electronics

MDPI

Article

Unified Object Detector for Different Modalities Based on Vision Transformers

Xiaoke Shen [1,*] and Ioannis Stamos [2,3]

[1] Mobi Systems, Inc., 48 Grove St., Somerville, MA 02144, USA
[2] Computer Science, Hunter College, The City University of New York, New York, NY 10065, USA; istamos@hunter.cuny.edu
[3] The Graduate Center, The City University of New York, New York, NY 10016, USA
* Correspondence: jimmy@takemobi.com

Abstract: Traditional systems typically require different models for processing different modalities, such as one model for RGB images and another for depth images. Recent research has demonstrated that a single model for one modality can be adapted for another using cross-modality transfer learning. In this paper, we extend this approach by combining cross/inter-modality transfer learning with a vision transformer to develop a unified detector that achieves superior performance across diverse modalities. Our research envisions an application scenario for robotics, where the unified system seamlessly switches between RGB cameras and depth sensors in varying lighting conditions. Importantly, the system requires no model architecture or weight updates to enable this smooth transition. Specifically, the system uses a depth sensor in low light conditions (night time) and both an RGB camera and a depth sensor or RGB camera only in well-lit environments. We evaluate our unified model on the SUN RGB-D dataset and demonstrate that it achieves a similar or better performance in terms of the mAP50 compared to state-of-the-art methods in the SUNRGBD16 category and a comparable performance in point-cloud-only mode. We also introduce a novel inter-modality mixing method that enables our model to achieve significantly better results than previous methods. We provide our code, including training/inference logs and model checkpoints, to facilitate reproducibility and further research.

Keywords: object detection; different modalities; vision transformers; unified model

Citation: Shen, X.; Stamos, I. Unified Object Detector for Different Modalities Based on Vision Transformers. *Electronics* **2023**, *12*, 2571. https://doi.org/10.3390/electronics12122571

Academic Editor: George A. Papakostas

Received: 5 May 2023
Revised: 26 May 2023
Accepted: 1 June 2023
Published: 7 June 2023

1. Introduction

Advances in computer vision and artificial intelligence have enabled the development of increasingly sophisticated robotic applications that enhance human lives. Autonomous vehicles, for instance, can transport individuals to their destination without the need for a human driver/operator, while autonomous mobile robots operating in warehouses can assist in order preparation. However, many robotic systems rely on multiple sensors, such as cameras and 3D sensors (LiDAR or depth), and not all sensors are equally effective in all scenarios. For instance, camera sensors may perform poorly in low-light conditions without supplementary lighting. Thus, the ability to operate in low-light conditions can significantly reduce electricity usage and promote environmentally friendly robot design.

The high accuracy achieved by camera-based vision systems in 2D detection owes much to the efficacy of feature extractors based on ConvNets [1] and Transformers [2]. Concurrently, CrossTrans [3] proposes that by converting 3D sensor data into pseudo images and applying cross-modality transfer learning, a 2D object detection system using identical networks to those used for RGB images can produce commendable results. This development prompts a natural question: can we further enhance performance by training a unified network with both RGB and 3D data, adopting an identical architecture and weights throughout? The proposed unified network accepts three types of sensor data, namely (1) RGB images, (2) pseudo images converted from 3D sensors, and (3) both RGB

images and pseudo images converted from 3D sensors. If a unified network can match or exceed the detection performance of separate networks, each optimized for a particular modality, it would make feasible the use of an eco-friendly system operating under natural lighting conditions during the day and without any extra lighting at night.

One notable advantage of implementing a unified system resides in its capacity to reduce memory consumption. The utilization of separate models for camera, depth, or both camera and depth necessitates approximately two to three times more storage space to accommodate the weight of the models. This advantage carries particular relevance for mobile robotics, which typically possess hardware with relatively limited capabilities when compared to consistently powered computers. Furthermore, as the size of models continues to grow (as exemplified by the Swin Transformer v2 [4] with 3 billion parameters, while the initial version, Swin Tiny [5], comprises less than 0.5 billion parameters), this benefit assumes a significant role in facilitating the integration of large models within robotics applications. This article examines the potential of such a unified system. Building on CrossTrans [3], which demonstrates the superior performance of a Vision-Transformer-based network over ConvNets-based networks, our study concentrates exclusively on the Vision Transformer network.

In summary, this article aims to address the following research questions:

1. Can a unified model achieve comparable or superior performance in processing both RGB images and pseudo images converted from point clouds?
2. If a unified model that processes both RGB and pseudo images is feasible, can the RGB and pseudo images be further fused to enhance the model's ability to process both RGB and point cloud data?

We conducted experiments that resulted in insightful observations and achieved state-of-the-art performance in 2D object detection. Our proposed unified model, named the Unified Object Detector for Different Modalities (UODDM), is capable of processing various types of images, including RGB images, pseudo images converted from point clouds, and inter-modality mixing of RGB images and pseudo images converted from point clouds. Figure 1 illustrates the differences between our model and other works. Furthermore, the performance comparison of different methods can be found in the "Results on the SUN RGB-D Dataset" session. Visualizations of UODDM outputs are presented in Figure 2. More samples can be found in our demo video accessed on 31 May 2023 https://youtu.be/PuRQCLLSUDI.

The key contributions of our work can be summarized as follows:

1. We propose two inter-modality mixing methods which can combine the data from different modalities to further feed to our unified model.
2. We propose a unified model which can process any of the following images: RGB images, pseudo images converted from point clouds or inter-modality mixing of RGB image, and pseudo images converted from point clouds. This unified model achieves similar performance to RGB-only models and point-cloud-only models. Meanwhile, by using the inter-modality mixing data as input, our model can achieve a significantly better 2D detection performance.
3. We open source our code, training/testing logs, and model checkpoints.

Code can be found https://github.com/liketheflower/UODDM, accessed on 31 May 2023.

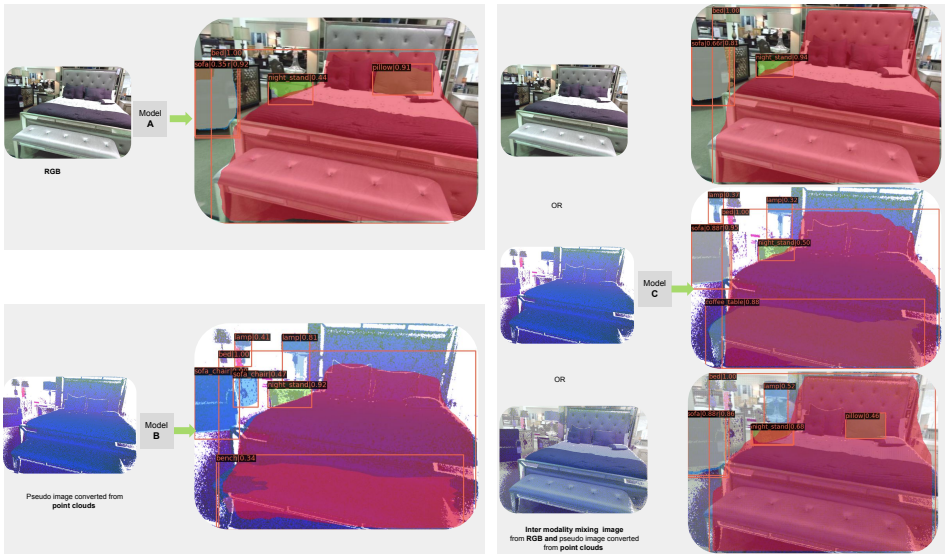

Figure 1. Model A exclusively processes RGB images, with the visualization generated solely from the RGB-trained model presented in this study. Model B operates on pseudo images converted from point clouds, and the visualization is derived from the CrossTrans [3] approach, which trains on these images. Model C is capable of processing RGB images, pseudo images converted from point clouds, or a combination of both. The visualization is based on UODDM with a Swin-T [5] backbone network.

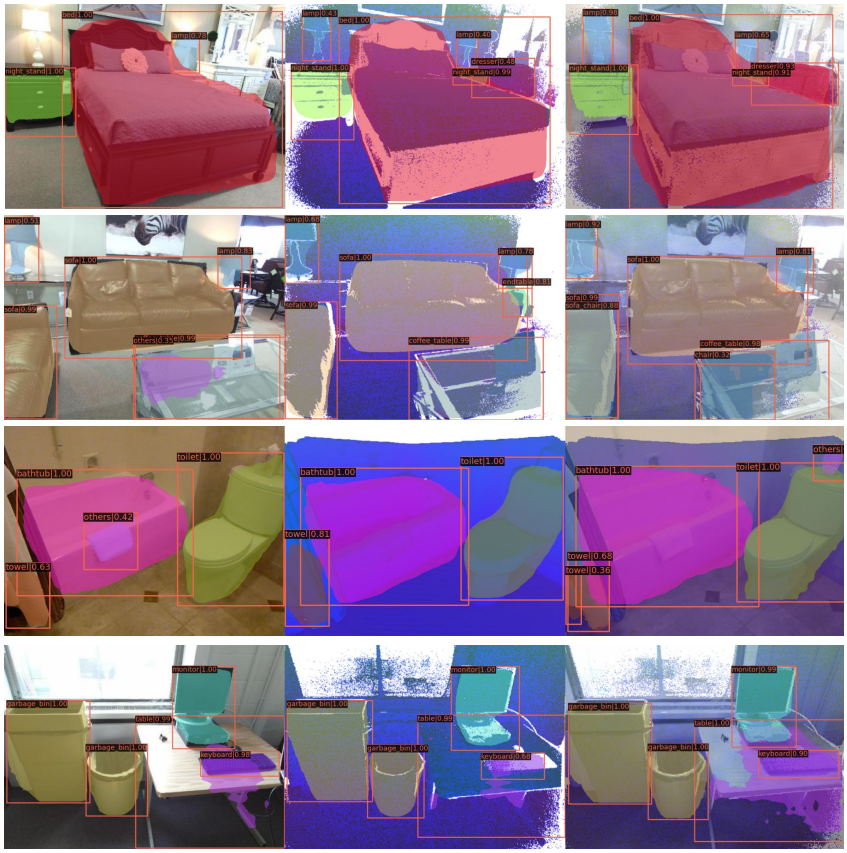

Figure 2. 2D detection results of UODDM, using the SUN RGB-D validation dataset. It showcases four examples of 2D detection visualization, with the left column showing RGB images, the middle column displaying pseudo images converted from point clouds, and the right column illustrating inter-modality mixing of RGB images and pseudo images converted from point clouds. The backbone network used in this study is Swin-T.

2. Related Work

Projecting 3D sensor data to 2D pseudo images: There are different ways to project 3D data to 2D features. HHA was proposed in [6], where the depth image is encoded with three channels: horizontal disparity, height above ground, and the angle of each pixel's local surface normal with gravity direction. The signed angle feature described in [7] measures the elevation of the vector formed by two consecutive points and indicates the convexity or concavity of three consecutive points. Input features converted from depth images of normalized depth (D), normalized relative height (H), angle with up-axis (A), signed angle (S), and missing mask (M) were used in [8]. DHS images are used in [9,10].

Object detection based on RGB images or pseudo images from point clouds by Vision Transformers: Object detection approaches can be summarized as two-stage frameworks (proposal and detection stages) and one-stage frameworks (proposal and detection in parallel). Generally speaking, two-stage methods such as R-CNN [11], Fast RCNN [12], Faster RCNN [13], FPN [14], and mask R-CNN [15] can achieve a better detection performance, while one-stage systems such as YOLO [16], YOLO9000 [17], and RetinaNet [18] are faster at the cost of reduced accuracy. For deep-learning-based systems, as the size of the network is increased, larger datasets are required. Labeled datasets such as PASCAL VOC dataset [19] and COCO (Common Objects in Context) [20] have played important roles in the continuous improvement of 2D detection systems. Most systems introduced here are based on ConvNets. Nice reviews of 2D detection systems can be found in [21]. When replacing the backbone network from ConvNets to Vision Transformer, the systems will be adopted to Vision-Transformers-backbone-based object detection systems. The most successful systems are Swin-transformer [5] and Swin-transformer v2 [4]. CrossTrans [3] explored the cross modality transfer learning by using both the ConvNets and Vision Transfomers based on the SUN RGB-D dataset based on the mask R-CNN [15] approach.

Inter-modality mixing: Ref. [22] learns a dynamical and local linear interpolation between the different regions of cross-modality images in a data-dependent fashion to mix up the RGB and infrared (IR) images. We explored both the static and dynamic mixing methods and found the static has a better performance. Ref. [23] uses an interpolation between the RGB and thermal images at the pixel level. As we are training a unified model supporting both the single modality image and multiple modality images as input, we do not apply interpolation to keep the original signal of each modality. We leverage the transformer architecture itself to automatically build up the gap between different modalities.

Multimodal data fusion: Multimodal data fusion can be performed using three different approaches: early fusion, late fusion, and deep fusion. Early fusion combines various modalities of data at a lower-dimensional common space, and a feature extractor is then employed to extract relevant information. Early fusion has been applied to object detection and audio–visual processing, as demonstrated in [24,25], respectively. Late fusion, on the other hand, employs independent feature extractors for different data sources and merges the extracted features in the final stage. Classical works on deep fusion for action recognition, gesture segmentation and recognition, and emotion recognition are demonstrated in [26–28], respectively. Deep fusion is characterized by fusing data at various stages of model training, transforming the input data into a higher-level representation through multiple layers, and allowing for the fusion of diverse modalities into a single shared representation layer. Various works such as [29–34] have applied deep fusion to object detection. The study in [35] explores all three fusion methods for indoor semantic segmentation. In our research, we have chosen to adopt the early fusion approach for multimodal data processing.

3. Methodology

In this section, we will describe our approach for converting structured point clouds to pseudo images and the methods we use for mixing various modalities, as well as our detection frameworks.

3.1. Convert Point Clouds to Pseudo 2D Images

In order to use pretrained models based on RGB images, we convert point clouds to pseudo 2D images with three channels. The point clouds can be converted to HHA or any three channels from DHASM introduced in [8].

For this work, we follow the same approaches in Frustum VoxNet [9] and CrossTrans [3] by using DHS to project 3D depth data to 2D images [8]. Here, we present a summary of the DHS encoding method. Similar to [6,8], we adopt Depth from the sensor and Height along the sensor-up (vertical) direction as two reliable measures. The Signed angle was introduced in [7] and summarized in [3] as the following: "For the Signed angle: Let us denote as $X_{i,k} = [x_{ik}, y_{ik}, z_{ik}]$ the vector of 3D coordinates of the k-th point in the i-th scanline. Knowledge of the vertical direction (axis z) is provided by many laser scanners, or even can be computed from the data in indoor or outdoor scenarios (based on line/plane detection or segmentation results from machine learning models) and is thus assumed known. Define $D_{i,k} = X_{i,k+1} - X_{i,k}$ (difference of two successive measurements in a given scanline i), and A_{ik}: the angle of the vector $D_{i,k}$ with the pre-determined z axis (0 to 180 degrees). The Signed angle $S_{ik} = sgn(D_{i,k} \cdot D_{i,k-1}) * A_{ik}$: the sign of the dot product between the vectors $D_{i,k}$ and $D_{i,k-1}$, multiplied by V_{ik}. This sign is positive when the two vectors have the same orientation and negative otherwise". Following [3], these three channel pseudo images are normalized to 0 to 1 for each channel. Some samples DHS images can be seen in Figures 3 and 4.

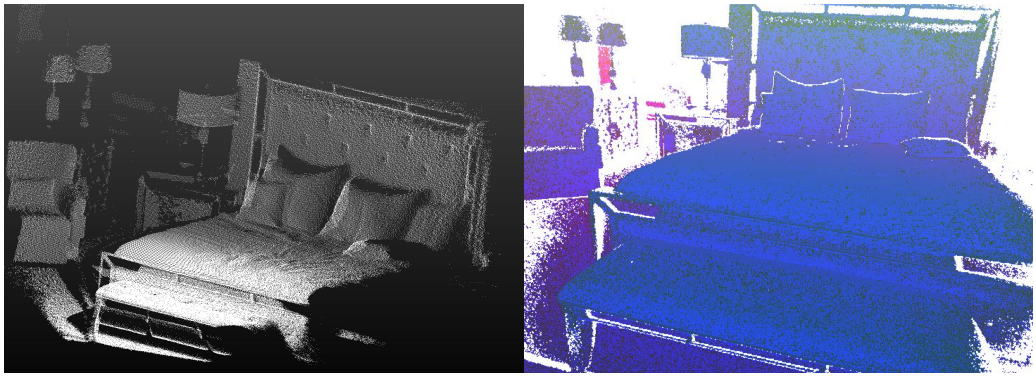

Figure 3. An example of converted pseudo three channel image from the point cloud. This example is one image selected from the validation set of the SUN RGB-D dataset. The corresponding RGB image can be found in Figure 1.

Figure 4. Inter-modality mixing. Left column: original RGB image (**up**) and the original DHS image (**bottom**). Middle column: the Chessboard Per Patch Mixing image. Patch size of 15 by 15 pixels (**up**) and 1 by 1 pixel (**bottom**). Right column: the Stochastic Flood Fill Mixing image. Edge connection probability of 0.5 and 0.5 for RGB and DHS, respectively (**up**), probability of 0.1 and 0.1 for RGB and DHS, respectively (**bottom**).

3.2. Inter-Modality Mixing

In order to expand the input options for our unified model, we introduce an inter-modality mixing approach that enables us to combine images from different modalities into a three-channel image for consumption by the model. This approach allows us to enhance the model's capabilities without modifying its architecture. By training a model using RGB images, DHS images, and the mixed RGB and DHS images, we can achieve a unified detector that is capable of processing different modalities as input.

When considering the fusion of different modalities, two approaches can be employed: mixing them into three channels, as implemented in our study, or utilizing a six-channel (RGBDHS) image configuration. However, we advocate for the former approach for the following reasons: Firstly, adopting a consistent channel number for RGB, DHS, and mixed RGB–DHS images enables the construction of a unified model. Secondly, leveraging pretrained weights from RGB images for the mixed RGB–DHS images confers notable benefits. The work [9] demonstrates that training a six-channel RGBDHS model from scratch yields a significantly inferior 2D detection performance.

Various techniques can be employed to fuse images from different modalities, and we propose two approaches:

- Per Patch Mixing (PPM): divide the whole image into different patches with equal patch size. Randomly or alternatively select one image source for each patch.
- Stochastic Flood Fill Mixing (SFFM): Using a stochastic way to mix the images from different modalities.

We implement the Per Patch Mixing approach with relative simplicity. Specifically, for each patch in the image, we alternatively selected a modality image to assign to that patch. Moreover, we opted to utilize square patches for our implementation. As a result, the mask for selecting the modality image for each patch resembles a chessboard pattern, leading us to refer to our implementation as Chessboard Per Patch Mixing (CPPM). Examples of CPPM are shown in the middle of Figure 4.

The Stochastic Flood Fill Mixing technique is an adaptation of the flood fill algorithm [36]. The approach involves establishing connections between neighboring pixels with a probability p, with separate probabilities for the RGB and DHS modalities. The algorithm can be implemented using four or eight neighbors to build the graph, with the latter including additional diagonal offsets. In our experiments, we used the four neighbor approach. The Python-style pseudocode for this algorithm is illustrated in Figure 5, while examples of SFFM are shown on the right side of Figure 4.

3.3. Two-Dimensional Detection Framework

For the purpose of 2D detection and instance segmentation, we adopt the conventional object detection framework, namely Mask R-CNN [15], which is implemented in MMDetection [37]. It follows a two-stage approach [21], namely region proposal and detection/segmentation, for accomplishing detection and segmentation tasks. During the fine-tuning of the model on the SUN RGB-D dataset, we disable the training of the mask branch. However, even with the default weights from the pre-trained model, the mask prediction branch can still generate acceptable mask predictions, as demonstrated in Figure 1. This observation aligns with the findings of the research on CrossTrans [3].

```
1   EXPLORED, UNEXPLORED = 1, -1
2   RGB, DHS, NOT_RGB_NOT_DHS = 0, 1, -1
3   connect_probs = {RGB: 0.5, DHS: 0.5}
4   neighbors = [(0, 1), (0, -1), (1, 0), (-1, 0)]
5
6
7   def opposite_image_type(img_type):
8       return RGB if img_type == DHS else DHS
9
10
11  def stochastic_flood_fill(i, j, status, mask, curr_img_type):
12      if (i, j) is not valid or status[i][j] == EXPLORED:return
13      fill_this_location = False
14      if mask[i][j] == curr_img_type:  # if same image type, mask it
15          fill_this_location = True
16      elif mask[i][j] == NOT_RGB_NOT_DHS:
17          if random_number <= connect_probs[curr_img_type]:  # stochastic part
18              fill_this_location = True
19          else:
20              mask[i][j] = opposite_image_type(curr_img_type)
21
22      if fill_this_location:
23          mask[i][j], status[i][j] == curr_img_type, EXPLORED
24          for di, dj in neighbors:
25              stochastic_flood_fill(i + di, j + dj, status, mask, curr_img_type)
26
27
28  def get_stochastic_flood_fill_mask(img_height, img_width):
29      H, W = img_height, img_width
30      status = UNEXPLORED * np.ones((H, W))
31      mask = NOT_RGB_NOT_DHS * np.ones((H, W))
32      mask[0][0] = RGB  # use first pixel as RGB for example
33      for i in range(H):
34          for j in range(W):
35              if status[i][j] == EXPLORED:continue
36              curr_img_type = mask[i][j]
37              status, mask = stochastic_flood_fill(i, j, status, mask, curr_img_type)
38      return mask
```

Figure 5. The pseudocode for Stochastic Flood Fill Mixing is presented in Python style, with line 17 to 20 representing the stochastic aspect that differentiates it from the original flood fill algorithm. The mask is used to determine which image's pixel value should be used in generating the mixing image.

3.4. Two-Dimensional Detection Backbone Networks

For the backbone network, we use Swin Transformer [5]; specifically, we explored Swin-Tiny's and Swin-Small's performance. The complexities of Swin-T and Swin-S are similar to those of ResNet-50 and ResNet-101, respectively. The window size is set to $M = 7$ by default. The query dimension of each head is $d = 32$, and the expansion layer of each MLP is $\alpha = 4$. The architecture hyper-parameters of these two models are:

- Swin-T: C = 96, layer numbers = $\{2, 2, 6, 2\}$.
- Swin-S: C = 96, layer numbers = $\{2, 2, 18, 2\}$.

C is the channel number of the hidden layers in the first stage. For details of the model architecture, please check the Swin Transformers [5] paper.

3.5. SUN RGB-D Dataset Used in This Work

The SUN RGB-D [38] dataset is an indoor dataset which provides both point cloud and RGB images. In this work, since we are building a 3D only object detection system, we only use the point clouds for fine tuning. The RGB images are not used during the fine tuning process. For the point clouds, they are collected based on four types of sensors:

Intel RealSense, Asus Xtion, Kinect v1, and Kinect v2. The first three sensors used an IR light pattern. Kinect v2 is based on the time of flight. The longest distance captured by the sensors is around 3.5 to 4.5 meters.

The SUN RGB-D dataset is split into a training set, which contains 5285 images, and a testing set, which contains 5050 images. For the training set, it is further split into a training only set, which contains 2666 images, and a validation set, which contains 2619 images. Similar to [9,10,39,40], we fine tune our model based on the training only set and evaluate our system based on the validation set.

3.6. Pre-Training

Both the Swin-T- and Swin-S-based networks (the pretrained weights are loaded from mmdetection [37]) are firstly pre-trained on ImageNet [41] and then pre-trained on the COCO dataset [20].

Data augmentation: When pre-training on the COCO dataset, image augmentations are applied during the training stage by randomly horizontally flipping the image with probability of 0.5; randomly resizing the image with a width of 1333 and a height from 480 to 800 (for details see the configure file from the github repository); randomly cropping the original image with a size of 384 (height) by 600 (width); and resizing the cropped image to a width of 1333 and a height from 480 to 800.

3.7. Fine-Tuning

Data augmentation: We follow the same augmentation as the pre-training stage. The raw input images have a width of 730 and a height of 530. These raw images are randomly resized and cropped during training. During testing, the images are resized to a width of 1120 and a height of 800, which can be divided by 32.

Hardware: For fine tuning, we use a standard single NVIDIA Titan-X GPU, which has a 12 GB memory. We fine tune the network for 133 K iterations for 100 epochs. It took about 29 h for the Swin-T-based network with a batch size of 2 (for 133 K iterations) for the RGB-only model. For the UODDM without the inter-modality mixing, it took about 2 days to train the model. For the UODDM with inter-modality mixing, the speed depends on the number of inter-modality mixing images added to the training data.

Fine-tuning subtasks: We are focused on the 2D object detection performance, so we fine tuned the model based on 2D-detection-related labels. Similar to CrossTrans [3], we did not train the mask branch to further verify whether reasonable mask detection can be created by using the weights from the pre-training stage.

4. Results on the SUN RGB-D Dataset

4.1. Experiments

The primary focus of our experiments centers around the training of the model using diverse input data and a comparison of performance differences. Specifically, we first trained a unified model on both RGB and DHS images for the UODDM without inter-modality mixing. In this procedure, during the training stage, RGB and DHS images were combined, resulting in a mixed dataset. When forming batches for training, both RGB and DHS images were selected randomly from this combined dataset. In contrast, for the UODDM with inter-modality mixing, we augmented the training data with inter-modality mixing images, in addition to the RGB and DHS images.

4.2. Evaluation Metrics

Following the previous works [3,6,9,40,42], we firstly used the AP50 (Average Precision at IoU = 0.5) as an evaluation metric. We also used the COCO object detection metric which is AP75 (Average Precision at IoU = 0.75) and a more strict one, AP at IoU = 0.50:0.05:0.95, to evaluate the 2D detection performance.

4.3. Evaluation Subgroups

We used the same subgroups as CrossTrans [3] to evaluate the performance. The subgroups are SUNRGBD10, SUNRGBD16, SUNRGBD66, and SUNRGBD79, which have 10, 16, 66, and 79 categories. A detailed list of these subgroups can be found in CrossTrans [3].

4.4. The Performance of UODDM without Inter-Modality Mixing

We first evaluated the performance of UODDM without inter-modality mixing. For this, the model was trained based on both the RGB and DHS images. Our model architecture is the same as the CrossTrans [3] work, which uses only DHS images to train the model. We traines a RGB-image-only model based on the same network to compare with the UODDM one's performance.

The performance evaluation of our proposed UODDM approach, measured in terms of mean average precision (mAP50), on the SUNRGBD79 dataset is presented in Figure 6. The results show that the UODDM model performs exceptionally well on both RGB and DHS images. Additionally, it is evident that the UODDM model significantly outperforms the DHS-only model in terms of performance on DHS images, which can be attributed to the inter-modality transfer learning from the RGB images. However, this performance improvement on DHS images comes at the slight cost of a performance reduction on RGB images. Nevertheless, the UODDM model's overall performance is promising as it is a single model that can handle different modalities, making it more efficient than maintaining two separate architectures or a single architecture with two different sets of weights. This efficiency is particularly valuable for robotics and edge devices, where a seamless perception system can be built, even when transitioning from daytime to nighttime scenarios. Table 1 provides additional results for our UODDM and single-modality models, reinforcing the same conclusions.

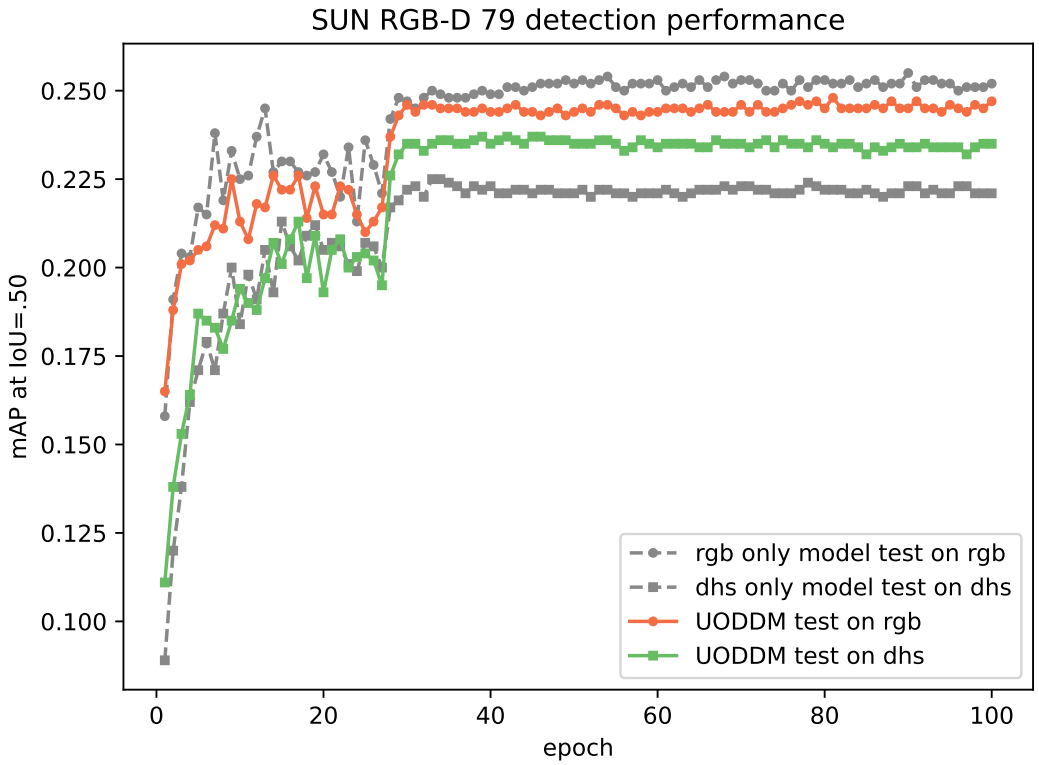

Figure 6. Comparison of the unified model's and separate models' performances with the training epochs. The RGB-only model is our new trained model based on RGB images. The DHS-only model is from CrossTrans [3]. The UODDM is trained based on both RGB and DHS images. The backbone for all these experiments is based on the Swin-T model.

Table 1. Results comparison based on mAP50 for different subgroups of UODDM and single modality only models. The bold ones represent the best results within each test modality category.

Model	Test on	Backbone	SUNRGBD10	SUNRGBD16	SUNRGBD66	SUNRGBD79
RGB only (ours)	RGB	Swin-T	**54.2**	52.3	**29.3**	**25.2**
UODDM (ours)	RGB	Swin-T	53.9	**52.5**	28.7	24.7
DHS only (CrossTrans [3])	DHS	Swin-T	55.8	52.7	26.1	22.1
UODDM (ours)	DHS	Swin-T	**56.6**	**53.4**	**27.7**	**23.5**

4.5. The Performance of UODDM with Inter-Modality Mixing

In our study, we investigated two different methods for inter-modality mixing, namely SFFM and CPPM. For SFFM, we generated six mixing images for each RGB and DHS image pair, with connection probabilities for RGB and DHS pixels being randomly selected from the range of 0.1 to 0.9. The first pixel's RGB and DHS masks were randomly initialized with equal probability. In contrast, for CPPM, we used square patches of size 1 by 1, resulting in one CPPM image for each RGB and DHS image pair. The performance of both approaches was evaluated and is presented in Table 2. Notably, the results suggest that the UODDM with CPPM outperforms the UODDM with SFFM. We attribute this to the generation of an excessive number of random images by SFFM, which can negatively impact the performance of the unified network on RGB and DHS images. Conversely, CPPM provides a comparable performance to the plain UODDM model. Furthermore, the use of the CPPM images generated from both RGB and DHS images led to the best 2D detection performance. Given the ability of UODDM with CPPM to support RGB, DHS, and CPPM images from RGB and DHS images, we propose it as a more powerful unified model.

Table 2. Results comparison based on mAP50 for different subgroups of UODDM and single-modality-only models. The bold ones represent the best results within each test modality category.

Model	Test on	Backbone	SUNRGBD10	SUNRGBD16	SUNRGBD66	SUNRGBD79
UODDM	RGB	Swin-T	53.9	52.5	**28.7**	**24.7**
UODDM + SFFM	RGB	Swin-T	24.6	17.5	19.2	20.1
UODDM + CPPM	RGB	Swin-T	54.2	51.9	27.7	23.7
UODDM + CPPM	RGB	Swin-S	**54.6**	**52.7**	27.5	23.6
UODDM	DHS	Swin-T	56.6	**53.4**	**27.7**	**23.5**
UODDM + SFFM	DHS	Swin-T	25.6	18.7	20.0	21.3
UODDM + CPPM	DHS	Swin-T	55.8	52.8	26.3	22.4
UODDM + CPPM	DHS	Swin-S	**57.4**	52.5	24.8	21.1
UODDM + CPPM	CPPM	Swin-T	58.1	55.8	**29.5**	**25.2**
UODDM + CPPM	CPPM	Swin-S	**58.4**	**56.1**	28.4	24.5

4.6. Influence of Different Backbone Networks

Table 2 presents the results obtained by using Swin-T and Swin-S as the backbone networks. It is observed that Swin-S is a more powerful network; however, the performance gain achieved is limited. Therefore, we propose the usage of the lightweight Swin-T as the backbone network to achieve a faster inference speed.

4.7. Comparison with Other Methods

In Table 3, we present a detailed comparison of per category results with previous works. Specifically, we evaluate the performance of our approach under three different input scenarios: RGB images, point cloud data, and a combination of RGB and point cloud data using our proposed inter-modality mixing method.

Table 3. Two-dimensional detection results based on the SUN RGB-D validation set. The evaluation metric is average precision, with a 2D IoU threshold of 0.5. The 10 categories SUNRGBD10 results are shown. The image source indicates the image type used during inference. During the training phase, the UODDM is the only system that employs images from various sources, as previously described. In contrast, the remaining systems utilize the same image source for both training and testing. The best two results across all image source are highlighted in **red** and **blue**, respectively. The best result for each image source is shown in **bold**.

Image Source (for testing)	Methods	Backbone	Bed	Toilet	Night Stand	Bathtub	Chair	Dresser	Sofa	Table	Desk	Bookshelf	SUNRGBD10 mAP_{50}
RGB	2D driven [40]	VGG-16	74.5	86.2	49.5	45.5	53.0	29.4	49.0	42.3	22.3	45.7	49.7
	Frustum PointNets [42]	VGG	56.7	43.5	37.2	**81.3**	**64.1**	**33.3**	**57.4**	**49.9**	**77.8**	**67.2**	**56.8**
	F-VoxNet [9]	ResNet 101	81.0	89.5	35.1	50.0	52.4	21.9	53.1	37.7	18.3	40.4	47.9
	RGB only model (ours)	Swin-T	83.2	**93.9**	51.8	54.2	60.4	23.7	51.3	46.3	22.5	**54.4**	54.2
	UODDM (ours)	Swin-T	**83.6**	87.1	**53.3**	58.8	62.5	22.6	54.2	46.8	22.0	48.0	53.9
	UODDM + CPPM (ours)	Swin-T	**83.6**	88.6	53.0	59.1	60.8	26.5	50.7	46.1	22.0	52.0	54.2
Depth/Point Cloud	F-VoxNet [9]	ResNet 101	78.7	77.6	34.2	51.9	51.8	16.5	48.5	34.9	14.2	19.2	42.8
	CrossTrans [3]	Swin-T	87.2	**87.7**	51.6	**69.5**	69.0	27.0	60.5	**48.1**	19.3	38.3	55.8
	UODDM (ours)	Swin-T	**88.1**	87.6	**53.8**	66.8	**69.5**	**28.7**	**62.2**	47.2	19.7	**41.9**	**56.6**
	UODDM + CPPM (ours)	Swin-T	**88.0**	85.6	51.8	68.3	**68.6**	26.9	**61.6**	45.5	**20.2**	41.7	55.8
RGB and Depth/Point Cloud	RGB-D RCNN [6]	VGG	76.0	69.8	37.1	49.6	41.2	**31.3**	42.2	43.0	16.6	34.9	44.2
	UODDM + CPPM (ours)	Swin-T	**86.5**	**91.0**	**54.4**	**70.2**	**67.2**	30.3	**57.5**	**48.7**	**22.8**	**52.7**	**58.1**

When considering RGB images as inputs, we observe that our best performing UODDM with CPPM or RGB-only model achieve slightly worse performance (54.2 mAP50 on SUNRGBD10) than the state-of-the-art Frustum PointNets [42]. On the other hand, when utilizing only point cloud data as inputs, our plain UODDM model (without inter-modality mixing) demonstrates a slightly better performance (56.6 mAP50 on SUNRGBD10) compared to the previous state-of-the-art method [3].

Remarkably, our proposed UODDM with CPPM significantly outperforms the previous best results obtained by RGB-D RCNN [6] (58.1 mAP50 on SUNRGBD10) in the scenario where both RGB and point cloud data are available. Notably, most prior works have focused on utilizing either RGB or point cloud data, with limited exploration of mixing methods for these modalities. Therefore, the proposed inter-modality mixing method constitutes a significant contribution to the field.

Moreover, our UODDM with CPPM method demonstrates a substantial performance gain compared to the strongest 2D detector for RGB images, i.e., Frustum PointNets [42]. Specifically, our approach achieves 58.1 mAP50 on SUNRGBD10, which is superior to the performance of Frustum PointNets (56.8 mAP50 on SUNRGBD10).

The results of SUNRGBD16, which has 16 categories, can be found is Table 4. From the results, we can see that our UODDM has a much better performance than previous state-of-the-art methods. Among the new methods, as expected, the UODDM + CPPM using both the RGB images and depth as inputs achieves the best result.

Table 4. Two-dimensional detection results based on the SUN RGB-D validation set for SUNRGBD16. The evaluation metric is average precision with a 2D IoU threshold of 0.5. Since the 10 categories are provided in Table 3, only the results of the remaining 6 categories are shown in this table. The bold ones represent the best results across all test modality categories.

Image Source (for Testing)	Methods	Backbone	Sofa Chair	Kitchen Counter	Kitchen Cabinet	Garbage Bin	Microwave	Sink	SUNRGBD16 mAP_{50}
RGB	F-VoxNet [9]	ResNet 101	47.8	22.0	29.8	52.8	39.7	31.0	43.9
	RGB-only model (ours)	Swin-T	60.4	**32.7**	39.8	67.0	48.1	47.3	52.3
	UODDM (ours)	Swin-T	63.7	28.9	38.4	67.1	57.4	46.2	52.5
	UODDM + CPPM (ours)	Swin-T	60.6	31.2	37.8	64.7	54.3	40.1	51.9
Depth/Point Cloud	F-VoxNet [9]	ResNet 101	48.7	19.1	18.5	30.3	22.2	30.1	37.3
	CrossTrans [3]	Swin-T	68.1	30.7	35.5	61.2	41.9	47.7	52.7
	UODDM (ours)	Swin-T	**68.5**	28.5	35.8	62.8	41.9	**51.5**	53.4
	UODDM + CPPM (ours)	Swin-T	67.6	29.2	33.0	61.6	47.4	47.5	52.8
RGB and Depth/Point Cloud	UODDM + CPPM (ours)	Swin-T	66.6	29.2	**41.2**	**68.6**	**57.9**	48.0	**55.8**

4.8. More Results Based on Extra Evaluation Metrics

More results based on mAP/mAP75 can be found in the Appendix A.

4.9. Number of Parameters and Inference Time

Table 5 presents the number of parameters and inference time for our proposed network architecture. The inference time reported for the Swin-T-based network is the same as that reported in the CrossTrans [3] paper, as we used the same network and hardware. However, since the Swin-S-based network is larger, the inference time is slower, which is expected.

Table 5. Number of parameters and inference time comparison. All speed tests are based on a standard single NVIDIA Titan-X GPU.

Method	Backbone Network	# Parameters (M)	GFLOPs	Inference Time (ms)	FPS
F-VoxNet [9]	ResNet-101	64	-	110	9.1
CrossTrans [3]	ResNet-50	44	472.1	70	14.3
CrossTrans [3]	Swin-T	48	476.5	105	9.5
UODDM (ours)	Swin-T	48	476.5	105	9.5
UODDM (ours)	Swin-S	69	419.7	148	6.8

5. Conclusions

This paper proposes novel inter-modality mixing methods and presents a unified model capable of processing various types of data modalities, including RGB images from cameras, DHS images from depth sensor, and inter-modality mixing images from both RGB and DHS sources. This unified model demonstrates a comparable performance to those of individual models trained on each modality. By eliminating the need to maintain distinct models for different modalities, this unified model exhibits a high memory efficiency and can be highly reliable in robotic perception systems, particularly in scenarios involving varying modalities such as day and night conditions.

Author Contributions: Conceptualization, I.S.; Methodology, X.S.; Software, X.S.; Formal analysis, X.S.; Investigation, X.S.; Resources, I.S.; Writing—original draft, X.S.; Writing—review & editing, I.S.; Visualization, X.S.; Supervision, I.S.; Funding acquisition, I.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by The National Science Foundation (NSF) of the United States Award CNS1625843.

Data Availability Statement: The data used in this research can be found and downloaded from this website https://rgbd.cs.princeton.edu.

Acknowledgments: We would like to express our gratitude to Zhujun Li and Jaime Canizales for their valuable comments and advice during the development of this work. We would also like to thank Zhujun Li for suggesting the name "chessboard" to describe the method of alternatively selecting a modality image based on square patches.

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

Appendix A. More Results

Besides the AP50, which was mainly used in previous works, we also use AP75 and AP to compare the results based on different methods. Meanwhile, we also report AP Across Scales of small, medium, and large by following the same standard of the COCO dataset. These results can be found in Table A1. From the results, we see that, in general, the UODDM with CPPM achieves the best performance on CPPM images. This is mainly due to the fact that both the RGB and DHS images are used for the system. When only using RGB images and when only using DHS images, the unified model UODDM with CPPM has a similar performance to the single-modality-based model.

Table A1. Further results comparisons based on AP@IoU = 0.75, AP, and AP of different scales. As most other works shown in Table 3 did not report the results AP@IoU = 0.75, AP, and AP of different scales, these works are not included in this table. The bold ones represent the best results across all test modality categories.

Method	Test on	Backbone Network	SUNRGBD10			SUNRGBD16			SUNRGBD66			SUNRGBD79					
			AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RGB only (Ours)	RGB	Swin-T	29.6	54.2	28.9	28.6	52.3	28.3	15.4	29.3	14.3	13.1	25.2	12.1	1.0	5.2	16.8
UODDM (Ours)	RGB	Swin-T	30.7	53.9	30.9	29.5	52.5	29.7	15.3	28.7	14.5	13.1	24.7	12.2	0.2	4.6	16.9
UODDM + CPPM (Ours)	RGB	Swin-T	30.9	54.2	30.5	29.2	51.9	28.9	14.5	27.7	13.2	12.4	23.7	11.1	0.7	4.1	15.7
UODDM (Ours)	RGB	Swin-S	31.8	54.7	32.3	30.0	52.5	29.6	15.1	28.2	13.7	12.9	24.4	11.6	0.4	4.2	16.1
UODDM + CPPM (Ours)	RGB	Swin-S	31.3	54.6	31.3	29.9	52.7	29.6	14.6	27.5	13.5	12.4	23.6	11.4	0.7	3.9	15.5
CrossTrans [3]	DHS	Swin-T	33.3	55.8	34.7	30.7	52.7	31.5	14.3	26.1	14.0	12.0	22.1	11.7	0.6	4.7	15.2
UODDM (Ours)	DHS	Swin-T	34.0	56.6	34.9	31.4	53.4	31.9	15.4	27.7	14.8	13.0	23.5	12.4	0.4	4.6	16.4
UODDM + CPPM (Ours)	DHS	Swin-T	33.8	55.8	35.9	31.3	52.8	32.5	14.9	26.3	14.7	12.6	22.4	12.4	0.6	4.4	16.0
UODDM (Ours)	DHS	Swin-S	**34.8**	57.6	**37.0**	31.7	53.7	32.4	14.8	26.6	14.2	12.5	22.6	12.0	0.7	4.2	15.7
UODDM + CPPM (Ours)	DHS	Swin-S	34.1	57.4	35.8	30.9	52.5	31.8	14.1	24.8	14.0	11.9	21.1	11.8	1.1	4.5	15.1
UODDM + CPPM (Ours)	CPPM images	Swin-T	34.2	58.1	35.0	32.6	55.8	33.2	**16.3**	**29.5**	**15.9**	**13.8**	**25.2**	**13.4**	0.4	5.0	**17.1**
UODDM + CPPM (Ours)	CPPM images	Swin-S	34.6	**58.4**	36.0	**33.0**	**56.1**	**34.4**	15.9	28.4	15.7	13.7	24.5	13.4	**1.2**	**5.4**	16.8

References

1. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
2. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
3. Shen, X.; Stamos, I. simCrossTrans: A Simple Cross-Modality Transfer Learning for Object Detection with ConvNets or Vision Transformers. *arXiv* **2022**, arXiv:2203.10456.
4. Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. Swin Transformer V2: Scaling Up Capacity and Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12009–12019.
5. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the International Conference on Computer Vision (ICCV), Montreal, Canada, 11–17 October 2021.
6. Gupta, S.; Girshick, R.B.; Arbeláez, P.A.; Malik, J. Learning Rich Features from RGB-D Images for Object Detection and Segmentation. In Proceedings of the Computer Vision-ECCV 2014-13th European Conference, Zurich, Switzerland, 6–12 September 2014; Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8695, pp. 345–360. [CrossRef]
7. Stamos, I.; Hadjiliadis, O.; Zhang, H.; Flynn, T. Online Algorithms for Classification of Urban Objects in 3D Point Clouds. In Proceedings of the 2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization Transmission, Zurich, Switzerland, 13–15 October 2012; pp. 332–339. [CrossRef]
8. Zelener, A.; Stamos, I. CNN-Based Object Segmentation in Urban LIDAR with Missing Points. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 417–425. [CrossRef]
9. Shen, X.; Stamos, I. Frustum VoxNet for 3D object detection from RGB-D or Depth images. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Village, CO, USA, 1–5 March 2020.
10. Shen, X.; Stamos, I. 3D Object Detection and Instance Segmentation from 3D Range and 2D Color Images. *Sensors* **2021**, *21*, 1213. [CrossRef] [PubMed]
11. Girshick, R.B.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv* **2013**, arXiv:1311.2524.
12. Girshick, R.B. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 December 2015; IEEE Computer Society: Washington, DC, USA, 2015; pp. 1440–1448. [CrossRef]
13. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
14. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 22–25 July 2017.
15. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 22–25 July 2017; pp. 2980–2988. [CrossRef]
16. Redmon, J.; Divvala, S.K.; Girshick, R.B.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *arXiv* **2015**, arXiv:1506.02640.
17. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. *arXiv* **2016**, arXiv:1612.08242.
18. Lin, T.; Goyal, P.; Girshick, R.B.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *arXiv* **2017**, arXiv:1708.02002.
19. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. 2012. Available online: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html (accessed on 31 May 2023).
20. Lin, T.; Maire, M.; Belongie, S.J.; Bourdev, L.D.; Girshick, R.B.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. *arXiv* **2014**, arXiv:1405.0312.
21. Shen, X. A survey of Object Classification and Detection based on 2D/3D data. *arXiv* **2019**, arXiv:1905.12683.
22. Huang, Z.; Liu, J.; Li, L.; Zheng, K.; Zha, Z. Modality-Adaptive Mixup and Invariant Decomposition for RGB-Infrared Person Re-Identification. *arXiv* **2022**, arXiv:2203.01735.
23. Ling, Y.; Zhong, Z.; Luo, Z.; Rota, P.; Li, S.; Sebe, N. Class-Aware Modality Mix and Center-Guided Metric Learning for Visible-Thermal Person Re-Identification. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 889–897.
24. Enzweiler, M.; Gavrila, D.M. A Multilevel Mixture-of-Experts Framework for Pedestrian Classification. *IEEE Trans. Image Process.* **2011**, *20*, 2967–2979. [CrossRef]
25. Barnum, G.; Talukder, S.; Yue, Y. On the Benefits of Early Fusion in Multimodal Representation Learning. *arXiv* **2020**, arXiv:2011.07191.

26. Simonyan, K.; Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, USA, 8–13 December 2014; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K., Eds.; Curran Associates, Inc.: Dutchess County, NY, USA, 2014; Volume 27.
27. Wu, D.; Pigou, L.; Kindermans, P.J.; Le, N.D.H.; Shao, L.; Dambre, J.; Odobez, J.M. Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1583–1597. [CrossRef]
28. Kahou, S.E.; Pal, C.; Bouthillier, X.; Froumenty, P.; Gülçehre, C.; Memisevic, R.; Vincent, P.; Courville, A.; Bengio, Y.; Ferrari, R.C.; et al. Combining Modality Specific Deep Neural Networks for Emotion Recognition in Video. In Proceedings of the 15th ACM on International Conference on Multimodal Interaction, ICMI '13, Sydney, Australia, 9–13 December 2013; Association for Computing Machinery: New York, NY, USA, 2013; pp. 543–550. [CrossRef]
29. Wu, Z.; Gobichettipalayam, S.; Tamadazte, B.; Allibert, G.; Paudel, D.; Demonceaux, C. Robust RGB-D Fusion for Saliency Detection. In Proceedings of the 2022 International Conference on 3D Vision (3DV), Prague, Czech Republic, 12–16 September 2022; IEEE Computer Society: Los Alamitos, CA, USA, 2022; pp. 403–413. [CrossRef]
30. Zhou, Z.; Wu, Z.; Boutteau, R.; Yang, F.; Demonceaux, C.; Ginhac, D. RGB-Event Fusion for Moving Object Detection in Autonomous Driving. *arXiv* **2023**, arXiv:2209.08323.
31. Wu, Z.; Allibert, G.; Meriaudeau, F.; Ma, C.; Demonceaux, C. HiDAnet: RGB-D Salient Object Detection via Hierarchical Depth Awareness. *IEEE Trans. Image Process.* **2023**, *32*, 2160–2173. [CrossRef]
32. Larsson, G.; Maire, M.; Shakhnarovich, G. FractalNet: Ultra-Deep Neural Networks without Residuals. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
33. Wang, J.; Wei, Z.; Zhang, T.; Zeng, W. Deeply-Fused Nets. *arXiv* **2016**, arXiv:1605.07716.
34. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-View 3D Object Detection Network for Autonomous Driving. *arXiv* **2016**, arXiv:1611.07759.
35. Li, Y.; Zhang, J.; Cheng, Y.; Huang, K.; Tan, T. Semantics-guided multi-level RGB-D feature fusion for indoor semantic segmentation. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 1262–1266. [CrossRef]
36. Wikipedia Contributors. Flood Fill—Wikipedia, The Free Encyclopedia. 2022. Available online: https://en.wikipedia.org/w/index.php?title=Flood_fill&oldid=1087894346 (accessed on 26 June 2022).
37. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv* **2019**, arXiv:1906.07155.
38. Song, S.; Lichtenberg, S.P.; Xiao, J. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In Proceedings of the The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
39. Song, S.; Xiao, J. Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images. *arXiv* **2015**, arXiv:1511.02300.
40. Lahoud, J.; Ghanem, B. 2D-Driven 3D Object Detection in RGB-D Images. In Proceedings of the the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
41. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
42. Qi, C.R.; Liu, W.; Wu, C.; Su, H.; Guibas, L.J. Frustum PointNets for 3D Object Detection From RGB-D Data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.