symmetry

MDPI

Article

Unsupervised Learning from Videos for Object Discovery in Single Images

Dong Zhao †, Baoqing Ding †, Yulin Wu, Lei Chen and Hongchao Zhou *

School of Information Science and Engineering, Shandong University, Qingdao 266237, China; zhaodong@mail.sdu.edu.cn (D.Z.); 201912470@mail.sdu.edu.cn (B.D.); yulinw@mail.sdu.edu.cn (Y.W.); lei.chen@sdu.edu.cn (L.C.)

* Correspondence: hongchao@sdu.edu.cn

† These authors contributed equally to this work.

Abstract: This paper proposes a method for discovering the primary objects in single images by learning from videos in a purely unsupervised manner—the learning process is based on videos, but the generated network is able to discover objects from a single input image. The rough idea is that an image typically consists of multiple object instances (like the foreground and background) that have spatial transformations across video frames and they can be sparsely represented. By exploring the sparsity representation of a video with a neural network, one may learn the features of each object instance without any labels, which can be used to discover, recognize, or distinguish object instances from a single image. In this paper, we consider a relatively simple scenario, where each image roughly consists of a foreground and a background. Our proposed method is based on encoder-decoder structures to sparsely represent the foreground, background, and segmentation mask, which further reconstruct the original images. We apply the feed-forward network trained from videos for object discovery in single images, which is different from the previous co-segmentation methods that require videos or collections of images as the input for inference. The experimental results on various object segmentation benchmarks demonstrate that the proposed method extracts primary objects accurately and robustly, which suggests that unsupervised image learning tasks can benefit from the sparsity of images and the inter-frame structure of videos.

Keywords: unsupervised learning; sparsity representation; object discovery; foreground; background; segmentation mask

Citation: Zhao, D.; Ding, B.; Wu, Y.; Chen, L.; Chen, H. Unsupervised Learning from Videos for Object Discovery in Single Images. *Symmetry* **2021**, *13*, 38. https://doi.org/10.3390/sym13010038

Received: 1 December 2020
Accepted: 23 December 2020
Published: 29 December 2020

1. Introduction

The unsupervised learning on images, which enables computers to learn the features of objects from unlabeled images, is an intriguing and challenging problem in computer vision. It has a potential impact on the development of the fundamentals of deep learning techniques and it may help to reduce the amount of required labeled data in many computer-vision tasks, such as classification, recognition, instance segmentation, etc. Our motivation for unsupervised learning on images is based on two observations: (1) images are naturally sparse: an image typically consists of multiple objects where each can be sparsely represented by a deep neural network, and each class of objects with similar appearances may appear in many images. Such a sparsity can be exploited in order to learn the features of the objects in an unsupervised manner. (2) Natural videos possess rich self-supervised information; for example, two objects could be distinguished when one object moves relative to another, even if their shapes change at the same time. In this paper, we explore the two ideas in a single unsupervised framework and propose a new method for training on videos and performing tasks on single images. In particular, we focus on the task of discovering primary objects from single images, and the method that was developed in this paper might be applied in many other computer-vision tasks.

Video understanding has been studied in computer vision for decades. In low-level vision, different methods have been proposed in order to find correspondences between pixels across video frames, which is known as optical flow estimation [1,2]. Camera motion and object motion can both result in optical flow, therefore these methods can not be aware of the *objects* in the scene. In high-level vision, object tracking and object discovery in videos have been well-studied [3–8], especially with the introduction of unsupervised challenge of the DAVIS dataset [9]. However, the unsupervising in that challenge only means that the supervision information is not required for the test phase, but it is still required for the training phase. Despite remarkable performance, these approaches benefit either from object labels [5,7], or from pre-trained models in order to generate proposals [3,10]. In this paper, we define that an unsupervised video learning module should not use any manual annotation or pre-trained models on the manual annotation.

The task of unsupervised object discovery in videos is strongly related to co-localization [11–15] and co-segmentation [16–23]. The task has been studied for more than a decade in computer vision, with initial works mainly being based on local feature matching and detection of their co-occurring patterns [24–27]. Recent approaches [12,15,18] discovered object tubes by linking candidate detection between frames with or without refining their location. Typically, the task of unsupervised learning from image sequences is formulated as an optimization problem for either feature matching, conditional random fields, or data clustering. However, it is inherently expensive, due to the combinatorial nature of the problem. Besides, it is time-consuming to perform object discovery in videos or in collections of images at test time. Different from that, our method moves the unsupervised discovery to the training stage, while, at test time, we apply the standard feed-forward processing in order to detect the object in single test images quickly.

Figure 1 illustrates our unsupervised object discovery (UnsupOD) framework. We formulate the problem of unsupervised learning in videos as the encoder–decoder structure that internally factors the video frame into a foreground, a background, and a mask, without direct supervision for any of these factors. However, without further assumptions, decomposing an image into these three factors is ill-posed. We construct the model that is based on the following assumptions.

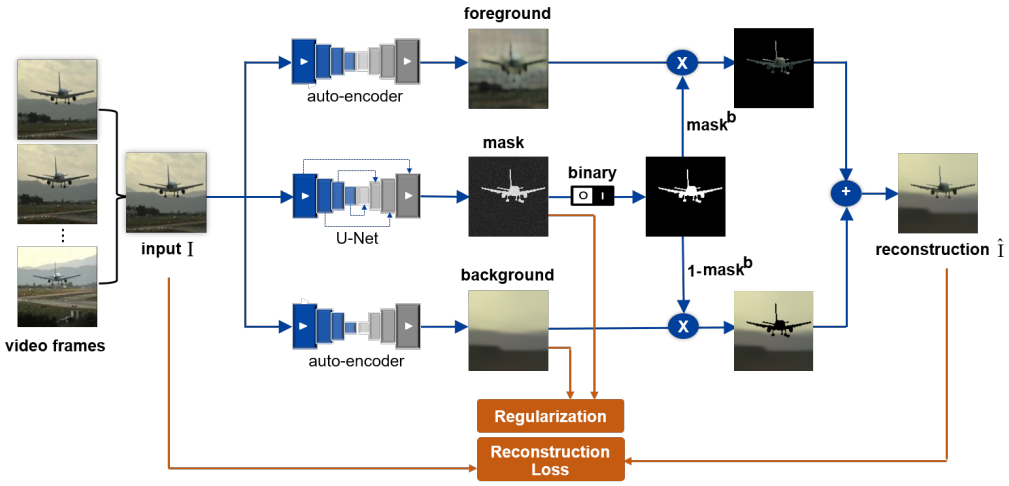

Figure 1. Overview of our unsupervised object discovery (UnsupOD) approach. Our model factors an input video frame F into the foreground, background, and segmentation mask. The video frames are sent into the network without being shuffled. It is trained to reconstruct the input without external supervision. Our objective function takes into account the reconstruction loss, and the regularization constraints for the background and mask, as represented by the orange line.

First, foreground objects are more difficult to model than their backgrounds, as their movements and appearance are more complex. With this assumption, we build a background model with two constraints. On the one hand, we construct a much narrower

bottleneck (i.e., the smallest layer in auto-encoder) to the background network than that to the foreground network. Consequently, a large amount of image information flows into the foreground network, and a small amount of information flows into the background network. On the other hand, we add a gradient constraint to the background in order to make it as "clean" as possible, which prevents foreground objects from appearing in the background.

Second, we define that a good mask should satisfy the following criteria: (i) it should present the object's shape and appearance as refined as possible. (ii) It has to be much closer to a binary image. (iii) It should display smooth contours without having holes, namely the closed region. We add constraints to the mask model based on the above criteria and show that the object masks produced by the mask model have nicer and smoother shapes, and capture well the figure-ground contrast and organization.

We combine these elements in an end-to-end learning formulation, where all of the components are only learned from raw RGB data. We demonstrate our method on various datasets. The experimental results show that our approach can produce high-quality segmentation masks without any manual annotation or pre-trained features, indicating that our unsupervised model learns a strong, high-level semantic feature representation for objects. Moreover, our model is able to discover and segment objects that belong to classes not in the training dataset, which further verifies the feasibility and generalization capability of our approach.

The main contributions are summarized, as follows:

- We propose a novel deep network architecture for unsupervised learning, which factors the image into multiple object instances that are based on the sparsity of images and the inter-frame structure of videos.
- We propose a method to discover the primary object in single images by completely unsupervised learning without any manual annotation or pre-trained features.
- Our segmentation quality tends to increase logarithmically with the amount of training data, which suggests the infinite possibilities of learning and generalization of our model. Besides, our model maintains a very high speed in testing and the experimental results demonstrate that it is at least two orders of magnitude faster than the related co-segmentation methods [4,27].

2. Related Work

Unsupervised learning. With the exponential growth of multimedia data in the Internet age, how to effectively utilize big data has attracted increasing attention [28,29], especially in the field of unsupervised learning without manual annotation. Previous works mainly fall into two categories: generative models and self-supervised approaches. The primary objective of generative models is to reconstruct the distribution of data as faithfully as possible. Classical generative models include Restricted Bolztmann Machines (RBMs) [30], Auto-Encoder (AE) [31], and Generative Adversarial Networks (GANs) [32]. Self-supervised learning exploits internal structures of data and formulates predictive tasks in order to train a model. Specifically, the model needs to predict either an omitted aspect or component of an instance given the rest. To learn a representation of images, the tasks could be: predicting the context [33], counting the objects [34], filling in missing parts of an image [35], recovering colors from grayscale images [36], or even solving a jigsaw puzzle [37]. For videos, self-supervised strategies include: leveraging temporal continuity via tracking [38,39], predicting future [40], or preserving the equivariance of egomotion [41]. Nevertheless, while self-supervised learning may capture relations among parts or aspects of an instance, it is unclear why a particular self-supervised task should help semantic recognition and which task would be optimal.

Object discovery from unlabeled data. Object discovery from unlabeled data is challenging due to the fact that it does not depend on any auxiliary information rather than a given unlabeled image. Thus, many methods are focused on solving the image co-localization problem [11–14,21,42,43]. Object discovery or co-localization is a process

for finding objects of the same class over multiple images or videos. Some earlier image co-localization methods [12–14,21] addressed this problem based on low-level features (e.g., SIFT, HOG). Recently, some works [3,42–44] learned a common object detector by using the features from some pre-trained CNN models, such as VGG-16 [45], which was used in [3,44]. However, the discovery module fully that is based on unsupervised learning should not use pre-trained features on manually labeled ground truth.

Co-segmentation. The co-segmentation task is aiming to discover an identical object within a collection of images or videos. The first co-segmentation method was proposed by Rother et al. [46], in which the same object is simultaneously segmented in two different images with histogram matching. Since then, a lot of works have begun to focus on co-segmentation to help further improve its performance [17,47–49], and these works either perform in a pair of images that contain the same object [47,49], or require some form of user interaction [48]. Recent years, Some researchers extended co-segmentation techniques in various ways. Joulin et al. [17] combined existing tools for bottom-up image segmentation with kernel methods commonly used in object recognition within a discriminative clustering framework. Inspired by the anisotropic heat diffusion, Kim et al. [16] proposed a distributed co-segmentation approach for a highly variable large-scale image collection. Vicente et al. [23] introduced the conception of "objectness" to the co-segmentation framework, suggesting foreground segment is required to be an object to improve co-segmentation results significantly. To match objects among images, Rubio et al. [50] applied region matching to exploit inter-image information by establishing correspondences between the common objects that appear in the scene. Lee et al. [51] proposed the notion of multiple random walkers on a graph, and then applied it to the image co-segmentation while using the repulsive restart rule. Generally, the co-segmentation techniques need a collection of images during testing. Different from that, each image is independently processed at test time in our model.

Video Foreground/Background Segmentation. Video foreground segmentation is the task of classifying every pixel in a video as foreground or background, separating all of the moving objects from the background. Early approaches [4,52–54] relied on heuristics in the optical flow field in order to identify moving objects, such as closed motion boundaries in [4]. These initial estimates were then refined by utilizing external cues, such as saliency maps [54] or object shape estimates [53]. Another line of work focused on building probabilistic models of moving objects while using optical flow orientations [55,56]. These methods are not based on a robust learning framework and they fail to generalize well to unseen videos. The recent introduction of a standard benchmark, DAVIS 2016 [9], has led to a renewed interest. More recent approaches [8,57,58] proposed deep models for directly estimating motion masks. For example, [8,57,59] adopted a learning-based approach and trained Convolutional Neural Networks (CNNs) that utilize RGB and optical flow as inputs for producing foreground segmentation. However, these approaches are based on either manually annotated mask labels as their supervising information, or a supervised pre-trained network in order to generate the object proposes. Our method directly estimates masks in a fully unsupervised manner.

3. Our Approach

Our goal is to train an unsupervised learning model from videos, and then the trained model can automatically discover the primary foreground object that appears in a single image, estimating both its bounding box and its segmentation mask.

The UnsupOD model is constructed, as illustrated in Figure 1. Given an unconstrained collection of videos, our target is to learn a model that receives a video frame as input and produces a decomposition of it into a foreground, a background, and a mask. Video frames are continuously sent to the network without being shuffled in order to combine both the appearance and motion information. The learning objective is reconstructive, as we have only raw videos to learn from: namely, the model is trained, so that the combination of the three factors gives back the input frame.

In order to learn such a decomposition without supervision for any of the components, we add constraints to each of these components separately. The following sections describe how this is done, looking first at the foreground and background model (Section 3.1), and then at how the segmentation mask is modeled (Section 3.2), followed by details of the reconstruction process (Section 3.3).

3.1. Foreground and Background Model

It is known that the object of interest usually has more complex and varied movements than its background scene: it has a distinctive appearance and usually causes occlusions and occupies less space. All of these differences make the foreground more difficult to model than the background. Because the background contains variations less than the foreground, we expect it to be better captured by the lower-dimensional subspace of the frames from a given video shot.

Under these observations, our solutions for constructing the foreground and background model are two-fold: (1) both of the networks are built on the fully convolutional auto-encoder architecture without skip connections between the encoder and decoder; (2) the settings for the foreground and background networks are the same, except that the bottleneck (i.e., the smallest layer in auto-encoder) of the background network is much narrower than that of the foreground network. The former separates the foreground and background, while the latter forces the network to model the background in the lower-dimensional subspace.

However, the powerful capability to model appearance of the auto-encoder makes a small amount of foreground pixels always remain in the background, which makes the object segmentation mask incomplete. To make the learned background clean to a great extent, we add a gradient regularization loss to the background:

$$R_{bg} = \nabla Bg. \tag{1}$$

Minimizing the background gradient forces the background to be as clean as possible, with no trace of the foreground (see the learned background shown in Figure 1). It is also important to note that minimizing the background gradient will also result in a blurred and unreal background. Fortunately, the foreground and background are just auxiliary tasks, and our ultimate goal is to obtain a good object mask.

3.2. Segmentation Mask Model

A good segmentation mask m satisfies the following criteria: (i) it should present the object shape and appearance as refined as possible. (ii) It has to be as close as possible to a binary image. (iii) It should display smooth contours without having holes, namely the closed region.

The first criterion is obtained by employing the U-Net [60] as the mask network. U-Net [60] is the most widely used structure in image segmentation, especially in medical image analysis [61–63], mainly because the encoder–decoder structure with skip connections allows for efficient information flow and helps to recover the full spatial resolution at the network output. After the last convolutional layer and the sigmoid activation function, the segmentation mask m will be a two-dimensional output map and its value is in the range of (0,1). The second criterion is enforced via a binary regularization loss, as expressed in Equation (2):

$$R_{binary} = -\sum_{x}\sum_{y}\left[m(x,y) - 0.5\right]^2, \tag{2}$$

where $m(x,y)$ is the value of the two-dimensional output map at position (x,y). The third criterion is enforced by a closed regularization loss, as defined in Equation (3):

$$\begin{aligned} R_{closed} = \frac{1}{4}\sum_{x}\sum_{y}\Big[&\big[m(x,y) - m(x-1,y)\big]^2 \\ &+ \big[m(x,y) - m(x+1,y)\big]^2 \\ &+ \big[m(x,y) - m(x,y-1)\big]^2 \\ &+ \big[m(x,y) - m(x,y+1)\big]^2\Big]. \end{aligned} \tag{3}$$

It is challenging to segment the primary objects from complex backgrounds in a fully unsupervised manner. Although the motion information of the object in video sequences can provide some clues, the camera moving in some shooting scenes can also result in ambiguities. For example, in the airplane taking off or landing scene, the background is moving and varying, while the primary object airplane is "still" in the middle of the video frames, as shown in Figure 2. Learning by motion clues will make the network mistakenly regard the moving background as the foreground object of interest. In order to overcome this problem, we introduce an area regularization constraint to guide the network, as expressed in Equation (4):

$$R_{area} = |\bar{m} - \alpha|, \tag{4}$$

where $\bar{m}$ is the mean value of m, and α is the proportion factor, $\alpha \in (0,1)$, which restrains the size of the learned mask and gives a penalty when the network regards the background as the foreground object of interest.

Figure 2. An example of the airplane landing scene.

To sum up, the final regularization loss for our mask model R_{mask} is therefore:

$$R_{mask} = \lambda_a R_{area} + \lambda_b R_{binary} + \lambda_c R_{closed}, \tag{5}$$

where λ_a, λ_b and λ_c are the weighting factors for corresponding regularization loss.

3.3. Image Reconstruction

Because the sigmoid nonlinearity is adopted in the last convolutional layer of the mask model, the value of output two-dimensional mask m is in the range of (0,1). We first round m to a binary mask m^b, and then the reconstruction $\hat{I}$ can be obtained by combining the foreground Fg and background Bg with the binary mask m^b:

$$\hat{I} = m^b \times Fg + (1 - m^b) \times Bg. \tag{6}$$

After binarization, the value of the two-dimensional map m is 0 or 1. A value of 1 indicates that this pixel belongs to the foreground; otherwise, this pixel belongs to the background. The reconstructed image is formed by selecting pixel values from the background and foreground according to the guidance of m^b.

The binarization is a rounding function, which is defined in Equation (7):

$$B = \begin{cases} 0, & m(x,y) < 0.5 \\ 1, & m(x,y) \geq 0.5 \end{cases} \tag{7}$$

However, the gradient of the binary function B is zero almost everywhere, except that it is infinite when $m(x,y) = 0.5$. In the back-propagation algorithm, the gradient is computed layer-by-layer with the chain rule in a backward manner. Thus, this will make any layer before the binary never be updated during training.

Based on the straight-through estimator on gradient [64] and inspired by [65], we introduce a proxy function $\tilde{B}$ in order to approximate B:

$$\tilde{B} = \begin{cases} 0, & m(x,y) < 0 \\ m(x,y), & 0 \leq m(x,y) \leq 1 \\ 1, & m(x,y) > 1 \end{cases} \tag{8}$$

Because the proxy function $\tilde{B}$ is differentiable, $\tilde{B}$ can be used in back-propagation. Importantly, we do not fully replace the binary function with a smooth approximation, but only its derivative, which means that B is still performed as usual in the forward pass.

Subsequently, the optimization loss for the whole model is therefore:

$$Loss = L_{recon} + \lambda_{bg} R_{bg} + \lambda_{mask} R_{mask}, \tag{9}$$

where $L_{recon} = |I - \hat{I}|^2$, R_{bg} and R_{mask} are the regularization for background and mask, respectively. λ_{bg} and λ_{mask} are the corresponding weighing factors for R_{bg} and R_{mask}.

4. Experiments and Analysis

In this section, we conduct extensive experiments in order to verify the performance of the proposed UnsupOD model. Firstly, we introduce the experimental setup (Section 4.1). Subsequently, we evaluate the UnsupOD on object discovery in video (Section 4.2). Finally, we verify the generalization ability of UnsupOD on object discovery in single images, including both seen classes (Section 4.3) and unseen classes (Section 4.4). We present both qualitative and quantitative results, and the comparisons with the previous methods.

4.1. Experimental Setup

4.1.1. Datasets

- The YouTube Objects (YTO) dataset [6]. The YTO dataset is a large-scale database that was collected from YouTube containing videos for each of 10 diverse object classes (aeroplane, bird, boat, car, cat, cow, dog, horse, motorbike, and train). The dataset has 5484 video shots for a total of 571,089 frames. The videos display significant clutter, with foreground objects coming in and out of focus and often out of sight, undergoing occlusions and significant changes in scale and viewpoint. The dataset also provides ground-truth bounding-boxes on the object of interest in one frame for each of 1407 video shots.
- The Object Discovery dataset [21]. The Object Discovery dataset is collected by automatically downloading images while using the Bing API, using queries for airplane, car, and horse. It contains 15k internet images: airplane (4542 images), car (4347 images), horse (6381 images), and it is annotated with high detail segmentation masks.
- The MSRC dataset [66]. The MSRC dataset is composed of 591 photographs of 21 object classes and hand-labeled with the assigned colors acting as indices into the list of object classes. All of the images were taken considering completely general lighting conditions, camera viewpoint, scene geometry, object pose, and articulation.
- The iCoseg dataset [48]. The iCoseg dataset is built from the Flickr® online photo collection and hand-labelled pixel-level segmentations in all of the images. It contains 38 challenging groups with 643 total images (∼17 images per group), consisting of animals in the wild, popular landmarks, sports teams, and other groups that contain a common theme or common foreground object.

4.1.2. Implementation Details

The unsupervised training is performed in training videos of the YouTube Objects (YTO) dataset [6]. The video frames are sent into the network without being shuffled, enabling the network to separate the object from the background that is based on the motion clue. Each video frame is resized to 128×128. We report the training details, including all of the hyperparameter settings in Table 1, and detailed network architectures in Tables 2 and 3. Our implementation is in PyTorch, and all of the experiments are conducted on a single GeForce GTX 2080ti GPU, and the whole training takes about 10 h.

Table 1. Training details and hyper-parameter settings.

Parameter	Value
Optimizer	Adam
Learning rate	1×10^{-5}
Number of epochs	10
Batch size	16
Proportion factor α	0.2
Loss weight λ_{bg}	1
Loss weight λ_{mask}	1
Loss weight λ_a	0.1
Loss weight λ_b	1
Loss weight λ_c	0.05
Input image size	128×128
Output image size	128×128

Table 2. U-Net architecture for mask. The arrows represent the corresponding skip connections.

Encoder	Output Size
Conv(3, 64, 3, 1, 1) $\times$ 2 $\rightarrow$ *1*	128
Down(2) + Conv(64, 128, 3, 1, 1) $\times$ 2 $\rightarrow$ *2*	64
Down(2) + Conv(128, 256, 3, 1, 1) $\times$ 2 $\rightarrow$ *3*	32
Down(2) + Conv(256, 512, 3, 1, 1) $\times$ 2 $\rightarrow$ *4*	16
Down(2) + Conv(512, 512, 3, 1, 1) $\times$ 2	8
Decoder	**Output Size**
Up(2) $\rightarrow$ *4*	16
Conv(1024, 256, 3, 1, 1) $\times$ 2 + Up(2) $\rightarrow$ *3*	32
Conv(512, 128, 3, 1, 1) $\times$ 2 + Up(2) $\rightarrow$ *2*	64
Conv(256, 64, 3, 1, 1) $\times$ 2 + Up(2) $\rightarrow$ *1*	128
Conv(128, 64, 3, 1, 1) $\times$ 2	128
Conv(64, 1, 3, 1, 1) + Sigmoid	128

Table 2 shows the U-Net [60] structure for our mask model. Arrows represent the corresponding skip connections, and each convolution is followed by batch normalization and ReLU (not shown in the table) except the output layer. In Table 3, the encoder networks for foreground and background are built on ResNet18 [67]. We add a convolution layer after each deconvolution layer and replace the last deconvolution layer with nearest-neighbor upsampling, followed by three convolution layers in order to mitigate checkerboard artifacts [68] in the decoder outputs. Abbreviations of the operators are defined, as follows:

- Conv(c_{in}, c_{out}, k, s, p): convolution with c_{in} input channels, c_{out} output channels, kernel size k, stride s, and padding p.
- Deconv(c_{in}, c_{out}, k, s, p): deconvolution with c_{in} input channels, c_{out} output channels, kernel size k, stride s, and padding p.
- Down(s): max-pooling downsampling with a scale factor of s.

- Up(s): nearest-neighbor upsampling with a scale factor of s.
- GN(n): group normalization with n groups.

Table 3. Network architecture for foreground and background. The output channel size c_{out} is 128 for the foreground and 32 for the background.

Encoder	**Output Size**
ResNet18.conv1	64
ResNet18.conv2_x	32
ResNet18.conv3_x	16
ResNet18.conv4_x	8
ResNet18.conv5_x	4
Conv(512, c_{out}, 3, 1, 1) + ReLU	4
Decoder	**Output Size**
Deconv(c_{out}, 512, 4, 2, 1) + ReLU	8
Conv(512, 512, 3, 1, 1) + ReLU	8
Deconv(512, 256, 4, 2, 1) + GN(64) + ReLU	16
Conv(256, 256, 3, 1, 1) + GN(64) + ReLU	16
Deconv(256, 128, 4, 2, 1) + GN(32) + ReLU	32
Conv(128, 128, 3, 1, 1) + GN(32) + ReLU	32
Deconv(128, 64, 4, 2, 1) + GN(16) + ReLU	64
Conv(64, 64, 3, 1, 1) + GN(16) + ReLU	64
Upsample(2)	128
Conv(64, 64, 3, 1, 1) + GN(16) + ReLU	128
Conv(64, 64, 5, 1, 2) + GN(16) + ReLU	128
Conv(64, 3, 5, 1, 2) + Sigmoid	128

4.1.3. Evaluation Metrics

- For the comparison of object localization bounding-boxes, we adopt the correct localization (CorLoc) metric following previous image localization works [11–13,43], which measures the percentage of images that were correctly localized according to the PASCAL criterion: an predicted box $\mathcal{B}_p$ is correct when compared with the ground-truth box $\mathcal{B}_{gt}$, when the intersection over union (IoU) overlap ratio $\frac{|\mathcal{B}_p \cap \mathcal{B}_{gt}|}{|\mathcal{B}_p \cup \mathcal{B}_{gt}|}$ is larger than 0.5.
- For the comparison of object segmentation masks, we evaluate, based on the P, the J metric, as described by Rubinstein et al. [21]—the higher P and J, the better. P refers to the per pixel precision (the ratio of correctly labeled pixels), while J is the Jaccard similarity (the intersection over union of the result and ground truth segmentation). Both measures of the are commonly used for evaluation in image segmentation.

4.2. Results on Video Dataset

We run experiments on the training videos of the YTO dataset and then evaluate the UnsupOD on its testing split. For fitting a box around our mask after training, we first binary the output mask, determine the connected components, filter out the small ones, and finally fit a tight box around each of the remaining components. To illustrate our method, we show qualitative results in Figure 3. We select one sample for each of the ten categories, and show the corresponding foreground, background, binary mask, and bounding box.

From Figure 3, we have the following observations:

(1) the primary object is completely separated from the background;
(2) the background model can automatically filled-in the "missing" image parts in its output;
(3) the produced object masks have smoother shapes, with very few holes, and capture the figure-ground contrast and organization well;

(4) the UnsupOD model is able to detect multiple objects (see the masks of the sixth column); and,
(5) the UnsupOD model boxes the parts that move with the object (see the bounding box in the penultimate column: only motorbike for ground truth, motorbike with a man in our bounding box); this is mainly because our motion-based approach groups pixels that share the same motion.

All of these observations reflect that the proposed UnsupOD model learns a strong, high-level semantic feature representation. This is because such unsupervised segmentation is difficult from low-level cues alone: objects are typically made of multiple colors and textures and, if occluded, might even consist of spatially disjoint regions. Therefore, to effectively do this segmentation is to implicitly recognize the object and understand its location and shape, even if it cannot be named.

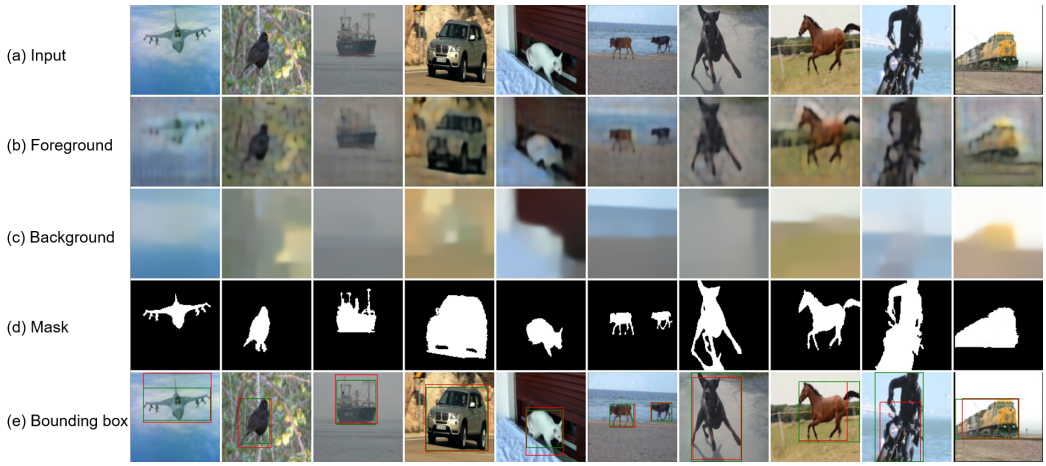

Figure 3. Visual results on the YTO dataset. (**a**): Input frame, choosing one frame for each of the ten categories; (**b**) foreground, output of our foreground network; (**c**) background, output of our background network; (**d**) binary segmentation mask, output of our binary mask model; and, (**e**) bounding box, green for the proposed method while red for the ground truth.

Next, we conduct experiments on all of the annotated frames and compare the UnsupOD approach with the typical methods on co-localization [6,12], object discovery [15,27], and video object segmentation [4] in Table 4. It is noticed that the compared methods utilize the whole video shots, while our UnsupOD method only needs a single image at test time. However, the UnsupOD outperforms the others in six out of 10 classes and it has the best overall performance. Moreover, our CNN feed-forward network processes each image in 0.03 sec, being at least two orders of magnitude faster than all other methods. It is also important to note that, in all of our comparisons, while our system is faster at test time, it takes much longer during its unsupervised training phase and requires large quantities of unsupervised training data. The results of the compared methods are reported from their corresponding papers.

Table 4. Comparisions of correct localization (CorLoc) (%) on YouTube Object dataset.

Method	Aero	Bird	Boat	Car	Cat	Cow	Dog	Horse	Mbike	Train	Avg	Time
Prest et al. [6]	51.7	17.5	34.4	34.7	22.3	17.9	13.5	26.7	41.2	25.0	28.5	N/A
Joulin et al. [12]	25.1	31.2	27.8	38.5	41.2	28.4	33.9	35.6	23.0	25.0	31.0	N/A
Stretcu et al. [27]	38.3	62.5	51.1	54.9	64.3	52.9	44.3	43.8	41.9	45.8	49.9	6.9s
Papazoglou et al. [4]	65.4	**67.3**	38.9	65.2	46.3	40.2	65.3	48.4	39.0	25.0	50.1	4s
Jun et al. [15]	64.3	63.2	73.3	**68.9**	44.4	62.5	**71.4**	52.3	**78.6**	23.1	60.2	N/A
(Ours) UnsupOD	**69.5**	60.5	**74.2**	60.8	**65.7**	**63.2**	65.8	**54.8**	43.7	**48.9**	**60.7**	**0.03 s**

4.3. Results on Single Images

We further compare the UnsupOD with previous methods that are designed for object discovery in collections of images. A representative current benchmark in this sense is the Object Discovery dataset [21]. In order to compare with previous method fairly, we use a subset of the dataset containing 100 images for each category. There are 18, 11, and seven outliers for three categories, respectively. We compare the UnsupOD with the previous methods on co-segmentaiton [16,17,19,21] and co-localization [11]. Note that these co-segmentation and co-localization methods utilize a set of images that contain the objects from the same category. These methods aim to either discover the bounding box of the main object in given image, or its fine segmentation mask. We evaluate our model on both the bounding box and segmentation mask of the primary object. For the bounding box comparison, we use the CorLoc metric following previous image localization works [13,43]; while, for the segmentation mask comparison, we adopt the P and J metric, which are commonly used for evaluation in image segmentation research. The CorLoc comparison is shown in Table 5 and the P, J comparison is available in Table 6.

Table 5. Comparisons of CorLoc (%) on Object Discovery dataset.

Method	Supervision	Airplane	Car	Horse	Avg
Kim et al. [16]	co-segmentation	21.95	0.00	16.13	12.69
Joulin et al. [17]	co-segmentation	32.93	66.29	54.84	51.35
Joulin et al. [19]	co-segmentation	57.32	64.04	52.69	58.02
Rubinstein et al. [21]	co-segmentation	74.39	87.64	63.44	75.16
Tang et al. [11]	co-localization	71.95	**93.26**	64.52	76.58
(Ours) UnsupOD	w/o	**82.93**	91.95	**67.05**	**80.64**

Table 6. Comparisons of P and J (%) on Object Discovery dataset.

Method	Airplane		Car		Horse	
	P	J	P	J	P	J
Kim et al. [16]	80.20	7.90	68.85	0.04	75.12	6.43
Joulin et al. [17]	49.25	15.36	58.70	37.15	63.84	30.16
Joulin et al. [19]	47.48	11.72	59.20	35.15	64.22	29.53
Rubinstein et al. [21]	88.04	55.81	85.38	64.42	82.81	51.65
(Ours) UnsupOD	**89.20**	**61.45**	**90.14**	**68.24**	**86.16**	**57.02**

In Table 5, we can observe that, as compared with the co-segmentation methods, the UnsupOD obtains a significant improvement, i.e., 5.48% [21], 22.62% [19], 29.29% [17], and 67.95% [16]. The UnsupOD also outperforms the co-localization method [11] by 4.06% (76.58% vs. 80.64%). In particular, on the horse category, which is the most challenging subcategory due to multi-objects and complex background, the UnsupOD achieves the best performance when compared with all other methods. The experimental results demonstrate that the UnsupOD is robust to the complex scenarios.

Table 6 summarizes the comparison on the segmentation mask. Generally, the segmentation task is more challenging than the localization task, because the former not only need to find the location of the objects, but also need to obtain their refined shape masks. When compared with the other co-segmentation methods, the UnsupOD obtains the best results in all three categories, according to both P and J metrics. The excellent performance on the segmentation task further verifies the capability of our UnsupOD model.

Figure 4 shows a qualitative comparison. We randomly select two images for each category and display their corresponding segmentation mask. We can see that the co-segmentation methods [16,19,21] can discover objects in relatively simple background, such as the car examples. While the background gets complicated, like the airplane

example, they are struggle to handle. On the contrary, the UnsupOD produces high-quality segmentation results with both simple and complex backgrounds. Moreover, our masks are more refined and they capture the shape of objects well, which is manly because the proposed motion-learning-based approach mines the high-level semantic features of objects and learns good visual features for appearance. Figure 5 illustrates more examples of our segmentation results.

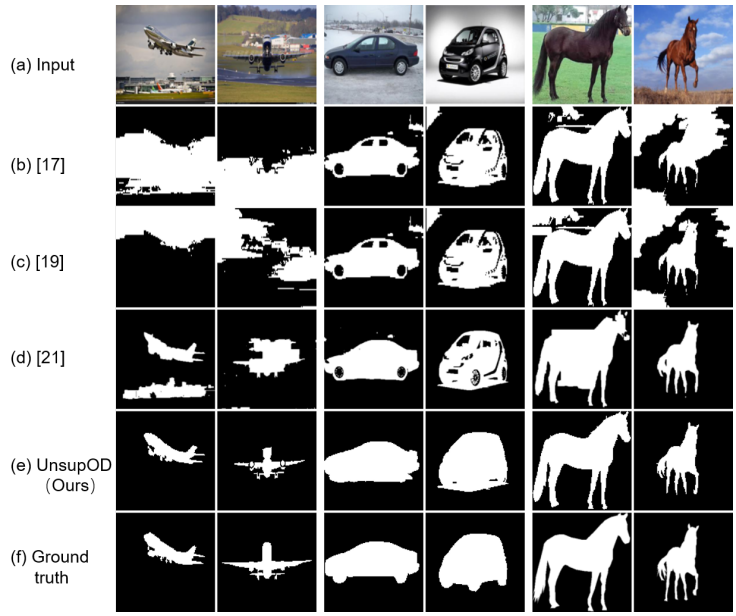

Figure 4. Qualitative results on Object Discovery dataset. (**a**) Input image; (**b**) segmentation mask, obtained by Joulin et al. [17]; (**c**) segmentation mask, obtained by Joulin et al. [19]; (**d**) segmentation mask, obtained by Rubinstein et al. [21]; (**e**) segmentation mask, obtained by our binary mask model; and (**f**) ground truth.

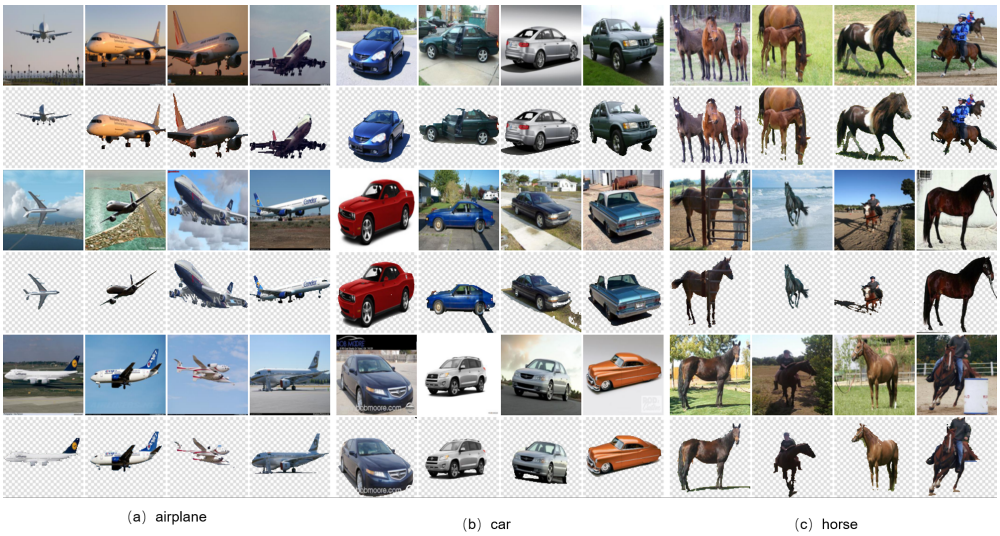

Figure 5. Sample results from three classes on Object Discovery dataset. For each image, we show a pair of the original (top) and our segmentation result (bottom).

4.4. Results on Single Images of Unseen Classes

We evaluate the UnsupOD on the images of unseen classes in order to demonstrate the generalization capabilities of the proposed approach. We conduct experiments on MSRC dataset [66] and iCoseg dataset [48], which have been widely used by previous work to evaluate co-segmentation performance. Most of the image classes in these two datasets

are unseen by our model. Both of the datasets include human-given segmentation that are only used for the evaluation.

We qualitatively compare the UnsupOD with the previous methods [16,19,21] on object segmentation in Figure 6. Six columns on the left show *hot-balloon*, *statue of liberty*, *windmill*, *pyramid*, *stonehenge* and *skate*, from iCoseg dataset, and six columns on the right show *chair*, *flower*, *sign*, *tree*, *bike*, and *house*, from the MSRC dataset. The papers for comparison are all co-segmentation methods, and they are allowed to see a collection of images within same class. However, our method has no other information besides the input image. Even so, the UnsupOD obtains decent segmentation mask in many challenging scenarios. Interestingly, for the *tree* and *bike*, our mask captures the internal contours well without bunching up, which is more consistent with human visual perception. The competitive results on the challenging unseen images demonstrate that incorporating deep learning techniques can efficiently mine target objects from unlabeled data in real-world applications.

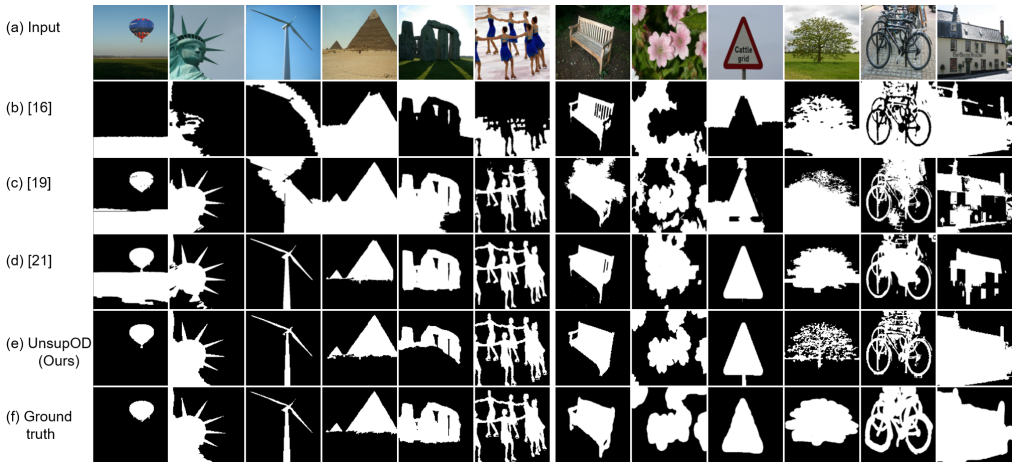

Figure 6. Qualitative results on MSRC dataset (left six columns) and iCoseg dataset (right six columns). (**a**) Input image; (**b**) segmentation mask, obtained by Kim et al. [16]; (**c**) segmentation mask, obtained by Joulin et al. [19]; (**d**) segmentation mask, obtained by Rubinstein et al. [21]; (**e**) segmentation mask, obtained by our binary mask model; and (**f**) ground truth.

5. Conclusions

We presented a novel deep network architecture for unsupervised learning, which learns to factor an image into multiple object instances that are based on the sparsity of images and the inter-frame structure of videos. We then performed the task of object discovery in this deep architecture by factoring an image into a foreground, a background, and an object mask. This is trained by an end-to-end learning without any supervision. The learning process is based on videos, but the generated model is able to discover objects from a single input image. The experimental results on various standard datasets show that the model is able to separate the foreground objects from their background completely, and obtain high-quality both segmentation masks and bounding-boxes. Our approach does not need any annotations, yet it still shows promising segmentation ability, which demonstrates that the proposed unsupervised model learns high-level semantic features and it might be applied to other computer-version tasks. For future work, we would like to extend our model to more than two object instances (not just foreground and background), which helps to solve the task of unsupervised object detection.

Author Contributions: Conceptualization, D.Z., B.D. and H.Z.; methodology, D.Z. and B.D.; software, B.D.; validation, D.Z.; formal analysis, D.Z. and B.D.; investigation, D.Z.; resources, D.Z. and B.D.; data curation, Y.W.; writing—original draft preparation, D.Z.; writing—review and editing, L.C. and H.Z.; supervision, H.Z.; project administration, B.D. and D.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant No. 62001267, and the Fundamental Research Funds of Shandong University under Grant No. 2020HW017.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Horn, B.K.; Schunck, B.G. Determining optical flow. Techniques and Applications of Image Understanding. *Int. Soc. Opt. Photonics* **1981**, *281*, 319–331.
2. Barron, J.L.; Fleet, D.J.; Beauchemin, S.S. Performance of optical flow techniques. *Int. J. Comput. Vis.* **1994**, *12*, 43–77. [CrossRef]
3. Zhang, R.; Huang, Y.; Pu, M.; Zhang, J.; Guan, Q.; Zou, Q.; Ling, H. Object discovery from a single unlabeled image by mining frequent itemsets with multi-scale features. *IEEE Trans. Image Process.* **2020**, *29*, 8606–8621. [CrossRef] [PubMed]
4. Papazoglou, A.; Ferrari, V. Fast object segmentation in unconstrained video. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 1777–1784.
5. Lu, X.; Wang, W.; Ma, C.; Shen, J.; Shao, L.; Porikli, F. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3623–3632.
6. Prest, A.; Leistner, C.; Civera, J.; Schmid, C.; Ferrari, V. Learning object class detectors from weakly annotated video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–24 June 2012; pp. 3282–3289.
7. Dave, A.; Tokmakov, P.; Ramanan, D. Towards segmenting anything that moves. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV), Seoul, Korea, 27–28 October 2019.
8. Jain, S.D.; Xiong, B.; Grauman, K. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2126.
9. Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; Sorkine-Hornung, A. A benchmark dataset and evaluation methodology for video object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 724–732.
10. Luiten, J.; Zulfikar, I.E.; Leibe, B. Unovost: Unsupervised offline video object segmentation and tracking. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 2000–2009.
11. Tang, K.; Joulin, A.; Li, L.J.; Fei-Fei, L. Co-localization in real-world images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 1464–1471.
12. Joulin, A.; Tang, K.; Fei-Fei, L. Efficient image and video co-localization with frank-wolfe algorithm. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 253–268.
13. Cho, M.; Kwak, S.; Schmid, C.; Ponce, J. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1201–1210.
14. Faktor, A.; Irani, M. "Clustering by composition"—Unsupervised discovery of image categories. In Proceedings of the European Conference on Computer Vision (ECCV), Florence, Italy, 7–13 October 2012; Springer: Cham, Switzerland, 2012; pp. 474–487.
15. Jun Koh, Y.; Jang, W.D.; Kim, C.S. POD: Discovering primary objects in videos based on evolutionary refinement of object recurrence, background, and primary object models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1068–1076.
16. Kim, G.; Xing, E.P.; Fei-Fei, L.; Kanade, T. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 169–176.
17. Joulin, A.; Bach, F.; Ponce, J. Discriminative clustering for image co-segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1943–1950.
18. Rochan, M.; Wang, Y. Efficient object localization and segmentation in weakly labeled videos. In Proceedings of the International Symposium on Visual Computing, Las Vegas, NV, USA, 8–10 December 2014; Springer: Cham, Switzerland, 2014; pp. 172–181.
19. Joulin, A.; Bach, F.; Ponce, J. Multi-class cosegmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 542–549.
20. Kuettel, D.; Guillaumin, M.; Ferrari, V. Segmentation propagation in imagenet. In Proceedings of the European Conference on Computer Vision (ECCV), Florence, Italy, 7–13October 2012; Springer: Cham, Switzerland, 2012; pp. 459–473.
21. Rubinstein, M.; Joulin, A.; Kopf, J.; Liu, C. Unsupervised joint object discovery and segmentation in internet images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 1939–1946.
22. Rubio, J.C.; Serrat, J.; López, A. Video co-segmentation. In Proceedings of the Asian Conference on Computer Vision (ACCV), Daejeon, Korea, 5–9 November 2012; Springer: Cham, Switzerland, 2012; pp. 13–24.

23. Vicente, S.; Rother, C.; Kolmogorov, V. Object cosegmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 2217–2224.
24. Leordeanu, M.; Collins, R.; Hebert, M. Unsupervised learning of object features from video sequences. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; Volume 1, p. 1142.
25. Liu, D.; Chen, T. A topic-motion model for unsupervised video object discovery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA, 18–23 June 2007; pp. 1–8.
26. Parikh, D.; Chen, T. Unsupervised identification of multiple objects of interest from multiple images: discovery. In Proceedings of the Asian Conference on Computer Vision (ACCV), Tokyo, Japan, 18–22 November 2007; Springer: Cham, Switzerland, 2007; pp. 487–496.
27. Stretcu, O.; Leordeanu, M. Multiple frames matching for object discovery in video. In Proceedings of the British Machine Vision Conference (BMVC), Swansea, UK, 7–10 September 2015; Volume 1, p. 3.
28. Thai, M.T.; Wu, W.; Xiong, H. *Big Data in Complex and Social Networks*; CRC Press: Boca Raton, FL, USA, 2016.
29. Stai, E.; Kafetzoglou, S.; Tsiropoulou, E.E.; Papavassiliou, S. A holistic approach for personalization, relevance feedback & recommendation in enriched multimedia content. *Multimed. Tools Appl.* **2018**, *77*, 283–326.
30. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef] [PubMed]
31. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
32. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
33. Doersch, C.; Gupta, A.; Efros, A.A. Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1422–1430.
34. Noroozi, M.; Pirsiavash, H.; Favaro, P. Representation learning by learning to count. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5898–5906.
35. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2536–2544.
36. Zhang, R.; Isola, P.; Efros, A.A. Colorful image colorization. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 649–666.
37. Noroozi, M.; Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 69–84.
38. Wang, X.; Gupta, A. Unsupervised learning of visual representations using videos. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2794–2802.
39. Wang, X.; He, K.; Gupta, A. Transitive invariance for self-supervised visual representation learning. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1329–1338.
40. Walker, J.; Doersch, C.; Gupta, A.; Hebert, M. An uncertain future: forecasting from static images using variational autoencoders. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 835–851.
41. Pathak, D.; Girshick, R.; Dollár, P.; Darrell, T.; Hariharan, B. Learning features by watching objects move. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2701–2710.
42. Li, Y.; Liu, L.; Shen, C.; van den Hengel, A. Image co-localization by mimicking a good detector's confidence score distribution. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 19–34.
43. Wei, X.S.; Zhang, C.L.; Li, Y.; Xie, C.W.; Wu, J.; Shen, C.; Zhou, Z.H. Deep descriptor transforming for image co-localization. *arXiv* **2017**, arXiv:1705.02758.
44. Wei, X.S.; Luo, J.H.; Wu, J.; Zhou, Z.H. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Trans. Image Process.* **2017**, *26*, 2868–2881. [CrossRef] [PubMed]
45. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
46. Rother, C.; Minka, T.; Blake, A.; Kolmogorov, V. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; Volume 1, pp. 993–1000.
47. Mukherjee, L.; Singh, V.; Dyer, C.R. Half-integrality based algorithms for cosegmentation of images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 2028–2035.
48. Batra, D.; Kowdle, A.; Parikh, D.; Luo, J.; Chen, T. icoseg: Interactive co-segmentation with intelligent scribble guidance. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3169–3176.

49. Hochbaum, D.S.; Singh, V. An efficient algorithm for co-segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Kyoto, Japan, 27 September–4 October 2009; pp. 269–276.
50. Rubio, J.C.; Serrat, J.; López, A.; Paragios, N. Unsupervised co-segmentation through region matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–24 June 2012; pp. 749–756.
51. Lee, C.; Jang, W.D.; Sim, J.Y.; Kim, C.S. Multiple random walkers and their application to image cosegmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3837–3845.
52. Faktor, A.; Irani, M. Video segmentation by non-local consensus voting. In Proceedings of the British Machine Vision Conference (BMVC), Nottingham, UK, 1–5 September 2014; Volume 2, p. 8.
53. Lee, Y.J.; Kim, J.; Grauman, K. Key-segments for video object segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 1995–2002.
54. Wang, W.; Shen, J.; Porikli, F. Saliency-aware geodesic video object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3395–3402.
55. Bideau, P.; Learned-Miller, E. It's moving! A probabilistic model for causal motion segmentation in moving camera videos. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 433–449.
56. Narayana, M.; Hanson, A.; Learned-Miller, E. Coherent motion segmentation in moving camera videos using optical flow orientations. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 1577–1584.
57. Tokmakov, P.; Alahari, K.; Schmid, C. Learning motion patterns in videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3386–3394.
58. Tokmakov, P.; Schmid, C.; Alahari, K. Learning to segment moving objects. *Int. J. Comput. Vis.* **2019**, *127*, 282–301. [CrossRef]
59. Tokmakov, P.; Alahari, K.; Schmid, C. Learning video object segmentation with visual memory. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4481–4490.
60. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
61. Mondal, A.K.; Dolz, J.; Desrosiers, C. Few-shot 3d multi-modal medical image segmentation using generative adversarial learning. *arXiv* **2018**, arXiv:1810.12241.
62. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
63. Isensee, F.; Petersen, J.; Kohl, S.A.; Jäger, P.F.; Maier-Hein, K.H. nnu-net: Breaking the spell on successful medical image segmentation. *arXiv* **2019**, arXiv:1904.08128.
64. Courbariaux, M.; Hubara, I.; Soudry, D.; El-Yaniv, R.; Bengio, Y. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv* **2016**, arXiv:1602.02830.
65. Li, M.; Zuo, W.; Gu, S.; Zhao, D.; Zhang, D. Learning convolutional networks for content-weighted image compression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 3214–3223.
66. Shotton, J.; Winn, J.; Rother, C.; Criminisi, A. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 1–15.
67. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
68. Odena, A.; Dumoulin, V.; Olah, C. Deconvolution and checkerboard artifacts. *Distill* **2016**, *1*, e3. [CrossRef]