BIFRNet: A Brain-Inspired Feature Restoration DNN for Partially Occluded Image Recognition

Jiahong Zhang [1,2] **, Lihong Cao** [1,2] [*] **, Qiuxia Lai** [1,2] **, Bingyao Li** [2] **, Yunxiao Qin** [1,2]

[1] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China
[2] Neuroscience and Intelligent Media Institute, Communication University of China, Beijing, China
{zhangjh, lihong.cao, qinyunxiao, qxlai}@cuc.edu.cn, l_bingyao@163.com

Abstract

The partially occluded image recognition (POIR) problem has been a challenge for artificial intelligence for a long time. A common strategy to handle the POIR problem is using the non-occluded features for classification. Unfortunately, this strategy will lose effectiveness when the image is severely occluded, since the visible parts can only provide limited information. Several studies in neuroscience reveal that feature restoration which fills in the occluded information and is called amodal completion is essential for human brains to recognize partially occluded images. However, feature restoration is commonly ignored by CNNs, which may be the reason why CNNs are ineffective for the POIR problem. Inspired by this, we propose a novel brain-inspired feature restoration network (BIFRNet) to solve the POIR problem. It mimics a ventral visual pathway to extract image features and a dorsal visual pathway to distinguish occluded and visible image regions. In addition, it also uses a knowledge module to store object prior knowledge and uses a completion module to restore occluded features based on visible features and prior knowledge. Thorough experiments on synthetic and real-world occluded image datasets show that BIFRNet outperforms the existing methods in solving the POIR problem. Especially for severely occluded images, BIRFRNet surpasses other methods by a large margin and is close to the human brain performance. Furthermore, the brain-inspired design makes BIFRNet more interpretable.

Introduction

Convolutional neural networks (CNNs) have achieved remarkable results in computer vision and even reached human-level performance in some tasks such as face recognition (Masi et al. 2018). However, the robustness of CNNs under partial occlusion is much lower than that of the human visual system (Geirhos et al. 2017), possibly because occluders can easily mislead CNNs to predict wrong results. Since occlusion occurs in many real-world images or videos, most computer vision tasks, such as human-computer interaction (Liu et al. 2020), instance segmentation (Ke, Tai, and Tang 2021) and human pose estimation (Das, Kishore, and Bhattacharya 2020) are troubled by the partial occlusion

---

problem. In this paper, we try to solve the partially occluded image recognition (POIR) problem.

A natural idea to solve the POIR problem is extracting robust features. For example, feature regularization (Liao et al. 2016) and convolution kernel regularization (Tabernik et al. 2016) were used to force feature activations to be disentangled for different objects. However, the paper (Kortylewski et al. 2021) indicated that these methods are not always robust to partial occlusion. Some works proposed that using generative models to restore the clean image from the partially occluded image before classification usually improves the classification accuracy (Nguyen et al. 2016; Yu et al. 2018; Yuan and Park 2019; Banerjee and Das 2020). Nevertheless, image-level restoration commonly results in costly computation, and training a generative model is challenging.

A recent popular direction to solve the POIR problem is removing occluded features (ROF). It removes the features that are corrupted by occluders and then utilizes the remaining visible features for classification. For example, an external sub-network can be used to generate mask-guided attention which discriminates occluded and visible features (Wan and Chen 2017; Zhang, Yang, and Schiele 2018; Pang et al. 2019; Qiu et al. 2021; Zou et al. 2020; Chen et al. 2021; Song et al. 2019a). Networks then classify images based only on the visible features.

Although ROF is effective to solve the POIR problem, it does not restore occluded features. In contrast, several works in neuroscience demonstrated that amodal completion which restores features in human brain is important for humans to recognize partially occluded objects (Weigelt, Singer, and Muckli 2007). Amodal completion extends edges behind occluders if a continuous smooth connection exists (Kellman and Shipley 1991) and fills missing surfaces (He and Nakayama 1992) and volumes (Tse 1999). In addition, the work (Peters and Kriegeskorte 2021) indicated that amodal completion relies on prior knowledge about image statistics.

Based on this, we suggest that an occlusion robustness model should be able to restore the occluded features. To obtain this ability, the model should firstly be able to distinguish visible and occluded parts and should have prior knowledge of the target objects.

We introduce a novel brain-inspired feature restoration network (BIFRNet) with all these capabilities, as shown in Figure 1. Inspired by human visual system, we construct a
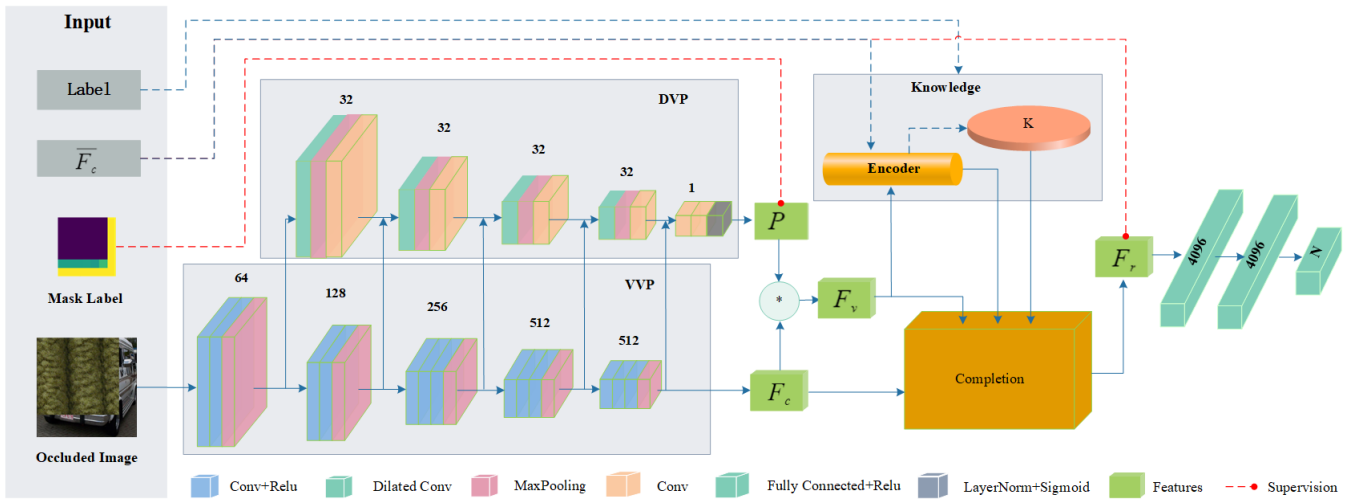
Figure 1: The architecture of BIFRNet. The ventral visual pathway (VVP) is based on VGG16, extracting image features F_c. The dorsal visual pathway (DVP) which consists of four D+MP+C blocks and a C+C+LayerNorm+Sigmoid block outputs spatial attention P. F_v is the visible feature obtained by the Hadamard product of F_c and P. The knowledge module forms prior knowledge $\mathcal{K}$ by training. $\mathcal{K}$ is input into the completion module, where restored features F_r is generated for classification. $\overline{F}_c$ is the clear image feature from the pre-trained VGG16 pool5 layer and is used as the label of F_r. The dotted line connections exist only in the training phase and do not exist in the testing phase. Red dotted lines denote the paths where we compute loss. Numbers in this figure denote the feature channels. $\overline{F}_c$ and F_v are the inputs of the encoder in training and testing, respectively.

ventral visual pathway (VVP) to extract image features and a dorsal visual pathway (DVP) to generate spatial attention for distinguishing occlusion (Maunsell and Cook 2002; Serences et al. 2004; Ungerleider and G. 2000). Amodal completion can be construed as an inference process which restores the occluded features based not only on the visible features but also prior knowledge about target objects (Peters and Kriegeskorte 2021). Therefore, a completion module is proposed to imitate this process and a knowledge module is proposed to provide the prior knowledge. Restoration here is feature-level and differs from the existing image-level restoration method (Nguyen et al. 2016; Yu et al. 2018; Banerjee and Das 2020) and is close to human brain. We review relevant neuroscience bases in Related Work and carry out ablation experiments of BIFRNet in Discussion, showing its interpretability and effectiveness. In conclusion, the main contributions of this paper include:

1) We propose a brain-inspired feature restoration network, BIFRNet. It mimics human ventral visual pathway to extract features and dorsal visual pathway to discriminate visible and occluded parts. It also has a knowledge module to provide prior knowledge and a completion module to restore the occluded features.

2) Studies of BIFRNet's components show their designated functions and effectiveness. The brain-inspired design provides BIFRNet with high interpretability and may benefit future research for partially occluded image recognition-like problems.

3) BIFRNet gets state-of-the-art performance on synthetic and real-world occluded images. On the synthetic dataset, it achieves human-like performance and outperforms the state-of-the-art model by 7% recognition accuracy for highly oc-

cluded images.

Related Work

CNNs for Occluded Image Recognition

Although CNN models are successful in image recognition, they suffer from the POIR problem. Data augmentation in terms of partial occlusion (DeVries and Taylor 2017; Yun et al. 2019) is a common method to solve this problem. However, the work (Kortylewski et al. 2021) indicated that it is not always robust. Recently, removing occluded features (ROF)-based methods become popular for POIR. It removes the features corrupted by occluders and then utilizes the remaining visible features for classification. Common ROF-based methods include attention-based and compositional models. There are many variants of attention mechanisms in dealing with occlusion (Zhang, Yang, and Schiele 2018; Xiao et al. 2019). Top-down feedback attention in TDAP-Net is composed of explainable part attention, which can reduce the contamination of occlusion (Xiao et al. 2019). Mask-guided attention is generated by an extra sub-network to distinguish occluded and non-occluded features. For example, MaskNet(Wan and Chen 2017) used mask attention to assign higher weights to the hidden units associated to the visible facial parts and achieves occlusion robustness in face recognition. Pairwise Differential Siamese Network (PDSN) (Song et al. 2019b) learned the correspondence between occluded facial regions and corrupted activations, and used the visible features to classify. Mask-guided attention was also used in other vision tasks such as occluded pedestrian detection (Zou et al. 2020; Chen et al. 2021). Compositional models achieved interpretable occlusion robustness by combining CNN and probability models (Wang et al. 2017; Zhang

et al. 2018; Kortylewski et al. 2021). Kortylewski et al. (Kortylewski et al. 2020) proposed a dictionary-based compositional model to learn the feature dictionary for better classification. They also presented a differentiable compositional model that can localize occluders and classify images by the visible features (Kortylewski et al. 2021). However, compositional models are complex and require manually designed constraints.

Neuroscience Basis for Occluded Image Recognition

It is generally believed that the human visual system includes ventral and dorsal visual pathways (Goodale and Milner 1992). The ventral pathway works as an object recognizer for objects at the center view. The dorsal pathway is to recognize objects' spatial positions and movements. When an object is partially occluded, the human visual system may have two strategies to restore the missing information: modal completion and amodal completion (Michotte, Thinès, and Crabbé 1991). Modal completion is a process of producing an illusory contour effect (Kanizsa 1979). Amodal completion is the ability to see an entire object despite parts of it being covered by another object in front of it, which is more suitable for solving the POIR problem than modal completion (Johnson and Olshausen 2005). Studies have shown that the completion of simple lines and shapes occurs early in the ventral visual pathway (Lee 2003; Doniger et al. 2000), and simple CNN networks such as AlexNet also have similar capabilities (Kim et al. 2021). However, complex completion needs to judge the relationship between the occluded object and the occluder (Johnson and Olshausen 2005). This completion requires more time and depends on the recurrent process (Sekuler and Palmer 1992; Shore and Enns 1997). In this paper, we model the relationship between the occluded object and the occluder by spatial attention and use the recurrent neural network to model the completion function. In addition, amodal completion must rely on prior knowledge about the statistics of images or about the shape of objects (Peters and Kriegeskorte 2021).

Proposed Method

The prediction of a normal CNN classification model can be fomulated as

$$p = f(x), \qquad (1)$$

where x is an image, f is the CNN model, and p is the predicted classification probability. In occlusion scene, an occluded image can be represented as:

$$x_{occ} = x \cdot o + m, \qquad (2)$$

where o is a binary occlusion position matrix with the same size as x. 0 and 1 in o denote the positions for occluded and non-occluded, respectively. m is the occluder, and $\cdot$ refers to element-wise production. Figure 2 visualizes this equation. Disturbed by the occluder, CNN models commonly misclassify x_{occ}, which is known as the POIR problem. This paper proposes BIFRNet to solve the POIR problem.
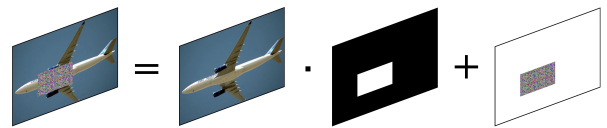

Figure 2: The visualization of Equation 2. Here, black elements in the matrix are 1, and white elements are 0.

Feature	Description
F_c	The features of the ventral visual pathway
$\bar{F}_c$	The clear image features from the pre-trained VGG16 pool5 layer, used as the label of F_r
P	The spatial attention from the dorsal visual pathway
F_v	The visible features obtained by multiplying F_c by P
F_r	The restored features
F_k	The knowledge coding
$\mathcal{K}$	The knowledge matrix

Table 1: Descriptions of the feature symbols in BIFRNet

Network Architecture

Figure 1 shows the proposed BIFRNet and Table 1 lists the descriptions of abbreviations. BIFRNet uses three main components to achieve occlusion robustness: 1) the visual pathways VVP and DVP, 2) the knowledge module, and 3) the completion module. VVP extracts image features and DVP generates the spatial attention for distinguishing occlusion. During training, prior knowledge of target objects will be formed in the knowledge module. The completion module restores occluded features with visible features and the knowledge. Finally, several fully connected layers are used for classification based on the restored features. The input of BIFRNet is an image with the size of $C \times H \times W$, and the features F_c, F_v and F_r are with size of $C_v \times H_v \times W_v$.

Visual Pathways Visual pathways of BIFRNet consist of VVP and DVP. VVP is VGG16 (Simonyan and Zisserman 2014) without fully connection layers. DVP is parallel to VVP, which consists of 4 D+MP+C and one C+C+LayerNorm+Sigmoid blocks in series, where D is dilated convolution layer with dilated rate 2, C is convolution layer, LayerNorm is layer normalization (Ba, Kiros, and Hinton 2016), and MP is Maxpooling layer. Dilated convolution is used to increase the receptive field and form good attention. Before every block in DVP, there are connections with VVP, where the features are concatenated. These connections allow DVP to consider the information from multiple layers in VVP. To get a trade-off between the performance and computing cost, we set the number of output channels in the middle layers of DVP to 32. The output of DVP is a spatial attention $P \in [0, 1]$ with the size of $H_v \times W_v$, which represents the feature occlusion probabil-
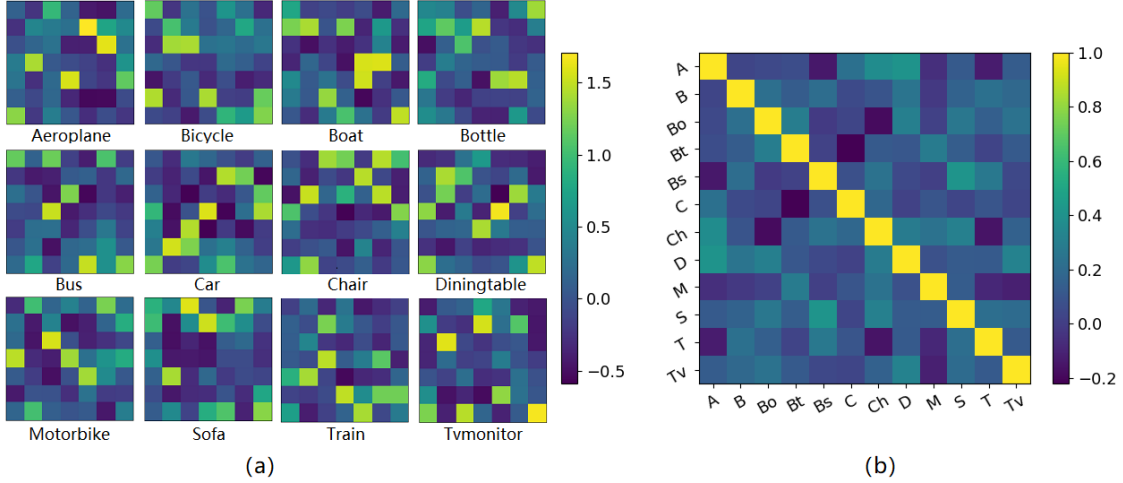
Figure 3: Illustration of $\mathcal{K}$ after training on the synthetic Occluded-Vehicles dataset (Kortylewski et al. 2021). Figures in (a) are the knowledge representations of 12 categories, aeroplane, bicycle, boat, bottle, bus, car, chair, diningtable, motorbike, sofa, train and tvmonitor in the dataset. (b) shows the similarity matrix of different category knowledge in $\mathcal{K}$, the names of which are abbreviated to A, B, Bo, Bt, Bs, C, Ch, D, M, S, T and Tv in sequence. Cosine similarity is used as the similarity measure. It is shown that $\mathcal{K}$ has well separable representations of the 12 categories.
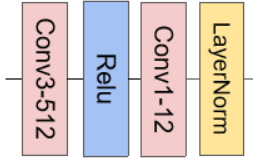

Figure 4: The architecture of the encoder. When training, the encoder receives clean image features and generates the knowledge code F_k. When testing, its input is the visible features F_v.

ity. This process can be modeled as a prediction problem:

$$P = DVP(F_{vvp}), \tag{3}$$

where F_{vvp} contains all the intermediate features of VVP. BIFRNet computes the Hadamard product of F_c and P to obtain the approximation of visible features F_v:

$$F_{vi} = F_{ci} * P, \tag{4}$$

where F_c is the $C_v \times H_v \times W_v$ feature from VVP. i is the channel index.

Prior-Knowledge When the occlusion is serious, only visible features of an occluded image cannot provide sufficient information. Thus prior knowledge of objects is required. The proposed knowledge module aims to encode image features to the object knowledge. It contains two components, an encoder and a knowledge matrix $\mathcal{K}$. Figure 4 shows the architecture of the encoder. $\mathcal{K}$ denotes the knowledge of all categories with size of $N \times H_v \times W_v$, where N denotes the number of categories. Each category has a $H_v \times W_v$ feature matrix, which is the knowledge representation.

$\mathcal{K}$ is randomly initialized. During training, the encoder encodes features ($\bar{F}_c$ in training and F_v in testing) of images

to knowledge coding, called F_k. We integrate F_k to form a representation space ζ, which contains knowledge of all categories. It is hoped that $\mathcal{K}$ is close to ζ.

Algorithm 1 shows the process of training $\mathcal{K}$. Each dimension in $\mathcal{K}$ is expected to represent the knowledge of an category. Firstly, F_k of an image is multiplied by the N dimensional one-hot vector converted from its classification label to get $\tilde{F}_k$. Values in $\tilde{F}_k$ are non-zero in the corresponding category dimension and are zero in the non-corresponding ones, which realizes to represent an object. Secondly, summing $\tilde{F}_k$ by batch forms a sample of ζ, noted as $\tilde{\zeta}$. If some classes are missing from one batch, the value of the corresponding classification dimensions in $\tilde{\zeta}$ are 0. It should be replaced with the corresponding dimension in $\mathcal{K}$ to prevent damage to the overall distribution. Finally, by minimizing the Kullback-Leibler divergence (KLD) between $\mathcal{K}$ and $\tilde{\zeta}$, $\mathcal{K}$ will represent knowledge of all categories.

After training, knowledge representations in $\mathcal{K}$ and the similarity matrix of them are shown in Figure 3, which demonstrates $\mathcal{K}$ does have prior knowledge of objects. In the supplementary material, we illustrate the formation process of knowledge. When testing, $\mathcal{K}$ is input into the completion module as a fixed matrix.

Completion The completion module restores occluded features, shown as Figure 5. The motivation to use the recurrent structure comes from neuroscience (Sekuler and Palmer 1992; Shore and Enns 1997) and the multi-view image generation network GQN (Eslami et al. 2018). GQN uses ConvLSTM (Shi et al. 2015) as the basic block to generate view-specific images based on the representations of other views. The completion module in BIFRNet also uses ConvLSTM as the backbone and receives F_v, F_k, $\mathcal{K}$ to complete F_c, get-

Algorithm 1: The training of the knowledge matrix $\mathcal{K}$

Require:

    The output of the encoder, F_k;

    The classification label, l;

    The number of categories, n;

Ensure: $\mathcal{K}$

    Initialize $\mathcal{K}$ with a random number matrix;

1: **while** not done **do**

2:     Initialize $\tilde{F}_k$ with a zero matrix;

3:     **for** each $k_i \in F_k$ **do**

4:         $l_i \Leftarrow$ One-hot($l[i]$);

5:         $\tilde{F}_k \Leftarrow l_i * k_i + \tilde{F}_k$;

6:     **end for**

7:     // b is the batch size

8:     $\tilde{\zeta} \Leftarrow \tilde{F}_k / b$;

9:     **for** $j \in n$ **do**

10:       **if** $\tilde{\zeta}[j]$ is 0 **then**

11:         $\tilde{\zeta}[j] \Leftarrow \mathcal{K}[j]$;

12:       **end if**

13:     **end for**

14:     $L_k \Leftarrow KLD(\tilde{\zeta}, \mathcal{K})$;

15:     minimize L_k;

16: **end while**

---

ting the restored feature F_r. It can be formulated by:

$$
\begin{aligned}
Initial\ state\quad & (F_k^1, \mathcal{K}^1, F_c^1) = (F_k, \mathcal{K}, F_c), \\
State\ update\quad & (F_k^{t+1}, \mathcal{K}^{t+1}, F_c^{t+1}) \\
& = ConvLSTM(F_k^t, \mathcal{K}^t, F_c^t, F_v), \\
Output\quad & F_r = Conv(F_c^7).
\end{aligned} \quad (5)
$$

The completion module contains 7 ConvLSTM blocks, where F_c^7 denotes the features of the last block.

Training Detail

BIFRNet can be trained end-to-end. Its loss function consists of four parts. For DVP, mask labels of occlusion are provided during training. We use four kinds of operations to generate occluded images and their corresponding mask labels for each clean image, which is shown in Figure 6. Because the convolution operation can maintain the spatial position, we use O_f in Figure6 as the mask label. The size of O_f is $H_v \times W_v$, which is the same size as the output of DVP. We use O_f to train DVP to assign high weights to visible parts. Thus the optimization object is to minimize the attention loss, where the Multi-Label Soft Margin Loss (MLSML) is adopted:

$$
\begin{aligned}
L_a(P, O_f) = & -\frac{1}{N} \times \sum_i O_f[i] \times \log((1 + \exp^{(-P[i])})^{-1}) \\
& + (1 - O_f[i]) \times \log(\frac{\exp^{-P[i]}}{1 + \exp^{-P[i]}})
\end{aligned} \quad (6)
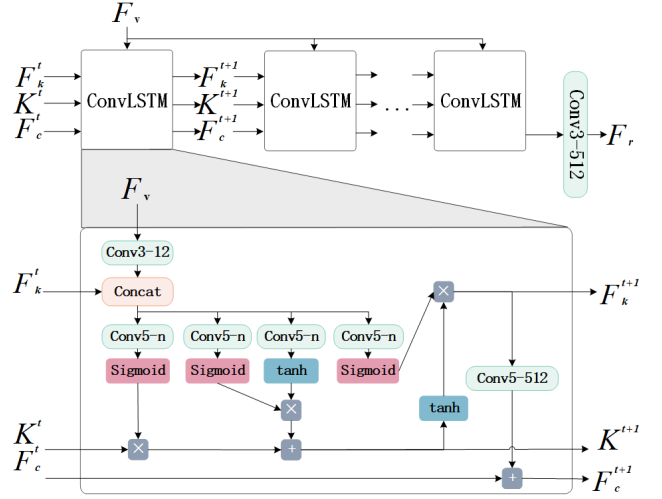$$

For the knowledge module, we minimize the KLD loss

Figure 5: The architecture of the completion module. Its core blocks are ConvLSTM.

L_k, shown in Algrithm 1.

$$
L_k(\tilde{\zeta} \| \mathcal{K}) = \sum \tilde{\zeta}(x) \log \frac{\tilde{\zeta}(x)}{\mathcal{K}(x)}. \quad (7)
$$

For the completion module, we suggest that its output F_r should be similar to the features of the corresponding clean image. Therefore, we use the clean image features $\bar{F}_c$ from the pre-trained VGG16 pool5 layer as the target of F_r and minimize the L2 loss of them:

$$
L_r = \frac{1}{2} \sum \| F_r - \bar{F}_c \|^2. \quad (8)
$$

Finally, cross-entropy loss function is used as the classification loss:

$$
L_c(p, y) = \frac{1}{M} \sum i \sum_{c=1}^{N} y_i^c log(p_i^c). \quad (9)
$$

where M is the number of samples, N is the number of classes, y_i^c is the label, and p_i^c is the predicted classification probability of the model. Totally, the loss function is as follows:

$$
L = L_k + L_r + L_c + L_a \times \alpha. \quad (10)
$$

Here, α is a weight parameter, and we set it 0.1. Discussion about it is presented in the supplementary material. After training, given an input image, BIFRNet will output the prediction result.

Experiments

Dataset

We use Occluded-Vehicles dataset and Occluded-COCO Vehicles dataset (Kortylewski et al. 2021) to evaluate the proposed BIFRNet. For Occluded-Vehicles dataset, images are from the PASCAL3D+ dataset (Xiang, Mottaghi, and Savarese 2014), and the occluded images are generated artificially by covering four different types of occluders (*i.e.,*

Occ.Area	L0: 0%	L1: 20-40%				L2: 40-60%				L3: 60-80%			
Occ.Type	-	w	n	t	o	w	n	t	o	w	n	t	o
VGG16	99.2	96.9	97.0	96.5	93.8	92.0	90.3	89.9	79.6	67.9	62.1	59.5	62.2
MaskNet	99.6	98.6	98.3	98.3	96.5	96.7	96.1	95.3	87.6	84.0	83.1	74.3	71.7
CoD	92.1	92.7	92.3	91.7	92.3	87.4	89.5	88.7	90.6	70.2	80.3	76.9	87.1
VGG+CoD	98.3	96.8	95.9	96.2	94.4	91.2	91.8	91.3	91.4	71.6	80.7	77.3	87.2
TDAPNet	99.3	98.4	98.9	98.5	97.4	96.1	97.5	96.6	91.6	82.1	88.1	82.7	79.8
CompNet-Multi	99.3	98.6	98.6	98.8	97.9	98.4	98.4	97.8	94.6	91.7	90.7	86.7	88.4
BIFRNet	**99.8**	**99.6**	**99.6**	**99.5**	**99.4**	**99.4**	**99.4**	**99.2**	**98.2**	**98.2**	**98.2**	**96.2**	**95.8**
Human	100.0	100.0				100.0				98.3			

Table 2: Classification results of BIFRNet and some state-of-the-art methods on the Occluded-Vehicles dataset with different levels of artificial occlusion (0%,20-40%,40-60%,60-80% of the object are occluded) and different types of occluders (w=white boxes, n=noise boxes, t=textured boxes, o=natural objects).
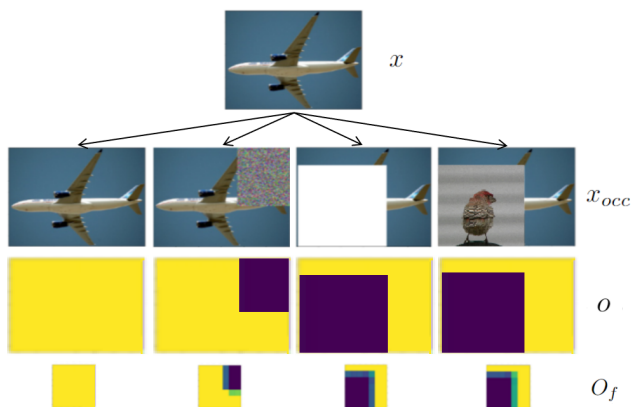

Figure 6: The generation method of the occluded images and its mask labels. Here x is a clean image from the training dataset, and x_{occ} is its occluded version. o and O_f are the occlusion masks of an occluded image and its feature, respectively. Elements in o are 0 when the corresponding postions are occluded and 1 when non-occluded. O_f is obtained by downsampling o. There are four occlusion types, unchanged, the noise block, the white block and the image occluders. Here, the image occluders we use are sampled from the first 20 categories in ImageNet valuation set (Deng et al. 2009) and the KTH-TIPS2-a dataset (Mallikarjuna et al. 2006). These images are not used to construct the test dataset. The occlusion position and occlusion rate are randomly selected.

white, noise, textures, and segmented objects) to the clean images. Four occlusion levels are used to represent the percentage of occluded area: Level-0 (0%), Level-1 (20-40%), Level-2 (40-60%), and Level-3 (60-80%). Images in Occluded-COCO Vehicles dataset are selected from MS-COCO (Lin et al. 2014) dataset and also has four occlusion levels. Especially, occluded images in this dataset are real-world. The two datasets both contain 12 different categories. Following (Xiao et al. 2019), we use all categories for training and use 6 categories including aeroplane, bicycle, bus, car, motorbike, and train, for testing. Detailed descriptions can be found in the supplementary material, accessible at

Occ.Area	L0	L1	L2	L3
VGG16	99.1	88.7	78.8	63.0
MaskNet	**99.5**	88.5	78.8	65.8
TDAPNet	99.4	88.8	87.9	69.9
CompNet-Multi	99.4	**95.3**	90.9	**86.3**
BIFRNet	**99.5**	94.8	**93.5**	**86.3**

Table 3: Classification results on the Occluded-COCO Vehicles dataset.

https://github.com/JiaHongZ/BiFRNet.

Experimental Settings

We use Adam (Kingma and Ba 2014) as the optimizer. Batch size and the initial learning rate are set to 64 and 0.0001, respectively, and the learning rate is cosine decayed. VGG16 used in this paper is pre-trained in ImageNet (Deng et al. 2009). The training and testing settings for Occluded-Vehicles and Occluded-COCO Vehicles datasets stay the same.

Experimental Results

We compare BIFRNet to VGG16 (Simonyan and Zisserman 2014) and some state-of-the-art models CoD (Kortylewski et al. 2020), TDAPNet (Xiao et al. 2019) and CompNet-Mul (Kortylewski et al. 2021) on Occluded-Vehicles and Occluded-COCO Vehicles. MaskNet (Wan and Chen 2017) was proposed to recognize human face and its backbone is ResNet (He et al. 2016). Here we re-implement it using VGG16 for a fair comparison. The human experimental data is from (Kortylewski et al. 2020).

Results on Occluded-Vehicles: Table 2 shows the comparison results on the synthetic dataset Occluded-Vehicles. We observe that BIFRNet significantly outperforms other methods at all occlusion levels, and achieves human-like performances. As for occlusion types, BIFRNet improves the classification accuracy on the object occlusion most.

Results on Occluded-COCO Vehicles: Here we evaluate BIFRNet on the real-world Occluded-COCO Vehicles dataset. Experimental results reported in Table 3 show that BIFRNet still outperforms the other methods on real-world occluded images.

Occ.Area	L0: 0%	L1: 20-40%				L2: 40-60%				L3: 60-80%			
Occ.Type	-	w	n	t	o	w	n	t	o	w	n	t	o
$\mathcal{K}+N(0,1)$	**99.84**	**99.61**	99.61	**99.52**	**99.40**	99.42	**99.38**	**99.22**	98.20	98.23	98.20	96.11	**95.90**
$\mathcal{K}+N(0,3)$	99.82	**99.61**	99.61	99.49	99.29	**99.45**	**99.38**	99.17	98.16	**98.25**	98.18	95.92	95.62
$\mathcal{K}+N(0,5)$	99.82	**99.61**	99.61	99.49	99.29	**99.45**	**99.38**	99.17	97.93	98.23	98.13	95.85	94.98
BIFRNet-K	99.72	99.54	**99.65**	99.38	99.22	99.24	99.29	98.76	97.86	97.51	97.77	94.15	93.73
BIFRNet	**99.84**	**99.61**	99.61	**99.52**	**99.40**	99.42	**99.38**	**99.22**	**98.23**	98.18	**98.25**	**96.15**	95.81

Table 4: Studies of Knowledge in BIFRNet on the Occluded-Vehicles dataset.

Occ.Area	L0: 0%	L1: 20-40%				L2: 40-60%				L3: 60-80%			
Occ.Type	-	w	n	t	o	w	n	t	o	w	n	t	o
Completion Ablation	99.4	98.8	99.0	98.9	98.5	98.5	98.6	98.4	97.1	96.1	96.1	92.9	92.6
BIFRNet-Completion	99.7	99.4	99.3	99.3	99.1	99.1	98.9	98.7	98.0	97.3	96.8	93.5	92.8
BIFRNet	**99.8**	**99.6**	**99.6**	**99.5**	**99.4**	**99.4**	**99.4**	**99.2**	**98.2**	**98.2**	**98.2**	**96.2**	**95.8**

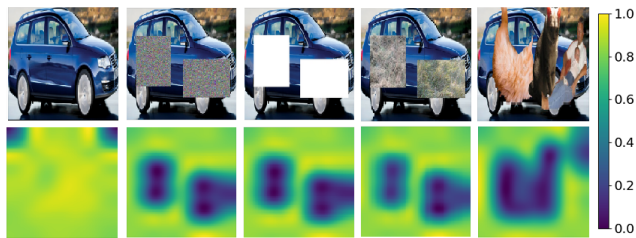Table 5: Studies of Completion in BIFRNet on the Occluded-Vehicles dataset.

Figure 7: Visualization of the attention maps P from the dorsal visual pathway, where the bright parts have higher attention than the darker ones. The upper row includes images from the Occluded-Vehicles dataset, and the lower row shows the corresponding attention map P.

Discussion

This section discusses how BIFRNet's three main components (visual pathways, the knowledge module, and the completion module) affect the performance.

DVP. DVP in BIFRNet generates the spatial attention, which distinguishes occluders from target objects. We visualize the attention map P in Figure 7. The actual size of the attention map is 7×7, and we upsample them to the same size of input images for convenient observation. Figure 7 visualizes a car image from the testing set of Occluded-Vehicles under different occluders at L2 occlusion level as well as the corresponding attention map. The bright regions in the attention map have higher weights than the darker regions, which indicates that DVP learns to discriminate occluded and non-occluded features. Moreover, DVP can be generalized to tackle real-world occlusion. Detailed study can be found in the supplementary material.

Knowledge. Here we conduct two experiments to evaluate the knowledge module. In the first experiment, we perturb $\mathcal{K}$ by a random matrix sampled from the normal distribution $N(\mu, \sigma)$. We denote the first experiment as $\mathcal{K} + N(\mu, \sigma)$ and report the results in Table 4. It is observed that the recognition accuracy gradually decreases as the random perturbation increases, which indicates that the knowledge

affects BIFRNet's recognition results. In the second experiment, we remove the knowledge module and retrain BIFRNet. This experiment is denoted as BIFRNet-K and the corresponding results in Table 4 demonstrate that the knowledge module is important for BIFRNet to achieve advanced performances, especially at high occlusion level (L3). Furthermore, we observe that changing $\mathcal{K}$ will achieve top-down knowledge modulated recognition , which is discussed in the supplementary material.

Completion. We studied the completion module from two aspects. At first, we tested BIFRNet by cutting off the completion module, that is, using F_v as the classification feature. The results are shown in Table 5, noted as Completion Ablation. It can be found that BIFRNet's classification accuracy decreases, and the more serious the occlusion exists, the more the accuracy decreases. BIFRNet-Completion in Table 5 denotes removing the completion module and retraining BIFRNet. Results show that the recognition performance reduces significantly when serious occlusion occurs. It indicates that the restored features is more robust for occlusion than the visible features.

Conclusion

This paper proposes a novel brain-inspired feature restoration network BIFRNet. It achieves higher performance than the previous methods for POIR problem on both the synthetic and real-world occluded image datasets. All the components of BIFRNet make contributions to this superior performance. The visual pathways provide BIFRNet with image features, the knowledge module provides prior object knowledge, and the completion module restores the occluded features. Although we use occluded images to train BIFRNet, experiments show that BIFRNet is also robust to the occluders it has never seen. BIFRNet can be trained and tested end-to-end, which is convenient for actual applications. We state that a robust occlusion framework should have the abilities of occlusion discrimination, prior knowledge, and feature restoration. This brain-inspired framework may provide instructive ideas for the future research on POIR and other POIR-like problems.

Acknowledgments

References

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Banerjee, S.; and Das, S. 2020. SD-GAN: Structural and Denoising GAN reveals facial parts under occlusion. *arXiv preprint arXiv:2002.08448*.

Chen, P.; Liu, W.; Dai, P.; Liu, J.; Ye, Q.; Xu, M.; Chen, Q.; and Ji, R. 2021. Occlude them all: Occlusion-aware attention network for occluded person re-id. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11833–11842.

Das, S.; Kishore, P. S. R.; and Bhattacharya, U. 2020. An End-To-End Framework For Pose Estimation Of Occluded Pedestrians. In *2020 IEEE International Conference on Image Processing (ICIP)*, 1446–1450. IEEE.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Doniger, G. M.; Foxe, J. J.; Murray, M. M.; Higgins, B. A.; Snodgrass, J. G.; Schroeder, C. E.; and Javitt, D. C. 2000. Activation Timecourse of Ventral Visual Stream Object-recognition Areas: High Density Electrical Mapping of Perceptual Closure Processes. *Journal of Cognitive Neuroscience*, 12(4): 615–621.

Eslami, S. A.; Jimenez Rezende, D.; Besse, F.; Viola, F.; Morcos, A. S.; Garnelo, M.; Ruderman, A.; Rusu, A. A.; Danihelka, I.; Gregor, K.; et al. 2018. Neural scene representation and rendering. *Science*, 360(6394): 1204–1210.

Geirhos, R.; Janssen, D. H.; Schütt, H. H.; Rauber, J.; Bethge, M.; and Wichmann, F. A. 2017. Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*.

Goodale, M. A.; and Milner, A. D. 1992. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1): 20–25.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

He, Z. J.; and Nakayama, K. 1992. Surfaces versus features in visual search. *Nature*, 359(6392): 231–233.

Johnson, J. S.; and Olshausen, B. A. 2005. The recognition of partially visible natural objects in the presence and absence of their occluders. *Vision research*, 45(25-26): 3262–3276.

Kanizsa, G. 1979. *Organization in vision: Essays on Gestalt perception*. Praeger Publishers.

Ke, L.; Tai, Y.-W.; and Tang, C.-K. 2021. Deep Occlusion-Aware Instance Segmentation with Overlapping BiLayers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4019–4028.

Kellman, P. J.; and Shipley, T. F. 1991. A theory of visual interpolation in object perception. *Cognitive psychology*, 23(2): 141–221.

Kim, B.; Reif, E.; Wattenberg, M.; Bengio, S.; and Mozer, M. C. 2021. Neural networks trained on natural scenes exhibit gestalt closure. *Computational Brain & Behavior*, 4(3): 251–263.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *Computer Science*.

Kortylewski, A.; He, J.; Liu, Q.; Cosgrove, C.; Yang, C.; and Yuille, A. L. 2021. Compositional Generative Networks and Robustness to Perceptible Image Changes. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, 1–8.

Kortylewski, A.; Liu, Q.; Wang, H.; Zhang, Z.; and Yuille, A. 2020. Combining compositional models and deep networks for robust object classification under occlusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1333–1341.

Lee, T. S. 2003. Computations in the early visual cortex. *Journal of Physiology-Paris*, 97(2-3): 121–139.

Liao, R.; Schwing, A.; Zemel, R. S.; and Urtasun, R. 2016. Learning deep parsimonious representations. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 5083–5091.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, L.; Xu, W.; Lu, C.; et al. 2020. FPHA-Afford: A Domain-Specific Benchmark Dataset for Occluded Object Affordance Estimation in Human-Object-Robot Interaction. In *2020 IEEE International Conference on Image Processing (ICIP)*, 1416–1420. IEEE.

Mallikarjuna, P.; Targhi, A. T.; Fritz, M.; Hayman, E.; Caputo, B.; and Eklundh, J.-O. 2006. The kth-tips2 database. *Computational Vision and Active Perception Laboratory, Stockholm, Sweden*, 11.

Masi, I.; Wu, Y.; Hassner, T.; and Natarajan, P. 2018. Deep Face Recognition: A Survey. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 471–478.

Maunsell, J. H. R.; and Cook, E. P. 2002. The Role of Attention in Visual Processing. *Philosophical Transactions: Biological Sciences*, 357(1424): 1063–1072.

Michotte, A.; Thinès, G.; and Crabbé, G. 1991. Amodal completion of perceptual structures. *Michotte's experimental phenomenology of perception*, 140–167.

Nguyen, A.; Yosinski, J.; Bengio, Y.; Dosovitskiy, A.; and Clune, J. 2016. Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space. *Retina-Vitreus*, 21(3): 166–177.

Pang, Y.; Xie, J.; Khan, M. H.; Anwer, R. M.; Khan, F. S.; and Shao, L. 2019. Mask-guided attention network for occluded pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4967–4975.

Peters, B.; and Kriegeskorte, N. 2021. Capturing the objects of vision with neural networks. *Nature Human Behaviour*, 5(9): 1127–1144.

Qiu, H.; Gong, D.; Li, Z.; Liu, W.; and Tao, D. 2021. End2End occluded face recognition by masking corrupted features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sekuler, A. B.; and Palmer, S. E. 1992. Perception of partly occluded objects: A microgenetic analysis. *Journal of Experimental Psychology: General*, 121(1): 95.

Serences, J. T.; Schwarzbach, J.; Courtney, S. M.; Golay, X.; and Yantis, S. 2004. Control of Object-based Attention in Human Cortex. *Cerebral Cortex*, 14(12): 1346–1357.

Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.

Shore, D. I.; and Enns, J. T. 1997. Shape completion time depends on the size of the occluded region. *Journal of Experimental Psychology: Human Perception and Performance*, 23(4): 980.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, L.; Gong, D.; Li, Z.; Liu, C.; and Liu, W. 2019a. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 773–782.

Song, L.; Gong, D.; Li, Z.; Liu, C.; and Liu, W. 2019b. Occlusion Robust Face Recognition Based on Mask Learning With Pairwise Differential Siamese Network. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 773–782.

Tabernik, D.; Kristan, M.; Wyatt, J. L.; and Leonardis, A. 2016. Towards deep compositional networks. In *2016 23rd international conference on pattern recognition (ICPR)*, 3470–3475. IEEE.

Tse, P. U. 1999. Volume completion. *Cognitive psychology*, 39(1): 37–68.

Ungerleider, S. K.; and G., L. 2000. Mechanisms of Visual Attention in the Human Cortex. *Annual Review of Neuroscience*, 23(1): 315–341. PMID: 10845067.

Wan, W.; and Chen, J. 2017. Occlusion robust face recognition based on mask learning. In *2017 IEEE international conference on image processing (ICIP)*, 3795–3799. IEEE.

Wang, J.; Xe, C.; Zhang, Z.; Zhu, J.; Xie, L.; and Yuille, A. L. 2017. Detecting Semantic Parts on Partially Occluded Objects. Technical report, Center for Brains, Minds and Machines (CBMM).

Weigelt, S.; Singer, W.; and Muckli, L. 2007. Separate cortical stages in amodal completion revealed by functional magnetic resonance adaptation. *BMC neuroscience*, 8(1): 1–11.

Xiang, Y.; Mottaghi, R.; and Savarese, S. 2014. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, 75–82. IEEE.

Xiao, M.; Kortylewski, A.; Wu, R.; Qiao, S.; Shen, W.; and Yuille, A. 2019. Tdapnet: Prototype network with recurrent top-down attention for robust object classification under partial occlusion. *arXiv preprint arXiv:1909.03879*.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2018. Generative Image Inpainting with Contextual Attention. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yuan, X.; and Park, I. K. 2019. Face De-Occlusion Using 3D Morphable Model and Generative Adversarial Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6023–6032.

Zhang, S.; Yang, J.; and Schiele, B. 2018. Occluded pedestrian detection through guided attention in cnns. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 6995–7003.

Zhang, Z.; Xie, C.; Wang, J.; Xie, L.; and Yuille, A. L. 2018. Deepvoting: A robust and explainable deep network for semantic part detection under partial occlusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1372–1380.

Zou, T.; Yang, S.; Zhang, Y.; and Ye, M. 2020. Attention guided neural network models for occluded pedestrian detection. *Pattern Recognition Letters*, 131: 91–97.