基于深度 CRF 网络的单目红外场景深度估计

王倩倩，赵海涛
（华东理工大学 信息科学与工程学院，上海 200237）

摘要：对单目红外图像进行深度估计，不仅有利于 3D 场景理解，而且有助于进一步推广和开发夜间视觉应用。针对红外图像无颜色、纹理不丰富、轮廓不清晰等缺点，本文提出一种新颖的深度条件随机场网络学习模型（deep conditional random field network, DCRFN）来估计红外图像的深度。首先，与传统条件随机场（conditional random field, CRF）模型不同，DCRFN 不需预设成对特征，可通过一个浅层网络架构提取和优化模型的成对特征。其次，将传统单目图像深度回归问题转换为分类问题，在损失函数中考虑不同标签的有序信息，不仅加快了网络的收敛速度，而且有助于获得更优的解。最后，本文在 DCRFN 损失函数层计算不同空间尺度的成对项，使得预测深度图的景物轮廓信息相比于无尺度约束模型更加丰富。实验结果表明，本文提出的方法在红外数据集上优于现有的深度估计方法，在局部场景变化的预测中更加平滑。

关键词：红外图像；深度估计；条件随机场；有序约束
中图分类号：TP391.9 **文献标识码**：A **文章编号**：1001-8891(2020)06-0580-09

Depth Estimation of Monocular Infrared Scene Based on Deep CRF Network

WANG Qianqian，ZHAO Haitao
(*School of Information Science and Engineering, East China University of Science and Technology, Shanghai* 200237, *China*)

Abstract：Depth estimation from monocular infrared images is required for understanding 3D scenes; moreover, it could be used to develop and promote night-vision applications. Owing to the shortcomings of infrared images, such as a lack of colors, poor textures, and unclear outlines, a novel deep conditional random field network (DCRFN) is proposed for estimating depth from infrared images. First, in contrast with the traditional CRF(conditional random field) model, DCRFN does not need to preset pairwise features. It can extract and optimize pairwise features through a shallow network architecture. Second, conventional monocular-image-based depth regression is replaced with multi-class classification, wherein the loss function considers information regarding the order of various labels. This conversion not only accelerates the convergence speed of the network but also yields a better solution. Finally, in the loss function layer of the DCRFN, pairwise terms of different spatial scales are computed; this makes the scene contour information in the depth map more abundant than that in the case of the scale-free model. The experimental results show that the proposed method outperforms other depth estimation methods with regard to the prediction of local scene changes.

Key words：infrared image, depth estimation, conditional random field, ordered constraint

0 引言

单目图像深度估计是计算机视觉中最基本的任务之一，已经在场景理解[1]、3D 建模[2]、机器人[3]、自动驾驶[4]等领域获得了广泛应用。图像深度即图像上某点像素到拍摄相机的距离。早期的单目深度研究主要关注几何假设，如箱体模型推测场景的空间布局[5]等，由于其严格的环境假设导致模型只在特殊场景下有效。

在单目场景深度估计的研究中，基于概率图模

收稿日期：2019-05-29；**修订日期**：2019-07-12.
作者简介：王倩倩（1993-），女，硕士研究生，主要从事计算机视觉方面的研究。
通信作者：赵海涛（1974-），男，博士，教授，主要从事模式识别、计算机视觉方面的研究。E-mail：haitaozhao@ecust.edu.cn。
基金项目：国家自然科学基金（61375007）；上海市科委基础研究项目（15JC1400600）。

型（probabilistic graphical models, PGM）的方法能够在输入和输出之间建立结构化联系，因此被广泛应用到单目深度估计中。马尔可夫随机场（Markov random field, MRF）和条件随机场（conditional random field, CRF）是应用最多的 PGM 方法，他们的关键点在于构造合理以及正确的特征表示。早期的研究[6]关注如何从单目线索中构造 MRF 的绝对特征（一元特征）和相对特征（成对特征），但是这些纹理变化、运动、遮挡、阴影、散焦等单目线索都是低维手工特征，缺乏场景的通用性。

为了解决特定场景的限制，结合语义等其他任务信息[7-8]和基于数据驱动的非参数方法[9-10]在一定程度上缓解了场景限制问题。Liu 等学者[7]以语义分类和几何先验为前提条件，首先通过 MRF 完成对图像的语义标注，之后使用预测的语义标签来指导深度估计模型。但在现实生活中，很难获得可靠的语义或其他额外信息。基于数据驱动的非参数方法是通过视觉相似性比较，从具有图像-深度的数据集中搜索候选图像的近似深度图。Liu 等学者[9]将深度估计当作连续离散的 CRF 优化问题，采用级联的 GIST、PHOG 等手工特征通过 K 近邻算法检索前 k 个候选深度图，但是 CRF 成对项只考虑了相邻超像素之间的遮挡关系。

Eigen 等学者[11]首先将卷积神经网络（convolutional neural networks, CNNs）应用在深度估计，从图像块上训练两个尺度的 CNNs，从而获得了较好的结果。之后的很多研究都是通过修改深度网络框架以期待获得更精确的结果。Laina 等学者[12]在 ResNet[13-15]中利用了反向 Huber 损失；Gu 等学者[13]基于 ResNet 估计红外图像的深度；Wu 等学者[14]在 CNNs 的基础上引入循环神经网络（recurrent neural network, RNN）的序列特征用于红外视频的深度估计。

上述基于深度学习的方法由于没有考虑优化结构损失，从而造成预测深度图的景物轮廓模糊等问题。CRF（或 MRF）明确地包含了结构约束，CNNs 与 CRF（或 MRF）的结合成为深度估计、语义分割等视觉任务的主流方法。Chen 等学者[16]采用全连接的 CRF[17]，通过平均场近似实现 CRF 推理，成对项特征来自预设的颜色和位置特征。Li 等学者[18]提出深度回归和深度精细框架，该模型 CRF 成对特征的设计来自文献[17]。Liu 等学者[19]提出了用于深度估计的 DCNF 模型，作者致力于在一个统一的框架下学习连续 CRF 和 CNNs。后来，Liu 等学者[20]进一步提出全卷积网络训练和超像素池化方法。文献[19-20]中 CRF 成对特征均来自颜色、颜色直方图和纹理这 3 个简单预设信息。Xu 等学者[21]将 Attention 机制与 CNN-CRF 结合用于深度估计，其 CRF 也是针对可见光图像设计的。

由于红外和可见光的成像原理差异，使得可见光图像适用于光线充足的环境，而红外图像不受光线条件的限制，这使其特别适用于夜间无人驾驶等夜视应用[22]，因而估计车载红外图像的深度信息意义重大。早前针对红外图像的研究[23-24]大多停留在手工特征的提取上，由于红外图像存在纹理、色彩信息不丰富等缺点，提取针对红外图像的特征具有一定的挑战性。

本文将深度估计表示为离散 CRF 学习问题，提出深度条件随机场网络学习模型，该模型主要由特征学习和 CRF 损失层组成。首先，一元项对单个像素的信息进行了估计与约束，成对项鼓励相邻像素拥有相似的深度信息。本文的一元特征是基于 Dense-Net[25]的深度卷积网络，充分利用了密集连接机制，能够通过特征在维度上的连接实现特征重用，使得模型的参数更少，性能更优。Noh 等学者[26]指出浅层 CNNs 可以获得某些低级特征及其非线性组合，本文成对特征来自四层密集连接的全卷积网络，为红外图像提供了更丰富的特征表示。其次，受文献[27-28]启发，并考虑车载图像的远近特性，本文将连续深度值离散到对数空间以获得有序标签，从而在 CRF 损失层中提出自适应 Huber Penalty 来增加标签有序性约束和多尺度信息交互约束。

1 基本原理

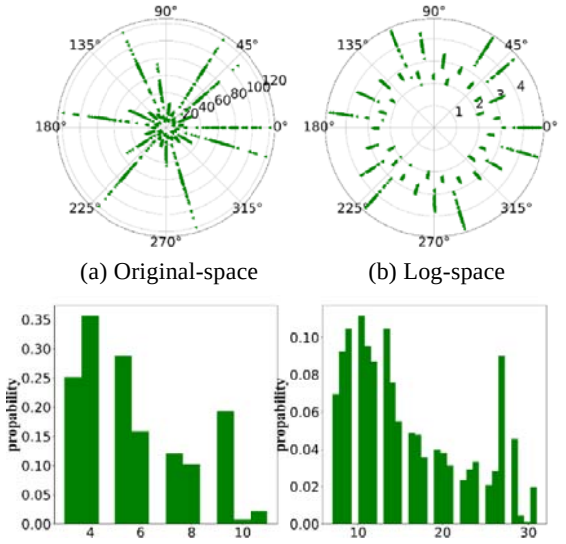
Cao 等学者[27]已经证明将传统的回归问题转化为有序多分类问题，不仅会增加预测结果的正确性，而且会加快预测效率。针对车载红外图像本身存在的特点，本文在对数空间里平均分割真实深度图获得各个深度层级的分类标签，用于深度估计实验：

$$l_i^{(k)} = \mathrm{floor}\left(\frac{\ln d_i^{(k)} - \ln d_{\min}}{\ln d_{\max} - \ln d_{\min}} \times T\right) \tag{1}$$

式中：$i \in \{1,2,\cdots,N\}$，$l_i^{(k)}$和 $d_i^{(k)}$分别表示第 k 张图像上像素 i 的真实深度标签和真实深度值；$d_{\min}$ 和 $d_{\max}$ 分别表示原深度图中所有像素点的最小和最大真实深度值；floor(·)表示向下取整；T 是深度标签的等级。

图 1(a)、(b)展示了数据集中任一深度图在原始空间和对数空间的分布情况，可以得出，在对数空间里分割原深度图可以获得更加均匀的深度分布。图 1(c)、(d)表明 T 值的选取影响整个分类的好坏，

如果选取过大，说明分割很密集与原深度图相差较少，则对提高预测效率贡献较小；反之，分割结果会很稀疏与原深度图相差较大，则会降低预测正确率。本文通过实验分别设置 T 值为 12、22、32、52、72，从数据分布图上可以得出结论：T=32 时，效果最佳。



(a) Original-space (b) Log-space

(c) T=12 (d) T=32

图 1 深度分布

Fig.1 Depth distribution

2 深度估计模型的设计

本文的目标是估计红外图像上每个像素点的深度标签信息，将单目深度估计问题当作 CRF 学习问题，从而提出深度条件随机场网络学习模型（deep conditional random field network, DCRFN）。DCRFN 的整体结构如图 2 所示，由特征学习和 CRF 损失层组成。在特征学习中，采用两个深度卷积模型分别学习 CRF 的一元特征和成对特征。其中，一元模型直接从输入图像学习每个像素到深度标签的映射，从而获得全局深度估计的结果；成对模型旨在获得邻域像素之间的约束关系，与传统 CRF 成对特征的简单预设不同，本文提出的成对特征可以从一个新颖的深度卷积网络中提取。在 CRF 损失层，其 CRF 一元项和成对项来自两个深度卷积模型的输出，通过最小化负对数似然获得最优的深度估计结果。

2.1 DCRFN 描述

假设 $I=\{I_1,\cdots, I_N\}$表示任意一张红外图像，其中 N 是输入图像的总像素；$Y=\{Y_1,\cdots, Y_N\}$表示每个输入像素的深度标签，其中 $Y_i\in\{l_1,l_2,\cdots, l_k\}$。CRF 为给定随机场 I 条件下，离散随机变量 Y 的马尔科夫随机场。条件随机场(I,Y)可以通过如下的 Gibbs 分布表示：

$$\Pr(Y|I)=\frac{\exp(-E(Y,I))}{Z(I)} \tag{2}$$

式中：Gibbs 能量函数 $E(I,Y)=\sum_{c\in C_G}\phi_c(Y|I)$，$G=(v,\varepsilon)$是建立在深度变量 Y 上；团 c 是 G 中节点 C_g 的子集，其中深度值是条件依赖关系，$\phi_c(Y|I)$ 是团 $c\in C_g$ 的势函数。

式 $Z(I)=\sum_{Y}\exp\left(-\sum_{c\in C_G}\phi_c(Y_c|I)\right)$ 用来保证所有概率加和为 1。本文将 Gibbs 能量表示为独立像素上的一元势和相邻像素上的成对势：

$$E(I,Y)=\sum_{i\in U}\Phi_i(Y_i|I)+\lambda\sum_{i\in U,j\in B_i}\Psi_{i,j}(Y_i,Y_j|I) \tag{3}$$

式中：λ是平衡参数。一元势 Φ_i 反映了每个像素分配标签的置信度，成对势 $\Psi_{i,j}$ 鼓励相邻像素被赋予相同的深度标签。U 表示输入红外图像的所有像素，B_i 是像素 $i(i\in U)$的邻域。

由于归一项和深度标签 Y 是无关的，预测一张新红外图像的深度信息 Y^*，最大化 $\Pr(Y|I)$ 等价于最小化能量函数 $E(I,Y)$：

$$Y^*=\underset{Y}{\operatorname{argmax}}\Pr(Y|I)\Leftrightarrow\underset{Y}{\operatorname{argmin}}E(I,Y) \tag{4}$$

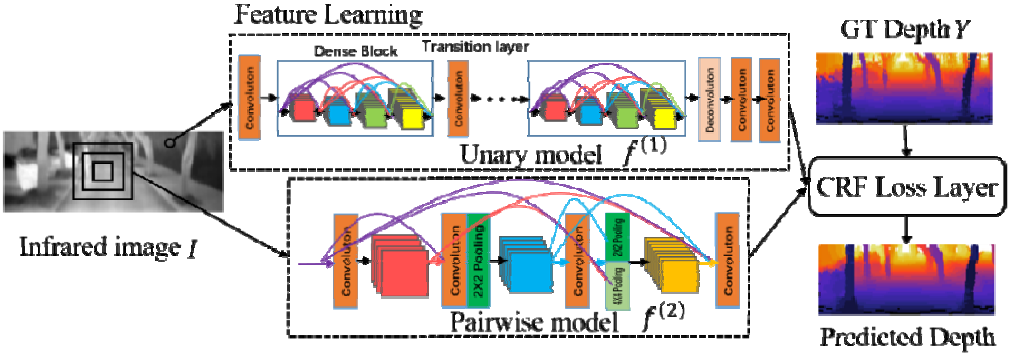


图 2 DCRFN 学习模型 Fig.2 DCRFN learning model

2.2 势函数

根据图 2 的两个深度卷积模型重新定义 CRF 的一元势和成对势。输入红外图像经过一元模型直接学习每个输入像素到深度标签的映射 $f^{(1)}$，经过成对模型 $f^{(2)}$在邻域上建立相关关系；通过特征学习的 $f^{(1)}$、$f^{(2)}$作为输入传递给 CRF 损失层。

2.2.1 一元势函数

CRF 的一元势 Φ_i，也就是等式(2)的第一项。本文引入密集连接模型作为一元分类器，其势函数定义：

$$\Phi_i\left(Y_i \mid I,\theta_1\right)=Y_i\ln\hat{Y}_i\left(\theta_1\right) \tag{5}$$

式中：$Y_i,\hat{Y}_i$ 分别表示像素 i 的真实深度标签和预测深度标签。由于 $f^{(1)}$为输入像素到深度标签的映射，本文认为 $\hat{Y}_i=f_i^{(1)}$。其中 $\theta_1=\{w_1,b_1\}$是一元模型的所有权重和偏置参数，通过 BP 算法来反向更新这些参数。

$$\Psi_{i,j}\left(Y_i|I_i,\theta_2\right)=\sum_{s\in S}\exp\left(-\frac{\|f_i^{(2)(s)}\left(\theta_2\right)-f_j^{(2)(s)}\left(\theta_2\right)\|^2}{\gamma_s^2}\right)h\left(Y_i^{(s)}-Y_j^{(s)}\right) \tag{6}$$

一元模型是基于 Dense-Net 的深度卷积网络，主要由密集块、上采样模块和卷积层组成，其密集块如图 2 所示。输入红外图像首先经过一个 7×7 的卷积层获得丰富的特征，之后通过 4 个密集块，每个密集块之间通过 3×3 的卷积连接实现尺寸的降低。通过对比实验确定各层输出通道数从前到后为 4、8、12、24。密集块充分利用了跨层连接操作，每个层都会与前面的层在维度上连接在一起，并作为下一层的输入，要求各层的特征图大小要一致。由于池化操作会损失大量的细节信息[28]，本文用卷积操作替代平均池化实现下采样。之后经过上采样模块恢复特征图尺寸至真实深度图尺寸。本文将最后两层全连接改为全卷积结构，输出特征图像尺寸为[40,144,32]，从而实现像素级别的分类预测。

2.2.2 成对势函数

CRF 的成对势 $\Psi_{i,j}$，也就是等式(2)的第二项。如图 3 所示，考虑与像素 i 相邻的八邻域，像素 i 的邻域标签信息对其产生约束。但是除了邻域像素存在相互作用外，由八邻域组成的块与邻域块有时也存在强烈的相互作用。本文在 S 个空间尺度上构造 CRF 使之形成尺度间的因果关系，粗尺度注重区域的各向同性，而较细尺度则注重物体边界的各项异性，尺度之间可构成马尔可夫链，较粗尺度的参数对较细尺度的参数有指导意义。通过成对模型获得各种相似特征，使得预测深度图更加平滑：

$$E\left(I,Y,\theta\right)=\sum_{i\in U}Y_i\ln\hat{Y}_i\left(\theta_1\right)+\lambda\sum_{s\in S}R_{ij}^{(s)}\left(\theta_2\right)h_{ij}^{(s)} \tag{7}$$

式中：$s=\{1,2,\cdots,S\}$表示在 S 个尺度上对邻域像素进行运算，较粗尺度的深度信息限制为较细尺度下深度信息的平均值，即 $f_i^{(2)(s+1)}=\frac{1}{9}\sum_{i\in U,j\in B_i}f_j^{(2)(s)}$，本文取 S=3（细节见章节 4.1）。$\theta_2=\{w_2,b_2\}$是成对模型的参数，γ_s 是超参数(细节见章节 4.1)。$R_{ij}^{(s)}=\exp\left(-\frac{\|f_i^{(2)(s)}\left(\theta_2\right)-f_j^{(2)(s)}\left(\theta_2\right)\|^2}{\gamma_s^2}\right)$衡量了两个相邻像素的光滑程度。$h$ 是改编后的 Huber penalty，用于有序分类的约束，其表达形式如下：

$$h\left(a\right)=\delta^2\left(\sqrt{1+\left(\frac{a}{\delta}\right)^2}-1\right) \tag{8}$$

式中：$a=\frac{\left|Y_i^{(s)}-Y_j^{(s)}\right|}{T}$，$T$ 是离散化后标签的数量；δ 是常数。对于相邻深度差异小的，误差为二次；而对于相邻深度差异大的，误差是线性。此有序约束函数能在训练过程中辅助更新网络的参数。

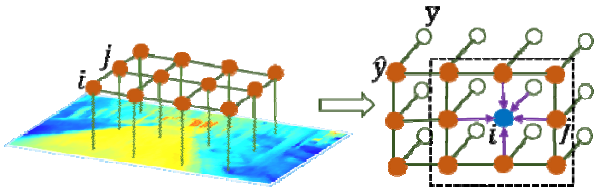


图 3 DCRFN 邻域约束关系

Fig.3 Neighborhood constraints of DCRFN

一元项只对单个像素的信息进行了估计与约束，成对项鼓励相邻像素拥有相似的标签信息。本文首先考虑红外图像无颜色且纹理信息不丰富，很难通过手工提取特征的缺点；其次考虑到 CRF 推断的可行性，成对模型在 Dense-Net 思想上引入跨密集块的密集连接。由于只有 4 层卷积，且卷积大小均为 3×3，大大减少了模型的参数量，使得 CRF 在邻域上实现精确推断。通过 2×2 和 4×4 的池化操作实现降采样，使得跨层连接不受特征图尺寸的限制，从而通过全密集连接获得红外图像丰富的成

对特征。

2.3 DCRFN的推理和学习

等式(5)、(6)分别定义了CRF的一元势和成对势，本文通过最小化能量函数实现对参数的更新，对参数求偏导：

$$\frac{\partial E(I,Y,\theta)}{\partial \theta}= \left[\sum_{i\in U}\frac{\partial \Phi_i(Y_i|I,\theta_1)}{\partial \theta_1},\lambda\sum_{i\in U,j\in B_i}\frac{\partial \Psi_{i,j}(Y_i,Y_j|I,\theta_2)}{\partial \theta_2}\right]^{\mathrm{T}} \tag{9}$$

$\theta=\{\theta_1,\theta_2\}$是DCRFN的所有训练参数。等式(9)的第一项是交叉熵损失函数对参数求偏导的过程，根据链式求导法则易推导。此处仅对第二项推理，假设 $\sum_{i\in U,j\in B_i}\partial\Psi_{i,j}(Y_i,Y_j|I,\theta_2)=L_2(\theta_2)$，则：

$$\frac{\partial L_2}{\partial \theta_2}=-\sum_{s=1}^{S}\sum_{i\in U,j\in B_i}h_{ij}^{(s)}\cdot\exp\left(-\frac{F}{\gamma_s^2}\right)\cdot F\cdot\frac{\partial F}{\gamma_s^2\partial\theta_2} \tag{10}$$

令 $\|f_i^{(2)(s)}(\theta_2)-f_j^{(2)(s)}(\theta_2)\|^2=F$，其中 $\exp\left(-\frac{F}{\gamma_s^2}\right)$不影响求偏导的过程，因此可当成常数；$h_{ij}^{(s)}$本身与变量$\theta_2$无关，因此也可做常数。

令 $\sum_{s=1}^{S}\sum_{i\in U,j\in B_i}h_{ij}^{(s)}\cdot\exp\left(-\frac{F}{\gamma_s^2}\right)\cdot F/\gamma_s^2=C$，则：

$$\frac{\partial L_2}{\partial \theta_2}=-\sum\nolimits_{s=1}^{S}\sum_{i\in U,j\in B_i}C\cdot\frac{\partial F}{\partial\theta_2} \tag{11}$$

令Z^l表示第l层神经元的净输入(未经过激活函数)，W^l表示第$l-1$层到第l层的权重，I^l是第l层神经元的输出(经过激活函数)即$I^l=f(Z^l)$，$f_l(\cdot)$是激活函数，那么第l层神经元的净输入$Z^l=W^l\otimes I^{l-1}+b_1^l$，其中⊗表示卷积运算。则第$l$层的误差项为：

$$\delta^l=\frac{\partial L}{\partial Z^l} \tag{12}$$

根据链式法则，得到第l层的误差项的计算公式：

$$\delta^l=\frac{\partial I^l}{Z^l}\cdot\frac{\partial Z^{l+1}}{\partial I^l}\cdot\frac{\partial L}{\partial Z^{l+1}}= f'(Z^l)\cdot\left((W^{l+1})^{\mathrm{T}}\cdot\delta^{l+1}\right) \tag{13}$$

进一步，通过BP对参数反向推导，L_2关于第l层权重W_2^l和偏置b_2^l梯度为：

$$\frac{\partial L_2}{\partial W_2^l}=\frac{\partial L_2}{\partial Z^l}\cdot\frac{\partial Z^l}{\partial W_2^l}=C\delta^l(I^{l-1})^{\mathrm{T}} \qquad \frac{\partial L_2}{\partial b_2^l}=\frac{\partial L_2}{\partial Z^l}\cdot\frac{\partial Z^l}{\partial b_2^l}=C\delta^l \tag{14}$$

3 网络训练

3.1 数据采集

NUSTMS数据集由南京理工大学无人车队拍摄所得，包含5098对由远红外摄像机拍摄的红外图像和测距雷达生成的深度图。红外图像和相应的深度图的分辨率分别为576×160和144×40。本文数据集被分成训练集（3488对），验证集（586对）和测试集（1024对），数据集的深度范围为3～100 m。其数据采集装置如图4所示。

3.2 训练过程

本文所有实验均在GeForce GTX 1080Ti显卡上采用深度学习库TensorFlow实现。在训练阶段，预先对红外图像进行归一化，通过截断正态分布来初始化权重，并将偏置初始化为零。采用计算每个参数自适应学习率的优化器Adam Optimizer来更新权重和偏置，初始学习率设置为1×10^{-4}。实验设置迭代次数为100000，batch-size为4。成对项的参数涉及$\{\lambda,\gamma_1,\gamma_2,\gamma_3,\delta\}$，其中参数$\lambda$用于平衡CRF一元势和成对势的贡献程度，本文在验证集上采用网格搜索，在区间[0,1]上确定表现最优的参数λ，取值为0.3时可以获得最优的rel和rms。$\gamma_s,s\in\{1,2,3\}$等价于在尺度S下相似特征的方差，它表达了相似特征的相似性和接近度。δ是自适应Huber Penalty的常数，首先将$\delta=1.5$作为初始化，在验证集上确定$\delta=1.2$最佳。整个网络在训练集上耗时约18 h，运行一张576×160的红外图像约0.04s（包括前端和后端）。

4 实验结果与分析

采用误差和正确率准则来定量评价本文提出的深度模型的表现，则有：

1）误差准则

平均相对误差（rel）：$\frac{1}{Q}\sum_{i=1}^{Q}\frac{\left|\hat{d}_i-d_i\right|}{d_i}$

均方根误差（rms）：$\sqrt{\frac{1}{Q}\sum_{i=1}^{Q}\left(\hat{d}_i-d_i\right)^2}$

平均 lg 误差（lg10）：$\frac{1}{Q}\sum_{i=1}^{Q}\left|\lg\left(\hat{d}_i\right)-\lg\left(d_i\right)\right|$

2）正确率

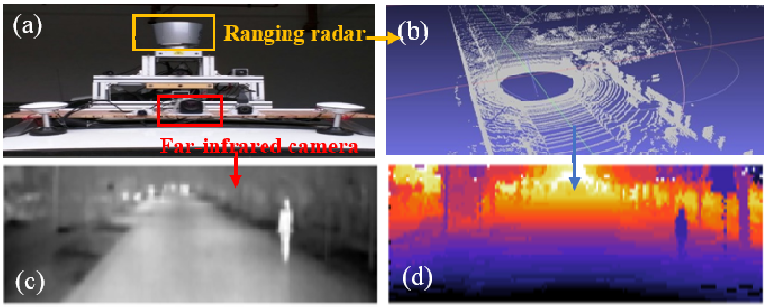
满足 $\max\left(\frac{d_i}{\hat{d}_i},\frac{\hat{d}_i}{d_i}\right)=\delta<t,t\in\left[1.25,1.25^2,1.25^3\right]$ 的 $\hat{d}_i$ 的百分比。

式中：Q 是测试集的所有像素数量的总和；$\hat{d}_i$ 和 d_i 分别表示像素 i 的预测深度和真实深度。

4.1 DCRFN 超参数的选择

参数 S 决定了成对势函数的能量约束原则，它不仅影响模型的复杂程度，而且影响预测深度图的平滑性。本文将参数 S 分别设置为 0、1、2、3、4、5，在评估指标和模型大小方面分析最优参数 S。表 1 展示了不同 S 值的表现，当 S 的值大于或等于 3 时，误差和正确率准则趋于稳定，最后一列表明模型尺寸随着尺度参数 S 的增大呈现上升趋势，为了实现更小的模型尺寸和更好预测性能的目标，本文设置 S=3。

除此之外，DCRFN 的超参数γ_1，γ_2，γ_3表达了在尺度参数 S=3 下相似特征的方差，决定了成对势函数的能量在像素的特征空间邻域的分布。若参数越小，则能量集中在特征空间较小的邻域内，使得模型对图像的边缘更敏感；反之，能量分散在特征空间较大的邻域内，使得模型对预测深度图的平滑性贡献更大。本文首先通过学习曲线缩小γ_1，γ_2，γ_3的搜索范围为[10,20]、[20,30]、[25,35]。实验发现γ_3 在区间[25,35]的变化几乎不影响模型的性能，故取γ_3 为 26。其次以 rel 指标为参考（如图 5），通过网格搜索在γ_1，γ_2，空间内搜索最优参数，其中颜色越深表示 rel 值越小，颜色越浅表示 rel 值越大，当γ_1=13，γ_2=24 时，rel 指标达到极小值，即寻得最优参数。



(a) 车载远红外相机及测距雷达；(b) 雷达散点图；(c) 原始红外图像；(d) 真实深度标签

(a) Far-infrared camera and ranging radar; (b) Radar scatter diagram; (c) Raw infrared image; (d) Ground-truth depth map

图 4 无人车数据采集装置

Fig.4 Unmanned vehicle data acquisition

表 1 定量评价指标在超参数 S 上的比较结果

Table 1 Comparisons results of the different evaluation indexes on hyper-parameter S

Method	Error(Lower is better)			Accuracy(Higher is better)			Model size
	rel	lg10	rms	δ<1.25	δ<1.25^2	δ<1.25^3	
Unary only(S=0)	0.212	0.077	3.210	0.694	0.877	0.917	-
DCRFN (S=1)	0.193	0.074	3.033	0.787	0.919	0.942	68.1MB
DCRFN (S=2)	0.200	0.073	3.060	0.781	0.923	0.947	72.8MB
DCRFN (S=3)	0.184	0.075	2.984	0.798	0.920	0.941	77.6MB
DCRFN (S=4)	0.200	0.074	3.072	0.777	0.925	0.945	82.3MB
DCRFN (S=5)	0.197	0.072	2.952	0.795	0.924	0.946	87.1MB

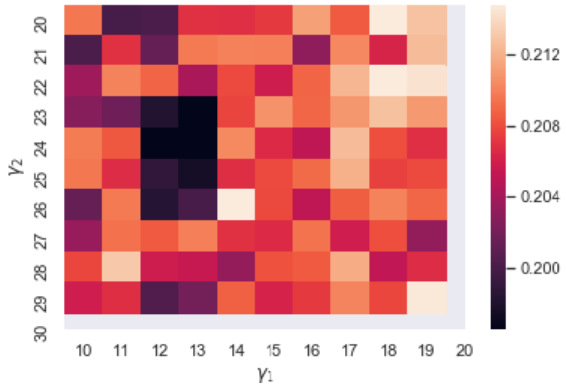


图 5 DCRFN 超参数寻优

Fig.5 Search for the optimal hyper- parameter of DCRFN

4.2 实验结果分析

在给定模型最优超参数的情况下，表 2 列出了本文提出的 DCRFN 与一些经典方法定量比较的结果。可以看出，在 NUSTMS 数据集上，本文方法优于有监督学习的方法。相比较于文献[6]，本文方法在误差和正确率两个评估指标上表现突出，说明通过深度卷积网络获得的一元和成对特征远优于文献[6]中通过手工提取的特征。文献[11]没有考虑邻域关系，这使得它们表现弱于将 CNNs 与 PGM 结合的方法（文献[16,19,27]和本文的 DCRFN）。在基于 PGM 的方法中，本文提出的方法优于其他 3 种方法，这主要由于本文针对红外图像的特点分别设计了两个深度特征模型，而文献[16,19,27]的成对特征来自简单的预设先验。图 6 列出了红外测试集中任意 4 个不同场景下的定性评估预测结果。可以看出，针对可见光设计的特征只能粗略的估计出红外场景的总体深度信息，缺乏了细节特征，如第一列树木的轮廓、第二列较深处的景物、第三列路上的行人等；而本文由于存在针对红外场景设计的深度特征网络，因此可以获得更加丰富的细节信息。

表 2 DCRFN 模型与其他经典方法的深度估计结果对比

Table 2 Comparison of depth estimation results between DCRFN model and other classical methods

Method	Error(Lower is better)			Accuracy(Higher is better)		
	rel	lg10	rms	δ<1.25	δ<1.25^2	δ<1.25^3
Make3D [6]	0.312	0.097	4.124	0.644	0.712	0.813
Eigen et al. [11]	0.230	0.087	3.318	0.716	0.885	0.919
DeepLabV1[16]	0.231	0.093	3.394	0.716	0.883	0.920
Liu-DCNF [19]	0.229	0.081	3.297	0.731	0.898	0.931
Cao et al. [27]	0.225	0.083	3.299	0.732	0.896	0.932
PRN [1]	0.210	0.087	3.252	0.741	0.898	0.932
DCRFN(Ours)	0.184	0.075	2.984	0.789	0.920	0.941

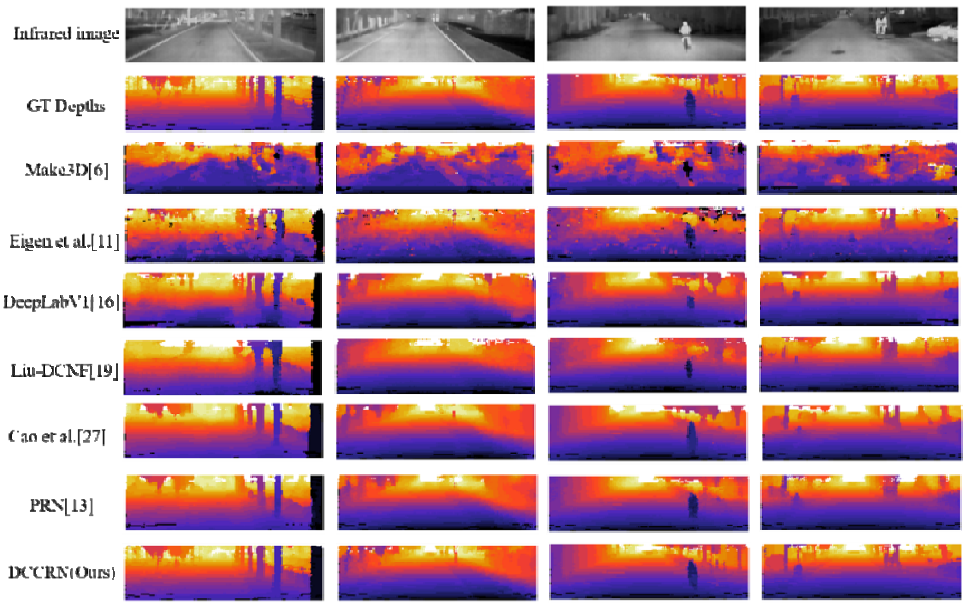


图 6 NUSTMS 数据集下预测深度图实例

Fig.6 Examples of predicted depth maps on NUSTMS dataset

最后，本文分析证明了成对项模型于整个模型的重要性。如图 7 所示，本文分别输出了仅一元模型，仅成对模型和总体输出结果的对比。通过定义一元项和成对项来模拟标签的联合分布，可以看出，仅由一元模型虽可以估计出的整体的深度图，添加了成对模型的约束后，使得预测结果更加平滑。与真实深度图相比，通过本文提出的 DCRFN 预测的深度图获得了更加良好的视觉体验。图 8 在指标 rms 和正确率上定量验证了上述说明。

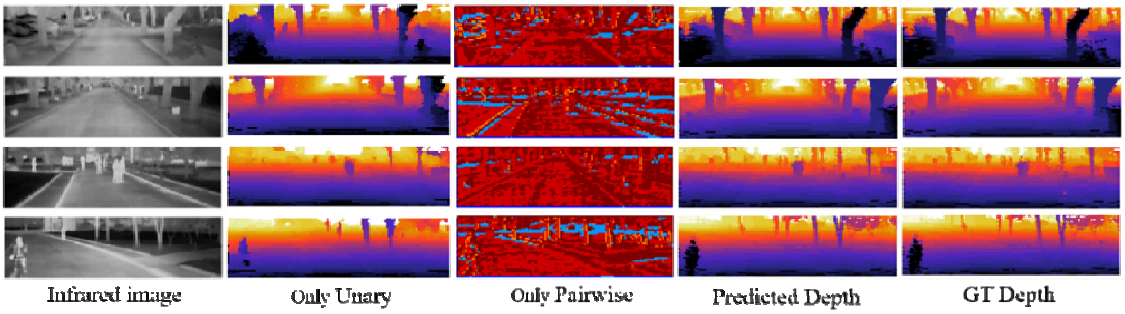


图 7　DCRFN 分解对比结果

Fig.7　The decomposition results of DCRFN

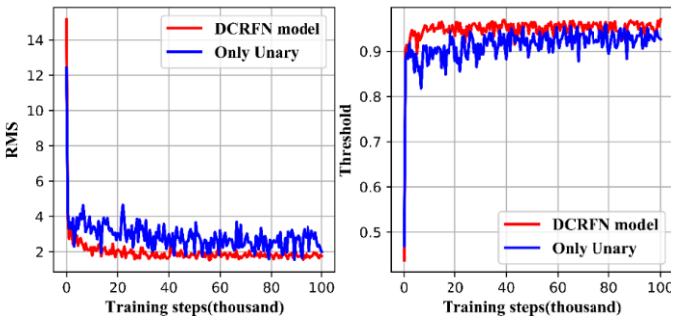


图 8　训练过程中 DCRFN 分解表现对比

Fig.8　The decomposition performance of DCRFN during training

5　结论

本文提出一种新颖的 DCRFN 模型来估计单目红外图像的深度信息，这有助于推进夜间视觉产品的应用。现有的红外图像深度估计方法虽基于 CNNs，由于忽略了优化结构损失，造成预测深度图模糊甚至错误。本文在充分考虑红外图像特点的条件下，将 CNNs 与 CRF 的优势结合在模型中，二者的联合优化提升了模型的泛化能力。值得注意的是，DCRFN 模型无需预先定义成对特征，可以实现全自主学习。同时深度的离散策略，使得有序约束能够融合到 DCRFN 损失函数中，从而获得更好的景物边缘预测。除此之外，DCRFN 不仅建立了原始红外图像和深度图之间的关系，而且还构造了场景不同尺度深度序列之间的关系。最后，实验评估指标证明了本文方法的可行性与准确性。

在未来的研究工作中，将考虑采用精简模型降低网络的参数规模，以便模型能在夜间自动驾驶领域实际应用。除此之外，考虑将红外场景深度估计任务用于如合成任务、目标追踪等夜视应用，从而辅助智能产品的夜间决策。

参考文献：

[1] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]//*IEEE Computer Vision and Pattern Recognition*, 2016: 3213-3223

[2] PENG X, SUN B, Ali K, et al. Learning deep object detectors from 3D models[C]//*IEEE Computer Vision and Pattern Recognition*, 2015: 1278-1286.

[3] Biswas J, Velos oM. Depth camera based indoor mobile robot localization and navigation[C]//*IEEE Robotics and Automation*, 2011: 1697-1702.

[4] Sivaraman S, Trivedi M M. Combining monocular and stereo-vision for real-time vehicle ranging and tracking on multilane highways[C]//*IEEE Intelligent Transportation Systems*, 2011: 1249-1254.

[5] Hedau V, Hoiem D, Forsyth D. Thinking inside the box: using appearance models and context based on room geometry[C]//*European Conference on Computer Vision*, 2010: 224-237.

[6] Saxena A, SUN M, Ng A Y. Make3D: learning 3D scene structure from a single still image[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2009, **31**(5): 824-840.

[7] LIU B, Gould S, Koller D. Single image depth estimation from predicted semantic labels[C]//*IEEE Computer Vision and Pattern Recognition*, 2010: 1253-1260.

[8] Russell B C, Torralba A. Building a database of 3D scenes from user annotations[C]//*IEEE Computer Vision and Pattern Recognition*, 2009: 2711-2718.

[9] LIU M, Salzmann M, HE X. Discrete-continuous depth estimation from a single image[C]//*IEEE Computer Vision and Pattern Recognition*, 2014: 716-723.

[10] Karsch K, LIU C, KANG S B. Depth extraction from video using non-parametric sampling[C]//*European Conference on Computer Vision*, 2012: 775-788.

[11] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C]//*International Conference on Neural Information Processing Systems*, 2014: 2366-2374.

[12] Laina I, Rupprecht C, Belagiannis V, et al. Deeper depth prediction with fully convolutional residual networks[J]. *International Conference on* 3*D Vision*, 2016: 239-248.

[13] 顾婷婷, 赵海涛, 孙韶媛. 基于金字塔型残差神经网络的红外图像深度估计[J]. **红外技术**, 2018, **40**(5): 21-27.
GU T T, ZHAO H T, SUN S Y. Depth estimation of infrared image based on pyramid residual neural networks[J]. *Infrared Technology*, 2018, **40**(5): 21-27.

[14] WU S C, ZHAO H T, SUN S Y. Depth estimation from infrared video using local-feature-flow neural network[J/OL]. *International Journal of Machine Learning and Cybernetics*, 2018: doi.org/10.1007/s13042-018-0891-9.

[15] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//*IEEE Computer Vision and Pattern Recognition*, 2016: 770-778.

[16] CHEN L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs[J]. *Computer Science*, 2015(4): 357-361.

[17] Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with Gaussian edge potentials[J]. *In Advances in Neural Information Processing Systems*, 2012(24): 109-117.

[18] LI N B, SHEN N C, DAI N Y, et al. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs[C]//*IEEE Computer Vision and Pattern Recognition*, 2015: 1119-1127.

[19] LIU F, SHEN C, LIN G. Deep convolutional neural fields for depth estimation from a single image[C]//*IEEE Computer Vision and Pattern Recognition*, 2015: 5162-5170.

[20] LIU F, SHEN C, LIN G, et al. Learning depth from single monocular images using deep convolutional neural fields[C]//*IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015: 5162-5170.

[21] XU D, WANG W, TANG H, et al. Structured attention guided convolutional neural fields for monocular depth estimation[C]//*IEEE Computer Vision and Pattern Recognition*, 2018: 3917-3925.

[22] Ibarra-Castanedo C, González D, Klein M, et al. Infrared image processing and data analysis[J]. *Infrared Physics and Technology*, 2004, **46**(1-2): 75-83.

[23] 张蓓蕾, 孙韶媛, 武江伟. 基于 DRF-MAP 模型的单目图像深度估计的改进算法[J]. **红外技术**, 2009, **31**(12): 712-715.
ZHANG B L, SUN S Y, WU J W. Depth estimation from monocular images based on DRF-MAP model[J]. *Infrared Technology*, 2009, **31**(12): 712-715.

[24] 席林, 孙韶媛, 李琳娜. 基于 SVM 模型的单目红外图像深度估计[J]. **激光与红外**, 2012, **42**(11): 1311-1315.
XI L, SUN S Y, LI L N. Depth estimation from monocular infrared images based on SVM model[J]. *Laser and Infrared*, 2012, **42**(11): 1311-1315.

[25] HUANG G, LIU Z, Laurens V D M, et al. Densely connected convolutional networks[C]//*IEEE Conference on Computer Vision and Pattern Recognition*, 2017: arXiv:1608.06993.

[26] Noh H, HONG S, HAN B. Learning deconvolution network for semantic segmentation[C]//*IEEE International Conference on Computer Vision*, 2015: 1520-1528.

[27] CAO Y, WU Z, SHEN C. Estimating depth from monocular images as classification using deep fully convolutional residual networks[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017(99): 1-1.

[28] FU H, GONG M,WANG C, et al. Deep ordinal regression network for monocular depth estimation[C]// *IEEE Computer Vision and Pattern Recognition*, 2018: 2002-2011.