Progressively Complementarity-aware Fusion Network for RGB-D Salient Object Detection

Hao Chen Youfu Li*
City University of Hong Kong, Kowloon, Hong Kong.
{hchen47-c@my., meyfli@}cityu.edu.hk

Abstract

How to incorporate cross-modal complementarity sufficiently is the cornerstone question for RGB-D salient object detection. Previous works mainly address this issue by simply concatenating multi-modal features or combining unimodal predictions. In this paper, we answer this question from two perspectives: (1) We argue that if the complementary part can be modelled more explicitly, the cross-modal complement is likely to be better captured. To this end, we design a novel complementarity-aware fusion (CA-Fuse) module when adopting the Convolutional Neural Network (CNN). By introducing cross-modal residual functions and complementarity-aware supervisions in each CA-Fuse module, the problem of learning complementary information from the paired modality is explicitly posed as asymptotically approximating the residual function. (2) Exploring the complement across all the levels. By cascading the CA-Fuse module and adding level-wise supervision from deep to shallow densely, the cross-level complement can be selected and combined progressively. The proposed RGB-D fusion network disambiguates both cross-modal and cross-level fusion processes and enables more sufficient fusion results. The experiments on public datasets show the effectiveness of the proposed CA-Fuse module and the RGB-D salient object detection network.

1. Introduction

The aim of salient object detection is to identify the object/objects attracting human beings most in a scene [2, 3]. Salient object detection is useful for a large range of computer vision and robotic vision tasks such as object recognition [4], image retrieval [5] and SLAM [6]. Traditional saliency detection models [3, 7-10] are performed merely on RGB images and can be categorized as bottom-up and top-down pipelines. Based on these two frameworks, various hand-crafted saliency features have been proposed. Recently, to overcome the lack of high-level contexts and the difficulty in exploring saliency-specific prior knowledge, a large body of deep convolutional neural networks (CNNs) [11-17] have been designed for RGB-induced salient object detection and have achieved appealing performance. However, when the salient object and background share similar appearance, these RGB-induced saliency detection models may be powerless to discriminate the salient object from background. In this case, the paired depth data, which contain affluent spatial structure and 3D layout information, can contribute a lot of additional saliency cues. Also, the robustness of depth sensors (e.g., Microsoft Kinect or Intel RealSense) to lighting changes will benefit a lot in extending the application scenarios of saliency detection. Accordingly, it is practically promising to involve the paired depth data in saliency detection. For the RGB-D saliency detection task, how to fuse the RGB and depth information sufficiently is the key issue. Most of previous models address this problem by directly concatenating RGB and depth features, or combining unimodal predictions. Recently, regarding that CNNs are more powerful in learning discriminative representations, a number of CNNs have been proposed for various RGB-D computer vision tasks, such as saliency detection [18, 19], semantic segmentation [20-23] and object recognition [1, 24-26]. Although encouraging performance has been achieved by these networks, there is

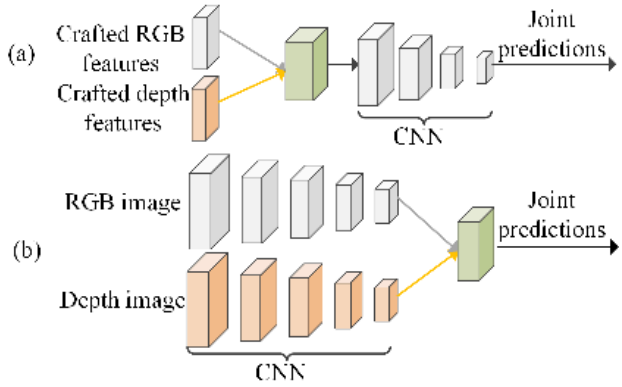

Figure 1: Traditional architectures of the CNN-based RGB-D salient object detection networks. (a) 'Early fusion' scheme adopted in [13] and (b) 'late fusion' scheme adopted in [14].

* Corresponding author

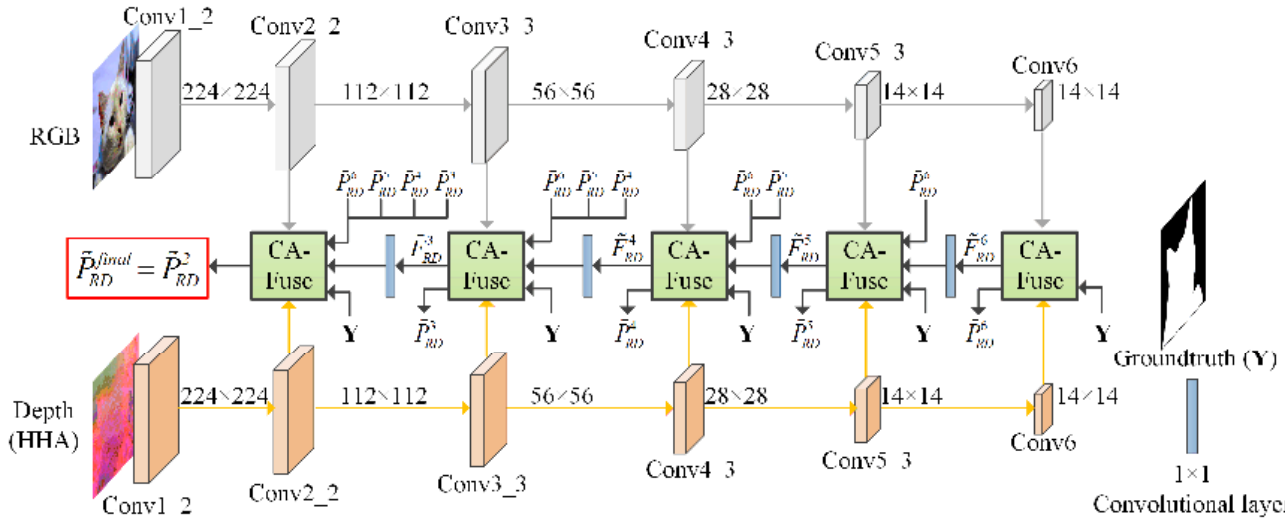

Figure 2: The architecture of the proposed progressively complementarity-aware fusion network for RGB-D salient object detection. Pooling layers are omitted for simplification. The 1×1 convolutional layers between neighboring CA-Fuse blocks are used for feature combination and dimensionality reduction (detailed parameters are shown in Table 1). Follow the practice in [1], we encode the depth image into 3-channel HHA representations.

still large room for further improvement in several key aspects: 1) How to formulate the complementary information between two modalities clearly and fuse it in a sufficient way. Most of previous RGB-D fusion networks explore the cross-modal complementarity by a two-stream architecture shown in Fig. 1 (a) and Fig. 1 (b), in which RGB and depth data are learnt separately by each stream, and then shared layers are appended at an early or late point to learn joint representations and cooperated decisions. However, the complementary information from the paired modality has not been explicitly formulated. As a result, the cross-modal complement is ambiguous and unlikely to be well-captured. 2) How to effectively exploit the useful cross-modal complement in multiple layers. Most of RGB-D fusion networks [19, 25, 27] combine RGB and depth modalities by only fusing their deep CNN features (i.e., late fusion), while we believe that the cross-modal complement for saliency detection exists across multiple levels, which are not well-explored by previous works. 3) It has been widely acknowledged that the features of different levels, which abstracts scenes in different scales, are also complementary. To be more specific, the deeper features typically carry more global contextual information and are more likely to locate the salient object correctly, while the shallower features supply more spatial details. Consequently, the issue of how to combine cross-level features should also be involved.

In our view, addressing these problems will enable the multi-modal fusion network to capture cross-modal and cross-level complement more sufficiently. To this end, in this work, we propose a progressively complementarity-aware fusion network (shown in Fig. 2). In this network, the complementarity-aware fusion ('CA-Fuse') module (see Fig. 3 (c)) is appended on the side of each CNN level and cascaded from deeper to shallower successively. In this way, multi-modal features from each level are selected and combined. Meanwhile, cross-level features and predictions are also chosen and fused progressively to make joint decisions. As shown in Fig. 3 (c), in the CA-Fuse module, the complementary features can be selected and incorporated with the paired modality adaptively via cross-modal residual connections and complementarity-aware supervisions. Introducing such a module recasts the problem of learning complementary information from the paired modality into asymptotically approximating the cross-modal residual function (see section 3.2 for details). Compared to directly concatenating multi-modal features, the proposed CA-Fuse module formulates the cross-modal complementarity explicitly, thus allowing more efficient multi-modal fusion. Besides, it also disambiguates the level-specific complementarity by cascading the CA-Fuse module successively with level-wise intermediate supervisions. Hence, the multi-modal fusion process will be complementarity-aware in terms of both cross-modal and cross-level views, resulting in sufficient multi-modal multi-level fusion. To our best knowledge, previous works mainly adopt residual connections to reuse features of preceding layers flexibly inside unimodal network streams [21, 28-30], while in this work, the residual connection is introduced in a cross-modal way in multiple levels for fusing RGB-D image pairs.

In summary, the proposed RGB-D salient object detection network enjoys several distinguished benefits:

1) The cross-modal complementarity can be explicitly

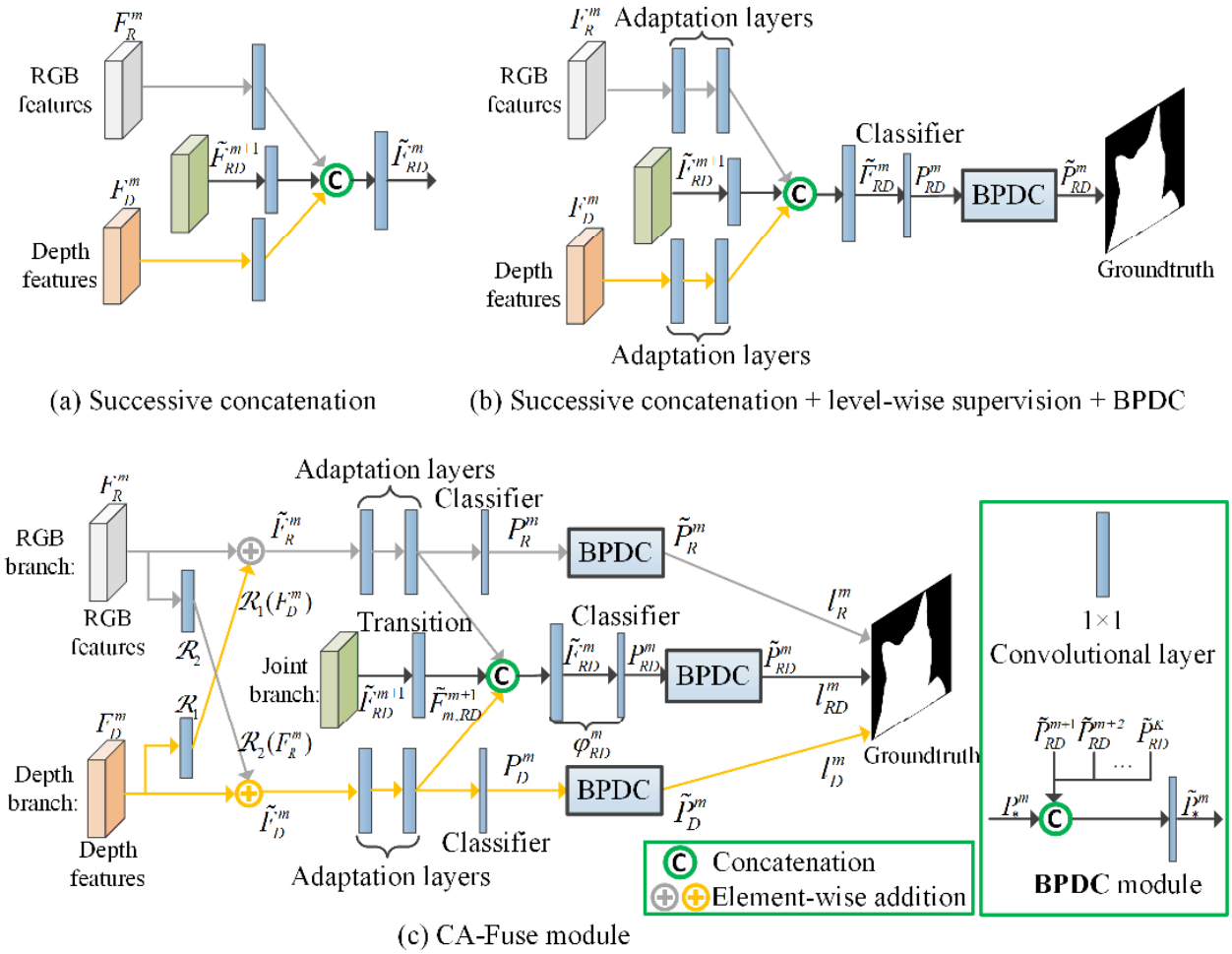

Figure 3: The architectures of different multi-modal fusion modules. See section 3.2 for more details.

encouraged, thus being explored more sufficiently and efficiently;

2) The cross-level complementarity will be exploited progressively from deep to shallow and the predictions will be enhanced in a coarse-to-fine manner;

3) The network does not rely on any pre-training process for each modality or post-processing stage. It is able to capture and fuse cross-modal and cross-level complementary information sufficiently to locate the salient object and meanwhile highlight its details in an end-to-end manner (only 0.06s testing time for each RGB-D pair).

2. Related work

Previous RGB-D salient object detection models [31-41] fuse RGB and depth information by three main modes: serializing RGB and depth as undifferentiated 4-channel input ('input fusion'), combining handcrafted RGB and depth saliency features ('feature fusion'), or performing unimodal predictions separately and then make joint decisions ('result fusion').

For example, Peng et al. [39] serialize a RGB-D pair as 4-channel input and feed it to a multiple-stage saliency inference model. Song et al. [38] use the 4-channel data to compute multi-scale saliency values. For the design of saliency features in the depth modality, Ju et al. [34] measure depth-induced saliency by evaluating the anisotropic center-surround difference. Feng et al. [37] leverage the angular density and size in depth distributions to quantify saliency. Result fusion methods include summation [35, 42], multiplication [31] and designed rules [33]. However, due to the lack of cross-modal interactions in the feature-extraction stage, the result fusion scheme is insufficient to leverage underlying cooperative information during the unimodal prediction course.

Recently, CNNs are adopted in RGB-D saliency detection to learn more discriminative RGB-D features. Qu et al. [18] combine the hand-designed low-level saliency features from RGB and depth modalities as the joint input and train a CNN from scratch to generate RGB-D hyper-features. However, owing to the loss of information

in the feature-crafting process and the limited training data, we argue that it may be hard to make full use of CNNs via learning a CNN from scratch to fuse the handcrafted RGB-D features. In contrast, Han et al. [19] use a two-stream late fusion architecture to fuse RGB-D deep features. The network is trained in a stage-wise manner and achieves encouraging performance. Nonetheless, in its multi-modal fusion stage, it still follows the paradigm of direct feature concatenation without any explicit formulation on the cross-modal complementarity. Besides, this method only focuses on fusing the high-level representations, while the complementary information in the shallower layers are ignored.

As for other RGB-D tasks, existing networks are designed with modeling the RGB-D correlation in the decision-making stage [22, 25, 27] or combining RGB-D features in a certain point [21, 23, 26] (i.e., early or late). However, none of these networks formulate the cross-modal complementarity explicitly in each level as done in this work.

3. Progressively Complementarity-aware Fusion Network

3.1. The overall architecture

The proposed CA-Fuse module can be incorporated in any basic network, e,g., the VGG-Net [43] and Res-Net [28]. Here we adopt the VGG-16 net as the trunk architecture for both RGB and depth streams for comparing fairly with previous works. The original VGG-16 net includes 5 convolutional blocks. To achieve global contextual reasoning, we append a 13×13 convolutional layer after the Conv5_3 layer as the 6-th convolutional block. Considering that the Conv1 block maybe too shallow to make reliable predictions, we will not incorporate the predictions of the Conv1 block.

3.2. The CA-Fuse module

To fuse multi-level features, the most straightforward method is to concatenate the features from different levels hierarchically (e.g., Fig. 3 (a)). However, it is cumbersome and ambiguous to fuse multi-level features by only minimizing the final prediction loss without any additional guidance. As a result, the characteristics of different levels may be unable to be well explored. Draw inspiration from unimodal networks [44] and [45], in which deep supervisions are introduced to facilitate convergence and generate hierarchical representations, we consider that an effective solution is to introduce intermediate supervisions on top of each multi-modal fusion level (Fig. 3 (b)). The added intermediate supervision can act as instruction to encourage multi-modal fusion in each level timely, thus reducing the multi-level fusion uncertainty.

Module	Adaptation layers		Transition layer
	1	2	
CA-Fuse 6	-	-	384, 1×1
CA-Fuse 5	384, 1×1	384, 1×1	384, 1×1
CA-Fuse 4	384, 3×3	384, 3×3	256, 1×1
CA-Fuse 3	192, 3×3	192, 3×3	128, 1×1
CA-Fuse 2	128, 3×3	128×3×3	-

Table 1: Illustration of the parameters of the intra-level adaptation layers inside the CA-Fuse module and the transition layer between two neighboring CA-Fuse modules.

Although this strategy is able to ease the multi-level multi-modal fusion process effectively, the multi-modal fusion component in each level still does not go beyond the traditional direct concatenation scheme, which in our view, is unlikely to sufficiently capture the cross-modal complementary information. To address this problem, we further tailor a complementarity-aware fusion (CA-Fuse) module (Fig. 3 (c)). In the CA-Fuse module, cross-modal residual connections (i.e., $\mathcal{R}_1(\bullet)$ and $\mathcal{R}_2(\bullet)$) along with complementarity-aware supervisions (i.e., l_R^m and l_D^m) are introduced to encourage the determination of complementary information from the paired modality. More specifically for the *m*-th level, the deep features from the depth branch (i.e., F_D^m) are firstly 'selected' by a 1×1 mapping layer and then the desired complementary features $\mathcal{R}_1(F_D^m)$ are added to the paired RGB branch via the cross-modal residual connection $\tilde{F}_R^m = F_R^m + \mathcal{R}_1(F_D^m)$. The enhanced RGB features $\tilde{F}_R^m$ are followed by two 1×1 convolutional layers to adapt the intermediate supervision and reduce the disturbance to the trunk stream during training. The detailed parameters of adaptation layers are shown in Table 1. Then a classifier is added to make predictions for the RGB branch $P_R^m = \varphi_R^m(\tilde{F}_R^m)$, where φ_R^m denotes the parameters of the adaption layers and the classifier. In this way, the objective that using $\mathcal{R}_1(\bullet)$ to extract complementary characteristic from the depth stream can be equivalently posed as approximating the residual function, i.e., $\tilde{F}_R^m - F_R^m$. This reformulation disambiguates the multi-modal combination, which means the solver may simply drives the residual mapping towards zero when F_R^m is sufficient to predict otherwise push $\mathcal{R}_1(\bullet)$ to abstract complementary information from F_D^m to help F_R^m for better predictions. Compared to concatenating F_R^m and F_D^m directly, such a preconditioning should ease the solver to

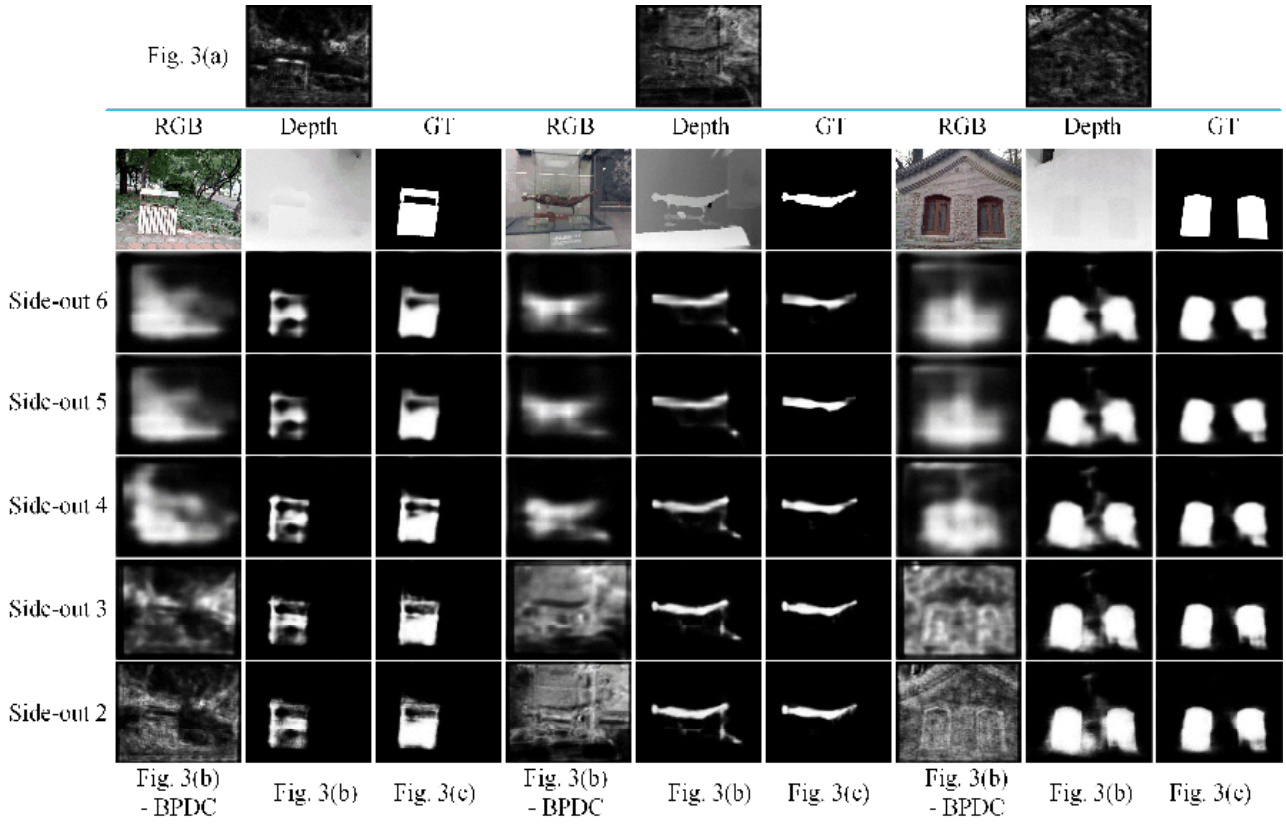

Figure 4: Visual comparison of using different multi-modal fusion modules shown in Fig. 3. 'GT' represents the ground-truth mask. 'Fig. 3(b) - BPDC' denotes the module in Fig. 3(b) without the BPDC component.

capture the complementary part from F_D^m. Then we add supervision on the RGB branch to encourage $\tilde{F}_R^m$ to be discriminative for saliency inference. The supervision will further enable informative extraction from F_D^m to complement F_R^m, and consequently drive the optimization of the residual function $\mathcal{R}_1(\bullet)$. Similarly, the residual connection and complementarity-aware supervision are also introduced for the depth branch to capture complementary information from F_R^m.

The combined RGB-D features from the m+1 layer $\tilde{F}_{RD}^{m+1}$ will be selected by a transition layer (detailed parameters are shown in Table 1). Then the enhanced features $\tilde{F}_R^m$ and $\tilde{F}_D^m$ along with the selected features $\tilde{F}_{m,RD}^{m+1}$ from the m+1 layer are concatenated and fused by one 1×1 convolutional layer to learn cooperative representations $\tilde{F}_{RD}^m$ and make integrated predictions

$$P_{RD}^m = \varphi_{RD}^m(\tilde{F}_R^m, \tilde{F}_D^m, \tilde{F}_{m,RD}^{m+1}), \quad (1)$$

where φ_{RD}^m denotes the parameters of the fusion layer and the classifier in the joint branch. Then level-wise supervision (i.e., l_{RD}^m) is added for learning multi-modal and multi-level combination.

However, due to the lack of global contextual reasoning in the shallow layers, directly minimizing the discrepancy between the predictions of any intermediate level and the ground-truth will be intractable. To encourage each level to learn desired level-specific representations, we propose to reuse the deeper features to supply high-level contexts for shallow layers. Inspired from the designs in [13] for unimodal problems, we add a backward prediction dense-connection (**BPDC**) module (shown in the right of Fig. 3(c)) to densely skip-connect the predictions from all the deeper layers (from $\tilde{P}_{RD}^{m+1}$ to $\tilde{P}_{RD}^K$ formally, where K=6 indicates the total number of convolutional blocks) to P_R^m, P_D^m and P_{RD}^m respectively. Concretely, the predictions from each level will be upsampled to 112×112 by fixed deconvolutional kernels firstly. Then all the predictions will be combined by a 1×1 convolutional layer to generate collective predictions $\tilde{P}_R^m$, $\tilde{P}_D^m$ and $\tilde{P}_{RD}^m$. By this way, the loss function of the m-th CA-Fuse block for one paired training sample includes the loss of RGB, depth and joint branches and can be represented as

Side-out	NLPR		NJUD		STEREO	
	Fig. 3(b)	Fig. 3(c)	Fig. 3(b)	Fig. 3(c)	Fig. 3(b)	Fig. 3(c)
2	0.836	**0.850**	0.845	**0.862**	0.864	**0.872**
3	0.839	**0.851**	0.843	**0.860**	0.864	**0.871**
4	0.838	**0.846**	0.837	**0.854**	0.863	**0.869**
5	0.813	**0.821**	0.809	**0.833**	0.848	**0.856**
6	0.808	**0.817**	0.813	**0.829**	0.846	**0.855**

Table 2: F-measure scores on three datasets with adopting different multi-modal fusion modules in Fig. 3.

$$
\begin{aligned}
l_{CAR}^m &= \lambda_1^m \cdot l_R^m + \lambda_2^m \cdot l_D^m + \lambda_3^m \cdot l_{RD}^m \\
&= \lambda_1^m \cdot d(\delta(\mathbf{w}_R^m P_R^m + \sum_{k=m+1}^{K} \mathbf{w}_1^k \tilde{P}_{RD}^k), \mathbf{Y}) \\
&+ \lambda_2^m \cdot d(\delta(\mathbf{w}_D^m P_D^m + \sum_{k=m+1}^{K} \mathbf{w}_2^k \tilde{P}_{RD}^k), \mathbf{Y}) \\
&+ \lambda_3^m \cdot d(\delta(\mathbf{w}_{RD}^m P_{RD}^m + \sum_{k=m+1}^{K} \mathbf{w}_3^k \tilde{P}_{RD}^k), \mathbf{Y}),
\end{aligned} \tag{2}
$$

where $\{\mathbf{w}_R^m, \mathbf{w}_D^m, \mathbf{w}_{RD}^m\}$ are the learnt weights for the predictions of three branches in the current m-th CA-Fuse block, while $\{\mathbf{w}_1^k, \mathbf{w}_2^k, \mathbf{w}_3^k\}$ are the weights for the k-th CA-Fuse block when we skip-connect the predictions of deeper layers to P_R^m, P_D^m and P_{RD}^m respectively. δ denotes the sigmoid function and $\{\lambda_1^m, \lambda_2^m, \lambda_3^m\}$ control the loss weights of each branch. We set all the weights $\lambda_1^m = \lambda_2^m = \lambda_3^m = 1$ without further tuning. d measures the cross-entropy loss between the predicted 2D saliency map $\tilde{P}$ ($\tilde{P}(x, y) \in [0,1]$ and (x,y) is the pixel location) and the ground-truth mask ($\mathbf{Y}$):

$$
d(\tilde{P}, \mathbf{Y}) = \mathbf{Y} \log \tilde{P} + (1 - \mathbf{Y}) \log(1 - \tilde{P}). \tag{3}
$$

We also involve a loss to encourage informative combination of all side outputs. Thus, the final loss function of the whole RGB-D salient object detection network is

$$
L_{final} = \sum_{m=2}^{6} l_{CAR}^m + d(\sum_{k=2}^{K} \tilde{\mathbf{w}}^k \tilde{P}_{RD}^k, \mathbf{Y}), \tag{4}
$$

where $\tilde{\mathbf{w}}^k$ is the weight for $\tilde{P}_{RD}^k$. By this way, the multi-modal features across different levels are explored and combined via successive cascade of CA-Fuse blocks and level-wise intermediate supervisions, and are reused via the skip-connections with each side output. As a result, the prediction of one CA-Fuse block incorporates the RGB-D features and predictions from all deeper levels. Therefore, we take the side output of the Conv2 CA-Fuse block as the final prediction, i.e., $\tilde{P}_{RD}^{final} = \tilde{P}_{RD}^2$.

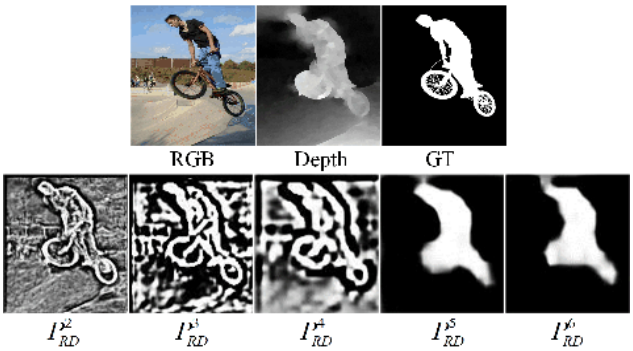

Figure 5: Outputs of each layer without combining with deeper outputs to show level-specific contributions visually.

4. Experiments

4.1. Datasets

We conduct our experiments on three most widely-used public benchmark datasets. **NLPR** [39] consists of 1000 image pairs, collected from indoor and outdoor scenes by Kinect. **NJUD** [34] and **STEREO** [46] datasets include 2003 and 797 stereoscopic images respectively. These images are mainly collected from the Internet and 3D movies. Depth images are generated by leveraging an optical method.

For fair comparison, we adopt the same training set as in [19], which contains 650 samples from the NLPR dataset and 1400 samples from the NJUD dataset. We also sample 50 image pairs from NLPR dataset and 100 image pairs from the NJUD dataset as the validation set. The training and validation sets are augmented by flipping and cropping in boundaries. The remaining samples and the STEREO dataset are used for testing. The mean values of HHA channels are computed by averaging the training samples.

4.2. Evaluation metrics

We adopt the standard metrics Precision-Recall (PR) curve and F-measure scores to evaluate the proposed method. Concretely, the saliency map will be binaried by using a series of thresholds and compared to the ground-truth. Then we will get a succession of Precision-Recall pairs and the PR curve. The formulation of F-measure is

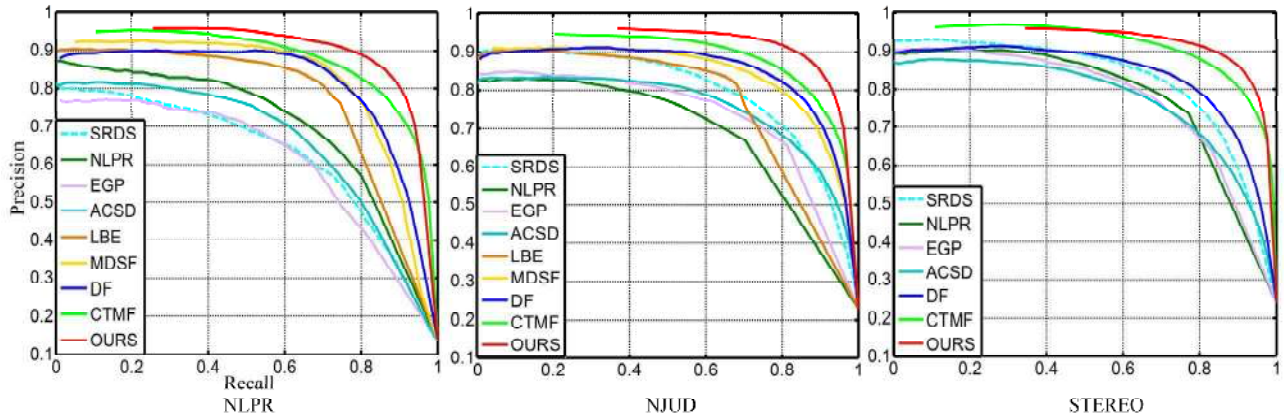

Figure 6: Quantitative comparisons to other models in terms of PR curves. The results of 'LBE' and 'MDSF' on the STEREO dataset are unavailable.

$$F_\beta = \frac{\left(1+\beta^2\right)\cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}, \quad (5)$$

where we set β^2=0.3 as suggested in [47].

4.3. Implementation details

We implement our experiments using the Caffe [48] toolbox on a PC with two 1070 GPUs. The learning rate, weight decay and mini-batch size are set as 1e-8, 0.0005 and 4 respectively. Due to limited GPU memory, we set "iter_size" as 2 to double the mini-batch size equivalently. The test time for each RGB-D image pair is merely **0.06s**.

4.4. On the effectiveness of the CA-Fuse module

We firstly investigate the effectiveness of the proposed CA-Fuse module. The saliency maps shown in the first row in Fig. 4 indicates that cascading multi-level features successively without intermediate level-wise supervisions results in ambiguous multi-level combination. The high-level contexts are not well incorporated into shallow layers. By adding intermediate supervisions (noted as 'Fig. 3(b) - BPDC' in Fig.4), the multi-modal fusion network is basically able to learn level-specific predictions. Visually, the shallow layers are capable of identifying edge information and the deep layers are able to learn global contexts to locate the salient object. Even so, the side outputs of the deep layers are badly irregular while the predictions of the shallow layers are too messy. This failure can be attributed to the deficiency of the cross-level interaction in the training phase, which probably results in *self-serving* optimization for each level individually rather than the desired collective convergence. Adding the BPDC module can remedy this shortcoming effectively (noted as 'Fig. 3(b)' in Fig. 4). With reference to the deeper side outputs, the shallow layers can enjoy high-level contexts

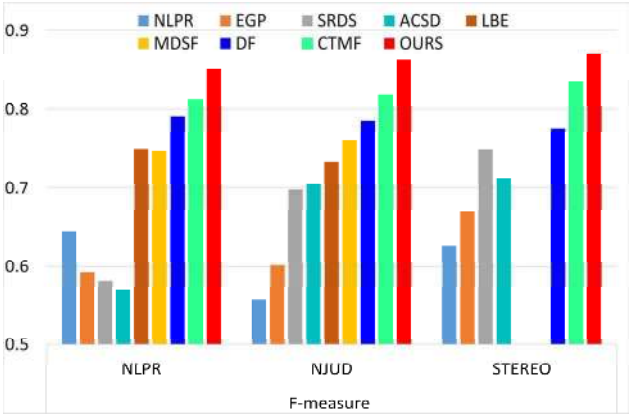

Figure 7: Quantitative comparisons to other models in terms of F-measure scores.

readily and the optimization objectives for shallow layers are degenerated into learning complementary low-level features only, thus easing the learning process and affording more cooperative multi-level fusion. As a result, the cross-level complement is better captured and incorporated and the saliency maps are enhanced from coarse to fine increasingly. Nonetheless, owing to that the multi-modal feature fusion component is still implemented by direct concatenation, the 'Fig. 3(b)' module also fails to utilize the cross-modal complement sufficiently to remove confusing background and refine salient details. In contrast, the CA-Fuse module, which involves cross-modal residual connections and complementarity-aware supervisions, is more likely to capture the cooperated information and boost better cross-modal combination, thus generating more precise saliency maps (see the columns indexed as 'Fig. 3(c)' in Fig. 4). The quantitative comparisons shown in Table 2 also demonstrate the basically step-wise accuracy gains by adopting the proposed CA-Fuse module. The F-measure score of each side output consistently outperforms the one generated without using the CA-Fuse

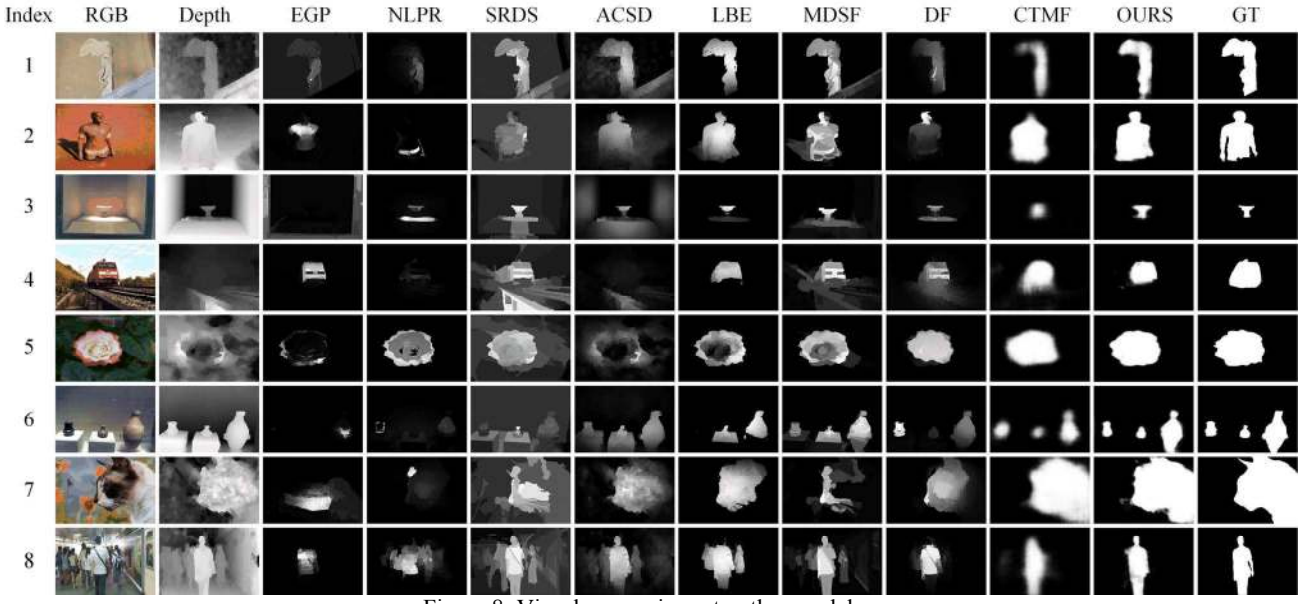

Figure 8: Visual comparisons to other models.

module. To further study the contribution of different layers, we visualize the individual output of each CA-Fuse block (i.e., P_{RD}^{m} without combining with deeper side outputs) to show level-specific inference. As shown in Fig. 5, the saliency inference scales decrease from global to local with the levels goes from deep to shallow. This visualization reveals the contribution of each layer clearly and verifies the effectiveness of the proposed cross-level fusion strategy.

4.5. Comparison with state-of-the-art

We compare our method with LBE [37], MDSF [38], EGP [36], NLPR [39], ACSD [34], SRDS [35] and two recent RGB-D saliency detection networks DF [18] and CTMF [19]. Fig. 6 and Fig. 7 indicate that our model achieves better performance than others with a large margin. More specifically, both the DF and the CTMF models are trained in a stage-wise manner. In contrast, our method is an end-to-end network, which involves no pre-training stage and post-processing operations. However, benefit from more sufficient fusion of multi-modal and multi-level features, our method still achieves much better performance than the DF and CTMF models. We also report saliency maps detected on various challenging scenes to show the advantages of the proposed method visually. Some representative samples are shown in Fig. 8 such as the appearance or depth of the salient object is not distinctive from the background (e.g., the 1st-3rd rows and the 4th-5th rows, respectively). Especially in the 4th row, the depth distribution introduces misleading saliency cues (i.e., the depth of the rail is more distinctive than the train). In the 5th row, the depth distribution is cluttered and carries little discrimination. In the 6th row, the depth of the salient object is locally-connected with some background objects. Also, the 6th row includes multiple disconnected salient objects. And in the 7th row, the appearance of the salient object is intra-variant. In the 8th row, the scene is crowded in terms of both appearance and depth distribution. In these challenging cases, most of other methods are unlikely to locate the salient object due to the lack of high-level contextual reasoning or robust multi-modal fusion strategy. Although the CTMF method is able to obtain more correct and uniform saliency maps than others, the fine details of the salient objects are lost severely due to the deficiency of cross-level fusion. By contrast, the proposed network is able to utilize both cross-modal and cross-level complementary information to learn cooperatively discriminative saliency cues and infer precise saliency values.

5. Conclusion

In this work, we propose an end-to-end RGB-D salient object detection network, which is complementarity-aware for fusing both cross-modal and cross-level features. The introduced cross-modal/level connections and modal/level-wise supervisions explicitly encourage the capture of complementary information from the counterpart, thus reducing fusion ambiguity and increasing fusion sufficiency. Comprehensive experiments demonstrate the effectiveness of the proposed multi-modal multi-level fusion strategies, which may also benefit other RGB-D systems and even other multi-model fusion problems.

Acknowledgement. This work was support by the Research Grants Council of Hong Kong (Project No CityU 11205015 and CityU 11255716).

References

[1] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from RGB-D images for object detection and segmentation. in *ECCV*, 2014.

[2] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI,* vol. 20, pp. 1254-1259, 1998.

[3] M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S. Hu. Global contrast based salient region detection. *TPAMI,* vol. 37, pp. 569-582, 2015.

[4] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang. Region-based saliency detection and its application in object recognition. *IEEE Trans. CSVT,* vol. 24, pp. 769-779, 2014.

[5] L. Shao and M. Brady. Specific object retrieval based on salient regions. *Pattern Recognit.,* vol. 39, pp. 1932-1948, 2006.

[6] A. Kim and R. M. Eustice. Real-time visual SLAM for autonomous underwater hull inspection using visual saliency. *IEEE Trans. Robot.,* vol. 29, pp. 719-733, 2013.

[7] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. in *NIPS*, 2006.

[8] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. SUN: A Bayesian framework for saliency using natural statistics. *Journal of vision,* vol. 8, pp. 32-32, 2008.

[9] J. Yang and M.-H. Yang. Top-down visual saliency via joint CRF and dictionary learning. in *CVPR*, 2012.

[10] D. Zhang, D. Meng, and J. Han. Co-saliency detection via a self-paced multiple-instance learning framework. *TPAMI,* vol. 39, pp. 865-878, 2017.

[11] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. in *CVPR*, 2015.

[12] N. Liu and J. Han. DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection. in *CVPR*, 2016.

[13] Q. Hou, M.-M. Cheng, X.-W. Hu, A. Borji, Z. Tu, and P. Torr. Deeply supervised salient object detection with short connections. *arXiv:1611.04849,* 2016.

[14] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu. Advanced Deep-Learning Techniques for Salient and Category-Specific Object Detection: A Survey. *IEEE Signal Processing Mag.,* vol. 35, pp. 84-100, 2018.

[15] N. Liu, J. Han, T. Liu, and X. Li. Learning to predict eye fixations via multiresolution convolutional neural networks. *TNNLS,* 2016.

[16] D. Zhang, J. Han, J. Han, and L. Shao. Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining. *TNNLS,* vol. 27, pp. 1163-1176, 2016.

[17] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu. Background prior-based salient object detection via deep reconstruction residual. *IEEE Trans. CSVT,* vol. 25, pp. 1309-1321, 2015.

[18] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang. RGBD Salient Object Detection via Deep Fusion. *TIP,* vol. 26, pp. 2274-2285, 2017.

[19] J. Han, H. Chen, N. Liu, C. Yan, and X. Li. CNNs-Based RGB-D Saliency Detection via Cross-View Transfer and Multiview Fusion. *IEEE Trans. Cybern.,* 2017.

[20] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik. Indoor scene understanding with RGB-D images: Bottom-up segmentation, object detection and semantic segmentation. *IJCV,* vol. 112, pp. 133-149, 2015.

[21] S.-J. Park, K.-S. Hong, and S. Lee. RDFNet: RGB-D Multi-Level Residual Feature Fusion for Indoor Semantic Segmentation. in *CVPR*, 2017.

[22] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang. Locality-Sensitive Deconvolution Networks with Gated Fusion for RGB-D Indoor Semantic Segmentation. in *CVPR*, 2017.

[23] D. Lin, G. Chen, D. Cohen-Or, P.-A. Heng, and H. Huang. Cascaded Feature Network for Semantic Segmentation of RGB-D Images. in *CVPR*, 2017.

[24] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard. Multimodal deep learning for robust rgb-d object recognition. in *IROS*, 2015.

[25] A. Wang, J. Cai, J. Lu, and T.-J. Cham. Mmss: Multi-modal sharable and specific feature learning for rgb-d object recognition. in *ICCV*, 2015.

[26] S. Song and J. Xiao. Deep sliding shapes for amodal 3D object detection in RGB-D images. in *CVPR*, 2016.

[27] H. Zhu, J.-B. Weibel, and S. Lu. Discriminative Multi-Modal Feature Fusion for RGBD Indoor Scene Recognition. in *CVPR*, 2016.

[28] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. in *CVPR*, 2016.

[29] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng. Dual path networks. *arXiv:1707.01629,* 2017.

[30] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *arXiv:1611.05431,* 2016.

[31] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao. Depth enhanced saliency detection method. in *ICIMCS*, 2014.

[32] A. Ciptadi, T. Hermans, and J. M. Rehg. An in depth view of saliency. in *BMVC*, 2013.

[33] K. Desingh, M. K. K, D. Rajan, and C. Jawahar. Depth really Matters: Improving Visual Salient Region Detection with Depth. in *BMVC*, 2013.

[34] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu. Depth saliency based on anisotropic center-surround difference. in *ICIP*, 2014.

[35] X. Fan, Z. Liu, and G. Sun. Salient region detection for stereoscopic images. in *DSP*, 2014.

[36] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Y. Yang. Exploiting global priors for RGB-D saliency detection. in *CVPRW*, 2015.

[37] D. Feng, N. Barnes, S. You, and C. McCarthy. Local background enclosure for RGB-D salient object detection. in *CVPR*, 2016.

[38] H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, and T. Ren. Depth-Aware Salient Object Detection and Segmentation via Multiscale Discriminative Saliency Fusion and Bootstrap Learning. *TIP,* vol. 26, pp. 4204-4216, 2017.

[39] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji. Rgbd salient object detection: a benchmark and algorithms. in *ECCV*, 2014.

[40] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. *IEEE Signal Processing Lett.,* vol. 23, pp. 819-823, 2016.

[41] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and C. Hou. Co-saliency Detection for RGBD Images Based on Multi-constraint Feature Matching and Cross Label Propagation. *TIP,* vol. 27, pp. 568-579, 2018.
[42] Y. Fang, J. Wang, M. Narwaria, P. Le Callet, and W. Lin. Saliency detection for stereoscopic images. *TIP,* vol. 23, pp. 2625-2636, 2014.
[43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556,* 2014.
[44] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-Supervised Nets. in *AISTATS*, 2015.
[45] S. Xie and Z. Tu. Holistically-nested edge detection. in *ICCV*, 2015.
[46] Y. Niu, Y. Geng, X. Li, and F. Liu. Leveraging stereopsis for saliency analysis. in *CVPR*, 2012.
[47] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. in *CVPR*, 2009.
[48] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. in *ACM MM*, 2014.