Application on the Loop of Multimodal Image Fusion: Trends on Deep-Learning Based Approaches

Gisel Bastidas-Guacho
ESPOL Polytechnic University, Ecuador
ESPOCH Polytechnic University, Ecuador
gis.bastidas@espoch.edu.ec

Patricio Moreno-Vallejo
ESPOL Polytechnic University, Ecuador
ESPOCH Polytechnic University, Ecuador
xavier.moreno@espoch.edu.ec

Boris Vintimilla
ESPOL Polytechnic University, Ecuador
boris.vintimilla@espol.edu.ec

Angel D. Sappa
ESPOL Polytechnic University, Ecuador
Computer Vision Center, Spain
sappa@ieee.org

2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS) | 979-8-3503-3337-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICPRS58416.2023.10179005

***Abstract*—Multimodal image fusion allows the combination of information from different modalities, which is useful for tasks such as object detection, edge detection, and tracking, to name a few. Using the fused representation for applications results in better task performance. There are several image fusion approaches, which have been summarized in surveys. However, the existing surveys focus on image fusion approaches where the application on the loop of multimodal image fusion is not considered. On the contrary, this study summarizes deep learning-based multimodal image fusion for computer vision (e.g., object detection) and image processing applications (e.g., semantic segmentation), that is, approaches where the application module leverages the multimodal fusion process to enhance the final result. Firstly, we introduce image fusion and the existing general frameworks for image fusion tasks such as multifocus, multiexposure and multimodal. Then, we describe the multimodal image fusion approaches. Next, we review the state-of-the-art deep learning multimodal image fusion approaches for vision applications. Finally, we conclude our survey with the trends of task-driven multimodal image fusion.**

***Index Terms*—multimodal, fusion, deep learning, task-driven**

I. Introduction

Nowadays, several sensors generate data allowing the acquisition of a large amount of rich multimodal images; some examples are Visible (VIS), Infrared (IR), Positron Emission Tomography (PET), Magnetic Resonance Imaging (MRI), Panchromatic (PAN), and Light Detection And Ranging (LiDAR), just to mention a few. These multimodal images may be fused in order to get all the information they provide in a single representation. Thus, multimodal image fusion refers to extracting relevant information from two or more images from different modalities and effectively combining them by a fusion strategy in order to reconstruct an enhanced image with richer and complementary information. Fused images are widely used in areas such as photography, medical science [1], remote sensing [2], semantic segmentation [3], [4], and computer vision.

Due to the traditional image fusion focus on mathematical transformation, which requires manual analysis and design of the fusion rules [5], several researchers have focused on deep learning techniques for multimodal image fusion. These methods can extract features and learn a fusion strategy by using a loss function that is designed for better extraction, fusion, and reconstruction. Deep learning techniques use variations of network architectures such as Convolutional Neural Networks (CNN), Generative Adversarial Networks (GAN), Autoencoders, and Transformers. [6]–[11] explore deep-learning solutions with multiple architectures. For instance, U2fusion [6] is an unsupervised image fusion network using DenseNet, a type of convolutional neural network. Vs et al. [7] propose a model based on transformers (IFT). It uses an encoder to extract multi-scale deep features from the input images, followed by a Spatio-Transformer (ST) fusion strategy and a nested decoder network. The general framework for multimodal image fusion consists of three main processes: feature extraction, feature fusion strategy, and feature reconstruction. Deep learning techniques are applied to one or all of these sub-processes of image fusion.

Since image fusion is intended to improve a further application, deep-learning approaches using multimodal image fusion on the loop of computer vision tasks and image processing applications have been proposed. Mees et al. [12] propose a method based on convolutional neural network experts using RGB, depth, and motion data for detecting objects in dynamic environments. In [13], the authors propose a model consisting of two networks: the first generates the 3D object proposal using a bird's eye view representation of point cloud while the second one is used for multi-view feature fusion. Liang et al. [14] propose an architecture that joins point-wise and ROI-wise feature fusion using LiDAR and camera data. Chen et al. [15] propose a salient object detector that embeds multimodal image pair (i.e., RGB-depth pair) into a modal agnostic structural representation and modal-specific content

space using encoders. Several surveys summarize approaches for image fusion based on deep learning [1], [3]. Besides, there exist surveys that describe applications that use multimodal image fusion [16], [17]. However, a review that summarizes deep learning-based multimodal image fusion in applications of computer vision or image processing is lacking. Thus, we propose a brief survey of multimodal image fusion using deep learning architectures in tasks such as object detection, salient object detection, semantic segmentation, etc. Since they have diverse applications such as autonomous vehicle navigation, surveillance, land cover classification, and robotics, to name a few [13], [14], [18]. Besides, we include recent studies about the use of application information in the loop of multimodal image fusion.

Furthermore, most image fusion approaches, including multimodal approaches, aim to achieve a better visual quality of a scene and higher metrics. However, they forget to consider whether the fused images are adequate to really improve the performance of specific computer vision tasks such as object detection, object tracking, edge detection, segmentation, etc. We analyze how existing approaches include multimodal image fusion in their frameworks since previous studies [19] demonstrate that considering only visual quality and quantitative metrics does not reflect the facilitation of applications. Thus, the existing deep learning-based approaches can be improved, including the feedback on the learning process between the fusion strategy and the application's performance.

The remainder of this paper is organized as follows, section II describes the addressed problem by introducing image fusion and general frameworks for image fusion tasks such as multifocus, multiexposure, and multimodal. Section III presents multimodal image fusion and the existing approaches. Section IV describes representative works done for deep learning-based multimodal image fusion in tasks such as object detection, semantic segmentation, and salient object detection; including the application on the loop approaches. Finally, section V presents the conclusions and trends about deep learning-based multimodal image fusion in vision applications.

II. Problem Formulation: Image Fusion

In this section, we present fundamental definitions involved in the image fusion problem.

A. Image Fusion

Image fusion refers to the integration of features from multiple images of the same scene to a single representation with more comprehensive information [20]. The main image fusion tasks are:

- **Multi-focus image fusion** consists on fusing several single-modality partially focused images to get an all-in-focus image.
- **Multimodal image fusion** consists on fusing several images from multiple modalities to generate a representation with richer information.
- **Multi-exposure image fusion** refers to fusing single-modality images with multiple exposure levels to generate a high-quality full-exposure image.

B. Image Fusion Categories

Most image fusion approaches can be classified according to the fusion strategy used or the fusion stage. The fusion stage refers to the phase within the framework pipeline where the fusion is performed. For instance, feature-level fusion is done after the feature extraction module (see Fig. 1). Kaur et al. [5] classify the image fusion strategies into spatial, frequency, and deep learning techniques. The spatial techniques work with the pixels by applying rules such as max-min, maximum, minimum, simple average, etc. Frequency techniques decompose the multiscale coefficients from the images. In this domain, there are methods such as Discrete Transform Fusion, Discrete Cosine Transform, Laplacian Pyramid Fusion, and so on. Finally, deep learning techniques use variations of network architectures such as Convolutional Neural Networks (CNN), Generative Adversarial Networks (GAN), Autoencoders, and Transformers.

In addition, according to the stage at which data is fused, image fusion techniques can be classified as Pixel Level, Feature level, and Decision Level [4]. Pixel-level fusion techniques combine the source images directly before feature extraction, that is, the input images are transformed into a signal, then a fusion of the transformed coefficients is performed, and an inverse transformation is done where the fused coefficients are transformed into the fused image. Feature-level techniques obtain refined characteristics from source images before fusing them. Besides, Zhang et al. [21] refer to feature-level fusion as deep fusion since it can use cross-modal information. Finally, Decision level techniques refer to dealing with information that is already been generated to represent some determination of a task [22].

Machine learning techniques are used for data fusion in the aforementioned levels. For instance, [23] proposes multisensor data fusion using SVM for fault detection (pixel-level). For feature-level fusion, Xu et al. [6] proposes an unsupervised image fusion network using feature extraction and information measurement. Besides, [24] proposes a fusion model based on SVM and Naive Bayes to fuse LIDAR and optical remote sensing data for land cover classification. Most recently, many machine learning techniques focus on feature-level fusion since it is not straightforward to extract the appropriate features from each modality in a traditional approach due to the resulting fusion image can involve redundant information. Specifically, deep learning methods have been proposed to overcome the drawbacks of traditional methods. [6]–[11]. Fig. 1 summarizes the classification of image fusion approaches.

C. General Frameworks for Image Fusion

There exist end-to-end general frameworks for image fusion tasks such as multi-focus, multi-modal, and multi-exposure. For instance, U2fusion [6] is an unsupervised image fusion network that preserves the adaptive similarity between the

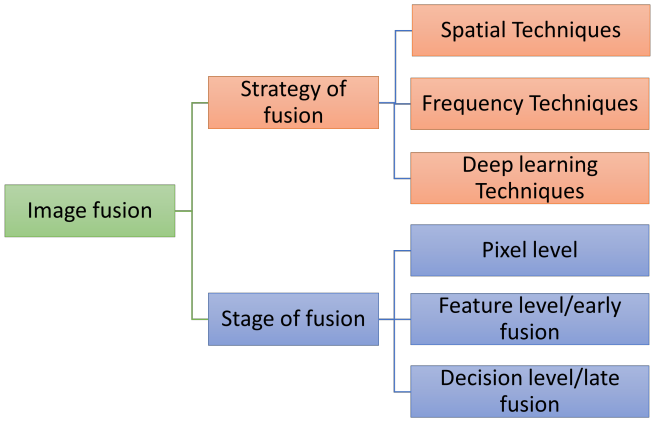

Fig. 1. Image fusion categories.

source images and the fusion result; that is, the method depends on the information preservation degree. In [20], the authors propose IFCNN, an end-to-end image fusion network that consists of three modules based on a CNN: feature extraction, feature fusion, and image reconstruction module. In this network, the feature extraction is done by two convolutional layers. Then, a fusion rule (elementwise-max, elementwise-min, elementwise-mean) is applied for fusing the features according to the type of source images. Although this network was trained for multi-focus images, authors mentioned that it could be used to fuse other multi-modal images without a finetuning procedure.

WaveFuse [25] combines multiscale discrete wavelet transform and a deep learning approach. This network architecture is based on an encoder-decoder where the encoder is used for feature extraction and consists of convolutional blocks. Then, the DWT-based module allows the transformation of the feature maps into the wavelet domain. The decoder obtains the fused image. In all the aforementioned general frameworks, the input consists of two or more images from one modality in the task of multifocus and multiexposure and from multiple modalities in the case of multimodal fusion. Then, the three main processes of image fusion (i.e., feature extraction, feature fusion, and feature reconstruction) are performed using deep learning architectures. These frameworks aim to be useful for several image tasks as a unified framework. However, the task is not considered in the fusion loop.

III. Multimodal Image Fusion

Multimodal image fusion aims to generate an image that keeps the best features from each modality. Thus, learning from multimodal images facilitates the possibility of capturing a richer representation from modalities. This representation can improve applications (e.g., security, object detection, remote sensing classification, semantic segmentation) because it takes advantage of the features of each modality. For instance, visible images can provide texture information as well as geometric details on the edges, while infrared images can provide thermal radiation information making it easy to detect the salient objects under low contrast or insufficient light. Besides, sometimes by fusing images, it is possible to obtain information that is not present if each modality is considered separately. The multimodal image fusion techniques proposed in the literature go from classical computer vision to machine learning-based techniques, including deep learning architectures. Most recently, researchers are proposing deep learning methods based on CNN [6], [8], [20], [26]–[29], GANs [30], and Transformers [7], [31], [32]. Multimodal image fusion follows the pipeline presented above for unimodal image fusion: feature extraction, in this case, from multiple modalities, feature fusion strategy to keep the best of each modality, and feature reconstruction to generate a single image.

There exist several modalities such as VIS, IR, NIR, LiDAR, and HSI, MS, PAN, to name a few. These modalities are generally fused in pairs to generate a representation with more comprehensive and complementary information. The most common fused modalities are VIS-IR; however, RGB-LiDAR, RGB-T, RGB-D, PAN-MS are some examples of other modalities pairs. Multiple approaches have been proposed based on deep learning techniques. Liu et al. [33] propose defining flexible priors and constraints depending on the features of the multi-modality images and then combining them into a bilevel optimization strategy and adaptive integration mechanics. Firstly, an edge-preserving module is applied to decompose the source images into base and detail layers. Then, the decomposed images are passed by the bilevel paradigm with adaptive integration. Similarly, [27] proposes using priors in an unsupervised image fusion method. Specifically, they use deep image priors (DIP) to exploit the ability of CNN to synthesize the prior in the source images.

In addition, RFNet [26] is a method that performs multimodal image registration and fusion. This method is based on a mutually reinforcing framework. The registration is performed in a coarse-to-fine fashion. The model consists of an image translation network (TransNet), an affine network (AffineNet), and a mutually reinforcing Fine Registration and Fusion Model (F2M). TransNet learns an image translation function to transfer multi-modal images into the same domain. AffineNet generates the affine transformation parameters. F2M performs texture-focused image fusion. Furthermore, there exist specific techniques for infrared and visible image fusion. RFN-Nest [10] is based on a residual fusion network (RFN), which contains an encoder, residual fusion network, and a decoder. This network is trained by a two-stage strategy, so the encoder and decoder are trained as an autoencoder based on Dense Fuse to extract multi-scale deep features, then the residual network is trained to fuse the features extracted at each scale for reconstructing the salient features.

Yang et al. [30] propose a model based on a texture conditional generative adversarial network (TC-GAN) for VIS-IR fusion. This network generates a combined texture map, that is, it generates a fused result with high-contrast information. The generator design is based on a codec structure with a SE-Net [34] attention module to improve the feature extraction by using the correlation between features. Moreover,

the discriminator uses convolution blocks to classify images at a pixel level. Besides, it can determine whether the texture distribution of the generated image is consistent with the visible image.

Current approaches are designed to achieve better visual quality and higher quantitative metrics of the fused image. Nevertheless, it can limit the use of the fused image to facilitate a specific computer vision application because each application may need the fusion of different features from each modality. The multimodal image fusion without a task-driven approach may not be efficient because this process would be separated from the application, and it is not guaranteed that application can achieve better performance. Despite this, several approaches have been proposed with a comparative performance. However, a better result can be achieved if the application is included in the loop of multimodal image fusion. Tang et al. [19] show the need for high-level vision tasks such as semantic segmentation on the loop of multimodal image fusion.

IV. Application of Multimodal Image Fusion in Computer Vision and Image Processing

In this section, we present state-of-the-art deep learning approaches using multimodal image fusion within the application frameworks and datasets.

The resulting image from multimodal image fusion contains more information to be used in tasks such as object detection, semantic segmentation, object tracking, edge detection, medical diagnosis, and so on [16]. Multimodal image fusion can be performed in an early or late fusion scheme. Early fusion is basically feature-level fusion; thus, the data is fused before being passed to the application module. While late fusion is decision-level fusion; thus, the decisions of the application for each modality are combined, which enables an easier decision fusion. Fig. 2 shows the pipelines of the classical fusion schemes.

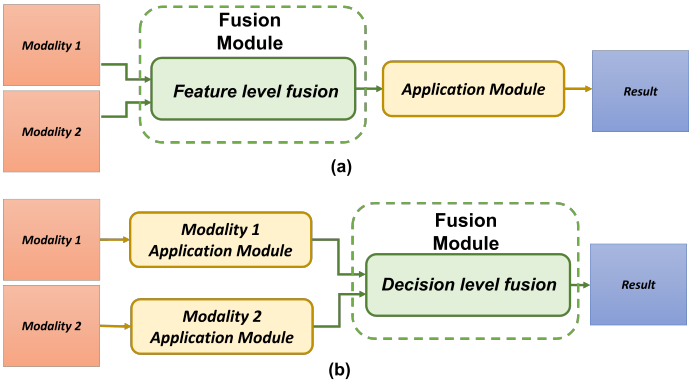

Fig. 2. Classical Fusion Schemes: (a) Early fusion; (b) Late fusion.

As shown in Fig. 2, the multimodal image fusion module is part of the application approach. However, the fusion process is performed before or after the application without considering the result of the application to improve the fusion or vice versa; that is, in the classical schemes, the modules (fusion and application) are separately and sequentially performed. On the contrary, recent studies based on deep learning techniques demonstrate that including feedback between networks for fusion and application can achieve better performance since these frameworks perform task-driven multimodal image fusion. Thus, deep learning-based vision applications on the loop of multimodal image fusion focus on including feedback in the learning process between the fusion and the application module to guarantee high-level task-driven multimodal image fusion. There are just a few approaches using the scheme of application on the loop of the fusion process, which mainly perform an early fusion. Fig. 3 shows the overall framework for vision applications on the loop of multimodal image fusion by using early fusion.

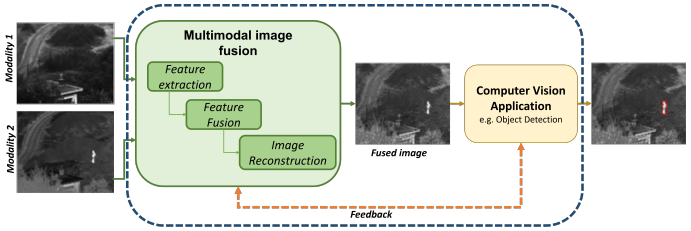

Fig. 3. Pipeline of task-driven multimodal image fusion.

In the following subsections, we describe representative approaches of multimodal image fusion within the application frameworks following classical and task-driven schemes of fusion.

A. 3D Object Detection

Since object detection is a fundamental task in computer vision and has diverse applications such as autonomous vehicle navigation, surveillance, and robotics, to name a few [13], [14], [18], it is currently a research focus. 3D object detection methods aim to identify an object as a class member. Most traditional object detection methods focus on detecting objects in RGB images. However, fusing different modalities such as RGB, depth, and LiDAR can improve object detection because each modality provides complementary scene information.

Several studies propose using multimodal image fusion within vision application frameworks to achieve end-to-end approaches based on deep learning techniques. Chen et al. [13] propose a Multi-View 3D object detection network using LiDAR point cloud and RGB images in order to predict 3D bounding boxes. This network consists of two parts: a 3D Proposal Network and a Region-based Fusion Network. In the 3D Proposal Network, the bird's eye view map is used to make the 3D location prediction. The network generates 3D box proposals and the fusion network combines multi-view features hierarchically. In addition, Sindagi et al. [35] propose an early fusion approach with PointFusion and a later fusion strategy with VoxelFusion; besides, this method extends VoxelNet architecture for multimodal inputs. In the PointFusion, LiDAR points are projected onto the plane then

image feature extraction is performed, while in VoxelFusion features are extracted from each voxel in the image using a pre-trained VGG-16.

Additionally, Gao et al. [9] propose a method for vehicle detection at nighttime that fuses infrared and visible images. The network is based on a GAN architecture. Thus, the generator fuses the features extracted from input images. The method consists of a visible branch, an infrared branch, and a self-attention fusion model. The fusion result is sent to the detection model. The detection model is based on RetinaNet. FusionPainting proposed by [36] consists of three modules: a multi-modal semantic segmentation module, an adaptive attention-based semantic fusion module, and a 3D object detector. Center fusion is proposed by [37]. This method uses a center point object detection network that generates a heat map for every object category. The extracted features for each modality are mapped to the center of the corresponding object on the image. A frustum association method is proposed that uses a 2D bounding box, estimated depth, and size in order to create a Region of Interest (ROI) frustum for the object in the 3D Space. Then, it maps the radar detection to the center of objects in the image.

In [38], a fusion framework with semantic understanding is proposed to assist the fusion process and enhance the depth estimation and object detection since it can improve the visibility of far and small objects in a scene. The method consists of cross-modal depth estimation and object detection framework. Besides, Lee et al. [38] propose Semantic-Guided Radar-Vision Fusion for Depth Estimation and Object Detection. This method proposes to integrate monocular RGB images, semantic information, and sparse radar point clouds. The method proposed by [31] is a Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. This method uses the self-attention mechanism of transformers for the fusion of intermediate feature maps between modalities. The network receives as input a single-view RGB image and LiDAR representation.

In [39], authors propose a network using infrared, visible, and polarization images for detecting defects in circuit boards. The backbone of the network is based on Convolutional Block Attention. It has three branches with downsampling for feature extraction, one branch for each modality. Besides, a modality-select attention module is implemented for image fusion. Then, the fused features are passed into a channel attention-path aggregation, which consists of a two-path pyramid structure with the subsequent attention structure to enhance features.Finally, YOLO architecture is used for defect classification

Recently, task-driven approaches have been proposed to include the application in the loop of multimodal image fusion since it is more beneficial to generate fused images that can enhance the performance of high-level vision tasks in real scenarios. DetFusion [40] is a recent detection-driven multimodal image fusion method that uses VIS-IR images. In this framework, a detection loss is used to guide and optimize the image fusion network; that is, it uses the object detection networks to guide the fusion. This framework comprises a shared attention-guided fusion network, a visible detection network, and an infrared detection network. An object-aware content loss is presented to guide the fusion network to learn information about contrast and texture from VIS and IR images. Furthermore, a shared attention module is proposed to learn object-specific information from the detection network. The architecture of the detection network uses Faster R-CNN. On the other hand, the feature extraction network uses ResNet-FPN.

Remarks: Most existing object detection approaches with multimodal image fusion use the classical fusion schemes, that is, early or late fusion, even hybrid fusion scheme that includes layers for early fusion followed by late fusion or a combination of them. These approaches perform the fusion before or after the application module. Thus, only pixel or feature information is used during fusion. However, recent studies demonstrate that object-related information used in multimodal image fusion is more helpful for object detection. Hence, leveraging application information in the fusion process allows for improving the performance of detection.

The trend of object detection approaches is to use detection-driven multimodal image fusion; that is, to create a connection between the learning process of the fusion module and the object detection module. On the other hand, the existing task-driven approaches for object detection only fuse VIS-IR modalities; therefore, using other modalities in detection-driven multimodal image fusion is still a research topic to be explored.

B. Semantic Segmentation

Semantic segmentation aims to semantically label each pixel of an image in order to have a better understanding of a scene [3]. The evolution of deep learning and image fusion promotes the improvement of semantic segmentation. Thus, deep learning-based approaches using multimodal image fusion for semantic segmentation have been proposed. Liu et al. [28] propose the use of contrastive losses in order to prevent only focusing on strong modalities instead, that is, the method ensures that the modality synergy and weak modalities (negative samples) are not ignored during fusion. This method is evaluated for semantic segmentation achieving comparable performance.

The network proposed in [41] fuses intramodality and intermodality features from LiDAR and RGB images to learn cross-modal interdependencies and contextual information. The encoder network consists of two branches one for RBG and another for LiDAR, besides, it receives intramodal features such as intensity, slope angle infrared-green-red bands data, normalized difference vegetation index, and so on. Moreover, the decoder network contains upsampling residual units to get feature maps. Additionally, the model contains multimodal attention fusion blocks, which focus on intramodal features fusion and intermodal features fusion. On the other hand, since remote sensing images have multiple scales, authors use cascade dilation convolution in a pyramid structure. In addition, Pfeuffer et al. [42] propose a network for semantic

segmentation in adverse weather conditions. This network is based on ICNet and uses Cascade Feature Fusion layers.

The approaches mentioned above follow classical fusion schemes; thus, they do not have a connection between the fusion and the semantic segmentation task. On the contrary, following the task-driven fusion scheme, Tang et al. [19] propose a framework where the image fusion network and the semantic segmentation network are concatenated and use the semantic loss to guide the fusion, which allows learning high-level semantic information. This study is the first approach to connect the application to multimodal image fusion through loss functions. The network is based on gradient residual dense block to extract fine-grained features.

Remarks: The more recent approaches for semantic segmentation with multimodal image fusion pay attention to putting the semantic loss in the loop of the fusion in order to enhance the segmentation result. Hence, semantic segmentation-driven multimodal image fusion is also an open research topic like detection-driven approaches. Furthermore, the existing framework is limited for VIS-IR image fusion; which motivates future works to attempt frameworks for other modalities such as LiDAR, Depth, and so on.

C. Salient Object Detection

Salient Object Detection (SOD) aims to distinguish the most visually attractive objects in an image. Although traditional methods have succeeded in the last years, they can fail when the background resembles the object. Thus, multimodal image fusion for SOD is a research focus because it comprehensively represents a scene when fusing multiple modalities. For instance, RGB images provide appearance features such as texture, while LiDAR point cloud provides information for depth estimation.

Duan et al. [43] propose a triple-diversity fusion network (TDSM), a diversity fusion model (DFM), and a dense decoder. For extracting features, the encoder is based on SwinTransformer since it is useful for locating salient objects due to it can define long-range intra-class dependencies of features. TDSM consists of three branches: RGB, Depth, and cross-modal branches. RGB and Depth branches generate saliency maps to reflect the most important regions of each modality; then, the features are integrated by a bi-directional interactive strategy. The cross-modal generates an edge map. These maps are integrated by DFM, identifying which regions need attention. Finally, DSD predicts saliency results.

Huang et al. [44] propose integrating features in a middle-level one time instead of multiple times using a fusion module that explores the mutual relation between modalities and utilizes the total, shared, and differential information of each modality during fusion. Moreover, Chen et al. [15] propose a cross-modal disentanglement framework that allows adaptive feature fusion and determines the complementary RGB and Depth features for saliency detection. This method uses encoders for extracting features, a fusion block, and a prediction layer.

Remarks: Although several approaches for salient object detection have been proposed using multimodal image fusion, approaches that follow a task-driven fusion scheme still need to be developed. Therefore, there is a significant opportunity to study the new fusion scheme (task-driven) for salient object detection.

D. Other Applications

Prakash et al. [31] propose Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. This method uses the self-attention mechanism of transformers for the fusion of intermediate feature maps between modalities. The network receives as input a single-view RGB image and LiDAR representation Li et al. [39] present a network using infrared, visible, and polarization images for detecting defects in circuit boards. The backbone of the network is based on Convolutional Block Attention. It has three branches with downsampling for feature extraction, one branch for each modality. This process is improved by Mobilenetv3. Besides, a modality-select attention module is implemented for image fusion. Then, the fused features are passed into a channel attention-path aggregation, which consists of a two-path pyramid structure with the subsequent attention structure to enhance features. FusAtNet proposed by [8] performs pixel-based classification for land-cover classification by using a cross-attention framework that uses attention from one modality to highlight features in another modality. It consists of a spectral and spatial attention module in order to take advantage of the spectral-spatial information of the Hyperspectral image (HSI) and spatial-elevation information of LiDAR.

Samal et al. [45] present a task-driven image fusion for object tracking and path planning. This approach uses RGB-LiDAR images by controlling the use of LiDAR images to detect the spatial and temporal regions from LiDAR where RGB fails. The network follows a decision-level fusion and consists of RGB object detector and LiDAR object detector. RGB object detector is based on Faster R-CNN and ResNet, while the LiDAR object detector is based on PointRCNN.

Remarks: Besides the applications mentioned above (object detection, semantic segmentation, and salient object detection), other computer vision tasks and image processing applications are focused on following the ”application on the loop of multimodal image fusion”, such as object tracking. Therefore, connecting the application information to the fusion process is a trend to be used in any vision application due to the improvements the task-driven fusion scheme offers.

E. Datasets

There are different datasets with multiple modalities containing registered images from the same scene. Despite of several datasets have initially been used only for multimodal fusion approaches. Nowadays, datasets are oriented toward applications, demonstrating that multimodal image fusion is required to enhance vision applications. Table I summarizes the datasets.

TABLE I
DATASETS FOR MULTIMODAL IMAGE FUSION

Dataset	Modalities	Year	Images	Application
People Unihall [46]	RGB-Depth	2011	3000 frames	3D Object Detection
KITTI [18]	RGB-LiDAR	2012	400 images - 11 classes	2D/3D Object Detection
NYUDv2 [47]	VIS-Depth	2012	1449 images	Semantic Segmentation
MUUFL Gulfport [48]	HSI, LiDAR	2013	53687 pixels - 11 classes	Land Cover Classification
NLPR [46]	RGB-Depth	2014	1000 images	Object Detection
MS COCO [49]	VIS-IR	2014	330K images - 80 classes	VIS-IR fusion
TNO [50]	VIS-NIR-LWR	2014	60 images	VIS-IR fusion
Vaihingen [51]	VHR - LiDAR	2014	33 images - 6 classes	Semantic Segmentation
SUN RGB-D [52]	RGB-Depth	2015	10K images	3D Object Detection
Cityscapes [53]	RGB-Depth	2015	5k images - 30 classes	3D Object Detection
KAIST [54]	VIS-IR	2015	1.6K images	VIS-IR fusion
SUN RGB-D [52]	RGB-Depth	2015	10K images	3D Object Detection
nuScenes [55]	RGB-LiDAR	2019	1000 scenes - 23 classes	2D Object Detection
RoadScene [56]	VIS-IR	2020	221 images	VIS-IR fusion
Aligned FLIR [57]	VIS-IR	2020	5142 images - 3 classes	2D Object Detection
LLVIP [58]	VIS-IR	2021	15488 images - 1 class	2D Object Detection

V. CONCLUSION

Multimodal image fusion is valuable in computer vision tasks and image processing applications since it generates a rich informative image from multiple modalities. Recently, deep learning approaches have been proposed for enhancing the fusion task. In this study, we review state-of-the-art multimodal image fusion approaches based on deep learning techniques for vision applications. These approaches follow a classical and recently task-driven fusion scheme. In the classical approaches, the fusion network is before or after the application, but it is an independent module within the application framework. On the other hand, task-driven approaches include the application in the loop of the multimodal image fusion module by using connections through loss functions to reinforce the application information in the fusion, which allows the improvement of the final result of the application. This demonstrates the potential of connecting low-level vision tasks with high-level vision tasks. Therefore, the trend of multimodal image fusion approaches based on deep learning techniques is to provide a connection between image fusion and vision tasks in order to provide information contained in the application to the multimodal image fusion network.

ACKNOWLEDGEMENTS

This work is partially supported by the Grant PID2021-128945NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe"; the "CERCA Programme / Generalitat de Catalunya"; and the ESPOL project CIDIS-12-2022.

REFERENCES

[1] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *Information Fusion*, vol. 33, pp. 100–112, 2017.

[2] D. Hong, J. Chanussot, and X. X. Zhu, "An overview of multimodal remote sensing data fusion: From image to feature, from shallow to deep," *International Geoscience and Remote Sensing Symposium*, pp. 1245–1248, 2021.

[3] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, "Deep multimodal fusion for semantic image segmentation: A survey," *Image and Vision Computing*, vol. 105, 2021.

[4] H. Ghassemian, "A review of remote sensing image fusion methods," *Information Fusion*, vol. 32, pp. 75–89, 2016.

[5] H. Kaur, D. Koundal, and V. Kadyan, "Image Fusion Techniques: A Survey," *Archives of Computational Methods in Engineering*, no. 0123456789, 2021.

[6] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A Unified Unsupervised Image Fusion Network," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 8828, no. c, pp. 1–1, 2020.

[7] V. Vs, J. Maria, J. Valanarasu, P. Oza, and V. M. Patel, "Image fusion transformer," *IEEE Int. Conf. on Image Processing*, pp. 3566–3570, 2022.

[8] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, "FusAtNet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and LiDAR classification," *Computer Vision and Pattern Recognition Workshop*, vol. 2020-June, pp. 416–425, 2020.

[9] P. Gao, T. Tian, T. Zhao, L. Li, N. Zhang, and J. Tian, "GF-Detection : Fusion with GAN of Infrared and Visible Images for Vehicle Detection in Nighttime," *Remote Sensing*, pp. 1–17, 2022.

[10] H. Li, X. J. Wu, and J. Kittler, "RFN-Nest: An end-to-end residual fusion network for infrared and visible images," *Information Fusion*, vol. 73, pp. 72–86, 2021.

[11] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Information Fusion*, vol. 45, pp. 153–178, 1 2019.

[12] O. Mees, A. Eitel, and W. Burgard, "Choosing smartly: Adaptive multimodal fusion for object detection in changing environments," *Int. Conf. on Intelligent Robots and Systems*, vol. 2016-November, pp. 151–156, 11 2016.

[13] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," *Computer Vision and Pattern Recognition*, vol. 2017-Janua, pp. 6526–6534, 2017.

[14] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-Task Multi-Sensor Fusion for 3D Object Detection," *Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 7337–7345, 12 2020.

[15] H. Chen, Y. Deng, Y. Li, T. Y. Hung, and G. Lin, "RGBD Salient Object Detection via Disentangled Cross-Modal Fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 8407–8416, 2020.

[16] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Information Fusion*, vol. 76, pp. 323–336, 12 2021.

[17] M. Gao, J. Jiang, G. Zou, V. John, and Z. Liu, "RGB-D-Based Object Recognition Using Multimodal Convolutional Neural Networks: A Survey," *IEEE Access*, vol. 7, pp. 43 110–43 136, 2019.

[18] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," *Computer Vision and Pattern Recognition*, pp. 3354–3361, 2012.

[19] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Information Fusion*, vol. 82, pp. 28–42, 6 2022.

[20] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Information Fusion*, vol. 54, pp. 99–118, 2 2020.

[21] C. Zhang, H. Wang, Y. Cai, L. Chen, Y. Li, M. A. Sotelo, Z. Li, and Z. Li, "Robust-FusionNet: Deep Multimodal Sensor Fusion for 3-D Object Detection Under Severe Weather Conditions," *Transactions on Instrumentation and Measurement*, vol. 71, 2022.

[22] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Information Fusion*, vol. 57, no. 2, pp. 115–129, 2020.

[23] T. P. Banerjee and S. Das, "Multi-sensor data fusion using support vector machine for motor fault detection," *Information Sciences*, 2012.

[24] B. Bigdeli, F. Samadzadegan, and P. Reinartz, "A decision fusion method based on multiple support vector machine system for fusion of hyperspectral and LIDAR data," *International Journal of Image and Data Fusion*, vol. 5, no. 3, pp. 196–209, 2014.

[25] S. Liu, M. Wang, and Z. Song, "WaveFuse: A Unified Deep Framework for Image Fusion with Discrete Wavelet Transform," *Int. Conf. on Neural Information Processing*, pp. 1–12, 2021.

[26] H. Xu, J. Ma, J. Yuan, Z. Le, and W. Liu, "RFNet: Unsupervised Network for Mutually Reinforcing Multi-modal Image Registration and Fusion," *Computer Vision and Pattern Recognition*, pp. 19 679–19 688, 2022.

[27] X. Ma, P. Hill, N. Anantrasirichai, and A. Achim, "Unsupervised Image Fusion Using Deep Image Priors," *Int. Conf. on Image Processing*, pp. 2301–2305, 11 2022.

[28] Y. Liu, Q. Fan, S. Zhang, H. Dong, T. Funkhouser, and L. Yi, "Contrastive Multimodal Fusion with TupleInfoNCE," *Int. Conf. on Computer Vision*, pp. 754–763, 2021.

[29] H. Li, X.-J. Wu, and J. Kittler, "Infrared and Visible Image Fusion using a Deep Learning Framework," *Int. Conf. on Pattern Recognition*, 2018.

[30] Y. Yang, J. Liu, S. Huang, W. Wan, W. Wen, and J. Guan, "Infrared and Visible Image Fusion via Texture Conditional Generative Adversarial Network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4771–4783, 12 2021.

[31] A. Prakash, K. Chitta, and A. Geiger, "Multi-Modal Fusion Transformer for End-to-End Autonomous Driving," *Computer Vision and Pattern Recognition*, pp. 7077–7087, 2021.

[32] S. Park, A. G. Vien, and C. Lee, "Infrared and Visible Image Fusion Using Bimodal Transformers," *IEEE Int. Conf. on Image Processing*, pp. 1741–1745, 11 2022.

[33] R. Liu, J. Liu, Z. Jiang, X. Fan, and Z. Luo, "A Bilevel Integrated Model with Data-Driven Layer Ensemble for Multi-Modality Image Fusion," *Transactions on Image Processing*, vol. 30, pp. 1261–1274, 2021.

[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," *Computer Vision and Pattern Recognition*, 2018.

[35] V. A. Sindagi, Y. Zhou, and O. Tuzel, "MVX-net: Multimodal VoxelNet for 3D object detection," *Int. Conf. on Robotics and Automation*, vol. 2019-May, pp. 7276–7282, 5 2019.

[36] S. Xu, D. Zhou, J. Fang, J. Yin, Z. Bin, and L. Zhang, "FusionPainting: Multimodal Fusion with Adaptive Attention for 3D Object Detection," *IEEE Conference on Intelligent Transportation System*, vol. 2021-September, pp. 3047–3054, 9 2021.

[37] R. Nabati and H. Qi, "CenterFusion: Center-based Radar and Camera Fusion for 3D Object Detection," *Winter Conference on Applications of Computer Vision*, pp. 1526–1535, 11 2021.

[38] W.-y. Lee, L. Jovanov, and W. Philips, "Semantic-Guided Radar-Vision Fusion for Depth Estimation and Object Detection," *British Machine Vision Conference*, 2021.

[39] M. Li, N. Yao, S. Liu, S. Li, Y. Zhao, and S. G. Kong, "Multisensor Image Fusion for Automated Detection of Defects in Printed Circuit Boards," *IEEE Sensors Journal*, vol. 21, no. 20, pp. 23 390–23 399, 10 2021.

[40] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "DetFusion: A Detection-driven Infrared and Visible Image Fusion Network," *ACM International Conference on Multimedia*, pp. 4003–4011, 10 2022.

[41] Y. Sun, Z. Fu, C. Sun, Y. Hu, and S. Zhang, "Deep Multimodal Fusion Network for Semantic Segmentation Using Remote Sensing Image and LiDAR Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022.

[42] A. Pfeuffer and K. Dietmayer, "Robust Semantic Segmentation in Adverse Weather Conditions by means of Sensor Data Fusion," *Int. Conf. on Information Fusion*, 7 2019.

[43] S. Duan, C. Xia, X. Gao, B. Ge, H. Zhang, and K. C. Li, "Multi-modality diversity fusion network with swin transformer for RGB-D salient object detection," *Int. Conf. on Image Processing*, pp. 1076–1080, 2022.

[44] N. Huang, Q. Jiao, Q. Zhang, and J. Han, "Middle-Level Feature Fusion for Lightweight RGB-D Salient Object Detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 6621–6634, 2022.

[45] K. Samal, H. Kumawat, P. Saha, M. Wolf, and S. Mukhopadhyay, "Task-Driven RGB-Lidar Fusion for Object Tracking in Resource-Efficient Autonomous System," *Transactions on Intelligent Vehicles*, vol. 7, no. 1, pp. 102–112, 3 2022.

[46] L. Spinello and K. O. Arras, "People detection in RGB-D data," *Int. Conf. on Intelligent Robots and Systems*, pp. 3838–3843, 12 2011.

[47] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images." *European Conference on Computer Vision*, vol. 7576, pp. 746–760, 2012.

[48] P. Gader, A. Zare, R. Close, J. Aitken, and G. Tuell, "Muufl gulfport hyperspectral and lidar airborne data set," 2013.

[49] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014, pp. 740–755.

[50] A. Toet, "TNO Image Fusion Dataset," 1 2014.

[51] F. Nex, M. Gerke, F. Remondino, H.-J. Przybilla, M. Bäumker, and A. Zurhorst, "Isprs benchmark for multi-platform photogrammetry," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2015.

[52] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," *Computer Vision and Pattern Recognition*, pp. 567–576, 2015.

[53] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," *Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 3213–3223, 4 2016.

[54] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral Pedestrian Detection: Benchmark Dataset and Baseline," 2015. [Online]. Available: http://rcv.kaist.ac.kr/multispectral-pedestrian/

[55] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "Nuscenes: A multimodal dataset for autonomous driving," *Computer Vision and Pattern Recognition*, pp. 11 618–11 628, 2020.

[56] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "FusionDN: A Unified Densely Connected Network for Image Fusion," *AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12 484–12 491, 4 2020.

[57] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks," *Int. Conf. on Image Processing*, vol. 2020-October, pp. 276–280, 9 2020.

[58] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "LLVIP: A Visible-infrared Paired Dataset for Low-light Vision," *IEEE Int. Conf. on Computer Vision*, vol. 2021-October, pp. 3489–3497, 8 2021.