Common Object Segmentation in Dynamic Image Collection using Attention Mechanism

by

Sana Baid
202011019

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY
in
INFORMATION AND COMMUNICATION TECHNOLOGY
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

June, 2022

Declaration

I hereby declare that

i) the thesis comprises of my original work towards the degree of Master of Technology in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,

ii) due acknowledgment has been made in the text to all the reference material used.

Saná Baid

Certificate

This is to certify that the thesis work entitled *Common Object Segmentation in Image Groups* has been carried out by **Sana Baid (202011019)** for the degree of Master of Technology in Information and Communication Technology at *Dhirubhai Ambani Institute of Information and Communication Technology* under my/our supervision.

Dr Avik Hati

Acknowledgments

I would like to take this opportunity to express deep gratitude to my supervisor, Prof. Avik Hati for providing constant guidance, support and motivation needed for my research work. Without his continuous help and trust, this thesis would not have been concluded.

I would like to thank my parents and all my peers who have boosted my morale at every step. It would have been a tedious journey without their constant motivation.

Lastly, I thank the Almighty for being always there for me.

Contents

Abstract

Semantic segmentation of image groups is a crucial task in computer vision that aims to identify shared objects in multiple images. This work presents a deep neural network framework that exhibits congruity between images, thereby co-segmenting common objects. The proposed network is an encoder-decoder network where the encoder extracts high-level semantic feature descriptors and the decoder generates segmentation masks. The task of co-segmentation between the images is boosted by an attention mechanism that leverages semantic similarity between feature descriptors. This attention mechanism is responsible for understanding the correspondence between the features, thereby determining the shared objects. The resultant masks localize the shared foreground objects while suppressing everything else as background. We have explored multiple attention mechanisms in 2 image input setup and have extended the model that outperforms the others for dynamic image input setup. The term dynamic image connotes that varying number of images can be input to the model, simultaneously, and the result will be the segmentation of common object from all of the input images. The model is trained end-to-end on image group dataset generated from the PASCAL-VOC 2012 [7] dataset. The experiments are conducted on other benchmark datasets as well and we can infer superiority of our model from the results achieved. Moreover, an important advantage of the proposed model is that it runs in linear time as opposed to quadratic time complexity observed in most works.

List of Tables

List of Figures

CHAPTER 1
Introduction

1.1 Image Segmentation

Image segmentation is one of the most fundamental problems in computer vision [3]. Image is segregated into smaller segments which are termed image groups. These segments are formed based on either similarity or discontinuity. There are various groups of image segmentation namely semantic segmentation, instance segmentation and panoptic segmentation. Semantic segmentation is a pixel wise classification of image. The aim is to assign a class or label to each pixel of an image by devising methodologies that partition the image into several semantic segments. It invariably draws a boundary around different objects present in an image. Foreground-Background segmentation is a special type of semantic segmentation that classifies the pixels belonging to specific class or classes as foreground and everything else as background. Segmentation finds its use-case in critical applications like autonomous driving, robotic navigation, localization, and scene understanding.

1.2 Common Object segmentation in Image group

The semantic segmentation of image groups [13] [22] is a task of segmentation of common objects from multiple images. It can be considered as a similarity measure between multiple images. Figure 1 shows segmentation of image pairs with and without common objects present. Many computer vision applications, such as interactive image segmentation, 3D reconstruction and object co-localization, to name a few, require this method to enhance their results.

The main idea behind joint segmentation of image groups is to identify the synergistic relation between the images and localise the common objects between them [13]. Significant work has been done on unsupervised models [18] [10] [8], however, their performance is dependent on the image feature selection and tun-

(a) An image pair with common object and their co-segmentation masks

(b) An image pair with no common object present and their co-segmentation masks

Figure 1.1: Co-segmentation results on PASCAL-VOC 2012 [7] dataset

ing methods. Image features are the most important descriptors that can model the commonalities between the images, and hence, it is vital that these features are as unerring as they can be. Traditional feature representation techniques like color histograms, SIFT descriptors [21, 12] cannot handle object scale variations and background clutter issues.

Recent years have seen a rapid surge in the use of deep learning models namely U-NET [6] and DeepLab [16] in segmentation architectures as well as feature extraction architectures and have shown superior results.Along with the success of deep learning models, the attention mechanism [20] has also boosted the performance of several deep learning models. The attention mechanism is based on the perceptual mechanism of human brain and eyes, and helps the model to focus on only the needed part of the image, which in this case is the common object.

1.3 Motivation

Segmentation of common objects in multiple images finds its application in various computer vision areas like 3D reconstruction, object co-localisation etc. Noteworthy work has already been done in common object segmentation from two images, however, for more than two images this area has been meekly explored. This forms the motivation of my thesis to devise a multiple model architecture

that aims at segmentation of common objects for a dynamic image collection. The term dynamic here implies that the model can take variable number of images in one group for joint segmentation.

1.4 Objective

The main objectives of my thesis are as follows:

- To develop a robust model that takes in variable number of images and generates segmentation masks that localise the common foreground objects.
- To draw comparison of proposed model with state-of-art baseline models and outperform them.

1.5 Main Contribution

The main contribution of my thesis are as follows:

- We have proposed four different attention models as bottleneck in a Siamese Encoder-Decoder model for a 2 image collection setup. We have drawn comparison between each model for segmentation results.
- We have extended the 2 image model to handle dynamic image input setup. We have proposed a recursive attention model to achieve this.
- Furthermore, we have extended the model with a classification head to handle negative image pair i.e. image group that has no common object.

1.6 Thesis Outline

This thesis is organised in 5 chapters

- Chapter 2 contains literature survey on single as well as multiple image segmentation.
- Chapter 3 presents the experimental setup and research methodology.
- Chapter 4 contains the results of the experiments and related discussion.
- Chapter 5 concludes the thesis and defines scope of future work.

CHAPTER 2
Related Work

This section entails the work done in the field of single image as well multiple image segmentation.

2.1 Single Image Segmentation

Many state-of-the-art deep learning architectures like U-net and DeepLab are built on the underlying encoder-decoder model to yield excellent image segmentation results. The encoder model exctracts high level semantic features of images at lower resolutions. These feature capture the semantic information of the image and thereby enhance the segmentation process of the image. The decoder scales back the lower resolution feature maps to the original image size along with segmentation results.U-Net [16] identifies that the spatial information is lost in the decoding phase in a basic encoder-decoder segmentation structure and so introduces skip connections to retain spatial information while decoding the image features.

DeepLab v3 [6] highlights two important challenges in segmentation tasks. First issue entails that due to mutliple pooling operations for the downsampling of image in deep convolutional networks the spatial information is lost which hampers segementation results. To overcome this issue, the model has removed downsampling from the last layers of the encoder layer and have upsampled the filter kernels by using atrous convolutions. The second issue is that objects exist at different scales in the image which makes the segmentation slightly difficult. To overcome this issue the paper has defined an approach called the spatial pyramid pooling, that applies filters at multiple atrous rates and multiple effective field-of-views on the feature maps, thereby identifying the objects at multiple scales. With these developments the DeepLab v3 model has acheived much enhanced results as compared to its previous models. It is also used in many state-of-the-

art segmentation and co-segmenatation model as feature extractor or backbone model.

2.2 Segmentation of Image Groups

Rother et al. [17] first proposed the term co-segmentation as the method of segmenting the common objects of an image pair simultaneously. They concluded that the segmentation achieved with more than one image was far more superior than single image segmentation. Post that, significant work have been done on multiple image segmentation, mostly 2 input image. Banerjee et al. [2] proposed a class-agnostic, twofold deep neural network (siamese network) architecture that segments common objects between two images. The masks generated for both the images have common objects classfied as foreground and rest of the image as background. The network is an encoder-decoder architecture, where VGG-16 architecture is adopted for the encoder to fetch high level semantic features from the images. Here, the decision network is of key importance as it tells whether there are common objects between the two images or not. Li et al. [13] proposed a robust, novel recurrent deep neural network that handle the variation of co-object in appearance.

Models proposed in [15] [14] and [22] also adopt similar encoder-decoder architecture, however, in [14] the image pairs are selected such that there always exists only one common object between them. For identification of common object, a mutual correlation network is introduced between encoder and decoder that performs feature matching over the encoded feature space produced. This correlation network helps to identify the common object in the image pairs.

Kim et al. [11] proposed a clustering approach that eliminates intra-class heterogeneity and builds intra and inter image connections. Hierarchial clustering is implemented that leverages multiple clustering to gain inter image connections and thereby co-segment images simultaneously. Joulin et al. [9] also proposed clustering based method for images with both low and high intra-class variations. They have used traditional bottom-up segmentation methods involving kernels, normalization cuts etc. within discriminitive clustering method based on positive definite kernels to obtain forground-background segmentation.

Chen et al. [4] proposed Siamese encoder-decoder architecture, however, in this

approach an attention learner is selected for the bottleneck layer. The main motivation behind this approach is to leverage the semantic and spatial similarity between the image pairs. They have proposed three different architectures for attention mechanism. The have claimed that the background and appearance of the objects in the image may not always be consistent and so the key to get accurate results is by leveraging semantic relations between images.

In [22] the authors have generated multi-resolution features for images. The coarse localisation of the common objects are achieved via supervised semantic and unsupervised spatial network modulators. The images are fed to multi resolution backbone structures to yield multi resolution features. These are passed to semantic and spatial modulators. The spatial modulator uses an unsupervised clustering approach to generate heat maps. The semantic modulator uses 2nd order pooling to get semantic relations between set of images. A classification module is also added to classify the presence of multiple classes in the images in order to handle mutliple common object case.

CHAPTER 3
Approach

3.1 Model Architecture

The architecture adopted for the segmentation task is an encoder-decoder architecture with the bottleneck of attention mechanism. Since, the segmentation task is for more than one images, we naturally require siamese encoder-decoder model. Siamese model refers to a model that has same weights when its working with more than one input images alongside each other. The overview of the general model is presented in Figure 3.1. The model can be divided into three main modules, Encoder, Attention and Decoder. Each of the modules are elaborated in the following subsections.

3.1.1 Siamese Encoder

The first part of the architecture is a siamese encoder that contains identical feature extraction convolution layers with shared parameters. The main objective of the encoder is to extract image features. The popularity of VGG-16 [19] motivated us to use its feature extraction module. The VGG-16 has 5 convolution blocks, comprising of combination of convolutions, ReLU and max pooling functions. Feature extractor module of VGG-16, pretrained on the ImageNet dataset, is used to initialize the encoder to produce low resolution, high level semantic information feature maps for the respective images. These feature maps are passed to a global average pooling module (GAP) to reduce the dimensions of the feature maps. This is achieved by pooling the channel information by spatially averaging each channel. These pooled features are then passed to the channel attention module.

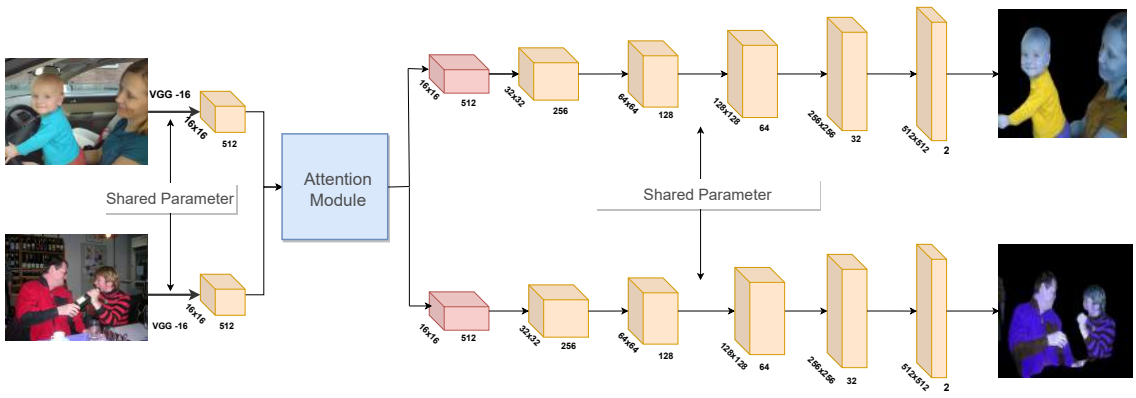

Figure 3.1: Base Architecture of the Segmentation model

3.1.2 Attention Module

The main motivation behind leveraging the attention mechanism as the bottleneck for encoder-decoder model is to recognise the semantic similarity between images. Attention mechanism will train the model to keep attention on feature channels that have high activation in all input images, and suppress other irrelevant feature channels. The high activation feature channels will invariably correspond to the image regions that have common objects. Thus, the model will learn to identify the properties of common objects.

There are two attention modules that are used, in different combinations, in the proposed model. First is the channel attention module that learns semantic similarity by activating those channels that are common between the image pairs. Second is the spatial attention module that learns to localise the common object in the respective image. The channel attention module is comprised of 2 fully connected layers separated by ReLU activation function. At last there is sigmoid activation which constricts the activation values in the range of 0-1. The spatial attention layer also has a similar structure with just one difference, the fully connected layers are replaced with convolution layers.

3.1.3 Siamese Decoder

As mentioned earlier, the output from the channel attention module contains the common object information. Hence for the task of common object extraction, we can use this to modulated the encoder features of the respective images. Thus, this output is up-sampled using bi-linear interpolation, and then pixel wise mul-

tiplied with the feature maps images. These fused feature maps are then passed to the Siamese decoders. The decoder consists of 5 blocks of deconvolution blocks, where each block has several deconvolution layers followed by ReLU activation except the last block. These blocks up-samples the output channel attention maps to localise the common objects as per the original image dimensions. Drop out layers are added after each up-sampling to avoid over fitting. After the last decoder block, there is a softmax layer with two channels that segments the foreground and background regions of the image group.

3.2 Image Pair Segmentation

Based on the base model, we have proposed four different kind of implementation for the attention modules for 2-image input. Chen et al. [4] have proposed three channel and spatial attention bottlenecks explained in the baseline section below. On those attention models we have proposed changes to boost the results. The proposed changes are elaborated after the baseline model.

3.2.1 Baseline

Chen et al. [4] have proposed encoder-decoder architecture with three combinations of channel and spatial attention as bottlenecks. We will elaborate on the Fused Channel Attention Module (FCA) which forms the basis of the modifications we have introduced.

Figure 3.2 outlines the FCA's channel attention architecture. The feature maps from the encoder for the two input images of spatial dimension of 16x16 and 512 channel features are passed to the channel attention layer. These are then passed to a global average pooling layer that performs spatial averaging for each channel feature to create 1x1x512 dimension pooled features. Then, they are individually passed to 2 fully connected neural networks and then finally to a sigmoid activation function that learns to activate features that are important and suppress those channels that are not important. These activated channel outputs of the two images, from the attention learners, are added with each other. The authors have claimed that the addition of the results will further activate only those channels or features that are common between the two images. This added channel output is then pixel wise multiplied to all the spatial pixels of the respective channels of the feature maps of the images. This is then passed to decoder to generate the segmentation maps.

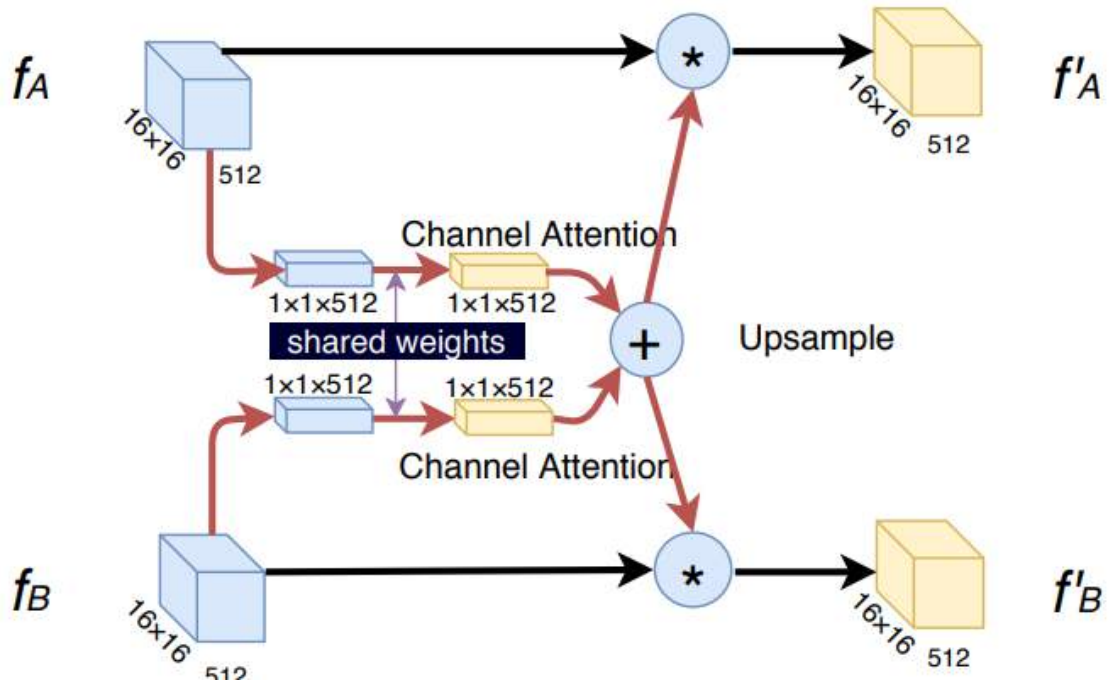

Figure 3.2: Fused Channel Attention (FCA) Model proposed by Chen et al. [4]

3.2.2 Proposed Changes

Proposed Method 1 with Channel Attention Module (M1C)

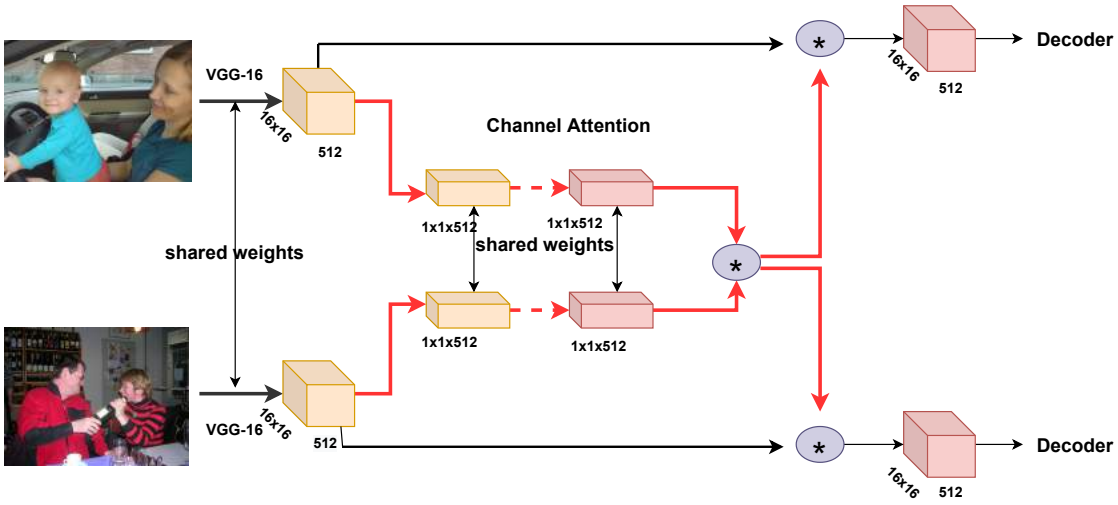

Figure 3.3: Architecture of M1C model

As the name suggests, M1C is similar to the M1C model proposed in [4] with a critical modification. After the pooled feature maps are passed to the fully con-

nected network and sigmoid activation function they are pixel wise multiplied instead of being added as in M1C. These multiplied activation values are the attention maps of the respective image. The reason for replacing addition with multiplication is to increase the gap between attention's activation values of the common features and the divergent features. The features present in one of the image will have higher value at the corresponding channel activation of that image but relatively low in the other image's channel activation. In case of addition, these values will add up to give relatively higher value, however, if we multiply them this value will be less. Therefore, intuitively, the values of common features in the channel activation will be closer to 1 and for the divergent features, it will be closer to 0. Finally, when this fused attention map will be multiplied with the feature maps the divergent features will be suppressed, thereby highlighting the common features with more effect.The entire architecture of M1C model is pictorially represented in Figure 3.3 and mathematically represented in Equation (3.1).

$$
\begin{aligned}
\boldsymbol{\alpha}_A^c &= \sigma\left(W^T * \text{ AvgPool}_{\text{channel}}\ (f_A) + b\right) \\
\boldsymbol{\alpha}_B^c &= \sigma\left(W^T * \text{ AvgPool}_{\text{channel}}\ (f_B) + b\right) \\
\boldsymbol{\alpha}_C &= \boldsymbol{\alpha}_A^c * \boldsymbol{\alpha}_B^c \\
f'_A &= \boldsymbol{\alpha}_C * f_A \\
f'_B &= \boldsymbol{\alpha}_C * f_B
\end{aligned}
\tag{3.1}
$$

where:

f_A, f_B = feature maps of image pair (A,B)
α_A^c, α_B^c = channel attention features maps
f'_A, f'_B = attended feature maps
W^T = weight
b = bias
σ = sigmoid

Proposed Method 1 with Channel and Spatial Attention Modules (M1CS)

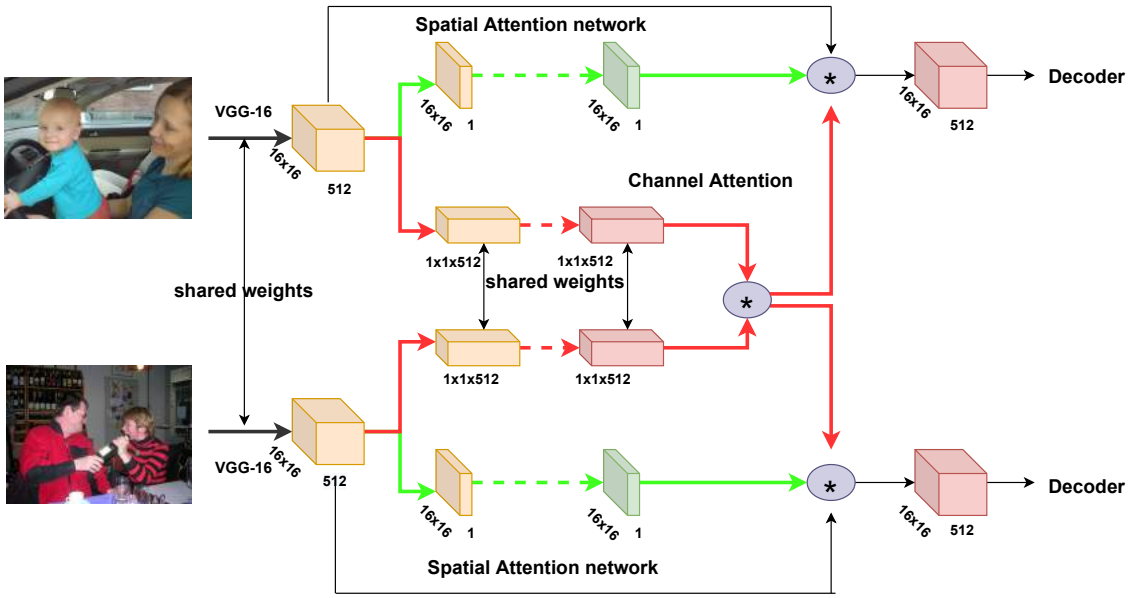

Figure 3.4: Architecture of M1CS model

In the M1C model the fused attention maps do not take into consideration the spatial information. They focus on finding only the semantic similarity between the images. Hence, a spatial attention layer, denoted by the green doted lines, is added. This attention layer constitutes of 2 convolutional modules followed by sigmoid activation function. They output spatial attention maps which is then pixel wise multiplied across all the channels. The channel attention activation maps are, as earlier, multiplied with all the spatial pixels of the respective channel. The entire architecture of M1CS model is pictorially represented in Figure 3.4 and mathematically represented in Equation (3.2).

$$
\begin{aligned}
\boldsymbol{\alpha}_A^c &= \sigma\left(W^T * \text{ AvgPool}_{\text{ channel}}\ (f_A) + b\right) \\
\boldsymbol{\alpha}_B^c &= \sigma\left(W^T * \text{ AvgPool}_{\text{ channel}}\ (f_B) + b\right) \\
\boldsymbol{\alpha}_A^s &= \sigma\left(Conv\left(\text{ AvgPool}_{\text{ spatial}}\ (f_A)\right)\right) \\
\boldsymbol{\alpha}_B^s &= \sigma\left(Conv\left(\text{ AvgPool}_{\text{ spatial}}\ (f_B)\right)\right) \\
\boldsymbol{\alpha}_C &= \boldsymbol{\alpha}_A^c * \boldsymbol{\alpha}_B^c \\
f_A' &= \boldsymbol{\alpha}_C * f_A * \boldsymbol{\alpha}_A^s \\
f_B' &= \boldsymbol{\alpha}_C * f_B * \boldsymbol{\alpha}_B^s
\end{aligned}
\tag{3.2}
$$

where:

f_A, f_B = feature maps of image pair (A,B)
α^c_A, α^c_B = channel attention features maps
α^s_A, α^s_B = spatial attention features maps
f'_A, f'_B = attended feature maps
W^T = weight
b = bias
σ = sigmoid

Proposed Method 2 with Channel Attention Module (M2C)

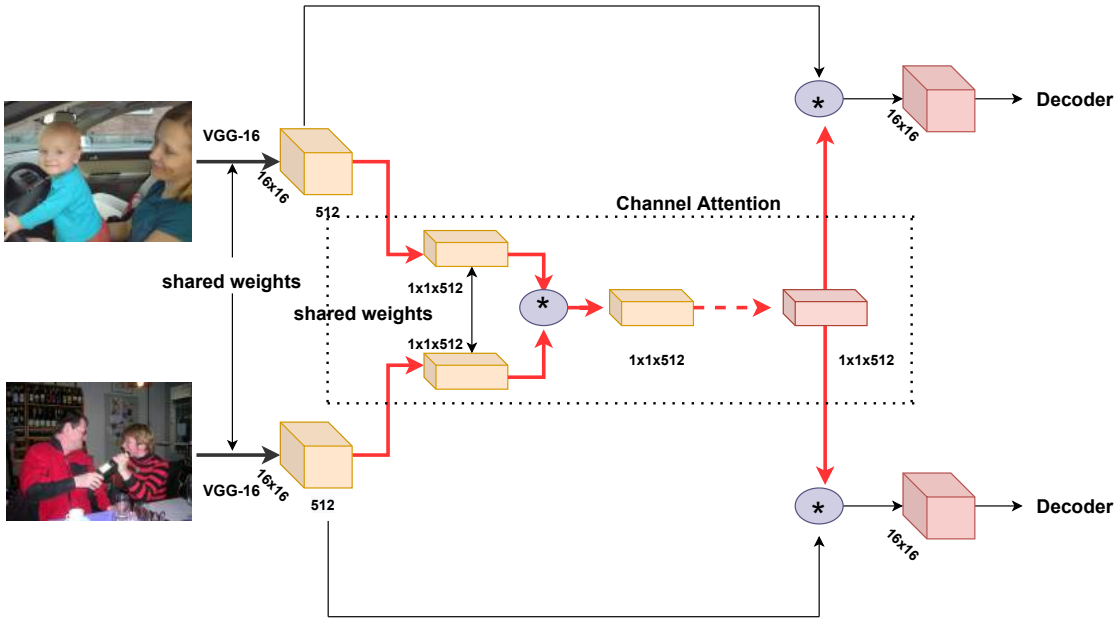

Figure 3.5: Architecture of M2C model

This approach focuses on reducing the network complexity of the channel attention without losing vital information. Instead of fusing the attention's activation features, the feature maps extracted from the VGG-16 encoder are pixel wise multiplied together and the resultant feature map is fed to the channel attention network. Apart from reduced complexity, another motivation behind this model is that concatenation of the feature maps will ensure that common semantic information's i.e. objects in the features maps value will be increased and the ones that are different will not add to the other, thereby boosting performance of the learner. The attention mechanism will produce a single channel activation map which is multiplied with all the spatial pixels of the respective channel of both the image's

encoded feature maps. The entire architecture of the M2C model is pictorially represented in Figure 3.5 and mathematically represented in Equation (3.3).

$$
\begin{aligned}
\boldsymbol{f} &= \text{AvgPool}_{\text{channel}}\left(f_A\right) * \text{AvgPool}_{\text{channel}}\left(f_B\right) \\
\boldsymbol{\alpha}^c &= \sigma\left(W^T f + b\right) \\
f'_A &= \boldsymbol{\alpha}^c * f_A \\
f'_B &= \boldsymbol{\alpha}^c * f_B
\end{aligned}
\tag{3.3}
$$

where:

f_A, f_B = feature maps of image pair (A,B)
f'_A, f'_B = attended feature maps
W^T = weight
b = bias
σ = sigmoid

Proposed Method 2 with Channel and Spatial Attention Modules (M2CS)

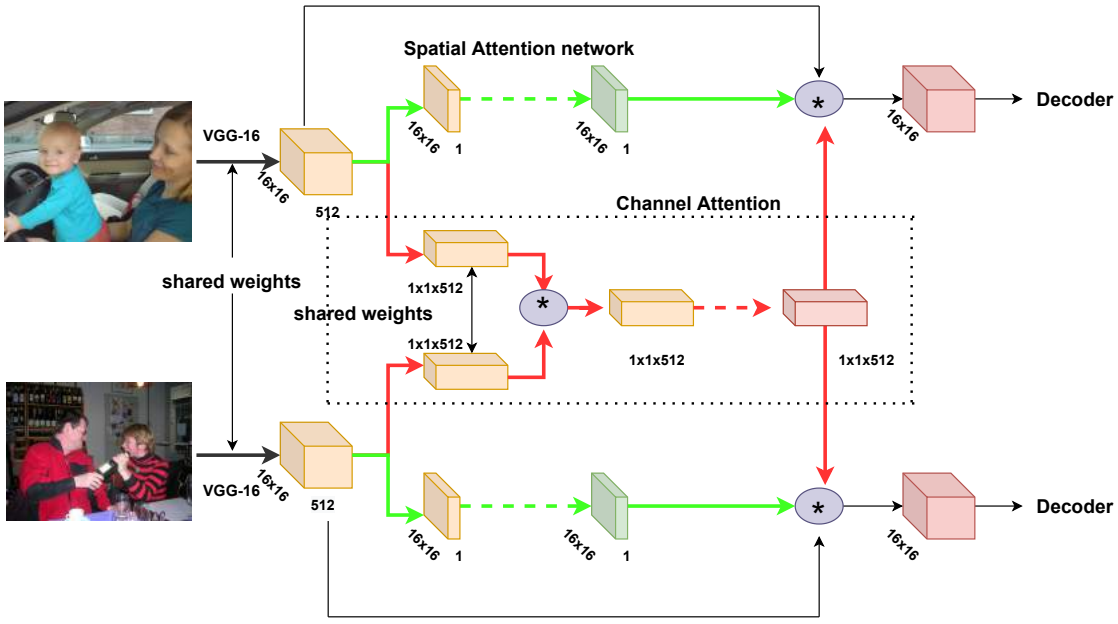

Figure 3.6: Architecture of M2CS model

In the M2C model the fused attention maps do not take into consideration the spatial information. They focus on finding only the semantic similarity between the images. Hence, a spatial attention layer, denoted by the green doted lines, is

added. This attention layer constitutes of 2 convolutional modules followed by sigmoid activation function. They output spatial attention maps which is then pixel wise multiplied across all the channels. The channel attention activation maps are, as earlier, multiplied with all the spatial pixels of the respective channel. The entire architecture of M2CS model is pictorially represented in Figure 3.6 and mathematically represented in Equation (3.4).

$$
\begin{aligned}
\boldsymbol{f} &= \text{AvgPool}_{\text{channel}}\ (f_A) * \text{AvgPool}_{\text{channel}}\ (f_B) \\
\boldsymbol{\alpha}^c &= \sigma\left(W^T f + b\right) \\
\boldsymbol{\alpha}^s_A &= \sigma\left(Conv\left(\text{AvgPool}_{\text{spatial}}\ (f_A)\right)\right) \\
\boldsymbol{\alpha}^s_B &= \sigma\left(Conv\left(\text{AvgPool}_{\text{spatial}}\ (f_B)\right)\right) \\
f'_A &= \boldsymbol{\alpha}^c * f_A * \boldsymbol{\alpha}^s_A \\
f'_B &= \boldsymbol{\alpha}^c * f_B * \boldsymbol{\alpha}^s_B
\end{aligned}
\tag{3.4}
$$

where:

f_A, f_B = feature maps of image pair (A,B)
α^s_A, α^s_B = spatial attention features maps
f'_A, f'_B = attended feature maps
W^T = weight
b = bias
σ = sigmoid

3.3 Semantic Modulator

Zhang et al. [22] proposed semantic modulator as channel attention mechanism along with an unsupervised spatial modualtor network for the segmentation task. We have integrated their channel attention mechanism with our encoder-decoder architecture as the channel attention module. This channel attention module is two fold, and the features here are directly taken from the encoder. The modulator, called SP(Spatial pooling) block, is shown in Figure 3.8. The features are passed to convolution layer to reduce the number of channels for better computations. The resultant feature map are passed onto a pooling layer that extracts higher order statistics. The resulting enhanced features from this SP layer for each image is concatenated. This concatenated enhanced features are again passed to the SP layer, in order to capture long range dependency between the features to

yield a activation value for each channel. This entire module is termed as HSP (Hierarchical second-order pooling) block and is pictorially represented in Figure 3.7

Additionally, in order to identify presence of multiple common classes in images, classification module is introduced here. The results from the channel modulator are passed to a fully connected layer followed by a sigmoid function to learn the co-category classes present. The motivation behind adding this classifier is to enable the model to learn presence of multiple classes in the images.

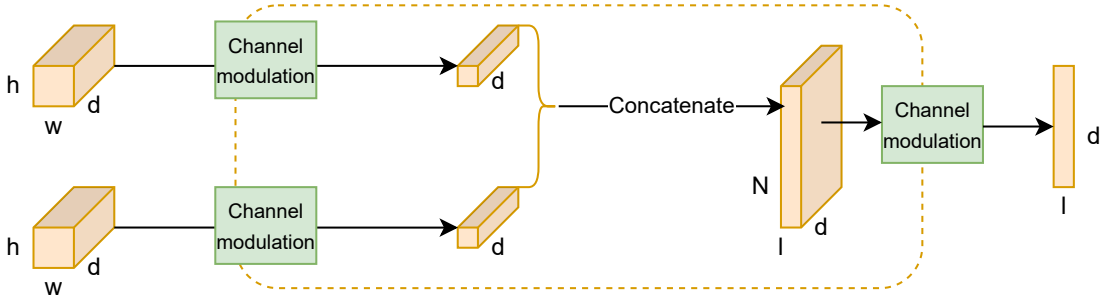

Figure 3.7: HSP(Hierarchical second-order pooling) model

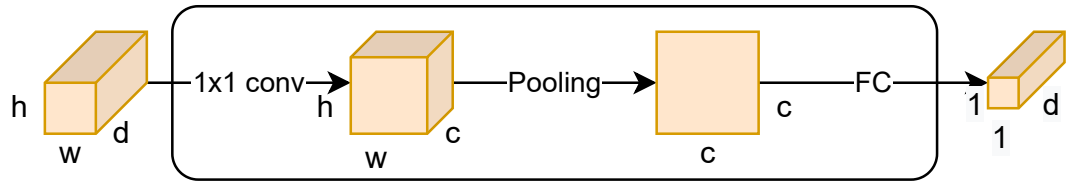

Figure 3.8: SP(Spatial pooling) model

3.4 Handling negative image pairs

Until now, the image pairs taken for training as well as for testing of the model have at least one object common amongst them. Such image pairs are termed positive image pairs. However, there can be cases where the image pairs have no object in common, which are called negative image pairs. To check whether our model can handle the negative image pairs, we have conducted an experiment. We have tested the M2C model with dataset comprising of 50:50 mix of positive and negative PASCAL-VOC 2012 image pairs. The jaccard score 4.2 obtained was

21 percent, which is deemed far too low. Therefore to take care of this scenario, we have proposed a modification in the model architecture.

We have introduced a classifier in M2C model. The Figure 3.9 shows a simplified diagram to understand the change in the architecture. We have introduced the classifier that takes input from the 4th block of the decoder and passes it to 2 fully connected layer followed by a sigmoid learning function. This classifier determines whether the image pairs have any common object between them or not. There are two learning processes in this architecture, first the segmentation task and second the classification task. The back-propagation for the segmentation task happens only if the 2 images have a common object otherwise only the classification path is back-propagated. This helps the network to learn attention weights only if the common object is present in the image.

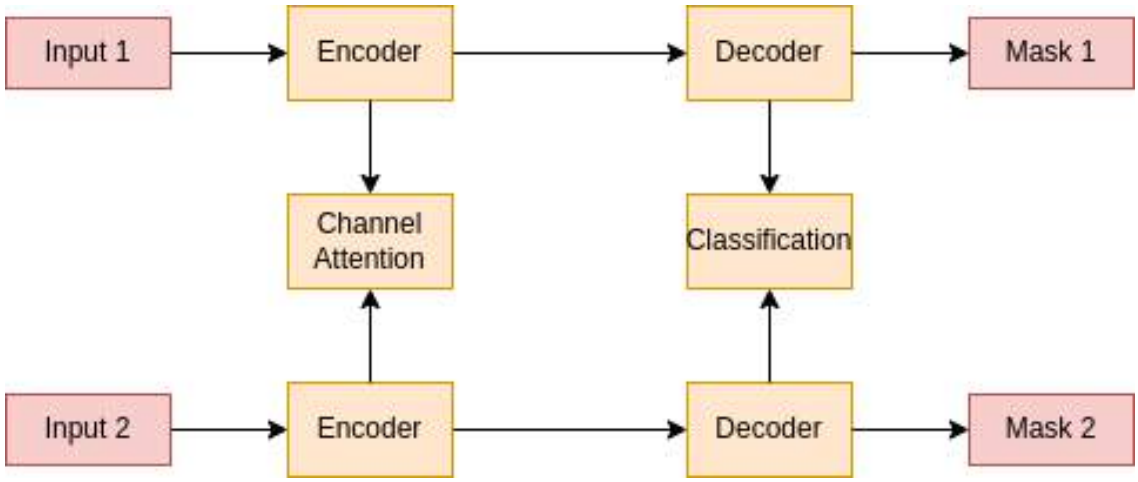

Figure 3.9: Segmentation Model with classifier

3.5 Multiple Image Segmentation Model

The multi-image segmentation model is an extension of the M2C model proposed in the previous section to incorporate dynamic number of images. It also follows the Siamese encoder-decoder architecture. The model takes a set of n images $\mathbb{I} = \{I_x\}_{x=1}^{n}$ as input, and produces corresponding masks $\{m_x\}_{x=1}^{n}$ with common objects segmented as the foreground. Figure 3.11 entails the overall architecture proposed for four images only. All the images are passed to VGG-16 based encoder to fetch feature maps F_1, F_2, F_3 and F_4 which are passed to global average pooling layer to get pooled feature maps $F_1^{'}, F_2^{'}, F_3^{'}$ and $F_4^{'}$. $F_1^{'}$ and $F_2^{'}$ are first passed to M2C attention module to get attention map G^2. This activation map is passed

as an input to the M2C along with the pooled feature map $F_3^{'}$. This gives attention map G^3, which is again passed as an input to the M2C with the pooled feature map $F_4^{'}$. The resultant attention map G^4 is then multiplied with all the feature maps of the 4 images and then passed to the decoder for further generation of the segmentation masks.

Generalised description of the above explained model is presented by the recursive model in Figure 3.10. The feature maps of n images F_n extracted from the encoder are individually passed to Global average Pooling Modules to extract $F_n^{'}$ pooled features. The recursive steps are followed with the pixel wise product of two pooled feature maps $F_1^{'}$, and $F_2^{'}$ of dimension as G_{in}^2. In every iteration x, G_{in}^x is passed to the fully connected network to produce the output G_{out}^x as given in Equation (3.5). The variables W_1 and W_2 represent the weight parameters of the two fully connected network respectively. Variables b_1 and b_2 represent the biases of the respective networks.

$$G_{out}^x = \sigma(W_2(ReLU(W_1 G_{in}^x + b_1)) + b_2), \tag{3.5}$$

The output from the fully connected network is, recursively, pixel wise multiplied with the pooled feature map of the images next in line. This is then passed again to the fully connected network. This process is continued till the feature maps of all images are incorporated. The recursive input to the module can be inferred as in Equation (3.6).

$$G_{in}^x = F_x^{'} * G_{out}^{x-1}. \tag{3.6}$$

The motivation for adopting this recursive channel attention mechanism is that there exists lot of noise in each image. This noise is due to the uncommon objects as well as background components in the images. Fusing all the feature maps as an input to the channel module will escalate the noise and result in poorly learned attention feature maps. Therefore in each iteration, we use the synergistic relationships between the explored image feature maps and the current image feature map to maximally suppress the noise data. The attention maps so learned will essentially capture common objects effectively resulting in more accurate segmentation masks. The final output from this recursive network is denoted as α and defined as $\alpha = G_{out}^n$, which is the final channel attention map for all the images in the collection. The entire algorithm is summarised in algorithm 1.

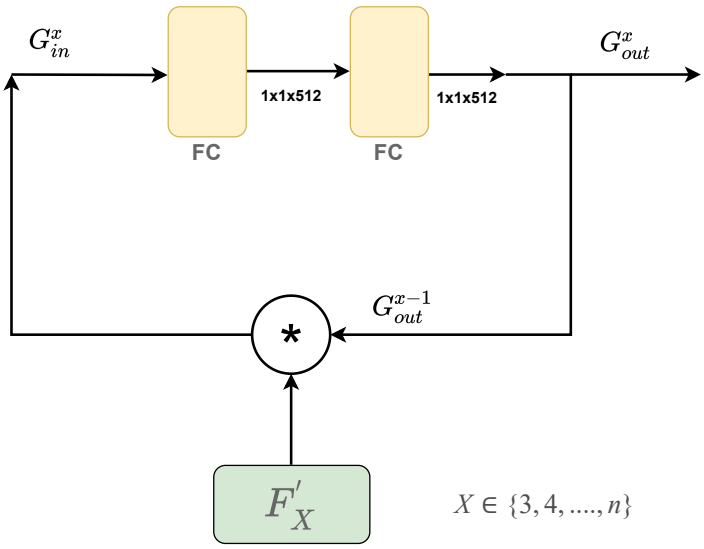

Figure 3.10: Channel Module

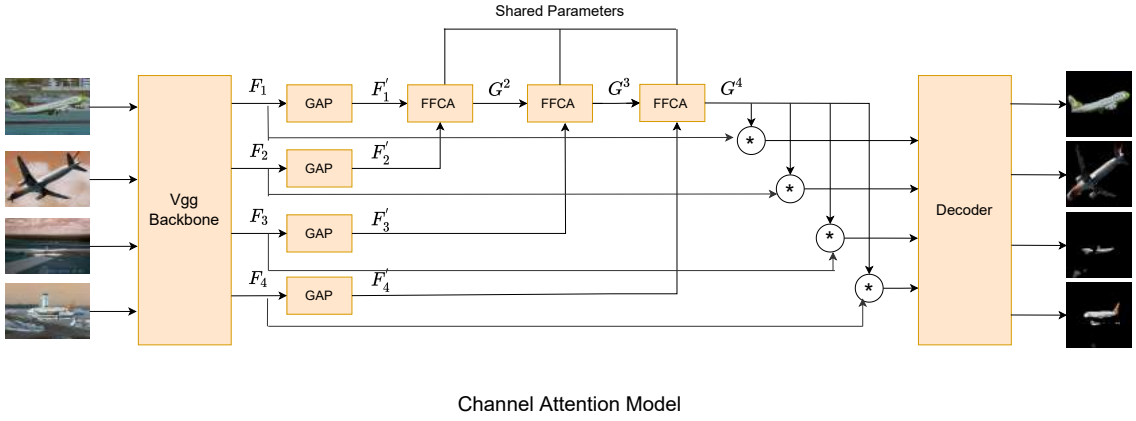

Figure 3.11: Multiple image segmentation model

Algorithm 1 Channel Attention Module

procedure CHANNELATTENTION($F'_1, F'_2, \ldots, F'_n$)
  Initialize $G^1_{out} \leftarrow F'_1$, $i \leftarrow 1$.
  for $i \leftarrow 1$ to n **do**
    $i \leftarrow i + 1$
    $G^i_{in} \leftarrow F'_i * G^{i-1}_{out}$
    $G^i_{out} \leftarrow \sigma(W_2(ReLU(W_1 G_{in} + b_1)) + b_2)$
  end for
  return G^n_{out}
end procedure

The number of images during training of the model is fixed. However, during testing stage we can give image group of any size as an input. Lets say that we have m number of input images $\mathbb{I} = \{I_x\}_{x=1}^{m}$. The entire method of generating their segmentation mask is outlined in the algorithm 2.

Algorithm 2 Multiple Image Segmentation (Test stage)

procedure MULTIPLESEGMENTATION($I_1, I_2 ... I_m$)
  for $i \leftarrow 1$ to m **do**
    Features $F_i \leftarrow VGG16(I_i)$
    Pooled features $F_i' \leftarrow GAP(F_i)$
  end for
  $\alpha = CHANNELATTENTION(F_1', F_2', \ldots, F_m')$
  for $i \leftarrow 1$ to m **do**
    Fused feature $FF_i \leftarrow F_i * \alpha$
    Masks $M_i \leftarrow DECODER(FF_i)$
  end for
  return masks
end procedure

CHAPTER 4
Experiment and Results

4.1 Dataset

The standard PASCAL VOC 2012 dataset [7] is used to train the models.There are 20 total distinguishable object classes: person, bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa and tv/monitor. There are total 2913 images with the class frequency as depicted in Figure 4.1. Image groups containing common classes are created from this dataset for training and testing purpose. For 2 image segmentation models there are two set of training dataset created. One with only positive image pairs with total of 25k image pairs and another with a 50:50 mix of positive and negative image pairs with total of 25k image pairs. The M1C, M1CS, M2C and M2CS are trained on positive image pair dataset while M2C with classifier is trained on mixed dataset. For testing purpose, positive image pair test data of size 169k and mix pair test data of size 50k is created from PASCAL VOC 2012 dataset. Moreover, these models are tested on MSRC sub dataset[1] as well. MSRC subdataset contains 7 classes: bird, car, cat, cow, dog, plane, sheep with total of 10 images in each class. Total of 315 image pairs are generated from these images for testing the models.

The multiple images segmentation model is trained on image groups of size 4. 30k sets of such image groups are created from PASCAL VOC 2012 dataset. The image groups contains images that have atleast one object common among each other. For purpose of testing we have created 7 datasets D1, D2, D3, D4, D5, D6, and D7 using images from PASCAL VOC 2012 dataset. Dataset D1, D2, D3, D4, D5, D6, and D7 comprises of image groups of size 2, 3, 4, 5, 6, 7 and 8 respectively. Each dataset has 50k sets of image groups. Individual and average jaccard score is calculated for these datasets.

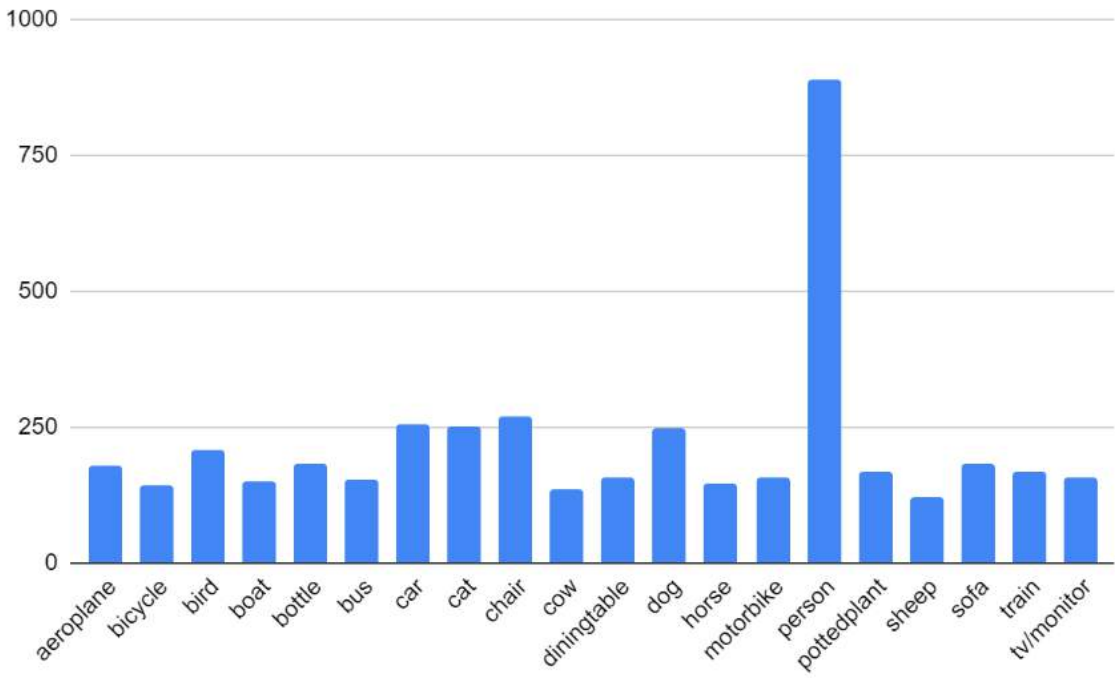

Figure 4.1: Class frequency graph of PASCAL-VOC 2012 [7] dataset.

4.2 Implementation details

The VGG-16 encoder implementation details (only convolution blocks for encoder) are given in Table 4.1. The channel attention network layers are 2 fully connected neural networks. The attended feature maps are up-sampled using deconvolution layers followed by ReLU activation units. The objective function is the cross entropy loss function Equation (4.1) and optimiser used is the Adams optimiser. The model is evaluated using the Jaccard score metric (Equation (4.2)).

Index	Layer Type	Filter
1	Convolution + ReLU	3x3x64
2	Convolution + ReLU	3x3x64
3	Max Pooling	
4	Convolution + ReLU	3x3x128
5	Convolution + ReLU	3x3x128
7	Max Pooling	
8	Convolution + ReLU	3x3x256
9	Convolution + ReLU	3x3x256
10	Convolution + ReLU	3x3x256
11	Max Pooling	
12	Convolution + ReLU	3x3x512
13	Convolution + ReLU	3x3x512
14	Convolution + ReLU	3x3x512
15	Max Pooling	

Table 4.1: Model architecture for VGG-16 convolutional blocks.

$$\mathcal{L} = -\frac{1}{2}\sum_{i=1}^{c} y_i \cdot \log\left(\hat{y}_i\right) \tag{4.1}$$

where:

c = no. of classes (foreground and background)

y_i = Probability

$$Jaccard = \sum_{i}^{\{n\}} \frac{1}{n}\frac{|m_i \cap g_i|}{|m_i \cup g_i|} \tag{4.2}$$

where:

n = Number of Images

m_i = predicted segmentation mask

g_i = ground truth segmentation mask

4.2.1 Results

Two Image Segmentation Models

The Table 4.2 compares the score for 2 image pairs for our proposed attention models with the baseline paper [4] and our implementation of the semantic mod-

Model	Pascal-VOC 2012	MSRC sub dataset
Chen [4] (CA)	74.24%	76.49%
Chen [4] (FCA)	76.41%	76.94%
Chen[4] (CSA)	74.36%	77.70%
Chen[5]	75.50%	94.30%
Semantic Modulator[22]	78.90%	77.7%
M1C	78.60%	75.45%
M1CS	78.49 %	70.67%
M2C	80.24 %	76.45%
M2CS	79.94 %	73.67%

Table 4.2: Comparison of Jaccard score between baselines Chen[4] , Chen[5] and our proposed attention models. The models are trained and on image pairs of PASCAL-VOC 2012 [7] dataset. Additionally, the models are also tested on MSRC sub dataset [1].

ulator inspired from [22]. The results prove to outperforms the baseline model results. The proposed channel models M2C, where the encoded features are multiplied and then passed onto the channel attention model outperforms the other models as well as the baseline models for PASCAL VOC 2012 dataset. For MSRC subset, the M2C model outperforms the other models and baselines except the Chen [5]. This could be because their model is trained on coco-stuff dataset which has 78 classes and more training dataset. Additionally, upon observation of the results, we collect that the presence of spatial attention mechanisms do not contribute in boosting the performance of the model.

M2C model with classifier

Attention model coupled with classifier handles the presence of negative image pairs as we can see from the results in Table 4.3. The classifier helps the model to learn to identify images that do not have common objects. This prevents the model from segmenting these images and we can directly infer the negative image pair.

Model	Jaccard Score
M2C model	21.00 %
M2C model with classifier	74.00 %

Table 4.3: Comparison of model trained with positive and positive-negative mix image pairs on PASCAL-VOC 2012 [7] dataset.

Model	Pascal-VOC 2012
Li[13]	63.00%
Zhang [22]	66.00%
Ours	68.53%

Table 4.4: Comparison of Jaccard score between Zhang[22], Li[13] and our proposed multiple image model trained and tested on PASCAL-VOC 2012 [7] image group dataset

Dataset	Jaccard Score
D1	70.23%
D2	68.43%
D3	68.51%
D4	68.33%
D5	68.54%
D6	68.78%
D7	66.90%

Table 4.5: Respective Jaccard score for D1, D2, D3, D4, D5 , D6 and D7 datasets.

Multiple Image Segmentation Model

The Table 4.4 compares the score for multiple image model with the models proposed by [22] and [13]. The results prove that our proposed multiple image segmentation model outperforms the [13] and [22] results on PASCAL VOC-2012 dataset. We have also presented the individual score for the aforementioned datasets D1, D2, D3, D4, D5, D6 and D7 in Table 4.5. Figure 4.2 presents visual results of segmentation for 4 input images. The multi image model successfully segments the common class amongst image groups. Along with the segmentation of common class objects the model also suppresses non-common objects into the background. For instance, in the Figure 4.2, the second input group the common class among the images is person and non-common class in table and car. The model has learned to put attention only to the common class in all images and therefore the generated masks have segmented the non-common objects as background.

Moreover, we have performed dynamic image input testing on our model. In Figure 4.3 the first row contains 8 images comprising the image group passed as an input to the model. First the model is tested for all the images and their resulting segmentation masks are superimposed on the images to detect common objects. The second row shows these results for 8 images. Then, one image is remove (last one) and the image group is again passed to the model to fetch seg-

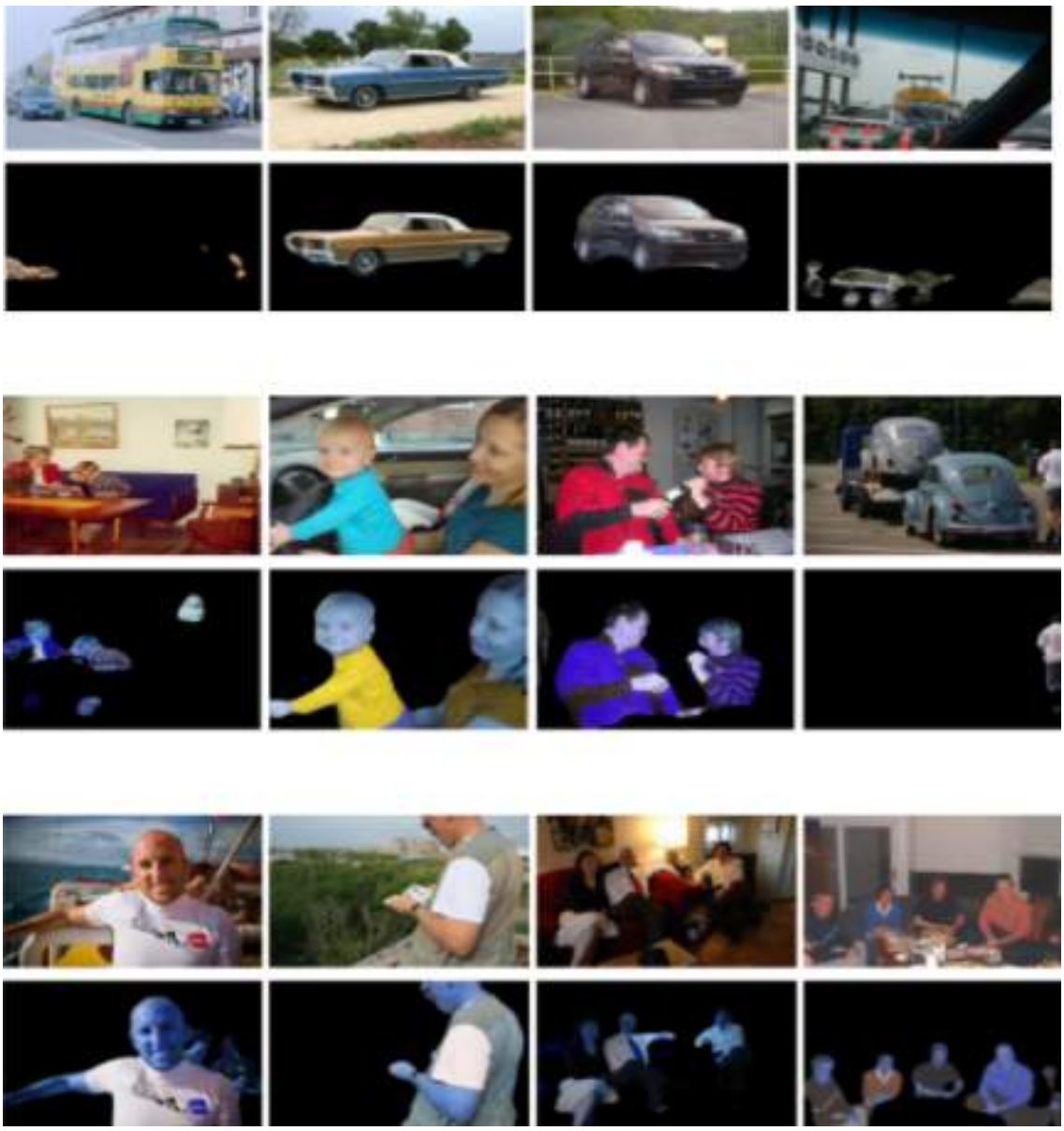

Figure 4.2: Segmentation results for input image group of size 4 to the proposed model. The model is trained on PASCAL-VOC 2012 [7] dataset. The test image group is also of the same dataset.

mentation masks. This process is repeated until we have only 2 image inputs. Row third to eighth displays the segmentation masks for input sizes 7, 6, 5 , 4, 3, and 2 respectively. As we can observe, the segmentation masks generated for the images are not affected by the number of input images. In all the cases the segmentation masks generated are almost same for the same input images. These results confirm the efficiency and accuracy of our model. More visual results for dynamic input images are given in Figures 4.4, 4.5, 4.6 and 4.7.

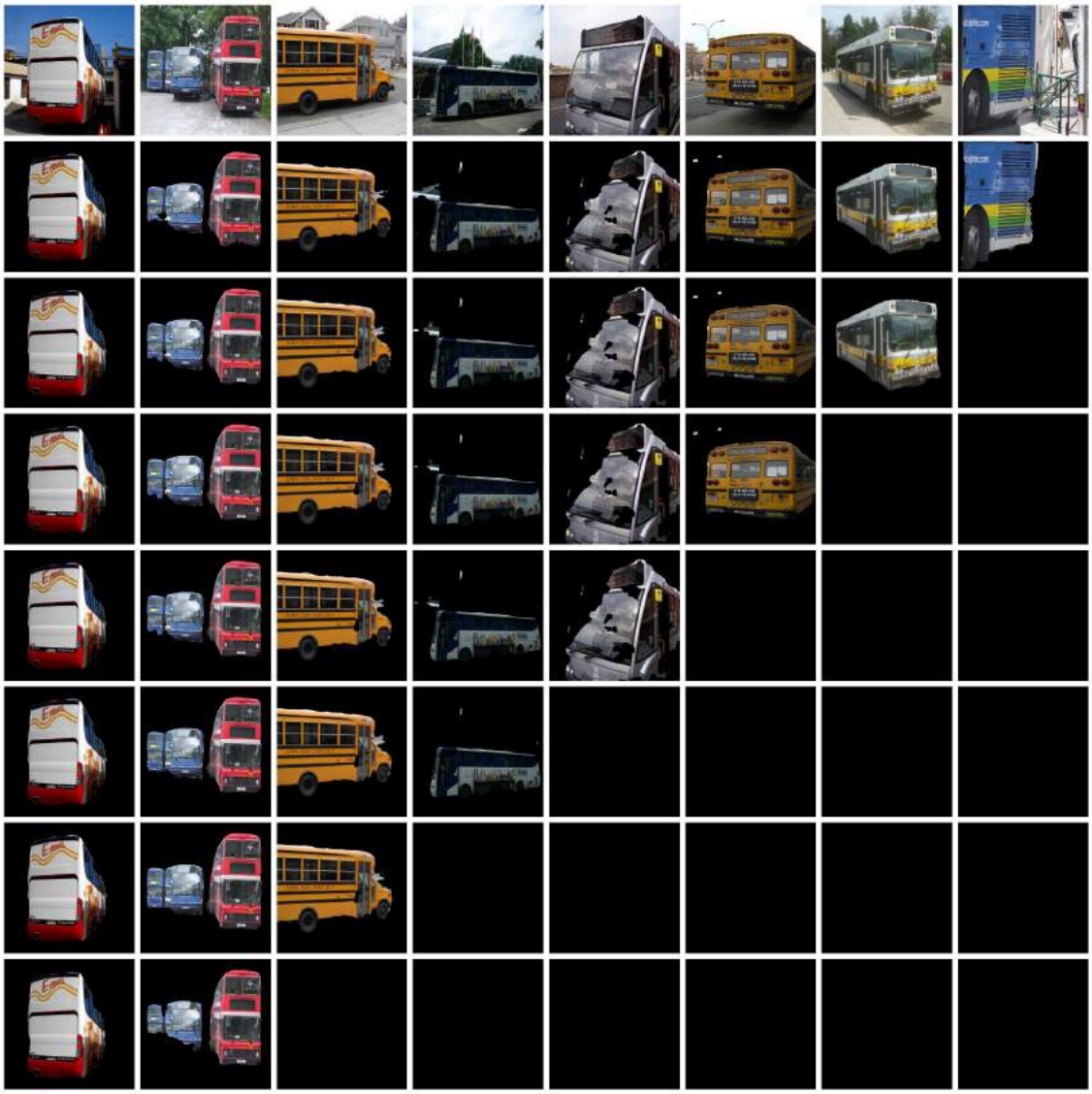

Figure 4.3: Comparison of segmentation between varying sizes of input image groups of PASCAL-VOC 2012 [7] dataset. The model is trained on image group of size 4 on the same dataset.

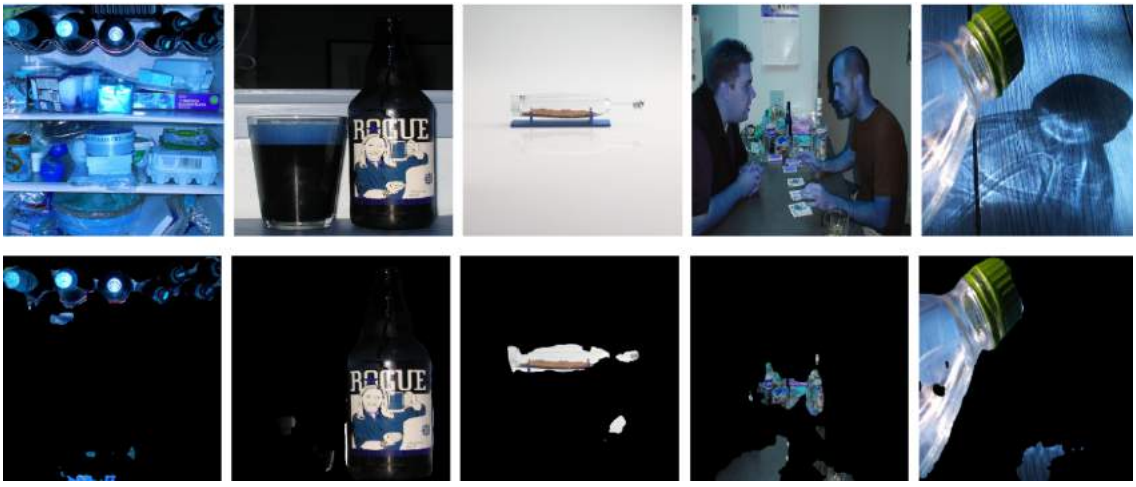

Figure 4.4: Segmentation results for 5-image input to the proposed model trained on image group of size 4 of PASCAL-VOC 2012 [7] data

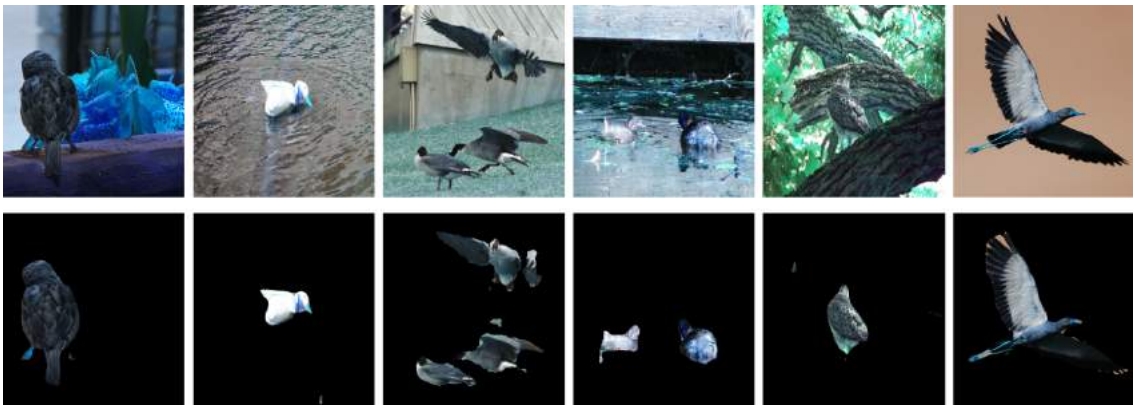

Figure 4.5: Segmentation results for 6-image input to the proposed model trained on image group of size 4 of PASCAL-VOC 2012 [7] data

Figure 4.6: Segmentation results for 7-image input to the proposed model trained on image group of size 4 of PASCAL-VOC 2012 [7] data

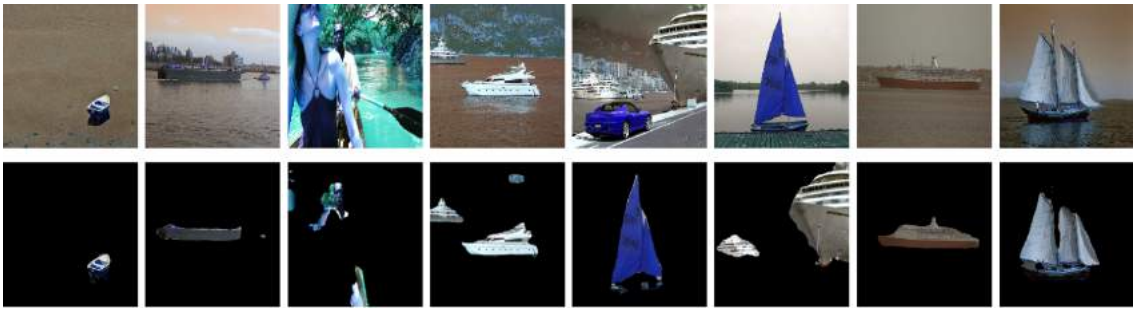

Figure 4.7: Segmentation results for 8-image input to the proposed model trained on image group of size 4 of PASCAL-VOC 2012 [7] data

4.3 Ablation Study

It could be argued that the segmentation model gives promising results due to the encoder-decoder mechanism or powerful VGG-16 for feature extraction and that the channel attention mechanism has little or no effect on the results acquired. Therefore, in order to gauge the importance of the channel attention mechanism we have conducted an experiment.

We have removed the channel attention mechanism from the encoder-decoder architecture. Instead, we have taken the pooled feature maps of the image pair, pixel wise multiplied them and passed them through sigmoid activation function. We have removed the attention mechanisms of the fully connected networks. Trained this model with the same dataset (PASCAL VOC 2012) of size 25k generated image pairs and tested with the same 169k test image pairs. The jaccard score obtained for the segmentation model without channel attention is 52.34% which is quite less than the score obtained with the channel as well as the spatial attention layers which thereby proves the power and importance of the attention mechanism in the segmentation task.

CHAPTER 5
Conclusion and Future work

In this thesis, we have proposed four different attention mechanisms that act as bottleneck for an encoder-decoder mechanism devised to perform common segmentation of image groups. The proposed models help to analyse the significance of channel and spatial attention in the segmentation task. The M2C model outperforms the other models as well as the baseline papers. We proposed modifications to this model to handle scenarios where there is no common object between a pair of image.

Furthermore, we have extended this attention mechanism, M2C, into a recursive model that handles multiple, dynamic number of images. This recursive, multiple image segmentation model module ensures that the model is capable of jointly segmenting a collection of an arbitrary number of images, which is a strong aspect of this method. Our proposed model has outperforms the baseline state-of-the-art models.

We intend to refine the multiple image segmentation model to handle various combinations of negative and positive image pair combinations. For this, we intend to create an appropriate dataset that handles all the possible combination of images in order to achieve a robust model. This summons the future scope of the thesis.

References

[1] N. Ali and B. Zafar. Msrc-v2 image dataset, Aug 2018.

[2] S. Banerjee, A. Hati, S. Chaudhuri, and R. Velmurugan. Cosegnet: Image co-segmentation using a conditional siamese convolutional network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 673–679, July 2019.

[3] X. H. Cao, Z. Obradovic, and K. Kim. A simple yet effective model for zero-shot learning. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 766–774, 2018.

[4] H. Chen, Y. Huang, and H. Nakayama. Semantic aware attention based deep object co-segmentation. In *Asian Springer Conference on Computer Vision*, pages 435–450, 2018.

[5] J. Chen, Y. Chen, W. Li, G. Ning, M. Tong, and A. Hilton. Channel and spatial attention based deep object co-segmentation. *Knowledge-Based Systems*, 211:106550, 2021.

[6] L. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[8] A. Hati, S. Chaudhuri, and R. Velmurugan. Co-segmentation of non-homogeneous image sets. In *25th IEEE International Conference on Image Processing (ICIP)*, pages 266–270, 2018.

[9] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1943–1950, 2010.

[10] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 542–549, 2012.

[11] E. Kim, H. Li, and X. Huang. A hierarchical image clustering cosegmentation framework. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 686–693, 2012.

[12] C. Lee, W.-D. Jang, J.-Y. Sim, and C.-S. Kim. Multiple random walkers and their application to image cosegmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3837–3845, 2015.

[13] B. Li, Z. Sun, Q. Li, Y. Wu, and H. Anqi. Group-wise deep object co-segmentation with co-attention recurrent neural network. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8518–8527, 2019.

[14] W. Li, O. Hosseini Jafari, and C. Rother. Deep object co-segmentation. pages 638–653, 2018.

[15] P. Mukherjee, B. Lall, and S. Lattupally. Object cosegmentation using deep siamese network. *CoRR*, abs/1803.02555, 2018.

[16] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. volume 9351, pages 234–241, 10 2015.

[17] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into MRFs. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 993–1000, 2006.

[18] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1939–1946, 2013.

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[21] Z. Yuan, T. Lu, and P. Shivakumara. A novel topic-level random walk framework for scene image co-segmentation. In *European Springer Conference on Computer Vision*, pages 695–709, 2014.

[22] K. Zhang, J. Chen, B. Liu, and Q. Liu. Deep object co-segmentation via spatial-semantic network modulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12813–12820, 2020.