Image Segmentation via L_1 Monge-Kantorovich Problem

Yupeng Li, Wuchen Li, and Guo Cao

Abstract—This paper provides a fast approach applying the Earth Mover's distance (EMD) (a.k.a optimal transport) for supervised and unsupervised image segmentation. The model globally cooperates the transportation costs (original Monge-Kantorovich type) among histograms of multiple dimensional features, e.g. gray intensity and texture, in image's foreground and background. The computational complexity is often high for the EMD between two histograms on Euclidean spaces with dimensions larger than one. We overcome this computational difficulty via the L_1 optimal transport. We rewrite the model into an L_1 type minimization with the linear dimension of feature space. We then apply a fast algorithm based on the primal-dual method. Compare to several state-of-the-art EMD models, the experimental results based on image data sets demonstrate that the proposed method has superior performance in terms of the accuracy and the stability of the image segmentation.

Index Terms—Optimal transport, Earth Mover's distance, Primal-dual algorithm, L_1 Monge-Kantorovich problem, Image segmentation.

I. INTRODUCTION

MAGE segmentation is a challenging fundamental problem in computer vision [1], [2], [3]. It is to divide an image into several meaningful regions. In the past few decades, many algorithms based on active contour have been proposed. In general, active contour models can be categorized into two groups: the edge-based and region-based models. Edgebased models use edge detector and evolve the curve towards sharp gradients of pixel intensity [4], [5], [6], [7]. The edgebased methods are often sensitive to the noise. To solve this issue, the region-based methods are introduced. The first one is proposed by Mumford-Shah [8], which approximates an image by a piecewise smooth function. Starting from this model, many other ones have been introduced [9], [10], [11], [12], [13], [14], [15], [16], [17], [18]. In particular, Chan and Vese proposed the famous model (CV) [11], [19] based on the level-set method, which deals with topological changes [20], [21], [22]. The initial choice in CV model is important. This is because its minimization problem is non-convex, different initializations will result at various segmentations. To deal with this problem, many models with convex properties [23], [24], [25] have been introduced. On the other hand, statistical segmentation [26], [27], [28] brings insights into

Y. Li, G. Cao are with School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (email: starli777@hotmail.com; caoguo@njust.edu.cn).

W. Li is with Department of Mathematics, University of California, Los Angeles, 90095, USA (e-mail: wcli@math.ucla.edu).

the problem. The statistical information, i.e. histograms over features, is introduced to measure the closeness between forward/background regions. In this direction, the functionals to measure closeness of histograms are needed. Typically, the widely used measurements are the L_1 distance [29], Kullback-Leibler divergence [30] and Bhattacharyya distance [31] etc.

Recently, the other distance/measurement among histograms, named Earth Mover's distance (EMD) [32], has been brought into attentions. Nowadays, it has been widely used in image retrieval problems [33], and in hand gesture recognition problems [34], [35]. The successful usage of EMD is because of its many desirable theoretical properties; see [36] and many references therein. A short review will be provided in section II.

The usage of EMD in image segmentation is initialized by the novel work of Chan et al. in [9]. They apply continuous cut model, and compare the patches (local) around pixels by EMD. Their model is convex within the framework of continuous cut, thus it is a unsupervised model and global minimizers can be computed very efficiently by existing algorithms. However, it relies on the choice of the size of patches. Tuning the value of size is not a straightforward task. Usually, it is required that the local histogram correctly approximate or close to the global ones. In addition, Peyre et al. [37], [38] consider Wasserstein active contour methods. Their minimization problems are nonconvex. So their methods are sensitive to the initializations and get struck at local minimizers. Swoboda et al. [39] and Rabin et al. [40] introduce the convex minimization problems without using patches, i.e. a global approach. Their models are built for supervised segmentation and co-segmentation.

Besides the above local and global modeling issues, there are one major difficulty limiting the application of EMD segmentation. The one dimensional feature space, usually the intensity of gray image, can hardly distinguish the foreground and background. Even on a gray image, more information, including texture, such as orientation and scale, should be taken into account to enhance the recognition ability. Thus the model needs to cooperate with multiple dimensional features, i.e. EMD with multiple dimensional histograms. However, the EMD in \mathbb{R}^d with d larger than 1 has no analytical solutions. Fast computation of EMD is a necessary but not simple task, since EMD between histograms supported on \mathbb{R}^d (n^d bins, where n is the number of discretization in one dimension) requires to solve a linear programming problem with n^{2d} variables, such as [34], [35].

In this paper, we overcome the limitation by applying key ideas in [41]. We introduce a global supervised and unsupervised variation models, which have the ability to handle

This work was funded in part by ONR grants N000141410683, N000141210838, DOE grant DE-SC00183838 and National Nature Science Foundation of China(Grant No.61371168).

multiple dimensional features easily; see Fig.1 for illustration. Our main idea inherits the formulation of [40], [39], and applies the EMD with homogeneous degree one ground metric $(L_1 \text{ optimal transport})$ [42]. Following the associated duality structure, we formulate the model into an L_1 type minimization with $O(n^d)$ variables, instead of $O(n^{2d})$ in the linear programming. By leveraging its L_1 minimization's structure, we apply a primal-dual method, in which the shrink operator plays vital roles in handling various histograms. The updates are very simple and explicit, and the convergence speed is fast. The iterations contain only 6 line codes.

The paper is organized as follows. In section II, we briefly review the segmentation models and L_1 optimal transport. Based on them, we introduce the proposed variation model. In section III, we design a fast algorithm toward the supervised and unsupervised image segmentation models, respectively. Evaluation of the experimental results are presented in section IV.

II. PROBLEM FORMULATION

In this section, we briefly review the related segmentation models, and connect the model with the theory of optimal transport, especially the L_1 type. Based on these, we introduce the proposed model.

Consider a two-phrase segmentation problem. Assume that an image $I: \Omega \to \mathbb{R}^1$ or \mathbb{R}^3 represents gray or color images. The segmentation seeks to find a meaningful partition Ω_1, Ω_2 in a spatial domain Ω of the image I, where $\Omega = \Omega_1 \cup \Omega_2$ and $\Omega_1 \cap \Omega_2 = \emptyset$. To achieve this goal, various models have been proposed in literature. Following Mumford-Shah, consider the following minimization problem:

$$\min_{\Omega_1,\Omega_2} \lambda \operatorname{Per}(\Omega_1,\Omega_2) + \operatorname{Dist}(\Omega_1,\operatorname{Ref}_1) + \operatorname{Dist}(\Omega_2,\operatorname{Ref}_2) \quad (1)$$

where regions Ω_1 , Ω_2 are minimization variables, Per is the perimeter of Ω_1 or Ω_2 's boundary with the weight constant $\lambda > 0$, Ref₁, Ref₂ are given references, known as the supervised terms, and Dist are functionals to estimate the closeness between region and references. The purpose of solving this minimization problem is as follows. The minimizer represents two regions whose boundaries are short, and whose areas are close to the references. The above minimization is not proposed explicitly. One need to find proper analytical representations for Ω_1 , Ω_2 . Many extensions have been discussed in this direction. Among them, a famous example is the Chan-Vese model. See [43] and many references therein for details.

The concept of distance between regions plays vital roles in above models. Recently EMD provides a particular metric among histograms, which can be used in this variational problem [9], [40]:

$$\min_{\Omega_1,\Omega_2} \lambda \operatorname{Per}(\Omega_1,\Omega_2) + \operatorname{EMD}(\operatorname{Hist}_{\Omega_1},\operatorname{Hist}_{\operatorname{Ref}_1}) + \operatorname{EMD}(\operatorname{Hist}_{\Omega_2},\operatorname{Hist}_{\operatorname{Ref}_2})$$
(2)

Here we can represent Ω_1 , Ω_2 by using their histograms Hist_{Ω_1} , Hist_{Ω_2} in the feature spaces $\mathcal{F} \subset \mathbb{R}^d$. \mathcal{F} represents the *d* dimensional features of the image. So EMD can be used to measure the closeness between the regions.

A. Review of Earth Mover's Distance

In history, the EMD is first introduced by Monge in 1781, and then relaxed by Kantorovich in 1940s as follows [44], [45], [46], [47]. Given two histograms ρ^0 , ρ^1 supported on a compact, convex feature set \mathcal{F} with equal total mass, and $c: \mathcal{F} \times \mathcal{F} \to \mathbb{R}_+$ a ground cost function. Consider

$$\operatorname{EMD}(\rho^0, \rho^1) := \min_{\pi} \quad \int_{\mathcal{F} \times \mathcal{F}} c(y, \tilde{y}) \pi(y, \tilde{y}) dy d\tilde{y} \quad (3)$$

where the infimum is taken among all joint measures (transport plans) $\pi(y, \tilde{y}) \ge 0$ having $\rho^0(y)$ and $\rho^1(\tilde{y})$ as marginals:

$$\int_{\mathcal{F}} \pi(y, \tilde{y}) d\tilde{y} = \rho^0(y) , \quad \int_{\mathcal{F}} \pi(y, \tilde{y}) dy = \rho^1(\tilde{y}) .$$
 (4)

Problem (3) is a well known linear programming, whose minimal value under suitable choice of c defines the EMD. It is also named Wasserstein metric. However, the computation of (3) is cumbersome for high-dimensional feature space \mathbb{R}^d , $d \ge 2$. If \mathcal{F} is discretized into n^d points, it requires to solve a linear programming problem with n^{2d} variables.

We focus on $c(y, \tilde{y}) = ||y - \tilde{y}||_2$ or $||y - \tilde{y}||_1$, the Euclidean or Manhattan ground distance. This choice of ground distance, homogeneous degree one type, is very close to Monge's original problem in 1781. Here the minimization (3) has an important reformulation [36], which settles the problem by

$$\operatorname{EMD}(\rho^0, \rho^1) := \inf_{m} \quad \int_{\mathcal{F}} \|m(y)\| dy , \qquad (5)$$

where the infimum is taken among all Borel flux functions $m(y) \in \mathbb{R}^d$, such that

$$\nabla_y \cdot m(y) = \rho^1(y) - \rho^0(y) . \tag{6}$$

Here ∇_y is the divergence operator and m(y) satisfies the zero flux condition: $m(y) \cdot n(y) = 0$, if $y \in \partial \mathcal{F}$, n(y) is the norm vector on the boundary of \mathcal{F} . Here $\|\cdot\|$ represents $\|\cdot\|_1$ or $\|\cdot\|_2$. In theory of optimal transport [42], it is well known that (3), (5) provide the equivalent metric for histograms. In particular, (5) enjoys many nice mathematical properties, whose minimizer satisfies the famous Monge-Kantorovich equation [48], [42].

In this paper, we focus on the modeling and computational advantages of (5). If \mathcal{F} is discretized by n^d bins, (5) only requires solving minimization with $O(n^d)$ variables and constraints. More importantly, (5) is very similar to the problems in compressed sensing, whose objective functional is homogeneous degree one and constraint is linear. Thus many fast algorithms are available. These two properties help to handle the computation of multiple dimension features model easily.

B. Proposed Model

We present the model by rewriting (2) explicitly. We form the variation problem in a continuous setting, and then provide more details about computations in next section.

Consider an image $I: \Omega \to \mathbb{R}^1$, and denote a binary segmentation $u: \Omega \to \{0, 1\}$. It is clear that the support of u represents the foreground Ω_1 or background Ω_2 . We



Fig. 1: Overall framework of the proposed method.

introduce histograms based on the image I. Following the ideas in [40], a feature space is introduced as follows. Consider $\mathcal{F} \subset \mathbb{R}^d$, where d is the dimension of feature space. A map is introduced by $M: \Omega \to \mathcal{I}$. Clearly the map M depends on the image I's information. A feature map is needed to transform the region into histograms (measures) of features. Denote Measure(\mathcal{F}) = { $\rho \in L^1(\mathcal{F}) : \rho(y) \ge 0$, $y \in \mathcal{F}$ }, and

$$H(u)(y) = \int_{\Omega} h(x, y)u(x)dx$$
(7)

where

$$h(x,y) = \begin{cases} 1 & \text{if } y = M(x) \\ 0 & \text{Otherwise.} \end{cases}$$

We remark that H is a linear operator from the function in Ω to Measure(\mathcal{F}). Thus the foreground and background can be represented by their measures H(u), $H(1-u) \in \text{Measure}(\mathcal{F})$, respectively. Similarly, the references are two fixed histograms a, b, where $a(y), b(y) \ge 0$ and $\int_{\mathcal{F}} a(y)dy = \int_{\mathcal{F}} b(y)dy = 1$. The total mass of H(u) and a may not be equal, and the classical EMD between unbalanced masses is infinity. It is worth mentioning that there are many ways to deal with this unbalanced issue; see results in partial optimal transport [49]. Here we consider the method proposed by Rabin et al. [40]. Construct reference measures a, b by $\hat{a}, \hat{b} \in \text{Measure}(\mathcal{F})$:

$$\hat{a}(u)(y) = a(y) \int_{\mathcal{F}} H(u)(y) dy ,$$

$$b(u)(y) = b(y) \int_{\mathcal{F}} H(1-u)(y) dy.$$
(8)

It is simple to check that \hat{a} , \hat{b} and H(u), H(1-u) have the equal total mass. Let us formulate the variational problem. By the co-area formula $u(x) = \mathbb{1}_{\Omega_1}(x)$, we form (2) as

$$\lambda \int_{\Omega} \|\nabla_x u(x)\| dx + \mathsf{EMD}(H(u), \hat{a}(u)) + \mathsf{EMD}(H(1-u), \hat{b}(u)) .$$

Then by relaxing $u(x) \in \{0,1\}$ to [0,1], and using (5), we arrive at

$$\inf_{u,m_1,m_2} \lambda \int_{\Omega} \|\nabla_x u(x)\| dx + \int_{\mathcal{F}} \|m_1(y)\| dy + \int_{\mathcal{F}} \|m_2(y)\| dy ,$$
(9a)

where the infimum is taken among u(x) and flux functions $m_1(y), m_2(y)$ satisfying

$$\begin{cases} 0 \le u(x) \le 1 \\ \nabla_y \cdot m_1(y) + H(u)(y) - a(y) \int_{\mathcal{F}} H(u)(y) dy = 0 \\ \nabla_y \cdot m_2(y) + H(1-u)(y) - b(y) \int_{\mathcal{F}} H(1-u)(y) dy = 0 \\ \end{cases}$$
(9b)

Variation (9) is the main problem considered in this paper. It is not hard to observe that the objective functional is convex, the constraint is linear and not empty. Thus there exists a minimizer for the segmentation.

III. ALGORITHM

In this section, we perform finite volume discretization on both spatial and feature domains. The discretized minimization problem of (9) is homogeneous degree one, and enjoys a suitable number of variables. We then design a fast algorithm via primal-dual methods [50].

To simplify presentation, consider $\Omega \subset \mathbb{R}^2$ and $\mathcal{F} \subset \mathbb{R}^2$, where Ω , \mathcal{F} are squares. We prepare several necessary notations. We discretize the spatial domain Ω uniformly by a step size Δx , and partition the features space \mathcal{F} uniformly with a step size Δy . The node of Ω is represented by (i, j), $i, j \in [1, n_1], |\Omega| = n_1^2$ while the partition of \mathcal{F} is denoted by $(k, l), k, l \in [1, n_2], |\mathcal{F}| = n_2^2$.

We first introduce the variables in the discretized model. Denote $u \in \mathbb{R}^{|\Omega|}$, where $u_{ij} \in [0, 1]$. A segmentation $\Omega = \Omega_1 \cup \Omega_2$ is determined by u and a small constant threshold $\delta > 0$:

$$(i,j) \in \Omega_1 \quad \text{if } u_{ij} < \delta \,, \quad \text{and} \quad (i,j) \in \Omega_2 \quad \text{if } 1 - u_{ij} < \delta \,.$$

So u and 1 - u provide the segmentation. For simplicity, we call Ω_1 foreground, and name Ω_2 background.

We then represent the foreground Ω_1 and background Ω_2 by histograms. Consider a map $M: \Omega \to \mathcal{F}$ and $H: \mathbb{R}^{|\Omega|} \to \mathbb{R}^{|\mathcal{F}|}$, which transfers a partition in spatial domain to a histogram in features spaces [40]. Here $H = (h_{ij,kl})_{(i,j)\in\Omega,(k,l)\in\mathcal{F}} \in \mathbb{R}^{|\Omega| \times |\mathcal{F}|}$ is a matrix, whose indexes

depend on both spatial and feature variables:

$$h_{ij,kl} = \begin{cases} 1 & \text{if } M(i,j) = (k,l) \\ 0 & \text{Otherwise} \end{cases}$$

More details of map H can be found in section IV. We derive the histogram of foreground by $H(u) \in \mathbb{R}^{|\mathcal{F}|}$:

$$H(u)_{kl} = \sum_{i,j} h_{ij,kl} u_{ij}$$

Similarly, we write the histogram of background by $H(1 - u) \in \mathbb{R}^{|\mathcal{F}|}$:

$$H(1-u)_{kl} = \sum_{i,j} h_{ij,kl} (1-u_{ij})$$

We also introduce the supervised terms, which are given by two fixed histograms $a, b \in \mathbb{R}^{|\mathcal{F}|}$, with $a_{kl}, b_{kl} \geq 0$, and $\sum_{k,l} a_{kl} = \sum_{k,l} b_{kl} = 1$. It is worth noting that histograms a and H(u) may not have the equal total mass during computations, which does not satisfy the requirement of EMD. To conquer this and keep the minimization linear, we replace the histogram a by $a \sum H(u)$, where $\sum H(u) = \sum_{i,j,k,l} h_{ij,kl} u_{ij}$. It is clear that

$$\sum_{k,l} a_{kl} \sum_{i,j,k,l} h_{ij,kl} u_{ij} = \sum_{i,j,k,l} h_{ij,kl} u_{ij} = \sum_{k,l} H(u)_{k,l} .$$

By the similar reason, the histogram b can be changed to $b\sum_{k,l} H(1-u)_{kl}$. From now on, we compare two histogram pairs, i.e. H(u), $a\sum_{k,l} H(u)_{kl}$, and H(1-u), $b\sum_{k,l} H(1-u)_{kl}$. Later on, we denote $\sum_{k,l}$ by \sum to simplify the notation.

We last introduce the discrete divergence and gradient operator. We illustrate them by \mathcal{F} . Denote $y = (y_1, y_2) \in \mathcal{F}$. For flux function $m = (m_{y_1}, m_{y_2}), m_{y_1}, m_{y_2} \in \mathbb{R}^{|\mathcal{F}|}$. Define the discrete divergence operator:

$$\operatorname{div}(m)_{kl} = \frac{1}{\Delta y} (m_{y_1,kl} - m_{y_1,(k-1)l} + m_{y_2,kl} - m_{y_2,k(l-1)})$$

And the zero flux condition is considered by

$$m_{y_1,1l} = m_{y_1,n_2l} = m_{y_2,k1} = m_{y_2,kn_2} = 0$$
, for $k, l \in [0, n_2]$

For $\Phi \in \mathbb{R}^{|\mathcal{F}|}$, define the discrete gradient operator. Denote $\operatorname{grad} \Phi = (\operatorname{grad}_{y_1} \Phi, \operatorname{grad}_{y_2} \Phi), \operatorname{grad}_{y_1} \Phi, \operatorname{grad}_{y_2} \Phi \in \mathbb{R}^{|\mathcal{F}|}$:

$$(\operatorname{grad}_{y_1} \Phi)_{ij} = (1/\Delta y) (\Phi_{i+1,j} - \Phi_{i,j}) ,$$

 $(\operatorname{grad}_{y_2} \Phi)_{ij} = (1/\Delta y) (\Phi_{i,j+1} - \Phi_{i,j}) .$

It is simple to check that grad is the adjoint operator of div.

A. Supervised Model

Based on above notations, we introduce the discretized variational problem of (9):

$$\begin{array}{ll} \underset{u,m_{1},m_{2}}{\text{minimize}} & \|m_{1}\|_{1,2} + \|m_{2}\|_{1,2} + \lambda \|\text{grad}u\|_{1,2} \\ \text{subject to} & 0 \leq u \leq 1 \\ & \operatorname{div}(m_{1}) + H(u) - a \sum H(u) = 0 \\ & \operatorname{div}(m_{2}) + H(1-u) - b \sum H(1-u) = 0 , \end{array}$$

$$(10)$$

where the objective function is

$$\sum_{k,l} \sqrt{m_{1y_1,kl}^2 + m_{1y_2,kl}^2} + \sum_{k,l} \sqrt{m_{2y_1,kl}^2 + m_{2y_2,kl}^2} + \lambda \cdot \sum_{i,j} \sqrt{\operatorname{grad}_x u_{kl}^2 + \operatorname{grad}_y u_{kl}^2} .$$

Since (10) is an L_1 type minimization, whose objective function is convex and constraints are linear, it admits a minimizer. Following the classical convexity analysis, we consider the saddle point problem to solve (10):

$$\min_{u,m_1,m_2} \max_{\Phi_1,\Phi_2,g} \|m_1\|_{1,2} + \|m_2\|_{1,2} + \lambda \langle u, \operatorname{div}(g) \rangle
+ \mathbb{1}\{\|g\|_2 \le 1\} + \langle \Phi_1, \operatorname{div}(m_1) + H(u)
- a \sum_{u} H(u) \rangle + \langle \Phi_2, \operatorname{div}(m_2) + H(1-u)
- b \sum_{u} H(1-u) \rangle,$$
(11)

where $\Phi_1, \Phi_2 \in \mathbb{R}^{|\mathcal{F}|}$ are the Lagrange multiplier of constraints, g is the other dual variable used to handle the total variation term, and $\langle \cdot, \cdot \rangle$ is the standard inner product in $\mathbb{R}^{|\mathcal{F}|}$:

$$\langle A, B \rangle = \sum_{k,l} A_{kl} B_{kl}$$

Saddle problem (11) can be computed easily via primaldual algorithms [50]. It applies the gradient descent in primal variables u, m_1, m_2 with the step size $\mu > 0$:

$$u^{K+1} = \arg\min_{u} \lambda \langle u, \operatorname{div}(g^{K}) \rangle + \langle \Phi_{1}^{K}, H(u) \\ - a \sum_{u} H(u) \rangle + \langle \Phi_{2}^{K}, H(1-u) \\ - b \sum_{u} H(1-u) \rangle + \frac{1}{2\mu} \|u - u^{K}\|_{2}^{2}$$

$$m_{1}^{K+1} = \arg\min_{m_{1}} \|m_{1}\|_{1,2} + \langle \Phi_{1}^{K}, \operatorname{div}(m_{1}) \rangle + \frac{1}{2\mu} \|m_{1} - m_{1}^{K}\|_{2}^{2}$$

$$m_{2}^{K+1} = \arg\min_{m_{2}} \|m_{2}\|_{1,2} + \langle \Phi_{2}^{K}, \operatorname{div}(m_{2}) \rangle + \frac{1}{2\mu} \|m_{2} - m_{2}^{K}\|_{2}^{2}$$
(12)

While the updates use the gradient ascent in the dual variable g, Φ_1, Φ_2 with step size $\tau > 0$:

$$\begin{split} g^{K+1} &= \arg \max_{g} \ \lambda \langle 2u^{K+1} - u^{K}, \operatorname{div}(g) \rangle \\ &+ \mathbbm{1} \{ \|g\|_{2} \leq 1 \} - \frac{1}{2\tau} \|g - g^{K}\|_{2}^{2} ; \\ \Phi_{1}^{K+1} &= \arg \max_{\Phi_{1}} \ \langle \Phi_{1}, \operatorname{div}(2m_{1}^{K+1} - m_{1}^{K}) \\ &+ H(2u^{K+1} - u^{K}) - a \sum H(2u^{K+1} - u^{K}) \rangle \\ &- \frac{1}{2\tau} \|\Phi_{1} - \Phi_{1}^{K}\|_{2}^{2} ; \\ \Phi_{2}^{K+1} &= \arg \max_{\Phi_{2}} \ \langle \Phi_{2}, \operatorname{div}(2m_{2}^{K+1} - m_{2}^{K}) \\ &+ H(1 - (2u^{K+1} - u^{K})) \\ &- b \sum H(1 - (2u^{K+1} - u^{K})) \rangle \\ &- \frac{1}{2\tau} \|\Phi_{2} - \Phi_{2}^{K}\|_{2}^{2} . \end{split}$$
(13)

In this case, optimization problems in primal and dual updates (12), (13) have explicit formulas depending on indices i, j, k, l. And the shrink operator plays vital roles in these computations, especially for updating flux functions m_1, m_2 . For example, consider

$$\min_{m_1} \left\{ \|m_1\|_{1,2} + \langle \Phi_1^K, \operatorname{div}(m_1) \rangle + \frac{1}{2\mu} \|m_1 - m_1^k\|_2^2 \right\}$$

=
$$\sum_{k,l} \min_{m_{1,kl}} (\|m_{1,kl}\|_2 - (\operatorname{grad}\Phi_1^K)_{kl} \cdot m_{1,kl} + \frac{1}{2\mu} \|m_{1,kl} - m_{1,kl}^K\|_2^2) .$$

Then m_1^{K+1} in (12) has a close form update:

$$m_{1,kl}^{K+1} = \operatorname{shrink}_2(m_{1,kl}^K + \mu \cdot (\operatorname{grad} \Phi_1^K)_{kl}, \mu)$$

where

shrink₂(v,
$$\mu$$
) =

$$\begin{cases}
(1 - \mu/\|v\|_2)v & \text{for } \|v\|_2 \ge \mu \\
0 & \text{for } \|v\|_2 < \mu
\end{cases}$$

is the shrink operator in \mathbb{R}^2 . Similarly update can be derived for m_2^{K+1} . In addition, the iterations for u, Φ_1 , Φ_2 , g are straightforward, because the minimizations there are linear.

We are now ready to state the proposed algorithm, which mainly contains the following 6 explicit iterations.

Algorithm for supervised model

Input: Two given histogram of features $a, b \in \mathbb{R}^{|\mathcal{F}|}$; Initial guess of $u^0, g^0 \in \mathbb{R}^{|\Omega|}, m_1^0, m_2^0 \in \mathbb{R}^{|\mathcal{F}|}$, step size $\mu, \tau > 0$, feature map $H \in \mathbb{R}^{|\Omega| \times |\mathcal{F}|}$. **Output**: Segmentation u

for $K = 1, 2, \cdots$ (Iterate until convergence)

1.
$$u_{ij}^{K+1} = u_{ij}^{K} - h \{ \lambda \operatorname{div}(g^{K})_{ij} - \sum_{k,l} \Phi_{1,kl} h_{ij,kl} + (\sum_{k,l} h_{ij,kl}) \cdot (\sum_{k,l} \Phi_{1,kl} a_{kl}) + \sum_{k,l} \Phi_{2,kl} h_{ij,kl} - (\sum_{k,l} h_{ij,kl}) \cdot (\sum_{k,l} \Phi_{2,kl} b_{kl}) \};$$

2.
$$m_{1,kl}^{K+1} = \text{shrink}_2(m_{1,kl}^K + \mu \cdot (\text{grad}\Phi_1^K)_{kl}, \mu);$$

3.
$$m_{2,kl}^{K+1} = \text{shrink}_2(m_{2,kl}^K + \mu \cdot (\text{grad}\Phi_2^K)_{kl}, \mu);$$

4.
$$g_{ij}^{K+1} = \operatorname{Proj}_{\{\|g_{ij}\|_2 \le 1\}} [g_{ij}^K - \tau \cdot (2(\operatorname{grad} u^{K+1})_{ij} - (\operatorname{grad} u^K)_{ij})];$$

$$\begin{aligned} & 5. \ \Phi^{K+1}_{1,kl} = \Phi^{K}_{1,kl} + \tau \cdot \left\{ \mathrm{div}(2m_{1}^{K+1} - m_{1}^{K})_{kl} \right. \\ & \left. - \sum_{i,j} h_{ij,kl}(2u_{ij}^{K+1} - u_{ij}^{K}) \right. \\ & \left. + a_{kl} \sum_{i,j,k,l} h_{ij,kl}(2u_{ij}^{K+1} - u_{ij}^{K}) \right\} ; \end{aligned}$$

$$\begin{aligned} 6. \ \Phi_{2,kl}^{K+1} &= \Phi_{2,kl}^{K} + \tau \cdot \left\{ \text{div}(2m_2^{K+1} - m_2^K)_{kl} \right. \\ &\quad \left. - \sum_{i,j} h_{ij,kl}(1 - 2u_{ij}^{K+1} + u_{ij}^K) \right. \\ &\quad \left. + b_{kl} \sum_{i,j,k,l} h_{ij,kl}(1 - 2u_{ij}^{K+1} + u_{ij}^K) \right\} ; \end{aligned}$$

In fact, there are many other interesting choices of ground metric, e.g. the Manhattan distance. The corresponding model (9) forms

$$\begin{array}{ll} \underset{u,m_{1},m_{2}}{\text{minimize}} & \|m_{1}\|_{1,1} + \|m_{2}\|_{1,1} + \lambda \|\text{grad}u\|_{1,1} \\ \text{subject to} & 0 \leq u \leq 1 \\ & \operatorname{div}(m_{1}) + H(u) - a \sum H(u) = 0 \\ & \operatorname{div}(m_{2}) + H(1-u) - b \sum H(1-u) = 0 , \end{array}$$

$$(14)$$

where the objective function is

$$\sum_{k,l} \left(|m_{1y_1,kl}| + |m_{1y_2,kl}| + |m_{2y_1,kl}| + |m_{2y_2,kl}| \right) + \sum_{i,j} \left(|\operatorname{grad}_x u_{kl}| + |\operatorname{grad}_y u_{kl}| \right) .$$

It is worth noting that (14) is the other L_1 optimization problem. Since its objective function is not strictly convex, it may have multiple minimizers. Motivated by the work in [41], we conquer this issue by adding a quadratic regularization:

$$\begin{array}{ll} \underset{u,m_{1},m_{2}}{\text{minimize}} & \|m_{1}\|_{1,1} + \|m_{2}\|_{1,1} + \lambda \|\text{grad}u\|_{1,1} \\ & + \frac{\epsilon}{2} \|m_{1}\|_{1,2}^{2} + \frac{\epsilon}{2} \|m_{2}\|_{1,2}^{2} \\ \text{subject to} & 0 \leq u \leq 1 \\ & \text{div}(m_{1}) + H(u) - a \sum H(u) = \end{array}$$

$$div(m_1) + H(u) - a \sum H(u) = 0$$

$$div(m_2) + H(1 - u) - b \sum H(1 - u) = 0$$

Here $\epsilon > 0$ is a small constant. Similar as the one with Euclidean ground metric, the primal dual algorithm for solving (14) has explicit and exact updates. Here we only point out the difference in the update. More details can be found in [41]. Let us illustrate the updates of m_1 . Consider

$$\begin{split} \min_{m_1} \{ \|m_1\|_{1,1} + \frac{\epsilon}{2} \|m_1\|_{1,2}^2 + \langle \Phi_1^K, \operatorname{div}(m_1) \rangle \\ + \frac{1}{2\mu} \|m_1 - m_1^k\|_2^2 \} \\ = & \sum_{k,l} \{ \min_{m_{1y_1,kl}} (|m_{1y_1,kl}| + \frac{\epsilon}{2} |m_{1y_1,kl}|^2 \\ - (\operatorname{grad}_{y_1} \Phi_1^K)_{kl} \cdot m_{1y_1,kl} + \frac{1}{2\mu} |m_{1y_1,kl} - m_{1y_1,kl}^K|^2) \\ + & \min_{m_{1y_2,kl}} (|m_{1y_2,kl}| + \frac{\epsilon}{2} |m_{1y_2,kl}|^2 - (\operatorname{grad}_{y_2} \Phi_1^K)_{kl} \\ \cdot m_{1y_2,kl} + \frac{1}{2\mu} |m_{1y_2,kl} - m_{1y_2,kl}^K|_2^2) \} . \end{split}$$

Thus the update of flux m_1 is very simple:

$$\begin{split} m_{1y_1,kl}^{K+1} = & 1/(1+\epsilon\mu) \text{shrink}_1(m_{1y_1,kl}^K + \mu \cdot (\text{grad}_{y_1}\Phi_1^K)_{kl}, \mu) \\ m_{1y_2,kl}^{K+1} = & 1/(1+\epsilon\mu) \text{shrink}_1(m_{1y_2,kl}^K + \mu \cdot (\text{grad}_{y_2}\Phi_2^K)_{kl}, \mu) \\ \end{split}$$
where

shrink₁(v,
$$\mu$$
) =
$$\begin{cases} (1 - \mu/|v|)v & \text{for } |v| \ge \mu \\ 0 & \text{for } |v| < \mu \end{cases}$$

is the shrink operator in \mathbb{R}^1 . The other updates are similar to the one in Euclidean ground metric.

end

We next demonstrate that the proposed algorithm converges to the minimizer of (10) and (14).

Theorem 1: Assume $\sqrt{\tau\mu} < 1/\max\{\lambda^2 \cdot \lambda_{\max}(\Delta_{\Omega}), \lambda_{\max}(\Delta_{\mathcal{F}})\}$, where $\Delta_{\Omega}, \Delta_{\mathcal{F}}$ denotes the discrete Laplacian operator in Ω , \mathcal{F} , and λ_{\max} denotes their largest eigenvalues. Then the iterations in (12) and (13) converges a minimizer of (10) or (14).

B. Unsupervised Model

In this part, we extend the framework into the unsupervised segmentation. We further modify the segmentation model (1) so that the input reference histograms are not required.

In this case, the proposed minimization is

$$\inf_{a,b,u} \lambda \int_{\Omega} \frac{1}{1 + \beta(\nabla I(x))^2} |\nabla u(x)| dx + \operatorname{EMD}(H(u), \hat{a}(u)) + \operatorname{EMD}(H(1-u), \hat{b}(u)) ,$$
(15)

where $u(x) \in [0,1]$ and reference histograms a, b are minimization variables. For better locating the boundary between foreground and background, we apply the weighted total variation, where β is a positive constant. For simplicity, we assume that a, b are delta histograms, which support at one point.

We propose to solve (15) with respect to references a, b, and partition variable u iteratively. For each iteration step k > 0, we first fix u^k , and solve minimization (15) only with variables a, b.

$$a^{k+1} = \arg \inf_{a \in D} \operatorname{EMD}(H(u^k), \hat{a}(u^k))$$

$$b^{k+1} = \arg \inf_{b \in D} \operatorname{EMD}(H(1 - u^k), \hat{b}(u^k)) , \qquad (16)$$

where D is the set of delta histograms.

We then fix a^{k+1} , b^{k+1} , and apply the supervised algorithm to solve (15), i.e.

$$u^{k+1} = \arg \inf_{u} \lambda \int_{\Omega} \frac{1}{1 + \beta(\nabla I(x))^2} |\nabla u(x)| dx + \text{EMD}(H(u), \hat{a}^{k+1}(u)) + \text{EMD}(H(1-u), \hat{b}^{k+1}(u))$$
(17)

In fact, the minimization in first step (16) has an explicit solution.

Lemma 1: Denote a joint histogram μ in \mathbb{R}^d , whose marginal distribution in each dimension is μ_i , $i = 1, \dots, d$. Let x_0^i be the median point of measure μ_i , i.e. $\mu_i(\tilde{y}_i > x_0^i) =$ $\mu_i(\tilde{y}_i \le x_0^i)$. Then the delta measure supported at the median point $x_0 = (x_0^i)_{i=1}^d$:

$$a^*(x) = \delta_{(x_0^i)_{i=1}^d}(x)$$

is the unique minimizer of

$$a^*(x) = \inf_{a \in D} \operatorname{EMD}(a, \mu)$$
.

Remark 1: Lemma 1 is only true for Earth Mover's distance with Manhattan ground distance, i.e. $c(x, y) = ||x - y||_1$.

Following Lemma 1, a^{k+1} , b^{k+1} are delta measures, which are supported at median points w.r.t. measure Hu, H(1-u)respectively. We now state the unsupervised model as follows. **Input**: Initial guess of u^0 ; Feature map $H \in \mathbb{R}^{|\Omega| \times |\mathcal{F}|}$. **Output**: Segmentation u

for
$$K = 1, 2, \cdots$$
 (Iterate until convergence)

1. Fix u^K and choose a^{K+1} as a delta histogram supported at the median point of $H(u^K)$.

2. Fix u^K and choose b^{K+1} as a delta histogram supported at the median point of $H(1 - u^K)$.

3. Fix a^{K+1} , b^{K+1} and apply the algorithm in supervised model to solve u^{K+1} in (17).

end

IV. EXPERIMENTAL RESULTS

In this section, we will show the implementation details and settings of the proposed methods.

A. Details on Feature Extractions

In experiments, we extract two kinds of features, the intensity and the texture features from the original image. Here the intensity features can be obtained from the image intensity directly; see Fig.2(a), and the texture features; see Fig.2(b), is obtained by texture function [51].



Fig. 2: The corresponding features map obtained from the test image. (a) shows the intensity features map. (b) is the texture features map.

Under these features, we consider the following map function

$$M: \Omega \to \mathcal{F} = [0, 256]^2$$
, $M(x_1, x_2) = (M_i(x_1, x_2))_{i=1}^2$.

Here

$$\begin{cases} M_1(x_1, x_2) = I(x_1, x_2) \\ M_2(x_1, x_2) = F(x_1, x_2) \end{cases}$$



Fig. 3: Exemplar regions with bounding boxes of the test images. The blue bounding box is the foreground and the red bounding boxes are the background.

where I is the intensity of the image at each pixel (x_1, x_2) , the texture function F is given by

$$\begin{split} F(x_1, x_2) &= \exp(-\frac{\det(g_{x_1 x_2})}{\sigma^2}) \ , \\ g_{x_1 x_2} &= \begin{pmatrix} 1 + (\partial_{x_1} p_{x_1, x_2})^2 & \partial_{x_1} p_{x_1, x_2} \partial_{x_2} p_{x_1, x_2} \\ \partial_{x_1} p_{x_1, x_2} \partial_{x_2} p_{x_1, x_2} & 1 + (\partial_{x_2} p_{x_1, x_2})^2 \end{pmatrix} \ , \end{split}$$

 p_{x_1,x_2} is the square patch of size $\tau \times \tau$ around the pixel (x_1,x_2) , i.e. $p_{x_1,x_2}(I) = \{I(x_1+t_1,x_2+t_2)\}, t_1 \in [-\frac{\tau}{2},\frac{\tau}{2}], t_2 \in [-\frac{\tau}{2},\frac{\tau}{2}], \tau$ is a scaling parameter which controls the degree of the image details. In experiments, we select $\tau = 10$. By the map M, we then construct the corresponding histogram in 2D feature spaces.

We next illustrate the supervised terms. They are derived from exemplary regions, which are defined by the user with bounding boxes. These regions are only used to build prior foreground and background histograms; see Fig.3. The blue bounding box is the foreground and the red bounding boxes are the background.

B. The Evolution Process of the Segmentation

In this sequel, we briefly explain the process of the segmentation for the proposed method. In Fig.4, we show the segmentation results for certain period of iterations. On the left, the figure shows the evolution of contour for the cheetah. On the right, it describes the corresponding movement of histograms. In details, Fig.4(a) illustrates the original image with exemplary regions. The corresponding 2D histogram contour of the blue bounding box is shown in Fig.4(b). The relationship between the image segmentation evolution process and the movement of the corresponding probability distribution contours is shown in Fig.4(c) and Fig.4(d), respectively. In Fig.4(d), these directions represent the flux function m_1 used in the algorithm, which gives the directions of movement for the histogram. In this example, one may note that the quality of the results on these images uses 64-bin quantized gray-scale intensities. Our tests are all based on a Matlab code, working on a Dell OPTIPLEX 990, with Intel(R) Core(TM) i5-2400 processor, 3.1GHZ, 8GB RAM.

C. Supervised Segmentation

We next exhibit the advantages of considering multidimension features over histograms. We present a comparison with the local histogram based segmentation (LHBS) model proposed by Ni et al [9] which uses 1D features (the grayintensity). Fig.5 shows the comparison results by using the LHBS model and the proposed method with 2D features (the gray-intensity and the texture), in which the first row shows the input images with bounding boxes, and from the second to the third rows are the segmentation results by applying the LHBS model and the proposed model, respectively. It can be clearly seen that the proposed method can achieve better segmentation results with more accurate foreground boundary. From our experimental results, it seems that the more features are used, the better segmentation can be derived. It is reasonable that 2D features are better than 1D features in segmenting the images since more information is taken into consideration to distinguish the foreground and background.

D. Unsupervised Segmentation

In this part, we present several examples to demonstrate the effectiveness of the unsupervised model.

We first apply 1D features (the gray-intensity) information to segment the images. Consider an initial partition of the image with u_0 , the reference histograms a, b obtained by calculating the median gray-intensity of the foreground region and background region, respectively. Fig.6 shows how we select the reference measure in each step. It is chosen as the red one, i.e. the delta measure supported at the median point of the blue histogram. The right figure demonstrates that the energy function decreases during each iterative steps. The segmentation results are presented in Fig.7. The number of iterations of the three images in Fig.7 (from left to right) is 50, 70 and 85, respectively, and the corresponding time is 42s, 63s, 74s. However, it can be seen that the usage of 1D information can hardly obtain the satisfying results. We then do segmentation the images with 2D features (the gray-intensity and the texture features). Fig.8 shows the segmentation results, the number of iterations is 40, 55, 80, respectively, and the corresponding time is 120s, 165s, 190s. The segmentation results in Fig.8 indicates that the performance of the 2D algorithm is better than the one in the 1D algorithm.

E. Qualitative and Quantitative Experiments

In this section, we use two popular public databases, Microsoft GrabCut database and Berkeley segmentation data set to evaluate the quality of the unsupervised model with 2D features (the gray-intensity and the texture).

Fig.9 illustrates some segmentation examples on the Microsoft GrabCut database, which demonstrates that the proposed method can work better since it combines the grayintensity and the texture feature together. The ratio of the number of correctly segmented pixels to the total number of pixels of the image which we called the accuracy rate is used as the accuracy measurement of the segmentation results. Fig.10 depicts the accuracy rate for each image between the LHBS model and our method. The comparison results show that the proposed descriptors can obtain the less segmentation error rate. The qualitative and quantitative comparison results indicates that the proposed model is able to segment the image more precisely.



Fig. 4: Dynamical segmentation map. (a) shows the original image with bounding boxes. (b) is the corresponding 2D histogram contour of the blue bounding box. (c) is the evolution segmentation process. (d) illustrates the corresponding movement of 2D contours.



Fig. 5: Supervised model with comparison to LHBS [9] model. Row 1st shows the input images with boundary boxes. Row 2nd shows the LHBS model. Row 3rd shows the proposed method based on 2D feature histograms.

We next give a further comparison on the Berkeley segmentation data set. In addition, we compare our method with the nonlocal active contours(NLAC) model [37]. We select 50 test images for experimental comparisons, and use Jaccard index and Hausdorff distance to quantitatively evaluate the performance of the comparing methods. Fig.12 and Fig.13 display the corresponding Jaccard index value and Hausdorff distance for each test images, respectively. It can be clearly seen that our method obtains the higher Jaccard index and lower Hausdorff distance, which means that our method has



Fig. 6: The left figure shows that how we select the reference measure in each step. It is chosen as the red one, i.e. the delta measure supported at the median point of the blue histogram. The right figure demonstrates that the energy function decreases during these iterative steps.



Fig. 7: Unsupervised model with 1D features (the gray-intensity).



Fig. 8: Unsupervised model with 2D features (the gray-intensity and the texture).



Fig. 9: unsupervised model with comparison to the LHBS [9] model. Row 1st shows the original images. Row 2nd shows the segmentation results achieved by the LHBS model. Row 3rd shows the proposed method segmentation results. Row 4th shows the binary results of the LHBS model. Row 5th shows the binary results of proposed method. Row 6th shows the ground truth.

more high-quality segmentation results. Fig.14 illustrates the average Jaccard index values of the three methods, the LHBS model is about 0.695, the NLAC is about 0.708, and our method is about 0.83. Their corresponding standard deviations are 0.119., 0.118. and 0.068, respectively. Furthermore, as shown by Fig.15, the average Hausdorff distances of the three methods are 9.4594, 8.7923, and 6.5533, respectively. The standard deviations of above methods are 2.3779, 2.1437, and 1.7408, respectively. Therefore, it demonstrates that our method has been proven more effective and higher stable segmentation accuracies among these methods.

V. CONCLUSION

In this paper, we propose a segmentation model inheriting key ideas in optimal transport theory and homogeneous degree one regularization. Compared to existing methods, our model leverages the structure of original Monge-Kantorovich problem [41], which allows us to compute the segmentation with multiple dimensional easily, and thus to enhance the accuracy of the image segmentation. It is worth mentioning that the model is a L_1 type minimization, whose minimization variable keeps the same number of feature spaces. This fact allows us to design fast algorithms via primal-dual approach and keep the computational complexity at a reasonable level. It is highly parallelizable [52]. The experimental results demonstrate the effectiveness of the method, comparing with several state-ofthe-art Earth mover's distance methods. Furthermore, due to the optimal transport theory, our technique can easily deal with a variety of image processing problems, such as image retrieval, video tracking and so on. Therefore, our future work will focus on applying the model on these challenging problems.

APPENDIX

Proof of Theorem 1: To prove the convergence, we only need to check the conditions in [50]. Denote the prime variable $X = (u, m_1, m_2)$, and dual variable $Y = (g, \Phi_1, \Phi_2)$. Rewrite the Lagrangian in (11) by

$$L(X,Y) = G(X) + X^T K Y - F(Y) ,$$

where G, F can be shown as convex functions and $K = \begin{pmatrix} \lambda \cdot \operatorname{div}_{\Omega} & 0 & 0 \end{pmatrix}$

 $\begin{pmatrix} 0 & \operatorname{div}_{\mathcal{F}} & 0 \\ 0 & 0 & \operatorname{div}_{\mathcal{F}} \end{pmatrix}, \text{ where } \operatorname{div}_{\Omega}, \operatorname{div}_{\mathcal{F}} \text{ represents the }$

discrete divergence operator in Ω and \mathcal{F} . Since

$$KK^T = \begin{pmatrix} \lambda^2 \cdot \Delta_{\Omega} & 0 & 0\\ 0 & \Delta_{\mathcal{F}} & 0\\ 0 & 0 & \Delta_{\mathcal{F}} \end{pmatrix}$$

and the algorithm converges for $\mu \tau \|KK^T\|_2^2 < 1$, we prove the result.

Proof of Lemma 1: The proof is from the definition of Earth Mover's distance. Consider

$$\operatorname{EMD}(a,\mu) := \min_{\pi} \quad \int_{\mathbb{R}^d \times \mathbb{R}^d} \|y - \tilde{y}\|_1 \pi(y, \tilde{y}) dy d\tilde{y}$$

s.t.

$$\int_{\mathbb{R}^d} \pi(y, \tilde{y}) d\tilde{y} = a(y) , \quad \int_{\mathbb{R}^d} \pi(y, \tilde{y}) dy = \mu(\tilde{y}) , \quad \pi(y, \tilde{y}) \ge 0$$

Since a(y) is the delta measure supported at x_0 , then $\pi(y, \tilde{y}) = 0$ a.e. if $y \neq x_0$, and $\int_{\{y=x_0\}} \pi(y, \tilde{y}) dy = \mu(\tilde{y})$. Thus

$$\begin{split} &\int_{\mathbb{R}^d} \Big(\int_{\mathbb{R}^d} \|y - \tilde{y}\|_1 \pi(y, \tilde{y}) dy \Big) d\tilde{y} \\ &= \int_{\mathbb{R}^d} \Big(\int_{\{y = x_0\}} \|x_0 - \tilde{y}\|_1 \pi(y, \tilde{y}) dy \Big) d\tilde{y} \\ &= \int_{\mathbb{R}^d} \|x_0 - \tilde{y}\|_1 \Big(\int_{\{y = x_0\}} \pi(y, \tilde{y}) dy \Big) d\tilde{y} \\ &= \int_{\mathbb{R}^d} \|x_0 - \tilde{y}\|_1 \mu(\tilde{y}) d\tilde{y} = \int_{\mathbb{R}^d} \sum_{i=1}^d \|x_0^i - \tilde{y}_i\|_1 \mu(\tilde{y}) d\tilde{y} \\ &= \sum_{i=1}^d \int_{\mathbb{R}^1} \|x_0^i - \tilde{y}_i\|_1 \mu_i(\tilde{y}_i) d\tilde{y}_i \ . \end{split}$$

Since μ_i is fixed and the minimization is taken among x_0^i , $i = 1, \dots, d$, then the optimal x_0^i is attached at the median point of μ_i .



Fig. 10: The segmentation accuracy tested on the Microsoft GrabCut database. The magenta contour, and green contour are the segmentation accuracy of the LHBS[9] model, and our method, respectively.



Fig. 11: Some comparison examples on Berkeley segmentation data set. Top to bottom: test images, results of the NLAC [37] model, the LHBS [9] model, and our method, respectively.



Fig. 12: The Jaccard values of segmentation results on the 50 Berkeley data set images. The magenta contour, the green contour, and the blue contour are the Jaccard values of the proposed method, the LHBS [9] model, and the NLAC [37] model, respectively.



Fig. 13: The Hausdorff distances of segmentation results on the 50 Berkeley data set images. The magenta contour, the green contour, and the blue contour are the Hausdorff distances of the proposed method, the LHBS [9] model, and the NLAC [37] model, respectively.



Fig. 14: The average Jaccard index values of the LHBS model, the NLAC model, and our method, respectively.



Fig. 15: The average Hausdorff distances of the LHBS model, the NLAC model, and our method, respectively.

REFERENCES

- Y. Lei, W. Yuan, H. Wang, Y. Wenhu, and W. Bo, "A Skin Segmentation Algorithm Based on Stacked Autoencoders," *Trans. Multi.*, vol. 19, no. 4, pp. 740–749, 2017.
- [2] L. Zhang, Y. Gao, Y. Xia, K. Lu, J. Shen, and R. Ji, "Representative Discovery of Structure Cues for Weakly-Supervised Image Segmentation," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 470–479, 2014.
- [3] H. Jiang, G. Zhang, H. Wang, and H. Bao, "Spatio-Temporal Video Segmentation of Static Scenes and Its Applications," *IEEE Transactions* on *Multimedia*, vol. 17, no. 1, pp. 3–15, 2015.
- [4] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic Active Contours," *International Journal of Computer Vision*, vol. 22, no. 1, pp. 61–79, 1997.

- [5] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.
- [6] S. Osher and N. Paragios, Geometric Level Set Methods in Imaging, Vision, and Graphics. Springer Science & Business Media, 2003.
- [7] G. Aubert, M. Barlaud, O. Faugeras, and S. Jehan-Besson, "Image Segmentation Using Active Contours: Calculus of Variations or Shape Gradients?" SIAM Journal on Applied Mathematics, 2006.
- [8] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," *Communications on Pure and Applied Mathematics*, vol. 42, no. 5, pp. 577–685, 1989.
- [9] T. Chan, S. Esedoglu, and K. Ni, "Histogram Based Segmentation Using Wasserstein Distances," in *Scale Space and Variational Methods in Computer Vision*, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2007, pp. 697–708.
- [10] S. C. Zhu and A. Yuille, "Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 9, pp. 884–900, 1996.
- [11] L. A. Vese and T. F. Chan, "A Multiphase Level Set Framework for Image Segmentation Using the Mumford and Shah Model," *International Journal of Computer Vision*, vol. 50, no. 3, pp. 271–293, 2002.
- [12] B. Sandberg, T. Chan, and L. Vese, "A Level-Set and Gabor-based Active Contour Algorithm for Segmenting Textured Images," UCLA Department of Mathematics CAM report, Tech. Rep., 2002.
- [13] C. Sagiv, N. A. Sochen, and Y. Y. Zeevi, "Integrated active contours for texture segmentation," *IEEE Transactions on Image Processing*, vol. 15, no. 6, pp. 1633–1646, 2006.
- [14] N. Paragios and R. Deriche, "Geodesic Active Regions: A New Framework to Deal with Frame Partition Problems in Computer Vision," *Journal of Visual Communication and Image Representation*, vol. 13, no. 1, pp. 249–268, 2002.
- [15] A. Dubrovina-Karni, G. Rosman, and R. Kimmel, "Multi-Region Active Contours with a Single Level Set Function," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 8, pp. 1585– 1601, 2015.
- [16] C. Li, X. Wang, S. Eberl, M. Fulham, and D. D. Feng, "Robust Model for Segmenting Images With/Without Intensity Inhomogeneities," *IEEE Transactions on Image Processing*, vol. 22, no. 8, pp. 3296–3309, 2013.
- [17] K. Zhang, L. Zhang, K. M. Lam, and D. Zhang, "A Level Set Approach to Image Segmentation With Intensity Inhomogeneity," *IEEE Transactions on Cybernetics*, vol. 46, no. 2, pp. 546–557, 2016.
- [18] R. Ronfard, "Region-based strategies for active contour models," *International Journal of Computer Vision*, vol. 13, no. 2, pp. 229–251, 1994.
- [19] T. F. Chan, B. Y. Sandberg, and L. A. Vese, "Active Contours without Edges for Vector-Valued Images," *Journal of Visual Communication and Image Representation*, vol. 11, no. 2, pp. 130–141, 2000.
- [20] S. Osher and J. A. Sethian, "Fronts propagating with curvaturedependent speed: Algorithms based on Hamilton-Jacobi formulations," *Journal of Computational Physics*, vol. 79, no. 1, pp. 12–49, 1988.
- [21] R. Kimmel, "Fast Edge Integration," in Geometric Level Set Methods in

Imaging, Vision, and Graphics. Springer, New York, NY, 2003, pp. 59–77.

- [22] R. Malladi, J. A. Sethian, and B. C. Vemuri, "Shape modeling with front propagation: A level set approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 2, pp. 158–175, 1995.
- [23] T. F. Chan, S. Esedoglu, and M. Nikolova, "Algorithms for Finding Global Minimizers of Image Segmentation and Denoising Models," *SIAM Journal on Applied Mathematics*, 2012.
- [24] E. S. Brown, T. F. Chan, and X. Bresson, "Completely Convex Formulation of the Chan-Vese Image Segmentation Model," *International Journal of Computer Vision*, vol. 98, no. 1, pp. 103–121, 2012.
- [25] A. Chambolle and J. Darbon, "On Total Variation Minimization and Surface Evolution Using Parametric Maximum Flows," *International Journal of Computer Vision*, vol. 84, no. 3, p. 288, 2009.
- [26] J. Kim, J. W. Fisher, A. Yezzi, M. Cetin, and A. S. Willsky, "A nonparametric statistical method for image segmentation using information theory and curve evolution," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1486–1502, 2005.
- [27] D. Cremers, M. Rousson, and R. Deriche, "A Review of Statistical Approaches to Level Set Segmentation: Integrating Color, Texture, Motion and Shape," *International Journal of Computer Vision*, vol. 72, no. 2, pp. 195–215, 2007.
- [28] D. Cremers and S. Soatto, "Motion Competition: A Variational Approach to Piecewise Parametric Motion Segmentation," *International Journal of Computer Vision*, vol. 62, no. 3, pp. 249–265, 2005.
- [29] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, "Cosegmentation of Image Pairs by Histogram Matching - Incorporating a Global Constraint into MRFs," in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1, 2006, pp. 993–1000.
- [30] D. Freedman and T. Zhang, "Active Contours for Tracking Distributions," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 518– 526, 2004.
- [31] O. Michailovich, Y. Rathi, and A. Tannenbaum, "Image Segmentation Using Active Contours Driven by the Bhattacharyya Gradient Flow," *IEEE Transactions on Image Processing*, vol. 16, no. 11, pp. 2787– 2801, 2007.
- [32] A. Adam, R. Kimmel, and E. Rivlin, "On Scene Segmentation and Histograms-Based Curve Evolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1708–1714, 2009.
- [33] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.
- [34] C. Wang, Z. Liu, and S. C. Chan, "Superpixel-Based Hand Gesture Recognition With Kinect Depth Camera," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 29–39, 2015.
- [35] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust Part-Based Hand Gesture Recognition Using Kinect Sensor," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1110–1120, 2013.
- [36] C. Villani, *Topics in Optimal Transportation*, ser. Graduate studies in mathematics. Providence, RI: American Mathematical Society, 2003, no. v. 58.
- [37] M. Jung, G. Peyré, and L. D. Cohen, "Texture Segmentation via Nonlocal Non-parametric Active Contours," in *Energy Minimization Methods* in *Computer Vision and Pattern Recognition*, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2011, pp. 74–88.
- [38] G. Peyré, J. Fadili, and J. Rabin, "Wasserstein active contours," in 2012 19th IEEE International Conference on Image Processing, 2012, pp. 2541–2544.
- [39] P. Swoboda and C. Schnörr, "Variational Image Segmentation and Cosegmentation with the Wasserstein Distance," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2013, pp. 321–334.
- [40] J. Rabin and N. Papadakis, "Convex Color Image Segmentation with Optimal Transport Distances," in *Scale Space and Variational Methods in Computer Vision*, ser. Lecture Notes in Computer Science. Springer, Cham, 2015, pp. 256–269.
- [41] W. Li, E. K. Ryu, S. Osher, W. Yin, and W. Gangbo, "A Parallel Method for Earth Mover's Distance," *Journal of Scientific Computing*, pp. 1–16, 2017.
- [42] L. C. Evans and W. Gangbo, "Differential equations methods for the Monge-Kantorovich mass transfer problem," *Mem. Amer. Math. Soc*, vol. 137, p. 653, 1999.
- [43] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.
- [44] M. Beckmann, "A Continuous Model of Transportation," *Econometrica*, vol. 20, no. 4, pp. 643–660, 1952.

- [45] J. Lellmann, D. A. Lorenz, C. Schönlieb, and T. Valkonen, "Imaging with Kantorovich–Rubinstein Discrepancy," *SIAM Journal on Imaging Sciences*, vol. 7, no. 4, pp. 2833–2859, 2014.
- [46] W. Li, "A study of stochastic differential equations and Fokker-Planck equations with applications," 2016.
- [47] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.
- [48] G. Buttazzo and F. Santambrogio, "A Model for the Optimal Planning of an Urban Area," *SIAM Journal on Mathematical Analysis*, vol. 37, no. 2, pp. 514–530, 2005.
- [49] E. K. Ryu, W. Li, P. Yin, and S. Osher, "Unbalanced and Partial l₁ Monge–Kantorovich Problem: A Scalable Parallel First-Order Method," *Journal of Scientific Computing*, pp. 1–18, 2017.
- [50] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," 2010.
- [51] N. Houhou, X. Bresson, A. Szlam, and T. F. Chan, Semi-Supervised Segmentation Based on Non-Local Continuous Min-Cut.
- [52] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman Iterative Algorithms for \$\ell_1\$-Minimization with Applications to Compressed Sensing," *SIAM Journal on Imaging Sciences*, vol. 1, no. 1, pp. 143–168, 2008.



Yupeng Li received the B.E. degree in Communication Engineering from the Nanjing University of Science and Technology, China, in 2011, the M.S. degree in Pattern Recognition and Intelligent System from the Jiangsu University, China, in 2014. He is currently a Ph.D. Candidate in the School of Computer Science and Engineering, Nanjing University of Science and Technology, China, and is a visiting student at the University of California, Los Angeles. His research interests include pattern recognition, machine learning, image analysis.



Wuchen Li Wuchen Li received the Ph.D. degree in mathematics from Georgia institute of technology in 2016. From 2016 to present, he was a computational and Applied mathematics assistant Professor in UCLA. His research interests include optimal transport, optimal control, stochastic oscillator, image processing and numerical analysis.



Guo Cao received the Ph.D. degree in pattern recognition and intelligence system from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2006. He has been an Associate Professor with the Department of Computer Science, Nanjing University of Science and Technology, Nanjing, China, since 2010. His research interests include remote sensing image processing, biometrics, and computer vision.