DDAC：面向卷积神经网络图像隐写分析模型的特征提取方法

王晓丹，李京泰，宋亚飞
（空军工程大学防空反导学院，陕西 西安 710051）

摘 要：针对基于卷积神经网络的图像隐写分析方法中使用人工设计的滤波器在特征提取过程中有效性低的问题，提出方向差分自适应组合（DDAC）特征提取方法。在计算中心像素与周围不同方向像素的差分后，使用 1×1 卷积对方向差分进行线性组合。根据损失对组合参数自适应更新来构建多样化的滤波器，使获取的嵌入信息残差特征更有效。使用截断线性单元提高嵌入信息残差和图像信息残差的比率，加快模型收敛速度并提高残差特征提取能力。实验结果表明，该方法使 Ye-net、Yedroudj-net 模型的准确率在 WOW 和 S-UNIWARD 数据集中提高 1.30%～8.21%。与固定和更新参数 SRM 滤波器方法相比，测试模型在不同隐写数据集中的准确率提高 0.60%～20.72%，并且训练过程更稳定。对比其他图像隐写分析模型，DDAC-net 具有更高的隐写分析效率。

关键词：图像隐写分析；卷积神经网络；特征提取；隐写分析富模型；截断线性单元
中图分类号：TP309
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2022089

DDAC: a feature extraction method for model of image steganalysis based on convolutional neural network

WANG Xiaodan, LI Jingtai, SONG Yafei
Air and Missile Defense College, Air Force Engineering University, Xi’an 710051, China

Abstract：To solve the problem that for image steganalysis based on convolution neural network, manual designed filter kernels were used to extract residual characteristics, but in practice, these kernels filter were not suitable for each steganography algorithm and have worse performance in application, a directional difference adaptive combination (DDAC) method was proposed. Firstly, the difference was calculated between center pixel and each directional pixel around, and 1×1 convolution was adopted to achieve linear combinations of directional difference. Since the combination parameters self-adaptively update according to loss function, filter kernels could be more effective in extracting diverse residual characteristics of embedding information. Secondly, truncated linear unit (TLU) was applied to raise the ratio of embedding information residual to image information residual. The model’s coveragence was accelerated and the ability of feature extraction was promoted. Experimental results indicate that substituting the proposed method could improve the accuracy of Ye-net and Yedroudj-net by 1.30%~8.21% in WOW and S-UNIWARD datasets. Compared with fix and adjustable SRM filter kernels methods, the accuracy of test model using DDAC increases 0.60%~20.72% in various datasets, and the training progress was more stable. DDAC-net was proved to be more effective in comparsion with other steganalysis model.

Keywords：image steganalysis, convolution neutral network, feature extraction, rich model of steganalysis, truncated linear unit

收稿日期：2021−12−23；**修回日期：**2022−03−23
基金项目：国家自然科学基金资助项目（No.61876189）
Foundation Item: The National Natural Science Foundation of China (No. 61876189)

0　引言

图像隐写是一种将信息隐藏在图像中进行隐秘传输的方法。由于使用常见的图像作为载体，攻击者难以发现隐藏的信息，从而保护了信道安全[1]。然而这项技术可能被用于恶意意图，例如将恶意代码隐写在移动应用的图像文件中以窃取私人信息等。为防止此类事件发生，需要使用各种方法检测图像中是否被嵌入隐藏信息，这类技术被称作隐写分析技术[2-3]。

近年来，深度学习已经在图像、视频、语音等各个领域发挥重要作用[4]。各类神经网络结构层出不穷[5-9]。一方面，基于生成对抗网络（GAN, generative adversarial network）[9]的图像隐写算法大幅提高了隐写容量[10-11]。另一方面，基于卷积神经网络（CNN, convolutional neural network）[12]的图像隐写分析方法快速发展。2014 年，Tan 等[13]首次将卷积神经网络用于图像隐写分析领域。随着研究者对该方法的理解不断加深，基于卷积神经网络的隐写分析方法已经全面超越传统算法并成为主流方案。

基于卷积神经网络的隐写分析算法主要包括3个部分：特征提取部分、特征融合部分和分类部分，其中特征提取部分的设计尤为重要，会极大地影响模型的收敛效果。一种方式是选择使用固定参数的滤波器。Qian-net[14]和 Xu-net[15]通过在网络的第一层使用隐写分析富模型（SRM, rich model of steganalysis）中的一个 5×5 滤波器来提取中心像素与周围像素的差异。Xu-net 使用绝对值层来使模型学习负向残差信息，并且使用全局池化来提取特征[15]。Yedroudj 等[16]使用 SRM[18]中全部的 30 个滤波器作为网络的第一层，并且在训练过程中不更新参数，结合绝对值层[15]和截断线性单元（TLU, truncated linear unit）[17]，极大地提高了模型性能。2018 年，Li 等[19]提出 ReST-net 结构，希望通过 3 种不同的子网获取不同特征。在预处理层初始化中，Subnet#1 使用 16 种不同参数的 6×6 Gabor 滤波器；Subnet#2 使用 16 种不同的 SRM 滤波器；Subnet#3 则先采用 SRM 滤波进行线性处理，再通过不同角度旋转后的 SRM 滤波器进行非线性处理，最后输出 14 个非线性特征图。

另一种方式是进行参数初始化后，使滤波器参数随模型训练自适应调整，获得更有效的滤波器。2014 年，Tan 等[13]设计了 3 种滤波器初始化方式，其中随机初始化滤波器得到的结果最差，而采用现有滤波器进行初始化得到的效果最好。Ye 等[17]提出对第一层卷积使用 SRM 滤波器参数初始化，并使滤波器参数参与神经网络训练，使用 TLU 激活函数替换 Tanh 激活函数[20]，使模型学习残差信息，并得到更有效的高通滤波器，提高隐写分析模型的收敛速度。HE 等[6]提出 SRnet，通过残差结构将浅层的特征传递至深层，缓解梯度消失的现象，自适应地学习滤波器的参数[21]。Zhang 等[23]提出 Zhu-net，通过结合 Inception[22]、金字塔池化[8]等多种结构，提出使用 SRM 中 3×3 的滤波器和部分 5×5 的滤波器初始化第一层卷积核的参数，并使参数随模型训练更新，相比于其他网络获得了更高的准确率。

为提高基于卷积神经网络的图像隐写分析模型中残差特征滤波器的多样性，本文提出方向差分自适应组合（DDAC, directional difference adaptive combination）特征提取方法。通过 1×1 卷积将中心像素与相邻像素的差分信息进行组合，使用带动量的随机梯度下降算法在模型训练过程中优化组合参数，使用 TLU 激活函数限制残差特征范围，降低图像信息的干扰，提高模型收敛速度。DDAC 方法在不同数据集中生成多样化的残差特征滤波器，从而更有效地提取嵌入信息的残差特征，提高模型检测准确率。实验分析了特征提取过程中不同结构、特征通道数量和截止函数的参数 T 对模型准确率的影响，并对比固定 SRM 和更新参数 SRM 这 2 种适用于卷积神经网络的特征提取方法。通过在现有模型中应用 DDAC 特征提取方法以及和现有模型的性能对比验证 DDAC 方法的有效性。

1　残差特征提取方法

1.1　残差特征提取模型

图像隐写分析主要通过建立残差模型，从图像中提取残差特征，利用统计模型强化特征并结合机器学习方法实现分类。对于一张长为 m 、宽为 n 的图像 $\boldsymbol{X}=\{X_{i,j}\}\in \mathrm{R}^{m\times n}$，其中像素值 $X_{i,j}$ 的取值范围为 0~255。嵌入信息为任意的多媒体信息，在嵌入过程中转化为比特数据流隐写在图像载体中。隐写信息为与图像 $\boldsymbol{X}$ 同样大小的矩阵 $\boldsymbol{I}=\{I_{i,j}\}\in \mathrm{R}^{m\times n}$，并且 $I_{i,j}=\{0,-1,+1\}$。通过对比原图，分析像素值改动的位置来提取隐写信息的编码。空域自适应隐写算

法通过设计失真函数，并使计算得到的失真值最小化来选择图像中的隐写位置[3] $\boldsymbol{P}=\{(i_0,j_0),(i_1,j_1),\cdots,(i_b,j_b)\}$，其中 b 为一张图像中隐写信息比特流的长度。隐写后的图像为 $\boldsymbol{S}=\{S_{i,j}\}\in\mathrm{R}^{m\times n}$，$\boldsymbol{S}=\boldsymbol{X}+\boldsymbol{I}$。对于图像中的一块 3×3 区域，隐写图像的组成为

$$\begin{bmatrix} S_{i-1,j-1} & S_{i,j-1} & S_{i+1,j-1} \\ S_{i-1,j} & S_{i,j} & S_{i+1,j} \\ S_{i-1,j+1} & S_{i,j+1} & S_{i+1,j+1} \end{bmatrix}= \begin{bmatrix} X_{i-1,j-1} & X_{i,j-1} & X_{i+1,j-1} \\ X_{i-1,j} & X_{i,j} & X_{i+1,j} \\ X_{i-1,j+1} & X_{i,j+1} & X_{i+1,j+1} \end{bmatrix}+\begin{bmatrix} I_{i-1,j-1} & I_{i,j-1} & I_{i+1,j-1} \\ I_{i-1,j} & I_{i,j} & I_{i+1,j} \\ I_{i-1,j+1} & I_{i,j+1} & I_{i+1,j+1} \end{bmatrix} \tag{1}$$

为使图像失真最小，嵌入的信息为+1 或−1，并且使嵌入信息的概率均值为 0，相当于人工添加的低频噪声。由于隐写信息与图像信息的比例极小，为提取隐写特征，模型通常要对图像进行预处理，降低图像信息的影响。Fridrich 等[16]提出隐写分析富模型 SRM，通过构建周围像素与中心像素之间的关系来获取隐写图像的残差，对残差进行统计获得隐写特征。残差 R_{res} 的计算方法为

$$R_{\mathrm{res}}=\hat{S}_{i,j}-cS_{i,j} \tag{2}$$

其中，$\hat{S}_{i,j}$ 为像素 $S_{i,j}$ 的 c 阶预测值，由像素 $S_{i,j}$ 的周围像素得到。SRM 中构建的关系有像素相减邻接模型（SPAM, subtractive pixel adjacency model）[24] 中的简单关系和其他复杂关系，分别为

$$\begin{aligned} R_{\mathrm{res}}&=S_{i,j+1}-S_{i,j}\\ R_{\mathrm{res}}&=(2S_{i,j-1}-S_{i-1,j-1}+2S_{i,j-1}-S_{i+1,j-1}+2S_{i+1,j})-4S_{i,j} \end{aligned} \tag{3}$$

minmax 残差通过选取 2 种像素关系中的较小值或较大值来引入非线性关系，期望增加特征的多样性。例如水平和垂直方向上的残差最小值为

$$R_{\mathrm{res}}=\min\{(S_{i,j-1}+S_{i,j+1}-2S_{i,j}),(S_{i-1,j}+S_{i+1,j}-2S_{i,j})\} \tag{4}$$

模型通过截断操作来限制残差的范围，方便使用共生矩阵来描述图像的残差特征；通过量化操作来提高嵌入信息对残差的影响，使隐写图像和载体图像的残差特征具有显著差异，即

$$R_{\mathrm{res}}=\mathrm{trunc}_T\left(\mathrm{round}\left(\frac{R_{\mathrm{res}}}{q}\right)\right) \tag{5}$$

其中，q 为量化的步长，当 $c>1$ 时 $q\in\{c,1.5c,2c\}$，当 $c=1$ 时 $q\in\{1,2\}$；round(·) 为取整操作；$\mathrm{trunc}_T(\cdot)$ 为截断操作。对提取得到的残差用共生矩阵进行统计，对统计结果进行整合获得残差特征，并用集成分类器等传统机器学习方法对样本进行分类，确定是否为隐写图像。

1.2 DDAC 特征提取方法

在载体图像中的一个像素区域 $\boldsymbol{X}'\in\mathrm{R}^{t\times t}$，$t\in\{t\mid t=2k+1,k\in Z\}$ 内，若中心像素为 $X_{i,j}$，则周围像素构成的向量为 $\boldsymbol{N}_{i,j}^{X'}$，满足条件 $\boldsymbol{N}_{i,j}^{X'}\subset\boldsymbol{X}'$，$X_{i,j}\notin\boldsymbol{N}_{i,j}^{X'}$。假设存在线性函数 h 可以根据周围像素得到中心像素的 c 阶值 $cX_{i,j}$，即

$$h(\boldsymbol{N}_{i,j}^{X'})=cX_{i,j} \tag{6}$$

载体图像区域 $\boldsymbol{X}'$ 中嵌入隐写信息 $\boldsymbol{I}'$ 后变为隐写图像区域 $\boldsymbol{S}'$，隐写区域内中心像素 $S_{i,j}$ 的残差 $\mathrm{res}\in\mathrm{R}$ 由中心像素的周围像素 $\boldsymbol{N}_{i,j}^{S'}$ 与 c 阶的中心像素 $cS_{i,j}$ 差分得到，即

$$R_{\mathrm{res}}=R(\boldsymbol{N}_{i,j}^{S'},S_{i,j})=h(\boldsymbol{N}_{i,j}^{S'})-cS_{i,j} \tag{7}$$

由式(1)可知，式(7)中隐写图像 $\boldsymbol{N}_{i,j}^{S'},S_{i,j}$ 的组成为

$$\begin{aligned} \boldsymbol{N}_{i,j}^{S'}&=\boldsymbol{N}_{i,j}^{X'}+\boldsymbol{N}_{i,j}^{I'}\\ S_{i,j}&=X_{i,j}+I_{i,j} \end{aligned} \tag{8}$$

根据函数 h 的线性性质得到

$$\begin{aligned} R_{\mathrm{res}}&=h(\boldsymbol{N}_{i,j}^{X'}+\boldsymbol{N}_{i,j}^{I'})-c(X_{i,j}+I_{i,j})=\\ &\quad h(\boldsymbol{N}_{i,j}^{I'})-cI_{i,j}+h(\boldsymbol{N}_{i,j}^{X'})-cX_{i,j} \end{aligned} \tag{9}$$

根据式(6)，最终残差值为

$$R_{\mathrm{res}}=h(\boldsymbol{N}_{i,j}^{I'})-cI_{i,j} \tag{10}$$

经过差分，载体图像中的像素获得的残差特征为零，而隐写图像中像素的残差特征则由隐写信息组成。残差特征对于隐写信息敏感，从而降低图像信息的影响。通过提取残差特征，增大载体图像与隐写图像的特征差异，从而更好地对样本进行分类。

在 SRM 方法的启发下，基于卷积神经网络的图像隐写分析算法，例如 Ye-net[17]、Yedroudj-net[18] 参考 SRM 的特征计算方法设计了各种高通滤波器提取图像残差。在滤波器的设计方式中，一种是直接使用 SRM 中的高通滤波器提取残差特征，另一种是先使用 SRM 滤波器的参数对卷积核进行初始化，

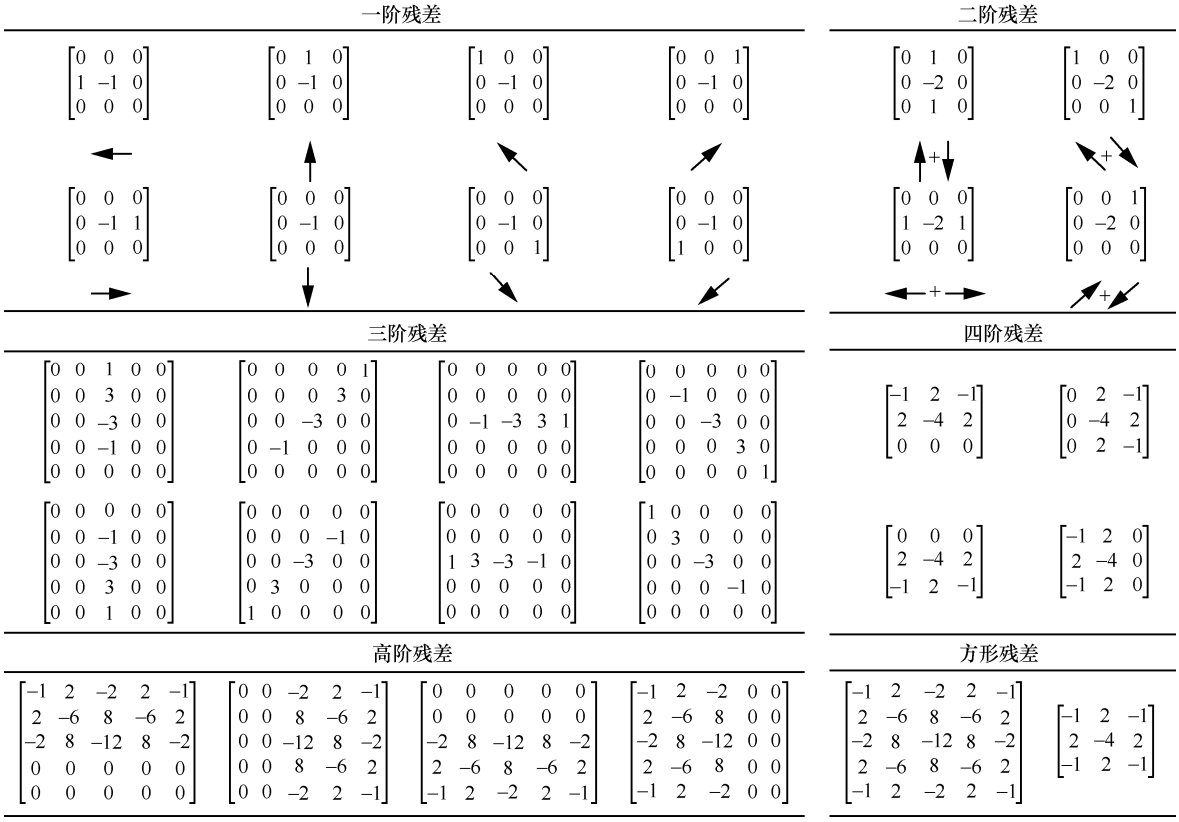


图 1　根据 SRM 方法设计的 30 个高通滤波器

然后在网络训练过程中优化滤波器参数。如果仅依赖训练获得滤波器则需要大量的训练轮次和时间[21]。根据 SRM 方法设计的高通滤波器共有 30 个，如图 1 所示。

图像卷积过程本质上是对像素进行加权求和。残差滤波器从图像的左上角开始计算得到第一个残差，然后根据步长逐行逐列地在图像上滑动，通过将像素值与对应的权重相乘求和获得残差，残差按中心像素的位置依次排列获得残差图。残差图的生成过程如图 2 所示。

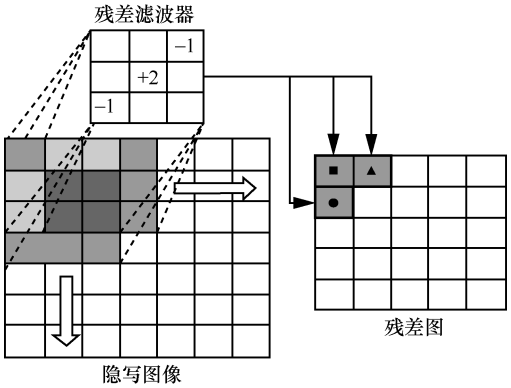


图 2　残差图的生成过程

一个大小为 $r\times r$ 的残差滤波器在一张大小为 $m\times n$ 的图像中进行图像卷积，无填充且步长为 $s\times d$ 时，生成特征图 R 的大小为

$$R_{\text{width}}=\frac{m-r}{s}+1$$
$$R_{\text{length}}=\frac{n-r}{d}+1 \tag{11}$$

在中心像素的周围像素中，相邻像素与中心像素相关性最高，因此使用相邻像素来构建残差提取函数。有 K 个大小为 3×3 的卷积核 $\boldsymbol{W}_k\in \mathrm{R}^{3\times 3}$，每个卷积核表示相邻像素与中心像素的线性关系。卷积核 $\boldsymbol{W}_k$ 对隐写图像的一块区域 $\boldsymbol{S}'\in \mathrm{R}^{3\times 3}$ 进行图像卷积的结果为

$$R_{\text{res}}=\boldsymbol{S}'*\boldsymbol{W}_k=\begin{bmatrix} S_{i,j}^{\nwarrow} & S_{i,j}^{\uparrow} & S_{i,j}^{\nearrow} \\ S_{i,j}^{\leftarrow} & S_{i,j} & S_{i,j}^{\rightarrow} \\ S_{i,j}^{\swarrow} & S_{i,j}^{\downarrow} & S_{i,j}^{\searrow} \end{bmatrix} * \begin{bmatrix} w_{\nwarrow} & w_{\uparrow} & w_{\nearrow} \\ w_{\leftarrow} & w_{i,j} & w_{\rightarrow} \\ w_{\swarrow} & w_{\downarrow} & w_{\searrow} \end{bmatrix} =$$
$$\sum_p w_p S_{i,j}^p + w_{i,j} S_{i,j} \tag{12}$$

其中，$p \in \{\leftarrow, \rightarrow, \uparrow, \downarrow, \nearrow, \swarrow, \nwarrow, \searrow\}$ 表示相对中心参数 $w_{i,j}$ 或像素 $S_{i,j}$ 的位置。卷积的计算过程与卷积核的参数表示相邻像素与中心像素的线性关系，$h(\boldsymbol{N}_{i,j}^{X}) = \sum_{p} w_p S_{i,j}^{p}, c = -w_{i,j}$，通过设计不同的卷积核构建多种关系来获取多样化的残差特征。为防止引入图像信息，使 $c = \sum_{p} w_p$ 消除多余的 $S_{i,j}$ 分量，对特征公式进行变换，即

$$R_{\text{res}} = \sum_{p} w_p (S_{i,j}^{p} - S_{i,j}) \tag{13}$$

其中，$S_{i,j}^{p} - S_{i,j}$ 为方向差分。残差的计算可以转化为对方向差分的线性组合，线性组合的参数 w_p 使用带动量的随机梯度下降算法[25]根据损失函数的值进行优化。参数的更新通过引入动量实现，在 pytorch 深度学习框架中的计算过程为

$$w_p{}' = w_p - v l_{\text{r}} \tag{14}$$

其中，w_p 为待优化参数；$w_p{}'$ 为优化后结果；l_{r} 为学习率，控制参数更新的快慢；v 为动量，计算方法为

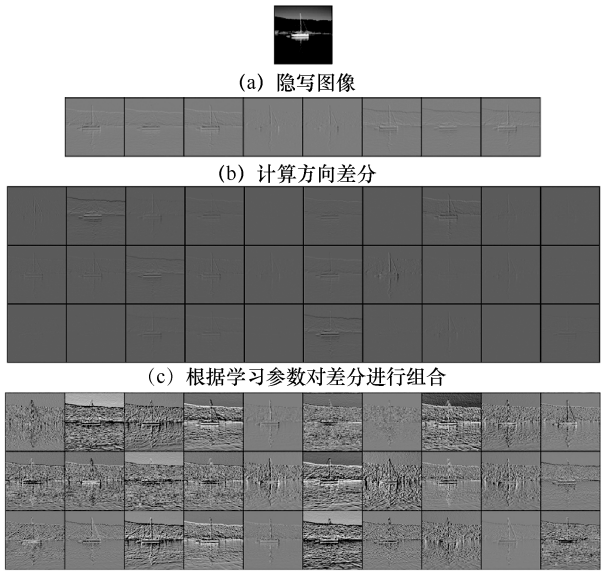
$$v = \alpha v' + \frac{\partial L_{\text{loss}}}{\partial w_p} \tag{15}$$

其中，L_{loss} 为当前参数下模型的损失，损失函数使用交叉熵损失函数；v' 与 v 分别为上一次参数更新时的动量与当前动量，通过引入参数在上一轮的更新值来提高模型的训练速度，使模型在优化过程中更容易越过鞍点，防止陷入局部最小值；α 为动量系数，控制上一轮参数更新值对本次更新的影响程度。学习率和动量系数在模型训练前设置。Sutskever 等[25]指出动量系数的设置会影响模型的优化过程，一般设置动量系数为 $\{0.9, 0.98, 0.995\}$ 中的数值；学习率则根据模型训练过程中损失的下降效果来确定。在 Ye-net 与 Yedroudj-net 中均使用在模型训练的不同阶段阶梯降低学习率的策略，通过降低模型训练后期参数更新的幅度，使模型的损失平稳地达到最小值。根据试错法对比不同学习率和动量的模型效果，在动量为 0.9、学习率为 0.1 并且每 50 轮训练学习率乘以 0.2 的参数设置条件下模型性能较高。

单一残差特征提取函数不能作用于所有像素，函数获取的特征中会掺杂其他特征。由式(10)可知，残差特征值由 0、−1、+1 的隐写信息组成，为使模型学习较小的隐写信息残差特征，降低其他特征影响，使用 TLU 激活函数对残差特征的大小进行限制[17]。激活函数的形式为

$$\text{TLU}(x) = \begin{cases} -T, & x < -T \\ x, & -T \leqslant x \leqslant T \\ T, & x > T \end{cases} \tag{16}$$

为使残差特征提取过程与卷积神经网络模型训练过程结合，首先获取中心像素与 8 个邻接像素的方向差分，分别产生 8 个特征图，然后使用不带偏置的 1×1 卷积将输入的 8 张特征图中同一位置的方向差分进行线性组合。一张使用 WOW 算法在嵌入率为 0.4 的条件下生成的隐写图像，残差在 DDAC 特征提取方法不同阶段的提取结果如图 3 所示。



(a) 隐写图像

(b) 计算方向差分

(c) 根据学习参数对差分进行组合

(d) 使用TLU激活函数降低图像信息

图 3 残差在 DDAC 特征提取方法不同阶段的提取结果

一种残差提取函数只能提取一幅图像中满足线性关系的部分像素的残差，为尽可能地提取残差特征，需要设计更多的残差滤波器，在模型训练过程中自适应地调节组合参数获取残差提取函数，使残差滤波器更具多样性和适应性。

本文在嵌入率为 0.4 的条件下用 WOW 算法和 S-UNIWARD 算法产生 2 个数据集，使用这 2 个数据集训练各得到 30 组参数，使用皮尔逊相关性系数来判断不同数据集产生的残差提取函数的相似性。对于两组变量 $\boldsymbol{U}, \boldsymbol{V} \in \mathrm{R}^{n}$，皮尔逊相关系数计算方式为

$$\rho_{U,V}=\frac{\operatorname{cov}(\boldsymbol{U},\boldsymbol{V})}{\sigma_U\sigma_V}=\frac{\sum_{i=1}^{n}(U_i-\bar{U})(V_i-\bar{V})}{\sqrt{\sum_{i=1}^{n}(U_i-\bar{U})^2}\sqrt{\sum_{i=1}^{n}(V_i-\bar{V})^2}} \tag{17}$$

其中，$\operatorname{cov}(\boldsymbol{U},\boldsymbol{V})$ 为变量 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的协方差，$\sigma_U\sigma_V$ 为 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 标准差的积。计算在嵌入率为 0.4 的条件下不同算法的数据集训练得到的两组组合系数的皮尔逊相关系数，结果如图 4 所示。

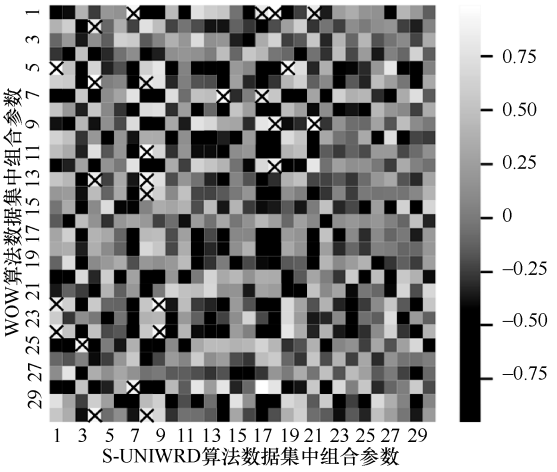


图 4　2 个数据集中残差提取函数的相关性

当皮尔逊系数值大于 0.8 时，说明两组变量具有极强相关性，可以认为两组变量相同。如图 4 所示，皮尔逊系数大于 0.8 的点用×标记，表示此点的横纵坐标对应序号的组合参数相同。在嵌入率为 0.4 条件下，S-UNIWARD 算法生成的数据集训练得到的 30 组组合参数中有 11 组与 WOW 算法生成的数据集训练得到的组合参数相同，另外产生了 19 组不同的组合参数。在残差提取函数学习过程中不仅获得了类似人工设计的方形残差滤波器，在不同数据集中还获得了多样化的组合参数，这些组合参数更适用于当前算法。使用 WOW 算法生成数据集中的组合参数在 S-UNIWARD 算法产生的数据集中验证模型效果，模型的准确率将会下降。S-UNIWARD 数据集中训练得到的残差滤波器如图 5 所示。

使用 DDAC 方法训练残差提取函数的参数更加稳定，可以提高模型拟合速度。如果通过直接对滤波器参数进行更新，由于缺乏限制，滤波器在建立过程中容易受图像信息干扰，导致模型训练过程不稳定。使用梯度优化算法对参数进行更新时要计算损失函数 Loss 相对参数 w_p 的偏导数 $\frac{\partial L_{\text{loss}}}{\partial w_p}$。若直接对滤波器的参数进行更新，已知第 k 个残差图 $\boldsymbol{R}_k=\boldsymbol{S}*\boldsymbol{W}_k$，根据链式法则和残差公式，损失对滤波器权值矩阵的偏导数为

$$\frac{\partial L_{\text{loss}}}{\partial \boldsymbol{W}_k}=\boldsymbol{S}*\frac{\partial L_{\text{loss}}}{\partial \boldsymbol{R}_k} \tag{18}$$

滤波器的大小为 3，$\frac{\partial L_{\text{loss}}}{\partial \boldsymbol{R}_k}=\boldsymbol{\delta}^k$，则滤波器 $\boldsymbol{W}_k$ 中参数 $w_{a,b}^k(a,b\in\{1,2,3\})$ 的偏导数为

$$\nabla w_{a,b}^k=\frac{\partial L_{\text{loss}}}{\partial w_{a,b}^k}=\sum_i\sum_j \delta_{i,j}^k S_{i+a-1,j+b-1} \tag{19}$$

从式(19)中可知，$\nabla w_{a,b}^k$ 与隐写图像的像素值呈正相关。在残差特征图中被激活的像素能够传递梯度用以更新滤波器参数，传递的梯度等于像素在残差特征图获得的梯度乘以像素值，从而放大像素值。DDAC 方法中残差图由图像的方向差分组合得到，即

$$\boldsymbol{R}_k=\sum_p\left(\boldsymbol{S}^p-\boldsymbol{S}\right)w_p^k \tag{20}$$

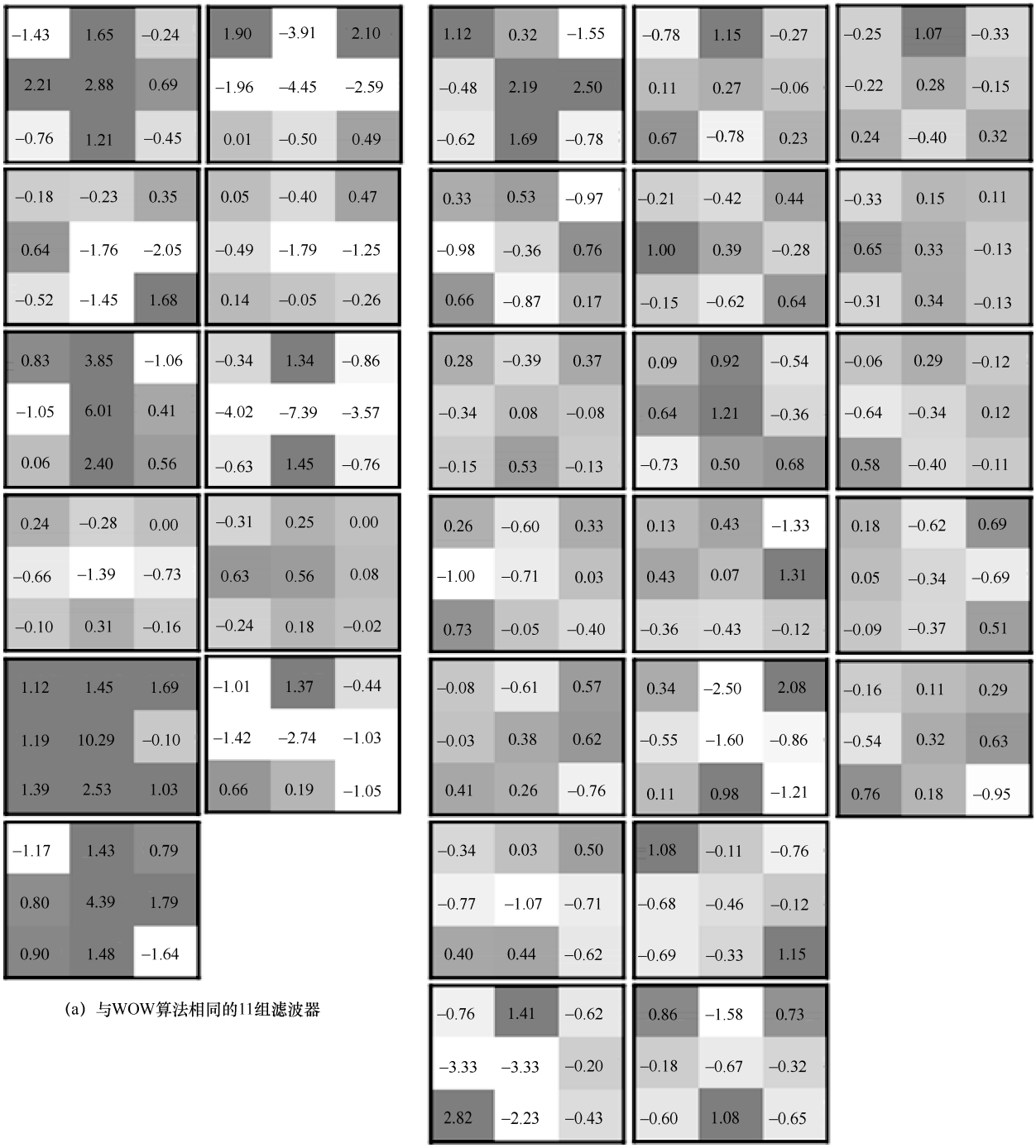
在对组合参数进行更新时，∇w_p^k 的计算方法为

$$\nabla w_p^k=\frac{\partial L_{\text{loss}}}{\partial w_p^k}=\frac{\partial L_{\text{loss}}}{\partial \boldsymbol{R}_k}\frac{\partial \boldsymbol{R}_k}{\partial w_p^k}=\sum_i\sum_j \delta_{i,j}^k\left(S_{i,j}^p-S_{i,j}\right) \tag{21}$$

DDAC 方法更新参数时与方向差分呈正相关。图像的像素值随图像内容变化，在不同的图像中参数的更新值变化较大。根据方向差分对组合参数优化，一方面可以从差分中提取固定模式，降低在不同图像中梯度的变化程度；另一方面图像经过方向差分后的像素值远远小于原图像的像素值，参数的更新幅度小，所以可以提高训练过程的稳定性。本文在 2.6 节中对 2 种方法进行比较，并对结论进行验证。

2　实验与分析

本节主要通过实验对比与分析论证 DDAC 方法的合理性和有效性。在构建测试模型后，通过替换不同结构的特征提取层，改变特征提取层中 1×1 卷积的通道数量，改变 TLU 激活函数中参数 T 的取值，分析特征提取结构的合理性。通过对比固定 SRM 滤波器、可变参数 SRM 滤波器和 DDAC 方法应用在测试模型的性能，以及分析 DDAC 方法应用在 Ye-net 与 Yedroudj-net 模型的效果来验证方法的有效性。最后在模型准确率、模型大小和模型时间



(a) 与WOW算法相同的11组滤波器

(b) 与WOW算法不同的19组滤波器

图 5　S-UNIWARD 数据集中训练得到的残差滤波器

复杂度上对比其他隐写分析模型。

2.1　测试模型结构

测试模型由 5 个卷积模块堆叠而成。每个卷积模块由卷积核大小为 3×3、步长为 1、填充为 0 的卷积层、pytorch 中参数默认的批归一化（BN, batch normalization）层[26]和修正线性单元（ReLU, rectified linear unit）[27]组成。首先，在卷积模块 3 和卷积模块 4 后使用平均池化，核大小为 3、步长为 2。然后，在卷积模块 5 后加入全局平均池化，将 128 个通道的特征缩减为 128 维向量。最后，经过三层全连接层进行分类，每 2 个全连接层间使用 ReLU 激活函数进行激活。测试模型是较简单的卷积神经网络结构，可以更好地对比特征提取方法对卷积神经网络的适用性。测试模型的结构如图 6 所示。

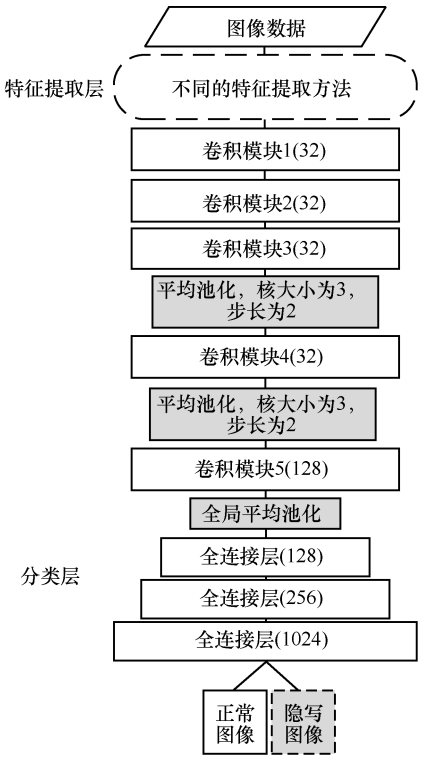


图 6　测试模型的结构

2.2　实验环境及评价指标

实验环境为 Ubuntu18，显卡型号为 RTX3080TI，CPU 型号为 AMD Ryzen 3600，使用数据集为 BOSSbase-v1.01，该数据集包含 10 000 张用不同相机拍摄的 512 像素×512 像素的灰度图片[28]。由于计算量的限制，使用 Pillow 库中 image 类的 resize 方法将 512 像素×512 像素的图片改变为 256 像素×256 像素，并且随机划分 4 000、1 000、5 000 张图像分别用作训练集、验证集和测试集，使用 WOW 和 S-UNIWARD 算法在嵌入率为 0.2 和 0.4 的条件下进行隐写，将载体图像和对应隐写图像作为一组样本，训练集、验证集、测试集中实际有 8 000、2 000、10 000 张图像。实验数据集处理方式和数据集划分方式与 Ye-net、Yedroudj-net、SRnet 保持一致以进行对比。

在分类问题中，混淆矩阵可以用来表示样本的分类结果，混淆矩阵为

$$\begin{pmatrix} \mathrm{TP} & \mathrm{FN} \\ \mathrm{FP} & \mathrm{TN} \end{pmatrix} \tag{22}$$

其中，TP（true positive）表示预测为隐写图像实际为隐写图像的样本数量，FP（false positive）表示预测为隐写图像实际为载体图像的样本数量，FN（false negative）表示预测为载体图像实际为隐写图像的样本数量，TN（true negative）表示预测为载体图像实际为载体的样本数量。

由于测试集中隐写图像和载体图像的数量相同，不存在不平衡情况，准确率可以有效表示模型的分类性能。准确率的计算方法为

$$\text{准确率} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{23}$$

另外，漏检率和虚警率也是重要的评价指标之一。漏检率表示所有隐写图像被错误识别为载体图像的比例，漏检率越低则模型检测隐写图像的能力越强。虚警率表示所有被模型识别为隐写图像的图像中载体图像的比例，虚警率越低则模型的可信度越高，两者的计算方法为

$$\begin{aligned} \text{漏检率} &= \frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}} \\ \text{虚警率} &= \frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}} \end{aligned} \tag{24}$$

2.3　不同结构的性能对比

为探究通过方向差分获得残差滤波器的组合方式并验证其合理性，设计 6 种不同的结构进行对比实验。初始结构的特征提取层由 30 个从 SRM 中提取的残差滤波器构成，参数不更新，作为实验的对照对象。6 种结构示意如图 7 所示。

结构 1 由两层 32 通道的 1×1 卷积构成，卷积层之间没有激活函数。结构 2 与结构 1 相比，在 1×1 的卷积层之间加入 TLU 激活函数用来获取非线性残差特征。结构 3 使用深度可分离卷积，使用分组卷积来结合 3×3 范围内的残差特征，验证前期在构建残差特征过程中获取残差特征间关系的必要性。结构 4 在结构 3 的卷积层间加入 TLU 激活函数，同样用于获取非线性特征。结构 5 在获得方向差分后使用传统 3×3 卷积核进行卷积，与初始预处理方式对比仅使用一阶残差特征的实验效果，同时对比结构 1 来验证使用 1×1 卷积对方向差分进行组合的有效性。结构 6 在结构 5 的基础上在卷积核间加入了 TLU 激活函数。

以上 6 种结构在最后一层后都使用 TLU 激活函数对输出的残差特征进行限制，可以加快模型的拟合速度。

在训练过程中采用随机梯度下降算法进行参数优化，动量设置为 0.9，学习率初始化为 0.1，并

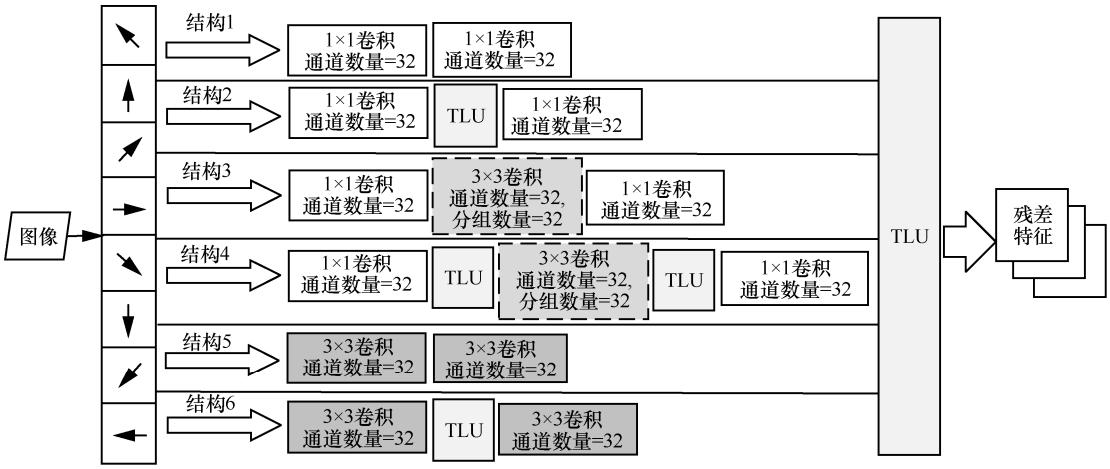


图 7　6 种结构示意

且在训练轮次为 100、150 时学习率乘以 0.2，总训练轮次为 200 次。batch_size 设置为 16，实际训练时每个 batch 中将包含 8 张原图像和相对应的 8 张隐写图像。实验将保存在验证集中效果最好的模型的参数，并在测试集中进行测试，不同结构的模型准确率如表 1 所示。

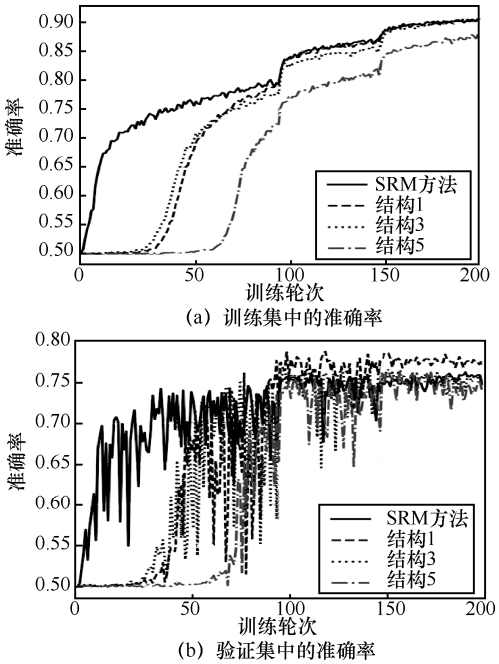
表 1　不同结构的模型在测试集的准确率

结构	WOW		S-UNIWARD	
	嵌入率为 0.2	嵌入率为 0.4	嵌入率为 0.2	嵌入率为 0.4
SRM 方法	0.738 3	0.844 2	0.691 3	0.803 1
结构 1	0.775 9	0.870 3	0.718 3	0.845 2
结构 2	0.761 5	0.845 1	0.708 3	0.835 4
结构 3	0.756 9	0.857 7	0.689 2	0.823 4
结构 4	0.730 8	0.832 0	0.695 1	0.820 7
结构 5	0.732 7	0.846 1	0.685 0	0.810 4
结构 6	0.730 5	0.831 2	0.689 7	0.808 9

实验结果表明，直接将方向差分进行线性组合的结构 1 获得的实验效果最好。结构 1 在 WOW 算法在嵌入率为 0.4 的测试集准确率为 87.03%，超过其他结构 1.26～3.92%。结构 1 引入激活函数后在 4 个测试集的准确率平均降低 1.48%，引入非线性关系对特征提取的效果提升不大，由于对训练集过拟合而导致性能降低。

结构 5、结构 6 相比 SRM 方法，模型在 WOW 嵌入率为 0.2 测试集的准确率降低 0.56%和 0.78%，S-UNIWARD 嵌入率为 0.2 测试集中降低 0.63%和 0.16%，在 WOW 嵌入率为 0.4 测试集和 S-UNIWARD 嵌入率为 0.4 测试集中，最多提高 0.73%，最多降低 1.30%。仅使用方向差分比使用全部 30 个残差滤波器在大部分数据集中的特征提取能力弱，但模型性能的损失较小，说明方向差分与其他滤波器相比在浅层特征提取中起关键作用。而使用线性组合的结构 1 在测试集的准确率相比 SRM 方法提高 2.62%～4.21%，说明对方向差分进行线性组合可以有效提高残差特征的提取和表达能力。

结构 1 比结构 3 和结构 4 在测试集中的准确率平均高 2.67%，说明过早地对 3×3 范围内的像素残差进行融合会影响残差特征的表达并最终影响模型的收敛，这同时也是结构 5 中直接使用 3×3 卷积对残差特征进行融合效果降低的原因。由于结构 1 中的 2 个 1×1 卷积层只提取线性关系，所以仅使用一层也能起到同样的效果。不同结构在训练集和验证集的准确率对比如图 8 所示。



(a) 训练集中的准确率

(b) 验证集中的准确率

图 8　不同结构在训练集和验证集的准确率对比

2.4　1×1 卷积的通道数量对模型的影响

相比于人工设计 SRM 滤波器，DDAC 方法可以设置更多的特征通道数量，从而获取更丰富的特征。1×1 卷积的通道数量决定残差滤波器的多样性，但过多的残差滤波器会导致模型冗余、过拟合等风险。当通道数量为 16、32、48、64、128 时测试模型在 S-UNIWARD 嵌入率为 0.2 和 WOW 嵌入率为 0.2 的数据集中验证集和测试集的准确率如图 9 所示。

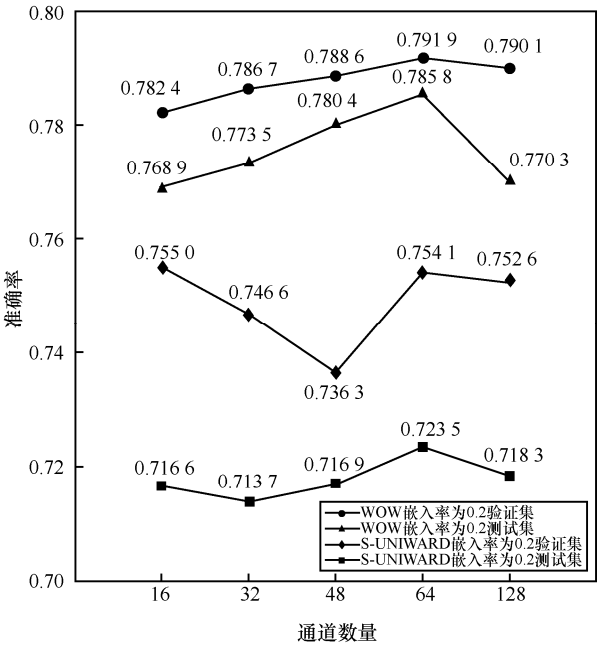


图 9　通道数量对模型准确率的影响

实验结果表明，随着通道数量的增加，模型获取的残差特征增多，测试集的准确率在 64 达到最大值，在 WOW 嵌入率为 0.2 和 S-UNIWARD 嵌入率为 0.2 的测试集中比通道数量为 16 时分别高 1.69%、0.69%。但通道数量从 64 增加至 128 时，模型在 WOW 嵌入率为 0.2 和 S-UNIWARD 嵌入率为 0.2 的测试集中准确率下降 1.55%、0.52%，说明过多的通道数量导致模型对训练集的样本过拟合使模型的泛化能力下降。根据实验结果，1×1 卷积层的通道数量设置为 64。

2.5　TLU 函数的参数 *T* 对模型的影响

使用 TLU 激活函数是为了限制残差特征的大小，提高嵌入信息和图像信息的比例，使模型在训练过程中更容易提取嵌入信息的残差特征。设置 TLU 的参数 *T* 为 1、4、7、10、13，模型在 S-UNIWARD 嵌入率为 0.4 和 WOW 嵌入率为 0.4 数据集中验证集和测试集的准确率如图 10 所示。

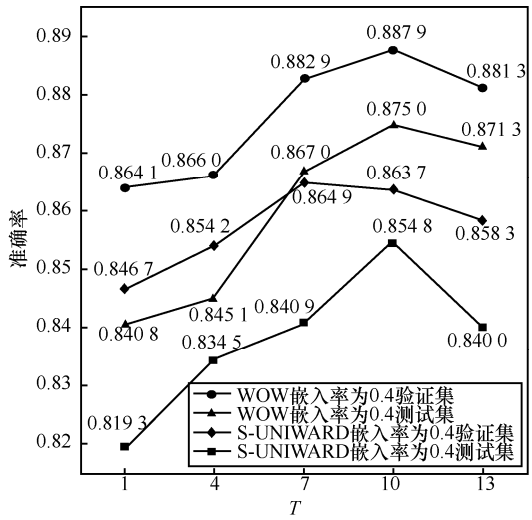


图 10　TLU 函数的参数 *T* 对模型准确率的影响

实验结果表明，随着 *T* 值的增加，模型在验证集和测试集中的准确率先逐渐提高，在 *T*=10 时达到最高。在 2 个数据集中，*T*=10 比 *T*=1 在测试集中的准确率提高 3.42%和 3.55%，而当 *T*=13 时准确率下降 0.37%和 1.48%，说明 *T* 值的选取会影响模型的效果。TLU 激活函数在输入大于 *T* 时输出 *T*，小于−*T* 时输出−*T*，梯度为 0。*T* 值过小时，会造成残差信息的损失以及梯度消失，无法有效更新方向差分组合的参数。*T* 值过大时，会降低嵌入信息残差和图像信息残差的比值，导致模型难以学习嵌入信息残差。

2.6　对比适用于卷积神经网络的特征提取方法

卷积神经网络模型一般使用固定 SRM 滤波器方法[30]和可变参数 SRM 滤波器方法[23]，将这 2 种方法与本文方法应用在测试模型中进行对比实验。为保持特征通道数量的一致性，将 DDAC 的通道数量设置为 30。模型的优化算法选用随机梯度下降算法，初始学习率为 0.1，动量为 0.9，训练轮数为 200 轮，学习率在 100 轮和 150 轮时乘以 0.2，batch_size 为 16，其中包含 8 组对应的载体图像和隐写图像。数据集使用在不同嵌入率下的 WOW 和 S-UNIWARD 隐写数据集，使用模型准确率作为评价指标。不同特征提取方法的准确率对比如表 2 所示。

表 2　不同特征提取方法的准确率对比

方法	WOW		S-UNIWARD	
	嵌入率为 0.2	嵌入率为 0.4	嵌入率为 0.2	嵌入率为 0.4
固定 SRM 滤波器	0.747 9	0.849 1	0.663 9	0.802 8
可变 SRM 滤波器	0.672 6	0.743 9	0.582 8	0.636 8
DDAC	0.753 9	0.872 3	0.694 7	0.844 0

实验结果表明，在嵌入率为 0.4 时，DDAC 方法在 WOW 数据集准确率为 87.23%，超过固定 SRM 滤波器方法 2.32%，超过可变 SRM 滤波器方法 12.84%；在 S-UNIWARD 数据集准确率为 84.40%，超过固定 SRM 滤波器方法 4.12%，可变 SRM 滤波器方法 20.72%。在嵌入率为 0.2 时，DDAC 方法在 WOW 数据集准确率为 75.39%，超过固定 SRM 滤波器方法 0.60%，超过可变 SRM 滤波器方法 8.13%；在 S-UNIWARD 数据集准确率为 69.47%，超过固定 SRM 滤波器方法 3.08%，可变 SRM 滤波器方法 11.19%。可变 SRM 滤波器方法由于受训练轮次的限制，检测效果不太理想，DDAC 方法比其他方法提升效果显著。

测试模型使用不同的特征提取方法在训练集中的收敛性如图 11 所示。当嵌入率降低时，残差信息降低，可变 SRM 滤波器方法会受图像信息干扰，导致准确率在训练过程中不稳定，提升较缓慢。DDAC 方法通过限制特征为方向差分来降低图像信息的影响，虽然在训练初期学习参数过程中准确率提升缓慢，但之后准确率快速提高，在各个数据集中获得最高准确率。实验结果表明，使用 DDAC 方法提取的特征更有效，相比于可变 SRM 滤波器特征提取方法能够更快收敛。

2.7　应用在现有模型的性能对比

将 Ye-net 与 Yedroudj-net 的特征提取部分替换为 DDAC 方法来验证适用性。Ye-net 使用 SRM 中的 30 个滤波器对卷积层参数进行初始化，参数随模型训练更新。而 Yedroudj-net 同样使用 SRM 中的 30 个滤波器对卷积层参数进行初始化，但不进行更新。为使对比条件一致，将 DDAC 方法中 1×1 卷积的通道数量设置为 30，并且用文献[17-18]中的训练方式对替换后的模型进行训练。具体的模型训练的超参数设置如表 3 所示。

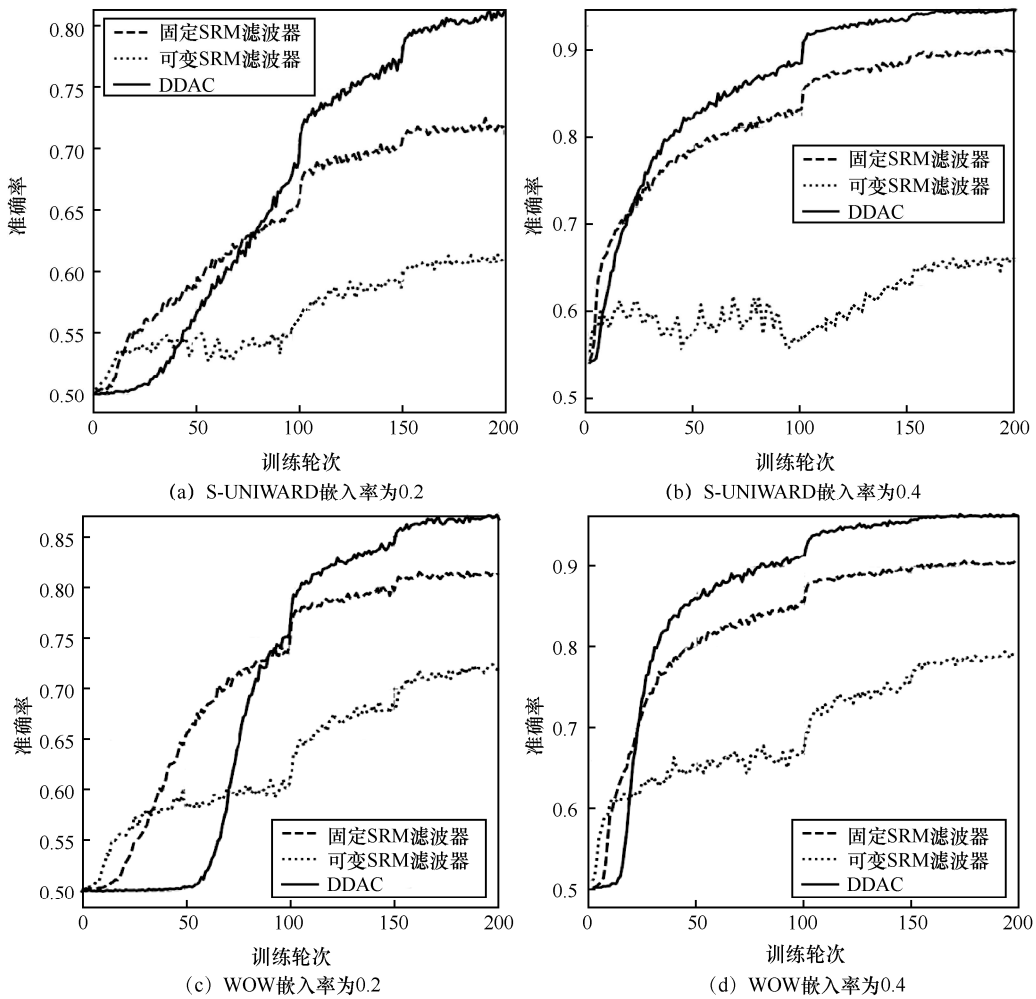


图 11　使用不同的特征提取方法在训练集中的收敛过程

表 3　　模型训练的超参数设置

模型名称	训练超参数
Ye-net	优化算法 Adadelta, 学习率为 0.4, 动量为 0.95, eps=1×10^{-8}, 权重缩减为 5×10^{-4}, 批次大小为 32, 学习率改变策略: 当训练轮次等于 100、150 时学习率乘以 0.2, 训练轮数为 200
Yedroudj-net	优化算法SGD, 学习率为0.01, 动量为0.95, 权重缩减 y=1×10^{-4}, 批次大小为 16, 学习率改变策略: 当训练轮次等于 100、150 时学习率乘以 0.1, 训练轮数为 200

训练结束后，保存验证集中效果最好的模型，并在测试集中进行测试获得模型的准确率、虚警率、漏检率。模型在不同数据集中的准确率如表 4 所示。

实验结果表明，使用方向差分自适应组合特征提取方法后，当嵌入率为 0.4 时 Ye-net 模型在 WOW 测试集的准确率提高 2.17%，嵌入率为 0.2 时提高 3.32%。在 S-UNIWARD 数据集嵌入率为 0.4 时提高 2.13%，嵌入率为 0.2 时提高 1.31%。而当嵌入率为 0.4 时 Yedroudj-net 模型在 WOW 测试集的准确率提高 1.30%，嵌入率为 0.2 时提高 2.48%。在 S-UNIWARD 数据集嵌入率为 0.4 时提高 8.21%，嵌入率为 0.2 时提高 7.55%。DDAC 方法有效提高了 2 种基于卷积神经网络的图像隐写分析模型，并且显著提高了 Yedroudj-net 对于 S-UNIWARD 算法的检测能力。

模型在各个数据集中的虚警率和漏检率分别如表 5、表 6 所示。实验结果表明，在嵌入率为 0.2 时 Yedroudj-net 在 WOW 和 S-UNIWARD 隐写算法数据集中的漏检率分别上升 0.72%和 4.82%，但虚警率分别下降 6.06%和 19.38%，使总体准确率得到提高。在嵌入率为 0.4 时，Ye-net 模型在 WOW 算法数据集中虚警率提高 2.16%，同时漏检率降低 9.28%，在 S-UNIWARD 算法数据集中漏检率提高 0.62%，虚警率降低 6.32%。除上述 4 种情况外，使用 DDAC 方法后 2 种模型的虚警率降低 1.04%~9.78%，漏检率降低 0.14%~13.1%，说明 DDAC 方法能够有效提高模型对隐写图像的检测能力，并且使决策结果的可信度更高。

2.8　DDAC-net 与其他隐写分析模型对比实验

本节将图像隐写分析模型 Ye-net[17]、Yedroudj-net[18]、SRnet[21]、Zhu-net[23]、SiaStegnet[31]、Hybrid-CNN[32]在 BOSSbase 数据集中进行训练和测试。实验根据文献[17-18,21,23,31-32]训练方法和条件进行复现。在 DDAC-net 中设置通道数量为 64，训练参

表 4　　Ye-net 与 Yedroudj-net 使用 DDAC 后的模型准确率

模型	WOW		S-UNIWARD	
	嵌入率为 0.4	嵌入率为 0.2	嵌入率为 0.4	嵌入率为 0.2
Ye-net	0.7918	0.6749	0.7432	0.6296
Ye-net+DDAC	0.8135(+0.0217)	0.7081(+0.0332)	0.7645(+0.0213)	0.6427(+0.0131)
Yedroudj-net	0.8573	0.7495	0.7540	0.6264
Yedroudj-net+DDAC	0.8703(+0.0130)	0.7743(+0.0248)	0.8361(+0.0821)	0.7019(+0.0755)

表 5　　Ye-net 与 Yedroudj-net 使用 DDAC 后的模型虚警率

模型	WOW		S-UNIWARD	
	嵌入率为 0.4	嵌入率为 0.2	嵌入率为 0.4	嵌入率为 0.2
Ye-net	0.1950	0.3272	0.2882	0.3570
Ye-net+DDAC	0.2166(+0.0216)	0.2470(−0.0802)	0.2250(−0.0632)	0.3466(−0.0104)
Yedroudj-net	0.1654	0.2978	0.2454	0.4980
Yedroudj-net+DDAC	0.1068(−0.0586)	0.2372(−0.0606)	0.1476(−0.0978)	0.3042(−0.1938)

表 6　　Ye-net 与 Yedroudj-net 使用 DDAC 后的模型漏检率

模型	WOW		S-UNIWARD	
	嵌入率为 0.4	嵌入率为 0.2	嵌入率为 0.4	嵌入率为 0.2
Ye-net	0.2518	0.3386	0.2474	0.4000
Ye-net+DDAC	0.1590(−0.0928)	0.3372(−0.0014)	0.2536(+0.0062)	0.3692(−0.0308)
Yedroudj-net	0.1654	0.2160	0.3086	0.2500
Yedroudj-net+DDAC	0.1422(−0.0232)	0.2232(+0.0072)	0.1776(−0.1310)	0.2982(+0.0482)

数设置与 2.3 节一致。隐写分析模型在 BOSSbase 数据集上的准确率如表 7 所示。

表 7 隐写分析模型在 BOSSbase 数据集上的准确率

模型	WOW		S-UNIWARD	
	嵌入率为 0.4	嵌入率为 0.2	嵌入率为 0.4	嵌入率为 0.2
Ye-net	0.791 8	0.674 9	0.743 2	0.629 6
Ye-net+DDAC	0.813 5	0.708 1	0.764 5	0.642 7
Yedroudj-net	0.857 3	0.749 5	0.754 0	0.626 4
Yedroudj-net+DDAC	0.870 3	0.774 3	0.836 1	0.701 9
SRnet	0.869 3	0.754 0	0.816 9	0.675 8
SiaStegnet	0.870 1	0.760 3	0.821 3	0.684 3
Zhu-net	0.881 6	0.767 1	0.847 3	0.715 0
Hybrid-CNN	0.774 0	0.547 0	0.919 0	0.503 0
DDAC-net	0.875 0	0.785 8	0.854 8	0.711 8

从表 7 可知，DDAC-net 在 WOW 与 S-UNIWARD 算法隐写的数据集中的准确率比 Ye-Net、Yedroudj-net、SRnet、SiaStegnet 高。在 S-UNIWARD 算法嵌入率为 0.4 下，比 Ye-Net 高 11.16%，比 Yedroudj-net 高 10.08%，比 SRnet 高 3.79%，比 SiaStegnet 高 3.35%。在 WOW 算法嵌入率为 0.2 下比 Zhu-net 高 1.87%，DDAC-net 在嵌入率为 0.2 下相比 Hybrid-CNN 方法具有明显优势。DDAC-net 对 S-UNIWARD 隐写算法的检测效果较好，在嵌入率为 0.4 时准确率为 85.48%，嵌入率为 0.2 时为 71.18%，在不同嵌入率下都有较高的准确率。

对比各个模型在训练集训练一轮的时间、参数量、浮点运算数（FLOP, floating point operations）和检测测试集的时间。测试集共有 5 000 对原图像和隐写图像，使用 Python 中的 thop 工具来计算模型的参数量和浮点运算数，使用 time 工具来计算模型的运行时间。实验结果如表 8 所示。

表 8 隐写模型的时间复杂度和参数量

模型	训练一轮时间/s	参数量/1×10^6	浮点运算数/1×10^9	测试集检测时间/s
Ye-net	16.1	0.107	1.94	6.01
Yedroudj-net	24.3	0.445	3.30	10.21
SRnet	49.2	4.777	5.26	16.15
DDAC-net	22.1	0.678	2.63	8.06

从表 7 和表 8 可知，DDAC-net 相比 SRnet 减少 50%的浮点运算数，相比 Yedroudj-net 减少 20%，但准确率上在各个数据集中都有提升，同时使模型检测速度提高。训练一轮的时间分别由 49.2 s 和 24.3 s 降低至 22.1 s，使模型训练速度加快。在文献[17,21]中 SRnet、Ye-net 需要训练 500 轮以上，而 DDAC-net 只需训练 200 轮，可以大大缩短训练时间。

3 结束语

特征提取环节在图像隐写分析模型中十分关键。本文提出方向差分自适应组合的特征提取方法，对中心像素的方向差分通过 1×1 卷积线性组合，在模型训练中自适应地更新组合参数，使用 TLU 激活函数提高隐写信息残差和图像残差的比率。实验对比不同的特征提取结构、1×1 卷积的通道数量、TLU 激活函数的参数 T 值对测试模型准确率的影响来论证方法的合理性。实验表明，应用 DDAC 特征提取方法可以有效提高隐写分析模型的准确率。对比固定 SRM 滤波器和可变 SRM 滤波器方法，应用 DDAC 方法的模型在不同隐写算法数据集中都取得良好效果。DDAC-net 相比于现有模型，可以提高检测准确率并降低参数量和运行时间，具备良好的应用前景。

Qian 等[33]提出在低嵌入率条件下训练模型时，可以使用在高嵌入率的数据集中训练得到的模型参数对模型进行初始化，来提高模型的训练效果。针对 DDAC 模型，可以进一步研究上述模型迁移策略对组合参数进行初始化的方法，使用在高嵌入率数据集中训练得到的组合参数，并在低嵌入率数据集的训练过程中进行微调，防止因隐写信息较少而难以训练有效的组合参数，提高模型在低嵌入率下的训练速度和准确率。后续可以基于 DDAC 方法设计更复杂的隐写分析模型，进一步提高隐写分析能力。

参考文献：

[1] CHEDDAD A, CONDELL J, CURRAN K, et al. Digital image steganography: survey and analysis of current methods[J]. Signal Processing, 2010, 90(3): 727-752.

[2] 付章杰, 王帆, 孙星明, 等. 基于深度学习的图像隐写方法研究[J]. 计算机学报, 2020, 43(9): 1656-1672.
FU Z J, WANG F, SUN X M, et al. Research on steganography of digital images based on deep learning[J]. Chinese Journal of Computers, 2020, 43(9): 1656-1672.

[3] 陈君夫, 付章杰, 张卫明, 等. 基于深度学习的图像隐写分析综述[J]. 软件学报, 2021, 32(2): 551-578.
CHEN J F, FU Z J, ZHANG W M, et al. Review of image steganalysis based on deep learning[J]. Journal of Software, 2021, 32(2): 551-578.

[4] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[5] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[6] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.
[7] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. Berlin: Springer, 2015: 234-241.
[8] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.
[9] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680.
[10] YANG J H, RUAN D Y, HUANG J W, et al. An embedding cost learning framework using GAN[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 839-851.
[11] TANG W X, TAN S Q, LI B, et al. Automatic steganographic distortion learning using a generative adversarial network[J]. IEEE Signal Processing Letters, 2017, 24(10): 1547-1551.
[12] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.
[13] TAN S Q, LI B. Stacked convolutional auto-encoders for steganalysis of digital images[C]//Proceedings of Signal and Information Processing Association Annual Summit and Conference (APSIPA). Piscataway: IEEE Press, 2014: 1-4.
[14] QIAN Y L, DONG J, WANG W, et al. Deep learning for steganalysis via convolutional neural networks[C]//Proceedings of SPIE - The International Society for Optical Engineering. Bellingham: SPIE Press, 2015: 171-180.
[15] XU G S, WU H Z, SHI Y Q. Structural design of convolutional neural networks for steganalysis[J]. IEEE Signal Processing Letters, 2016, 23(5): 708-712.
[16] YEDROUDJ M, COMBY F, CHAUMONT M. Yedroudj-net: an efficient CNN for spatial steganalysis[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2018: 2092-2096.
[17] YE J, NI J Q, YI Y. Deep learning hierarchical representations for image steganalysis[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2545-2557.
[18] FRIDRICH J, KODOVSKY J. Rich models for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.
[19] LI B, WEI W H, FERREIRA A, et al. ReST-net: diverse activation modules and parallel subnets-based CNN for spatial image steganalysis[J]. IEEE Signal Processing Letters, 2018, 25(5): 650-654.
[20] KALMAN B L, KWASNY S C. Why tanh: choosing a sigmoidal function[C]//Proceedings of International Joint Conference on Neural Networks. Piscataway: IEEE Press, 1992: 578-581.
[21] BOROUMAND M, CHEN M, FRIDRICH J. Deep residual network for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(5): 1181-1193.
[22] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. New York: ACM Press, 2017: 4278-4284.
[23] ZHANG R, ZHU F, LIU J Y, et al. Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 1138-1150.
[24] PEVNY T, BAS P, FRIDRICH J. Steganalysis by subtractive pixel adjacency matrix[J]. IEEE Transactions on Information Forensics and Security, 2010, 5(2): 215-224.
[25] SUTSKEVER I, MARTENS J, DAHL G, et al. On the importance of initialization and momentum in deep learning[C]//International Conference on Machine Learning. Saarland: DBLP, 2013: 1139-1147.
[26] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//Proceedings of the 32nd International Conference on Machine Learning. Saarland: DBLP, 2015: 448-456.
[27] NAIR V, HINTON G E. Rectified linear units improve restricted boltzmann machines[C]//Proceedings of the 27th International Conference on Machine Learning. Saarland: DBLP, 2010: 807-814.
[28] BAS P, FILLER T, PEVNÝ T. “Break our steganographic system”: the ins and outs of organizing BOSS[C]//Information Hiding. Berlin: Springer, 2011: 59-70.
[29] KODOVSKY J, FRIDRICH J, HOLUB V. Ensemble classifiers for steganalysis of digital media[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(2): 432-444.
[30] 沈军, 廖鑫, 秦拯, 等. 基于卷积神经网络的低嵌入率空域隐写分析[J]. 软件学报, 2021, 32(9): 2901-2915.
SHEN J, LIAO X, QIN Z, et al. Spatial steganalysis of low embedding rate based on convolutional neural network[J]. Journal of Software, 2021, 32(9): 2901-2915.
[31] YOU W K, ZHANG H, ZHAO X F. A siamese CNN for image steganalysis[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 291-306.
[32] ARIVAZHAGAN S, AMRUTHA E, SYLVIA L J W, et al. Hybrid convolutional neural network architecture driven by residual features for steganalysis of spatial steganographic algorithms[J]. Neural Computing and Applications, 2021, 33(17): 11465-11485.
[33] QIAN Y L, DONG J, WANG W, et al. Learning and transferring representations for image steganalysis using convolutional neural network[C]//Proceedings of 2016 IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2016: 2752-2756.

[作者简介]

王晓丹（1966− ），女，陕西汉中人，博士，空军工程大学教授，主要研究方向为智能信息处理、机器学习。

李京泰（1998− ），男，重庆人，空军工程大学硕士生，主要研究方向为图像隐写分析。

宋亚飞（1988− ），男，河南汝州人，博士，空军工程大学副教授，主要研究方向为机器学习及其在目标识别、入侵检测等领域中的应用。