有理论上的意义，它无法计算。

最优的高频函数的引入使我们不需要再关心不同超分辨率算法的细节，因为它达到了期望风险的下界，即任何超分辨率算法的期望风险都不可能比它的更低。因此我们只需要估计“最优”超分辨率算法的期望风险。容易看出，

$$V(\mathbf{x})=\int_{\mathbf{y}}\left\|\mathbf{V}\mathbf{y}\right\|^2\tilde{p}_y(\mathbf{y}\mid\mathbf{x})\mathrm{d}\mathbf{y}-\left\|\phi_{opt}(\mathbf{x};\tilde{p}_y\right\|^2.\tag{5}$$

另一方面，从一般自然图像的统计特性（$p_h(\mathbf{h})$的性质）我们可以论证以下的保守估计：

$$\left\|\phi_{opt}(\mathbf{x};\tilde{p}_y\right\|^2\le\frac{3}{4}\frac{\int_{\mathbf{y}}\left\|\mathbf{V}\mathbf{y}\right\|^2 p_{x,y}\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right)\mathrm{d}\mathbf{y}}{p_x(\mathbf{x})}.\tag{6}$$

其证明细节可参见[27]。综合（4）–（6）我们得到

$$\begin{aligned}\tilde{g}_s(N,m) &= \int_{\mathbf{x}} p_x(\mathbf{x})\left(\int_{\mathbf{y}} \|\mathbf{Vy}\|^2\ \tilde{p}_y(\mathbf{y}\,|\,\mathbf{x})\mathrm{d}\mathbf{y} - \left\|\phi_{opt}(\mathbf{x};\tilde{p}_y\right\|^2\right)\mathrm{d}\mathbf{x} \\ &\geq \frac{1}{4}\int_{\mathbf{x}} p_x(\mathbf{x})\int_{\mathbf{y}} \|\mathbf{Vy}\|^2\ \tilde{p}_y(\mathbf{y}\,|\,\mathbf{x})\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \\ &= \frac{1}{4}\int_{\mathbf{x},\mathbf{y}} \|\mathbf{Vy}\|^2\ p_{x,y}\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\ &= \frac{1}{4}\int_{\mathbf{h}} \|\mathbf{VQh}\|^2\ p_h(\mathbf{h})\mathrm{d}\mathbf{h} \\ &= \frac{1}{4}tr\left[(\mathbf{I}-\mathbf{UD})\mathbf{\Sigma}(\mathbf{I}-\mathbf{UD})^T\right] + \frac{1}{4}\left\|(\mathbf{I}-\mathbf{UD})\bar{\mathbf{h}}\right\|^2 \\ &\equiv \tilde{b}(N,m),\end{aligned}$$

其中 $\mathbf{\Sigma}$ 和 $\bar{\mathbf{h}}$ 分别是高分辨率图像的协方差矩阵和均值。以上推导中我们利用了等式 $\mathbf{VQ}=\mathbf{I}-\mathbf{UD}$，这可以从 $\begin{pmatrix}\mathbf{U} & \mathbf{V}\end{pmatrix}\begin{pmatrix}\mathbf{D}\\ \mathbf{Q}\end{pmatrix}=\mathbf{I}$ 得到。$\tilde{b}(N,m)$ 就是我们估计的期望风险的下界。

3.3 基于学习的超分辨率算法的极限

3.1 节里已经论证过，有效的放大倍数要使得超分辨率算法得到的高分辨率图像和真实图像接近。这个接近程度可以选作均方误差，即每像素的平均误差。虽然均方误差和人的视觉并不完全吻合，但目前并没有更好的量化准则。虽然如此，如果均方误差很大，还是可以肯定超分辨率效果不好。因此，我们可以选一个比较大的阈值 T，如果对某放大倍数 m_0，均方误差的下界 $b(N,m)=\left(\frac{1}{K}\tilde{b}(N,m)\right)^{\frac{1}{2}} > T$，其中 K 是高分辨率图像的像素数，则有效的放大倍数就不会超过 m_0。因此，基于学习的超分辨率算法的极限可以估计为 $b^{-1}(T)$。

3.4 下界的计算与阈值的选取

上节已经说明了估计基于学习的超分辨率算法的极限的基本思路，但是有两个问题没有解决。一是如何计算下界 $\tilde{b}(N,m)$；二是如何选取阈值 T。

要计算 $\tilde{b}(N,m)$，我们需要知道高分辨率图像的协方差矩阵和均值。虽然对一般自然图像的统计模型的研究已经有了一定的时间，比如[29]，但是所获得的模型都不准确：一般自然图像的统计特性符合该模型，但符合该模型的图像却未必是自然图像。虽然没有准确的模型可用，我们可以尝试从真实数据里估计需要的协方差矩阵和均值，即采集足够多的高分辨率图像，然后从这些样本里估算协方差矩阵和均值。于是就产生一个问题：采多少才够？下面的定理回答了这个问题。

定理 1：如果我们独立地采集 $M(p,\varepsilon)=\dfrac{(C_1+2C_2)^2}{16p\varepsilon^2}$ 张高分辨率图像，那么估计的下界 $\hat{\tilde{b}}(N,m)$ 偏离真实下界 $\tilde{b}(N,m)$ 的误差不超过 ε 的概率至少是 $1-p$，其中

$$C_1=\left\{E\left(\left\|(\mathbf{I}-\mathbf{UD})(\mathbf{h}-\bar{\mathbf{h}})\right\|^4\right)-tr^2\left[(\mathbf{I}-\mathbf{UD})\mathbf{\Sigma}(\mathbf{I}-\mathbf{UD})^T\right]\right\}^{\frac{1}{2}},$$

$$C_2=\left(\bar{\mathbf{b}}^T\mathbf{\Sigma}\bar{\mathbf{b}}\right)^{\frac{1}{2}},\ \bar{\mathbf{b}}=(\mathbf{I}-\mathbf{UD})^T(\mathbf{I}-\mathbf{UD})\bar{\mathbf{h}}.$$

其证明可在[27]里找到。在实验中，我们取 $p=0.01$，$\varepsilon=\dfrac{1}{2}Kb(N,m)$，使得均方误差的估计偏差 $\left|\hat{b}(N,m)-b(N,m)\right|\le 0.25$ 的概率高于 99%。我们选 0.25 作为阈值是因为它是灰度量化误差的均值。

我们对不同的 N 和 m 的取值按照定理 1 采集了相应数目的高分辨率图像，计算得下界曲线如图 2 所示。实验细节请参考[27]。可以看到，对于不同的 N，各曲线吻合得相当好，而且都按 $(m-1)^{\frac{1}{2}}$ 的方式增长。

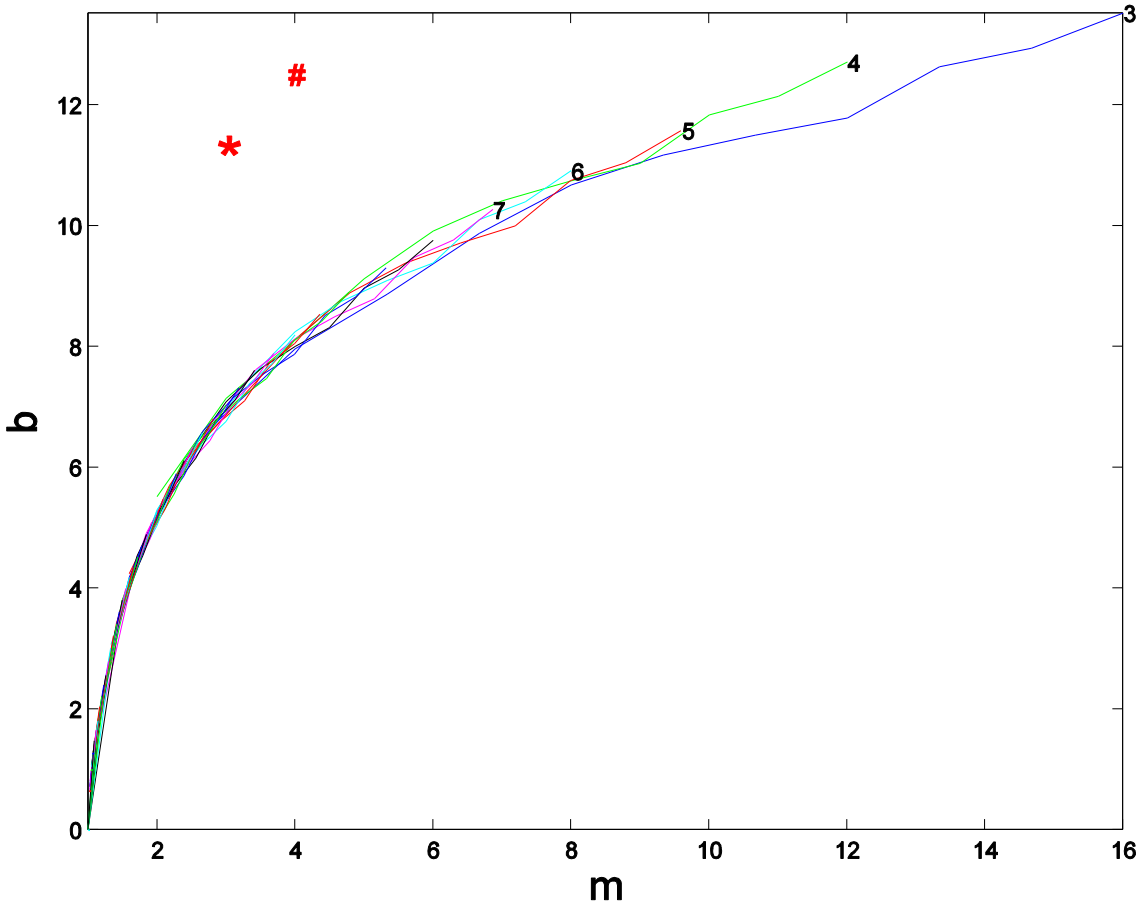

图 2. 从实际数据中计算得的均方误差下界曲线。对每条曲线，N 是固定的，它的值标在曲线末端（为了避免图太拥挤，大的 N 值没有标出）。‘*’和‘#’分别表示[13]和[5]里的算法的平均均方误差。

对于阈值T的选取，很不幸它并没有公认的数值，我们也只好从实验中估计。我们计算了用[13]和[5]里的算法获得的高分辨率图像和真实图像之间的均方误差，分别是 11.1 和 12.6，它们都在我们估计的下界上方（图 2）。从图 3 中我们看到，[13]和[5]里的算法的结果和真实图像的差别较大，因此取$T=11.1$是可以接受的。

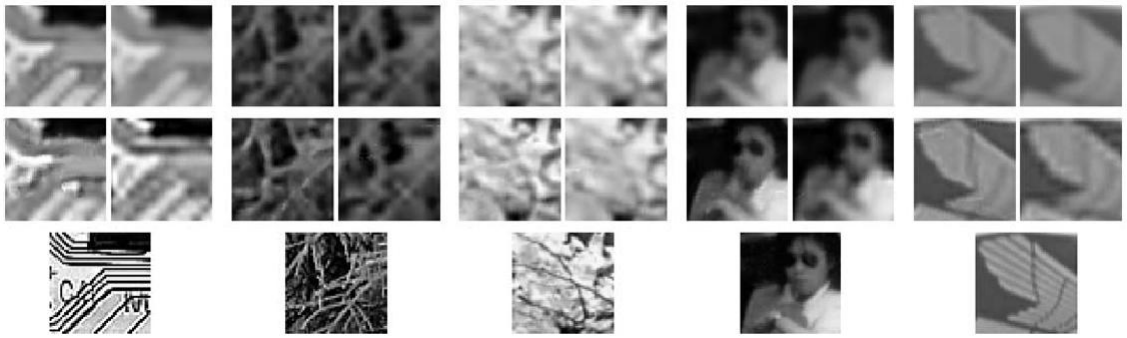

图 3. 用[13]（放大倍数为 3）和[5]（放大倍数为 4）算得的部分超分辨率结果。在每一组图像里，左上角是16×16的低分辨率图像，插值到48×48；左中是[13]的超分辨率结果；右上角是12×12的低分辨率图像，插值到48×48；右中是[5]的超分辨率结果；最下面是真实的高分辨率图像。

基于图 2 中的均方误差下界曲线和通过图 3 估计的阈值，我们可以粗略估计：基于学习的**通用**超分辨率算法的极限是 10 倍。当然这个估计有点保守，我们希望将来能做得更精细些。

3.5 讨论

上节估计的基于学习的通用超分辨率算法的极限有点保守，要得到更好的估计，一个可能的途径是考虑噪声。噪声的来源很多，一部分是成像过程中产生的，还有一部分是来源于我们数学模型的虚拟噪声（注意（3）里的经验风险的定义假定了不同的图像的成像过程是一样的）。如果把噪声考虑在内，经验风险应当定义为：

$$\tilde{g}_s^{'}(N,m)=\int_{\mathbf{h},\mathbf{n}}\left\|\mathbf{h}-s(\mathbf{Dh}+\mathbf{n})\right\|^2 p_{h,n}\left(\begin{pmatrix}\mathbf{h}\\ \mathbf{n}\end{pmatrix}\right)\mathrm{d}\mathbf{h}\mathrm{d}\mathbf{n},$$

其中$p_{h,n}$是高分辨率图像和噪声的联合分布。这里的噪声包含了数学模型的虚拟噪声。在新模型下，如果高分辨率图像和噪声是独立的，我们可以类似得到经验风险的下界：

$$\tilde{b}^{'}(N,m)=\frac{1}{4}tr\left[(\mathbf{I}-\mathbf{UD})\mathbf{\Sigma}(\mathbf{I}-\mathbf{UD})^T\right]+\frac{1}{4}tr\left(\mathbf{U}\mathbf{\Sigma}_n\mathbf{U}^T\right)+\frac{1}{4}\left\|(\mathbf{I}-\mathbf{UD})\bar{\mathbf{h}}-\mathbf{U}\bar{\mathbf{n}}\right\|^2,$$

其中$\mathbf{\Sigma}_n$和$\bar{\mathbf{n}}$分别是噪声的协方差矩阵和均值。很显然$\tilde{b}^{'}(N,m)>\tilde{b}(N,m)$，因为$\bar{\mathbf{h}}$基本上是常值向量，导致$(\mathbf{I}-\mathbf{UD})\bar{\mathbf{h}}\approx\mathbf{0}$。所以，如果我们能够计算出$\tilde{b}^{'}(N,m)$，我们就能得到超分辨率算法的极限的更精确的估计。不幸的是，$\tilde{b}^{'}(N,m)$的计算比$\tilde{b}(N,m)$困难得多，因为我们暂时无法通过分离噪声来估计它的协方差矩阵和均值，特别是这里的噪声还包含着虚拟噪声。

4 结语

基于学习的超分辨率算法是新兴的超分辨率算法，比起传统算法它既有优点也有缺点。为了达到更好的性能，可以从很多方面来考虑，比如：

1. 可否更好地表达和利用先验知识？
2. 可否对低分辨率图像进行简单的分析，使得不同部分可以使用不同的先验知识？
3. 可否刻画出超分辨率算法的性能和样本库之间的关系？
4. 可否设计出在一定放大倍数范围之内都可工作的算法？

本人希望将来能看到对超分辨率算法的更多的理论分析。

参考文献

[1] Tsai R Y, Huang T S. Multiple frame image restoration and registration. In: *Advances in computer vision and image Processing*, Greenwich: JAI Press, 1984, 317–339

[2] Borman S, Stevenson R L. Spatial resolution enhancement of low-resolution image sequences: a comprehensive review with directions for future research. *Technical Report*, University of Notre Dame, 1998

[3] Farsiu S, Robinson D, Elad M, Milanfar P. Advances and challenges in superresolution. *International Journal of Imaging Systems and Technology*, 2004, 14(2):47-57

[4] Park S C, Park M K, Kang M G. Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, 2003, 20(3):21-36

[5] Freeman W T, Pasztor EC. Learning low-level vision. In: *Proceedings of Seventh International Conference on Computer Vision*, Corfu, Greece, 1999, 1182-1189

[6] Bégin I, Ferrie F R. Blind super-resolution using a learning-based approach. In: *Proceedings of International Conference on Pattern Recognition*, 2004, 1:549-552

[7] Bishop C M, Blake A, Marthi B. Super-resolution enhancement of video. In Bishop CM and Frey B (Eds.), *Proceedings of Artificial Intelligence and Statistics*. Society for Artificial Intelligence and Statistics, 2003

[8] Candocia F M, Principe J C. Super-resolution of images based on local correlations. *IEEE Transactions on Neural Network*, 1999, 10(2):372-380

[9] Hertzmann A, Jacobs C E, Oliver N, Curless B, Salesin DH. Image analogies. *SIGGRAPH*, 2001

[10] Kursun O, Favorov O. Single-frame super-resolution by inference from learned features. *Istanbul University Journal of Electrical & Electronics Engineering*, 2003, 3(1):673-681

[11] Miravet C, Rodríguez F de B. A hybrid MLP-PNN architecture for fast image superresolution. In: *Proceedings of International Conference on Artificial Neural Network*, 2003, 417-424

[12] Pickup L, Roberts S J, Zisserman A. A sample texture prior for image superresolution. In: *Proceedings of Advances in Neural Information Processing Systems*, 2003, 1587-1594

[13] Sun J, Tao H, Shum H-Y. Image hallucination with primal sketch priors. In: *Proceedings of Computer Vision and Pattern Recognition*, 2003, II:729-736

[14] Zhang L, Pan F. A new method of images super-resolution restoration by neural

networks. In: *Proceedings of Ninth International Conference on Neural Information Processing*, 2002, 5:2414-2418
[15] Baker S, Kanade T. Limits on super-resolution and how to break them. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2002, 24(9):1167-1183
[16] Capel D, Zisserman A. Super-resolution from multiple views using learnt image models. In: *Proceedings of Computer Vision and Pattern Recognition*, 2001, II:627-634
[17] Dedeoğlu G, Kanade T, August, J. High-zoom video hallucination by exploiting spatio-temporal regularities. In: *Proceedings of Computer Vision and Pattern Recognition*, 2004, II:151–158
[18] Gunturk B K, Batur A U, Altunbasak Y, Hayes III M H, Mersereau R M. Eigenface-domain super-resolution for face recognition. *IEEE Transactions on Image Processing*, 2003, 12(5):597-606
[19] Li Y, Lin X. An improved two-step approach to hallucinating faces. In: *Proceedings of Third International Conference on Image and Graphics*, 2004, 298-301
[20] Li Y, Lin X. Face hallucination with pose variation. In: *Proceedings of Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2004, 723-728
[21] Liu C, Shum H-Y, Zhang C S. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In: *Proceedings of Computer Vision and Pattern Recognition*, 2001, 192-198
[22] Lin D, Liu W, Tang X. Layered local prediction network with dynamic learning for face super-resolution. In: *Proceedings of IEEE International Conference on Image Processing*, Genova, Italy, 2005
[23] Liu W, Lin D, Tang X. Hallucinating faces: TensorPatch super-resolution and coupled residue compensation. In: *Proceedings of Computer Vision and Pattern Recognition*, 2005, II:478-484
[24] Liu W, Lin D, Tang X. Neighbor combination and transformation for hallucinating faces. In: *Proceedings of IEEE International Conference on Multimedia and Expo*, Amsterdam, Netherlands, 2005
[25] Liu W, Lin D, Tang X. Face hallucination through dual associative learning. In: *Proceedings of IEEE International Conference on Image Processing*, Genova, Italy, 2005
[26] Roweis S, Saul L. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000, 290:2323–2326
[27] Lin Z, He J, Tang X, Tang C-K. Limits of learning-based superresolution algorithms. *International Journal of Computer Vision*, 2008, 80(3):406-420
[28] Lin Z, Shum H-Y. Fundamental limits of reconstruction-based superresolution algorithms under local translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, 26(1):83-97
[29] Srivastava A, Lee A B, Simoncelli E P, Zhu S-C. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 2003, 18:17-33