Double Structured Nuclear Norm-Based Matrix Decomposition for Saliency Detection

JUNXIA LI (Member, IEEE), ZIYANG WANG, AND ZEFENG PAN
Jiangsu Key Laboratory of Big Data Analysis Technology (B-DAT), Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China
Corresponding author: Junxia Li (junxiali99@163.com)

ABSTRACT

Saliency detection aims at identifying the most important and informative area in a scene. Recently low rank matrix recovery (LR) theory becomes an effective tool for saliency detection. The existing LR-based methods all work under the popular low rank and sparsity pursuit framework and perform well for images with small or homogeneous objects. However, if the image is with heterogeneous objects, the sparsity property of the object cannot be guaranteed. Moreover, as a useful tool for depicting a spatially structured matrix variable, nuclear norm (corresponding to the low rank) considers only the global structure but overlooks the inherent local structure of the data. We address these problems by proposing a double structured nuclear norm-based matrix decomposition (DSNMD) model for saliency detection. In the model, a tree-structured nuclear (TSN) norm is firstly introduced to constraint both the background and foreground regions. We also empirically demonstrate that TSN norm provides stronger performance at capturing the underlying structural information of the image regions including global structure, local structure, and internal structure of each node of the tree, and it deservedly inherits the advantages of both nuclear norm and sparsity-related norms (e.g., \( \ell_1 \)-norm, group sparsity norm) for saliency detection. Comprehensive evaluations on six benchmark datasets indicate that our method universally surpasses state-of-the-art unsupervised methods and performs favorably against supervised approaches.

INDEX TERMS

Saliency detection, tree structure, matrix decomposition, double structured nuclear norm, structural information.

I. INTRODUCTION

Saliency detection aims to detect distinctive regions in an image that draw human attention. Recently, it has attracted much attention in the vision community for its wide range of applications such as image foreground annotation [2], content-aware image/video retargeting [4] and compression [5], objects-of-interest segmentation [6], image/video summarization [7], person re-identification [8], etc.

With the aim to facilitate various applications such as those mentioned above, many computational models for saliency detection have been proposed in recent years [18], [19], [56], [62] [9], [10], [12], [41], [50], [51]. They include biological plausible model-based approaches to simulate the visual attention mechanism and data-driven methods, in order to compute a saliency map that indicates salient regions in an image. The commonly adopted mechanism for saliency estimation is based on local, global and regional contrasts with different forms, trying to extract regions that show distinct characteristics from their surroundings. Besides the widely exploited contrasts based mechanism, there are various formulations for saliency measurement based on different theories and principles such as graph theory [20], [54], [55], information theory [21], [22], spectral analysis [5], [23], [64], statistical model [30], and deep learning [1], [3], [26]–[29]. Most of these approaches may work well the images with relatively simple background and homogeneous objects. However, they may falter in distinguishing between salient objects and irrelevant clutter which are not part of salient object when both regions have high local stimuli. Deep learning-based
approaches achieve a relatively high accuracy in saliency detection. However, as we known, their training step is often time-consuming.

Recently, low rank matrix recovery (LR) theory becomes a powerful tool for saliency detection and achieves impressed results [24], [25]. The aim of the LR is to recover a low rank matrix from corrupted observations, in which the corrupted entries are unknown and the errors can be arbitrarily large, but are assumed to be sparse [46]. Yan et al. [31] firstly introduce the LR theory into the field of saliency detection, which derives from the consideration that the background regions usually lie in a low-dimensional subspace in a certain feature space and thus can be approximated by a low rank matrix, while the object can be regarded as the noise. In [31], sparse coding is used as an intermediate representation of image features and then fits to LR model to recover salient objects. Shen and Wu [44] propose to modulate the image features with learnt transform matrix to meet the low rank and sparse properties. To capture the underlying structure of the object matrix cannot be guaranteed. Moreover, all these methods use nuclear norm to depict the low rank matrix, which considers only the global structure but overlooks the inherent local structure of the data.

To address these problems, in this paper we propose a double structured nuclear norm based matrix decomposition (DSNMD) model for saliency detection. In DSNMD, a tree-structured nuclear (TSN) norm is firstly introduced to constraint both the background and foreground regions. We demonstrate that TSN norm provides stronger performance in capturing the underlying structural information of image regions including global structure, local structure, and internal structure of each group, and it deservedly inherits the advantages of both nuclear norm and sparsity-related norms (e.g., $\ell_1$-norm, group sparsity norm) for saliency detection.

The main contributions of this work can be summarized as follows:

- A tree-structured nuclear norm is introduced to constraint both the background and foreground matrices.
- We establish a novel double structured nuclear norm-based matrix decomposition (DSNMD) model for salient object detection.
- An effective alternating direction method of multipliers (ADMM) based optimization algorithm is presented to solve the proposed DSNMD model.
- We develop a DSNMD-based salient object detection framework, which can seamlessly integrate low-level bottom-up saliency and high-level priors in a unified way.
- Experiments on six datasets show that the proposed method achieves better performance compared with 23 state-of-the-art methods.

The remainder of the paper is organized as follows: Section II reviews and analyzes the previous work, especially the LR-based approaches. In Section III, we describe the proposed DSNMD-based saliency detection model in details. The experimental results are listed in Section V. Finally, Section VI gives the conclusions of the paper.

II. RELATED WORK

A. SALIENT OBJECT DETECTION

From the perspective of information processing mechanisms, existing saliency detection methods can be roughly grouped into data-driven bottom-up, task-oriented top-down and combination of bottom-up and top-down approaches. Bottom-up methods mainly rely on low level cues such as color, intensity, orientation, texture, etc. and try to extract regions that show distinct characteristics from their surroundings [10], [56], [58], [61]–[63]. Bottom-up approaches are fast to implement and require no prior knowledge of the image, but the methods may falter in distinguishing between salient objects and irrelevant clutter which are not part of salient object when both regions have high local stimuli. On the contrary, top-down approaches are task driven and make use of high-level human perceptual priors knowledge about the scene or the context to identify the salient regions [11], [13]–[15]. However, top-down methods demand more complete understanding of the context of the image which results in high computational costs. Moreover, the high diversity of object types limits the generalization and scalability of these approaches. Integration of top-down and bottom-up approaches has been discussed in [16], [17].

B. LR-BASED SALIENCY

Our work in this paper is related to recent methods that consider the sparsity prior in saliency detection. Compared with the deep learning based methods, the LR-based approaches could be treated as shallow learning models with shallow-structured architectures. Yan et al. [31] adopt an over-complete dictionary to encode image patches and then feed the coding vector to the LR model to recover salient objects. After that, Shen and Wu [44] devise a supervised method to exploit feature transformation with the high-level semantic, color, and center priors to meet the low-rank and sparse assumption. To better match the LR model, Zou et al. [42] leverage the segmentation prior driven from the connectivity between each region and image border to guide matrix recovery. Most of these methods use RPCA model for low-rank matrix recovery:
objects, and $B$ represents the low-rank matrix corresponding to the non-salient background. $\| \cdot \|_s$ denotes the matrix nuclear norm (sum of the singular values of a matrix), $\| \cdot \|_1$ is the $\ell_1$-norm which promotes sparsity, and the parameter $\lambda > 0$ balances the effects between two parts. Peng et al. [32], [41] introduce the structured sparsity into the RPCA model and present a structured matrix decomposition (SMD) model, which formulates the task of salient object detection as a problem of low-rank and structured sparse matrix decomposition.

$$\min_{B,F} \|B\|_s + \lambda \Omega(F), \quad s.t. \quad X = B + F,$$

(2)

where $\Omega(\cdot)$ is a structured sparsity regularization to capture the spatial and feature relations of patches in $F$. SMD model is actually an extension of the RPCA model on sparse constraints. To exploit the weak sparsity in the case of the salient objects with big size, Tang et al. [45] estimate a high-level background prior by employing the location, color, and boundary priors to weight the image feature matrix.

Different from the previous LR-based methods, this paper first introduces a tree-structured nuclear (TSN) norm to constrain both the background and foreground regions, which has better performance at capturing the underlying structural information of image regions including global structure, local structure, and achieves better performance in the task of saliency predication.

### III. SALIENCY DETECTION VIA DOUBLE STRUCTURED NUCLEAR NORM-BASED MATRIX DECOMPOSITION MODEL

As discussed in the related work, the state-of-the-art LR-based saliency detection models are work under the low rank and sparsity pursuit framework. However, if the image is with large-scale or heterogeneous objects, the sparsity property of the object cannot be guaranteed. In this paper, we introduce a tree-structured nuclear (TSN) norm to constrain both the background and foreground regions, and propose a DSNMD model for saliency detection. Fig. 1 shows the framework of DSNMD-based salient object detection.

#### A. PROBLEM FORMULATION

As shown in the left column of Fig. 1, the given image $I$ is first decomposed into a set of super-pixels $\{P_i\}_{i=1}^{N}$, where $N$ is the number of super-pixels. For each super-pixel $P_i$, a $D$-dimension feature vector is extracted and denoted as $x_i \in R^D$. By arranging these $N$ vectors into a matrix, we get the matrix representation of the original image $X = [x_1, x_2, \ldots, x_N] \in R^{D \times N}$. The next problem is to design an effective model to decompose the feature matrix $X$ into two parts, the matrix $B$ corresponding to the background regions and the matrix $F$ corresponding to the salient foreground regions.

$$X = B + F.$$

(3)

Eq. 3 represents a severely under-constrained problem since it is virtually impossible to infer matrices $B$ and $F$ without additional information. In other words, without imposing any restrictions on Eq. 3, there are infinite number of solutions for $B$ and $F$. To seek a suitable solution that is benefit for our saliency detection, some criteria for characterizing matrices $B$ and $F$ are needed.

#### B. DSNMD MODEL

1) TREE-STRUCTURED NUCLEAR NORM

Before detailing the structured nuclear norm, we first give the definition of an index tree [33]. An index tree is a hierarchical structure, and each node contains a group of indices. In this paper, an index corresponds to a super-pixel. For an index tree $T$ with depth $m$ including $N$ indices $\{1, 2, \ldots, N\}$, let $G_j^i$ be the $j$th node at the $i$th level, $T_i = \{G_1^i, \ldots, G_m^i\}$ is the set of all nodes in layer $i$, where $n_i$ represents the number of nodes. In particular, for the root node, there is $T_1 = G_1^1 = \{1, 2, \ldots, N\}$. The nodes in the index tree satisfy two conditions: 1) there is no overlap between the indices of
nodes from the same level, i.e., for \( \forall 2 \leq i \leq m, 1 \leq j, k \leq n_i, j \neq k, G_j^i \cap G_k^i = \emptyset \); let \( G_j^{i-1} \) be the parent node of a non-root node \( G_j^i \), there is \( G_j^i \subseteq G_j^{i-1} \) and \( \bigcup_j G_j^i = G_j^{i-1} \).

An example tree drawn from a hierarchical segmentation of an image is given in the “Index Tree Construction” block of Fig. 1. Here, we use an index tree to encode the spatial relation of image super-pixels.

Unlike Group Sparsity regularization, they constrain each group vector with mixed \((\ell_1, \ell_2)\) or \((\ell_1, \ell_{\infty})\). Tree-Structured Nuclear (TSN) norm uses nuclear norm to depict the structure between each group vector (i.e., tree nodes). Given matrix \( M \in R^{n \times d} \), the TSN norm is defined as [34]:

\[
\Omega(M) = \sum_{i=1}^{d} \sum_{j=1}^{n_i} w_{i,j} \| M_{G_j^i} \|_2 = \| M_{G_j^i} \|_{\Omega_{i,j}}
\]  

(4)

where \( w_{i,j} \geq 0 (i = 1, \ldots, m; i = 1, \ldots, n_i) \) is the weight for the node \( G_j^i \), \( M_{G_j^i} \) is the sub-matrix of \( M \) corresponding to the node \( G_j^i \). Obviously, the \( \Omega(M) \) is still a norm.

2) MATRIX RECOVERY FOR SALIENCY DETECTION

Through the observation and experiment of a large number of images, the following conclusions are obtained:

- Similar to the existing methods [31], [41], [44], we also consider that the feature vectors of background super-pixels have strong correlation and thus lie in a low-dimensional subspace;
- Different from the existing methods, we consider that the object super-pixels also lie in a low-dimensional subspace independent of the background subspace.

Based on the above conclusions, we use TSN norm to describe both the foreground matrix \( F \) and the background matrix \( B \). In addition, inspired by Peng et al. [41], a Laplacian regularization is introduced to the model, which can enlarge the gaps between salient objects and the background in feature space.

\[
\Theta(B, F) = \frac{1}{2} \sum_{i,j} \| f_i - f_j \|_2^2 V_{i,j}
\]  

(5)

where \( f_i \) is the \( i^{th} \) column of \( F \). The similarity matrix (i.e., affinity matrix) \( V = (V_{i,j}) \in R^{N \times N} \) is defined as:

\[
V_{i,j} = \begin{cases} 
\exp(-\frac{\| x_i - x_j \|_2^2}{2\sigma^2}), & \text{if } (P_i, P_j) \in \Delta; \\
0, & \text{otherwise},
\end{cases}
\]  

(6)

where \( x_i \) and \( x_j \) correspond to the feature vectors of super-pixel \( P_i \) and \( P_j \), respectively. \( \Delta \) represents the set of adjacent super-pixel pairs.

Based on the above considerations, the following initial model is established:

\[
\min_{B,F} \| B_{G_j^i} \|_{w_{ij},x} + \lambda \| F_{G_j^i} \|_{w_f,x} + \frac{\alpha}{2} \sum_{i,j} \| f_i - f_j \|_2^2 V_{i,j},
\]  

s.t. \( X = B + F \),

(7)

where \( w_{ij} \) is the weight of \( B_{G_j^i} \), and \( w_{ij} \) is the weight of \( F_{G_j^i} \).

Because (8) is formally a matrix decomposition model, it is called Double Structured Nuclear Norm-based Matrix Decomposition (DSNMD).

3) DISCUSSION

Compared with the RPCA model (1) and SMD model (2), the similarity of the three models lies in the assumption that the matrix corresponding to the background regions is low rank. Both RPCA and SMD use nuclear norm to depict the background matrix, which can be understood as they only use “global low rank”, while DSNMD model uses structured nuclear norm to describe the global structure and local structure of background matrix. Another key difference is that both RPCA and SMD assume that the foreground matrix is sparse. However, when the object scale is large or the internal characteristics of the object are inconsistent, the assumption that the object has sparse property will no longer hold. DSNMD considers that the object matrix is also low rank, and feature vectors of foreground regions and background regions are lie in two independent low dimensional subspaces, so DSNMD also uses structured nuclear norm to describe the foreground matrix. Although SMD uses structured sparse norm to describe the target matrix, compared with \( \ell_1 \) norm in RPCA, the structured sparse norm takes into account the spatial relationship between image regions, but ignores the internal structure of each tree node. Structured nuclear norm has the advantages of nuclear norm, \( \ell_1 \) norm and structured sparse (also known as group sparse) norm.

C. SALIENCY MAP PREDICTION

Following [41], [44], we extract three kinds of features from the original image including RGB color (5-dimension), steerable pyramids (12-dimension) and Gabor filter (36-dimension), to construct a 53-dimension feature representation. Then, we use the simple linear iterative clustering (SLIC) [48] to over-segment the image into \( N \) super-pixels, and each super-pixel is represented by a 53-dimension feature vector. In the index tree construction, the similarity of every adjacent super-pixel pair is first calculated by Eq. 6. Then, the spatially neighboring super-pixels are merged according to their similarity by a graph-based image segmentation algorithm [49]. After several rounds of region merging, a series of segmentations with increasing granularity (i.e., super-pixel scale) are obtained, in which the granularity is determined by a threshold \( T \). In each granularity layer, the segments correspond to the nodes at the corresponding layer in the index tree. Finally, a hierarchical segmentation
with granularity from fine to coarse is obtained, and a multi-layer index tree structure is established.

In solving DSNMD, we default \( w_{ij}^{bg} = 1 \) and \( w_{ij}^{fg} = 1 \). Let \( F^* \) be the optimal solution (with respect to \( F \)) to problem (8). To obtain a saliency value for each super-pixel \( P_i \), we define a simple assignment function on the low rank matrix \( F^* \):

\[
S(P_i) = \sum_{j=1}^{D} |F^*(j, i)|.
\]

(9)

A larger response of \( S(P_i) \) means a higher saliency rendered on the corresponding super-pixel \( P_i \). The resulting saliency map is obtained through merging all super-pixels together. After normalizing, we can get the final saliency map \( S \).

1) EXTENSION TO INTEGRATE HIGH-LEVEL PRIORS

Inspired by [41], we further extend the proposed DSNMS-based saliency detection to integrate high-level priors. We fuse three types of priors, i.e., location, color and background prior, to generate a high-level prior map. These three priors are finally multiplied together to produce the high-level prior map.

For each super-pixel \( P_i \), its high-level prior, denoted by \( p_i \in [0, 1] \), indicates the likelihood that \( P_i \) belongs to a salient object based on high-level information. Therefore, \( 1 - p_i \) denotes the likelihood that \( P_i \) belongs to the background region. They are encoded into the DSNMD by weighting each component in the tree-structured nuclear norm differently. In particular, we define \( w_{ij}^{bg} \) and \( w_{ij}^{fg} \) as

\[
\begin{align*}
  w_{ij}^{bg} & = \max((p_k \mid k \in G^*_j)), \\
  w_{ij}^{fg} & = 1 - \max((p_k \mid k \in G^*_i)).
\end{align*}
\]

(10)

In this way, the high-level prior knowledge is seamlessly encoded into the DSNMD model to guide the matrix decomposition and enhance the saliency detection. It is worth noting that if we fix \( w_{ij}^{bg} = 1/2 \) and \( w_{ij}^{fg} = 1/2 \) for each node \( G^*_j \), the proposed model is degraded to the pure low-level saliency detection model.

D. OPTIMIZATION VIA ADMM

The DSNMD problem (8) is convex and could be solved by the popular Alternating Direction Method of Multipliers (ADMM) [35]. We first introduce an auxiliary variable \( E \) to make the objective function separable, and thus convert Eq. (8) into the following equivalent problem:

\[
\begin{align*}
\min_{B, F, E, Y_1, Y_2} & \quad \|B_{G^*_j}\|_{w_{ij}^{fg},*} + \lambda \|F_{G^*_i}\|_{w_{ij}^{fg},*} + atr(EE^T), \\
\text{s.t.} & \quad X = B + F, \quad F = E.
\end{align*}
\]

(11)

The augmented Lagrangian function of problem (11) is

\[
\begin{align*}
L_\mu(B, F, E, Y_1, Y_2) &= \|B_{G^*_j}\|_{w_{ij}^{fg},*} + \lambda \|F_{G^*_i}\|_{w_{ij}^{fg},*} + atr(EE^T) \\
&+ tr(Y_1^T(X - B - F)) + tr(Y_2^T(F - E)) \\
&+ \frac{\mu}{2} \|X - B - F\|^2_F + \|F - E\|^2_F,
\end{align*}
\]

(12)

where \( Y_1 \) and \( Y_2 \) are the Lagrange multipliers, \( \mu > 0 \) is the penalty parameter. The standard augmented Lagrange multiple method minimizes \( L_\mu \) with respect to variables \( B, F \) and \( E \) simultaneously. However, to exploit the property that the variables \( B, F \) and \( E \) in objective function are separable, ADMM decomposes the minimization of \( L_\mu \) into three sub-problems which minimizes \( B, F \) and \( E \), respectively.

Then, the ADMM for problem (11) consists of the iterations as follows:

(i) Given \( F = F^k, E = E^k, Y_1 = Y_1^k, Y_2 = Y_2^k \), updating \( B \) by

\[
B^{k+1} = \arg\min_B L_\mu(B, F^k, E^k, Y_1^k, Y_2^k),
\]

(13)

(ii) Given \( B = B^{k+1}, E = E^k, Y_1 = Y_1^k, Y_2 = Y_2^k \), updating \( F \) by

\[
F^{k+1} = \arg\min_F L_\mu(B^{k+1}, F, E^k, Y_1^k, Y_2^k),
\]

(14)

(iii) Given \( B = B^{k+1}, F = F^{k+1}, Y_1 = Y_1^k, Y_2 = Y_2^k \), updating \( E \) by

\[
E^{k+1} = \arg\min_E L_\mu(B^{k+1}, F^{k+1}, E, Y_1^k, Y_2^k),
\]

(15)

(iv) Given \( B = B^{k+1}, F = F^{k+1}, Y_1 = Y_1^k \), updating \( Y_1 \) by

\[
Y_1^{k+1} = Y_1^k + \mu(X - B^{k+1} - F^{k+1}),
\]

(16)

(v) Given \( F = F^{k+1}, E = E^{k+1}, Y_2 = Y_2^k \), updating \( Y_2 \) by

\[
Y_2^{k+1} = Y_2^k + \mu(F^{k+1} - E^{k+1}).
\]

(17)

The key steps are to solve problems (13), (14) and (15). Simple manipulation shows that problem (13) is equivalent to

\[
B^{k+1} = \arg\min_B \|B_{G^*_j}\|_{w_{ij}^{fg},*} + \frac{1}{2} \|B - J_B\|^2_F,
\]

(18)

where \( \tau = 1/\mu \) and \( J_B = X - B^k + Y_1^k/\mu \). The optimal solution can be computed via the singular value thresholding algorithm [36]. Specifically, given a matrix \( Q \in R^{p \times q} \) of rank \( r \), its skinny singular value decomposition (SVD) of \( Q \) is

\[
Q = U_{p \times \Sigma} \Sigma V_{q \times r}^T, \quad \Sigma = diag(\sigma_1, \ldots, \sigma_r),
\]

(19)

where \( \sigma_1, \ldots, \sigma_r \) are positive singular values, and \( U_{p \times \Sigma} \) and \( V_{q \times r} \) are corresponding matrices with orthogonal columns. For a given \( \tau > 0 \), the singular value shrinkage operator \( D_\tau(\cdot) \) is defined as follows

\[
D_\tau(Q) = U_{p \times \Sigma(diag([\max(0, \sigma_j - \tau)]_{1 \leq j \leq r}))} V_{q \times r}^T.
\]

(20)

*Theorem 1 [36]: For each \( Q \in R^{p \times q} \) and \( \tau > 0 \), the singular value shrinkage operator obeys

\[
D_\tau(Q) = \arg\min_H \|H\|_\Sigma + \frac{1}{2} \|H - Q\|_F^2.
\]

(21)

Clearly, (18) is an extension for the objective of (21) from the viewpoint of hierarchical structure. Fortunately, we
Algorithm 1 Solving Problem (18)

Input: Matrix \( J_B \), the index tree \( T \) with nodes \( G_j^i \) \( (i = 1, \ldots, d; j = 1, \ldots, n_i) \), weights \( w_{ij}^g \geq 0 \) \( (i = 1, \ldots, d; j = 1, \ldots, n_i) \), \( \tau > 0 \), \( \tau_{ij} = \tau w_{ij} \).

1: Set \( B^d = J_B \).
2: For \( i = d \) to 1 do
3:  \( \text{For } j = 1 \) to \( n_i \) do
4:      Compute \( B^i_{G_j} = D_{\tau_{ij}}(B^{i+1}) \)
5:  End for
6: End for

Output: The optimal solution \( B^1 \).

Algorithm 2 Solving DSNMD by ADMM

Input: Data matrix \( X \), index tree \( T = \{ G_j^i \} \), parameters \( \lambda, \alpha, w_{ij}^g, w_{ij}^r, \epsilon_1, \epsilon_2 \).

1: Initializing: \( B^0 = 0, F^0 = 0, E^0 = 0, Y_1^0 = 0, Y_2^0 = 0, \mu = 0.1, \mu_{max} = 10^6, \rho = 1.2, k = 0 \).
2: While not converged do
3:  Updating \( B^{k+1} \) by Algorithm 1;
4:  Updating \( F^{k+1} \) by Algorithm 1;
5:  Updating \( E^{k+1} \) by Eq. (24);
6:  Updating the multipliers \( Y_1^{k+1} \) and \( Y_2^{k+1} \) by Eq. (16) and Eq. (17), respectively;
7:  Updating the parameter \( \mu = \min(\rho \mu, \mu_{max}) \).
8: If Eq. (25) is not satisfied go to Step 3.
9: End while

Output: The optimal solution \( F^* \).

still can obtain the closed form solution of (18) by Algorithm 1. Furthermore, we have proved that \( S^1 \) returned by Algorithm 1 is the unique solution to (18).

Problem (14) is equivalent to

\[
F^{k+1} = \arg \min_{F} \gamma \| F_{G_j^i} \|_{\psi_{\rho \epsilon}}^{\rho} + \frac{1}{2} \| F - J_F \|_{F}^2,
\]

where \( \gamma = \lambda/2\mu \) and \( J_F = (X - B^{k+1} + E^{k} + (Y_1^{k+1} - Y_2^{k+1})/\mu)/2 \). Similarity, (22) can be calculated by Algorithm 1 with the ‘input’ changed correspondingly.

From problem (15), we derive the following problem:

\[
E^{k+1} = \arg \min_{E} \alpha tr(EE^T) + tr((Y_2^{k+1})^T (F^{k+1} - E)) + \frac{\mu}{2} \| F^{k+1} - E \|_{F}^2.
\]

Taking derivative of the objective function in (23), we get

\[
E^{k+1} = (\mu F^{k+1} + Y_2^{k+1})(2\alpha I + \mu I)^{-1}.
\]

Boyd et al. [35] give the optimality conditions and stopping criteria of the ADMM algorithm. Based on the results in [35], we use the following termination conditions:

\[
\| X - B^{k+1} - E^{k+1} \|_{F}^2 / \| X \|_{F}^2 < \epsilon_1,
\]

\[
\max(\| B^{k+1} - B^k \|_{F}^2 / \| X \|_{F}^2, \| F^{k+1} - F^k \|_{F}^2 / \| X \|_{F}^2) < \epsilon_2.
\]

where \( \epsilon_1 \) and \( \epsilon_2 \) are the termination condition parameters.

The detailed ADMM algorithm for DSNMD is summarized in Algorithm 2.

1) CONVERGENCE ANALYSIS

There have been many studies focusing on the convergence of ADMM. Especially, utilizing the properties of the saddle points, Boyd et al. [35] analyzed convergence of ADMM with two variables. He et al. [37], [38] presented some significant convergence results by virtue of variational inequalities. What’s more, He et al. [39] illuminated that the ADMM owns a convergence rate of \( O(1/k) \), where \( k \) is the iteration number. Nonetheless, for more than two blocks of variables, it still has not been done to provide an affirmative convergence proof. Recently, Chen et al. [40] presented a sufficient condition to ensure the convergence of the direct extension of ADMM, and they obtained an important result: the direct extension of ADMM is not necessarily convergent. Considering the above results, it is enough that we use Eq. (25) as a stopping criterion.

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENTAL SETUP

1) DATASETS

To comprehensively evaluate the performance of our proposed model, we conduct extensive experiments on six publicly-available datasets: ASD [63], THUS [67], DUT-OMRON [54], iCoSeg [68], SOD [66], and ECSSD [69]. ASD contains 1,000 images with accurate human-marked labels for salient objects. The images in this dataset are relatively simpler than the other five datasets that we consider. THUS is the largest dataset which consists of 10,000 images. DUT-OMRON contains 5,168 images carefully labeled by five users. The images in ASD, THUS, and DUT-OMRON datasets are with a single salient object. The iCoSeg is a publicly-available co-segmentation dataset, including 38 groups of totally 643 images. Each image in iCoSeg may contain one or multiple salient objects. In this paper, we use it to evaluate the performance of salient object detection. The SOD dataset is composed of 300 images from Berkeley segmentation dataset. Some of the images in SOD include more than one salient object. ECSSD dataset contains 1,000 images. In ECSSD, many images contain multiple objects with various locations and scales, and highly cluttered background, which makes it very challenging for saliency detection.
2) EVALUATION METRICS
To evaluate the performance of our proposed method, we test our results based on four metrics including the precision-recall (PR) curve, the F-measure curve, the mean absolute error (MAE) score [61], and the recently proposed weighted F-measure (WF) score [71] 

The precision is defined as the ratio of correctly detected salient pixels to all the detected pixels, and the recall is the fraction of correctly detected salient pixels belonging to the salient pixels in ground truth. The F-measure is obtained as
\[
\text{F-measure} = \frac{1 + \beta^2 \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}} \quad (\text{P = precision, R = recall}).
\]

We set \(\beta = 0.3\) which is the same as in [51], [58], [63]. The PR curve and F-measure curve are created by varying the saliency threshold from 0 to 255.

As complementary to precision and recall rates, we also report the MAE score to evaluate the performance of our proposed method. The MAE calculates the average difference between a continuous saliency map \(S\) and the binary ground truth \(G\) in pixel level and is defined as 
\[
\text{MAE} = \frac{1}{|S - G|} \sum |S - G|.
\]

3) PARAMETER SETTINGS
We perform SLIC [48] to over-segment the original image, where the super-pixel number \(N\) is set to 200. In tree construction, we set the affinity threshold as \(T = \{300, 600, 200\}\), producing three granularity-increasing segmentations.

By adding the initial over-segmentation and the whole image, we build up a five-layer index tree. The tradeoff parameters \(\lambda\) and \(\alpha\) in our model (8) are empirically set as \(\lambda = 0.7\) and \(\alpha = 1.1\) throughout the experiments.

DSNMD is implemented in mixed Matlab and C++ on a desktop machine with 15-3470 3.20 GHz CPU and 8 GB RAM. It takes on average 3.24 seconds to process one image with size of 400 × 300.

B. COMPARISON WITH LR-BASED METHODS
We compare the proposed model with other existing LR-based saliency detection methods including ULR [44], SLR [42], and SMD [41]. Note that ULR, SLR and SMD share a common assumption that an image can be represented as a low-rank matrix corresponding to the background, plus a sparse matrix that relates to foreground objects. Specifically, all these three methods exploit nuclear norm to delineate the low-rank matrix. For sparse matrix, ULR and SLR use \(\ell_1\)-norm to promote sparsity, while SMD introduces a structured sparsity-inducing norm to capture the underlying structure of data. In our model, to well capture the global structure and local structure of the data, and meanwhile to well characterize the internal structure of each group, we introduce the tree-structured nuclear norm to constraint both the background and foreground matrices.

Note that all the LR-based methods have two versions: without and with high-level priors. In the case of pure low-level saliency detection, i.e., without high-level priors, we can see from Fig. 2 that our method consistently outperforms other LR-based methods under all metrics. It demonstrates that the tree-structured nuclear norm constraint is much more effective than the \(\ell_1\)-norm, nuclear norm, and structured sparsity norm for salient object detection. When taking high-level priors into account, the performance of all the LR-based models are all further improved as illustrate in Fig. 2. The proposed method again obtains the best performance in terms of PR and F-measure curves. This well indicates that both the tree-structured nuclear norm constraint and high-level priors are beneficial for saliency detection.

Figure 3 gives some results of our method against ULR, SLR, and SMD. We can clearly see that, compared with other methods, the proposed method can not only completely extract the entire salient object from each image without many scattered patches, but also produce nearly equal saliency values of the pixels within the salient object. This further demonstrates the effects of the structured nuclear norm constraint and high-level priors for saliency detection.

\[1\] We use the code provided by Peng et al. [41] and Jiang et al. [47] to compute these four metrics.

\[2\] Both ULR and SLR use robust principal component analysis (RPCA) [46] model to recovery the low rank matrix, here we do not report the results of SLR.
C. COMPARISON WITH THE STATE-OF-THE-ART METHODS

1) QUANTITATIVE COMPARISON

We compare our method with 23 state-of-the-art approaches, including four LR-based methods (ULR [44], LRR [43], SLR [42] and SMD [41]), and 19 recently proposed prominent methods (FCB [19], MSGC [18] BL [50], BSCA [51], DRFI [47], RBD [52], DS [53], AMC [55], MR [54], HS [56], PCAS [57], GC [59], RC [58], GS [60], SF [61], CA [62], FT [63], SR [64], and LC [65]). We use author's implementation with default parameters or results for evaluation except for LRR and SLR which are reported by Peng et al. [41] 3. It is important to note that besides DRFI, all other methods are unsupervised.

Figure 4 shows the quantitative results of the proposed method against other approaches in terms of PR and F-measure.

3In [41], the authors do not report the results of LRR on ASD dataset. In this paper, we also do not give the LRR results on ASD.
F-measure curves on the six datasets, with corresponding MAE and WF scores reported in Table 1. Note that the MAE scores of the FCB [19] and MSGC [18] are reported by the authors. From the left half of Fig. 4, we see that the proposed approach achieves the first or second highest precision rate when the recall rate is fixed. On the DUT-OMRON and SOD datasets, the PR curve of our method is the second best, while DRFI obtains the best. It is worth noting that DRFI is an unsupervised method and it requires a large amount of training (around 24 hours for training [47]), whereas our method is fully unsupervised, which skips the training process and therefore enjoys more flexibility. As shown in the right half of Fig. 4, our method obtains high F-measure scores in a wide range on all the six datasets, indicating less sensitivity to the selection of a threshold. From Table 1, we can see: i) on the THUS, iCoSeg and ECSSD datasets, our method achieves the best performance in terms of MAE and WF; ii) on the ASD dataset, our method performs the best in terms of MAE, the second in WF; iii) on the DUT-OMRON and SOD, our method obtains the best in terms of WF, the second in MAE. The best MAE score is achieved by DRFI both in these two datasets. Overall, our method achieves superior performance with respect to all the previous unsupervised saliency detection models for all of the six datasets. Also, our method is highly competitive when compared to DRFI on all the six datasets. It’s worth mentioning that the proposed method significantly outperforms all the 23 salient object detection methods in terms of all the four evaluation metrics on the ECSSD dataset. This well demonstrates that our model has strong abilities in handling images with complex scenes.

Our evaluation does not include some of the latest deep-learning based methods. The crux of this paper is to propose a novel unsupervised model which is able to achieve similar or superior performance to supervised methods like DRFI without preparing expensive training data. This provides simplicity and easy-to-use generality in many practical applications where computing power is limited and ground truth annotations are very expensive or impossible to acquire.
2) VISUAL COMPARISON

We present some results of saliency maps generated by some best methods for qualitative comparison in Fig. 5. As can be seen, our method generates more accurate saliency maps in various challenging cases compared to other methods. For images containing heterogeneous objects (e.g., row 3, 6 and 8 in Fig. 5), having a cluttered background (e.g., row 2, 4 and 11 in Fig. 5) and having a low contrast between objects and background (e.g., row 1, 7, 10 and 12 in Fig. 5), most of existing saliency methods cannot effectively highlight the salient objects. For example, in the row 6 and 12 of Fig. 5, some salient regions do not pop out from the background, and in the row 2 and 10 of Fig. 5 some parts of background regions also stand out along with the object regions. Our model can suppress background regions and highlight the complete salient object regions with well-defined boundaries more effectively than other methods.

Besides, our model can highlight both small-scale salient object (e.g., row 5 in Fig. 5) and large-scale salient object (e.g., row 9 in Fig. 5) more effectively compared to other saliency models. These results demonstrate the robustness of our model, and conform the effectiveness of the proposed structured nuclear norm constraint in separating the two low rank subspaces.

D. LIMITATION AND ANALYSIS

The images in Fig. 6 show failure cases where the proposed method is unable to detect the salient object in some scenarios. In this paper, we utilize the tree-structured nuclear norm to delineate the structural characteristics of both the salient object regions and background regions. Therefore, it may be difficult to suppress some small background regions with distinctive appearances, such as column 1 and 2 in Fig. 6. The main reason is that the feature vectors of those regions do not lie in the low dimensional subspace and thus may be incorrectly highlighted as salient object. Besides, the proposed method may fail when the salient objects with partial occlusion, such as the third column in Fig. 6, due to that the constructed index-tree is not precise enough in this case. However, we believe that investigating more effective region merging algorithms for the index-tree construction would be greatly beneficial. We will leave it as the starting point of our future research.
V. CONCLUSION

In this paper, we formulate the task of salient object detection as a problem of structured nuclear norm-based matrix decomposition and propose a double structured nuclear norm-based matrix decomposition (DSNMD) model. In DSNMD, we utilize the tree-structured nuclear norm to delineate the underlying structure of both the salient object regions and background regions in the feature space. Moreover, high-level prior knowledge is seamlessly integrated into our model to enhance the saliency detection performance. Experiments on six datasets demonstrate that the proposed method can achieve superior performance in terms of different evaluation metrics, compared with the state-of-the-arts.

REFERENCES

[1] S. M. Kennedy, W. Williamson, J. D. Roth, and J. W. Scrofani, “Cluster-based spectral-spatial segmentation of hyperspectral imagery,” IEEE Access, vol. 8, pp. 140361–140391, 2020.

[2] X. Cao, C. Zhang, H. Fu, X. Guo, and Q. Tian, “Saliency-aware nonparametric foreground annotation based on weakly labeled data,” IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 6, pp. 1253–1265, Jun. 2016.

[3] C. Duanmu and J. Zhu, “The image super-resolution algorithm based on the dense space attention network,” IEEE Access, vol. 8, pp. 140599–140606, 2020.

[4] A. Shamir and S. Avidan, “ Seam carving for media retargeting,” Commun. ACM, vol. 52, no. 1, pp. 77–85, Jan. 2009.

[5] C. Guo and L. Zhang, “A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression,” IEEE Trans. Image Process., vol. 19, no. 1, pp. 185–198, Jan. 2010.

[6] B. C. Ko and J.-Y. Nam, “Object-of-interest image segmentation based on human attention and semantic region clustering,” J. Opt. Soc. Amer. A, Opt. Image Sci., vol. 23, no. 10, pp. 2462–2470, Oct. 2006.

[7] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, “Summarizing visual data using bidirectional similarity,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2008, pp. 1–8.

[8] R. Zhao, W. Ouyang, and X. Wang, “Unsupervised salience learning for person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2013, pp. 3586–3593.

[9] X. Zhou, Z. Liu, C. Gong, and W. Liu, “Improving video saliency detection via localized estimation and spatiotemporal refinement,” IEEE Trans. Multimedia, vol. 20, no. 11, pp. 2993–3007, Nov. 2018.

[10] J. Kim, D. Han, Y.-W. Tai, and J. Kim, “Saliency region detection via high-dimensional color transform,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 883–890.

[11] J. Yang and M.-H. Yang, “Top-down visual saliency via joint CRF and dictionary learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 2296–2303.
B. He and H. Yang, “Some convergence properties of a method of multipliers for linearly constrained monotone variational inequalities,” Oper. Res. Lett., vol. 23, nos. 3–5, pp. 151–161, Oct. 1998.

B. He and X. Yuan, “On the O(1/n) convergence rate of the Douglas–Rachford alternating direction method,” SIAM J. Numer. Anal., vol. 50, no. 2, pp. 700–709, 2012.

C. Chen, B. He, Y. Ye, and X. Yuan, “The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent,” Math. Program., vol. 155, nos. 1–2, pp. 57–79, Jan. 2016.

H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, “Salient object detection via structured matrix decomposition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 818–832, Apr. 2017.

W. Zou, K. Kpalma, Z. Liu, and J. Konzsin, “Segmentation driven low-rank matrix recovery for saliency detection,” in Proc. Brit. Mach. Vis. Conf., Jun. 2013, pp. 1–14.

C. Lang, G. Liu, J. Yu, and S. Yan, “Saliency detection by multitask sparsity pursuit,” IEEE Trans. Image Process., vol. 21, no. 3, pp. 1327–1338, Mar. 2012.

X. Shen and Y. Wu, “A unified approach to salient object detection via low rank matrix recovery,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 853–860.

C. Tang, P. Wang, C. Zhang, and W. Li, “Salient object detection via weighted low rank matrix recovery,” IEEE Signal Process. Lett., vol. 24, no. 4, pp. 490–494, Apr. 2017.

J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, “Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization,” in Proc. Adv. Neural Inf. Process. Syst., 2009, pp. 2080–2088.

H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, “Salient object detection: A discriminative regional feature integration approach,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., Jun. 2013, pp. 2083–2090.

R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, “Slic superpixels compared to state-of-the-art superpixel methods,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 11, pp. 2274–2282, 2012.

P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient graph-based image segmentation,” Int. J. Comput. Vis., vol. 59, no. 2, pp. 167–181, Sep. 2004.

Y. Li, H. Lu, X. Ruan, and M.-H. Yang, “Salient object detection via bootstrap learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1884–1892.

Y. Qin, H. Lu, Y. Xu, and H. Wang, “Saliency detection via cellular automata,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 110–119.

W. Zhu, S. Liang, Y. Wei, and J. Sun, “Saliency optimization from robust background detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2014, pp. 2814–2821.

X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, “Saliency detection via dense and sparse reconstruction,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2013, pp. 2976–2983.

C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, “Saliency detection via graph-based manifold ranking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2013, pp. 3166–3173.

B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, “Saliency detection via absorbing Markov chain,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2013, pp. 1665–1672.

Q. Yan, L. Xu, J. Shi, and J. Jia, “Hierarchical saliency detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2013, pp. 1155–1162.

R. Margolin, A. Tal, and L. Zelnik-Manor, “What makes a patch distinct?” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2013, pp. 1139–1146.

M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, “Global contrast based salient region detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2011, pp. 409–416.

M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, “Efficient salient region detection with soft image abstraction,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2011, pp. 136–152.

Z. Zhai and M. Shah, “Visual attention detection in video sequences using spatiotemporal cues,” in Proc. ACM Multimedia, Oct. 2006, pp. 815–824.

V. Movshovitz and J. H. Elder, “Design and perceptual validation of performance measures for salient object segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, Jun. 2010, pp. 49–56.

M. Cheng, N. Mitra, X. Huang, P. Torr, and S. Hu, “Global contrast based salient region detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 409–416, Mar. 2015.

D. Batra, A. Gowdla, D. Parikh, J. Luo, and T. Chen, “Interactively co-segmenting topically related images with intelligent scribble guidance,” Int. J. Comput. Vis., vol. 93, no. 3, pp. 273–292, Jul. 2011.

J. Shi, Q. Yan, L. Xu, and J. Jia, “Hierarchical image saliency detection on extended CSSD,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 4, pp. 717–729, Apr. 2016.

X. Li, Y. Li, C. Shen, A. Dick, and A. V. D. Hengel, “Contextual hypergraph modeling for salient object detection,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 3328–3335.

R. Margolin, L. Zelnik-Manor, and A. Tal, “How to evaluate foreground maps,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 248–255.