SDM: Semantic Distortion Measurement for Video Encryption

Yongquan Hu, Wei Zhou, Shuxin Zhao, Zhibo Chen, Weiping Li
CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System,
University of Science and Technology of China, Hefei 230027, China
chenzhibo@ustc.edu.cn

***Abstract*— Semantic information is important in video encryption. However, existing image quality assessment (IQA) methods, such as the peak signal to noise ratio (PSNR), are still widely applied to measure the encryption security. Generally, these traditional IQA methods aim to evaluate the image quality from the perspective of visual signal rather than semantic information. In this paper, we propose a novel semantic-level full-reference image quality assessment (FR-IQA) method named Semantic Distortion Measurement (SDM) to measure the degree of semantic distortion for video encryption. Then, based on a semantic saliency dataset, we verify that the proposed SDM method outperforms state-of-the-art algorithms. Furthermore, we construct a Region Of Semantic Saliency (ROSS) video encryption system to demonstrate the effectiveness of our proposed SDM method in the practical application.**

I. Introduction

With the development of multimedia, visual privacy protection becomes increasingly significant in our life. Video encryption is one of the most important means [1]. In general, traditional image quality assessment (IQA) metrics are used to evaluate the security effect of video encryption systems [2]. For example, the encrypted region is taken as safe if the peak signal to noise ratio (PSNR) of this region is around 10dB [3]. However, existing IQA metrics are usually associated with image signal instead of visual contents, which can not reflect the semantic information distortion caused by encryption due to the well-known semantic gap [4]. Meanwhile, few studies focus on visual security evaluation which has a crucial impact in measuring the semantic effectiveness of video encryption [5]. Therefore, we propose a semantic-level full-reference (FR-IQA) image quality assessment method, which we name Semantic Distortion Measurement (SDM), to measure the loss of semantic information for video contents caused by encryption.

According to the existence of reference images, existing IQA approaches are classified into three categories, namely full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA) and no-reference IQA (NR-IQA). The NR-IQA refers to automatic quality assessment of distorted images without the corresponding reference images. Most of these algorithms estimate the human perceptual quality by extracting discriminative features from distorted images based on natural scene statistic (NSS) [6], [7], [8]. In the RR-IQA mode, the comparison is limited to the partial representation of reference images. For instance, a set of reduced reference entropic differencing (RRED) algorithms for IQA based on information theory are proposed [9]. In this paper, we focus on FR-IQA, where the quality of the distorted test image is obtained based on the comparison with the non-distorted reference image. A top-down framework is usually adopted in FR-IQA [10], [11], [12], which tries to model the human visual system (HVS) [13] based on some global assumptions. Examples of classic FR-IQA algorithms include the structure similarity index (SSIM) [10] and the visual signal-to-noise ratio (VSNR) [11]. In general, these FR-IQA methods are presented according to perceptual image signal level. However, our proposed SDM method is at a higher semantic level which can reflect the semantic understanding of images more accurately. Additionally, the basic ideal of the proposed SDM method is based on image to text domain transformation which has not been tried by other similar studies as far as we know [14], [15], [16].

It should be emphasized that we design our method based on the assumption that some studies show that the most of time human focus on object-like regions when looking at an image [17], [18]. Therefore, in most practical applications such as video surveillance, we can reasonably assume that semantic information mainly exists in foreground objects and the relationship between them rather than the background, which is a key idea of our paper. Based on this assumption, we can achieve semantic encryption by encrypting objects in the foreground. As we know that the region of interest (ROI) encryption, an important type of encryption schemes, refers to encrypting only the region of interest manually selected by the administrator [19], [20]. Nevertheless, the ROI encryption cannot guarantee the semantic target of encryption because the background may be chosen to encrypt. Thus, we construct the Region Of Semantic Saliency (ROSS) video encryption system to overcome this disadvantage. Specifically, we replace the manual tracking module in ROI encryption system with the object detection module to realize the automatic detection and semantic target encryption in the foreground. Meanwhile, we also need to consider how to evaluate the validity of our proposed SDM method. Specifically, because of the same subjective visual distortion caused by encryption and obscur, we obtain distorted images by obscuring objects. First, we separately obscure all the semantic objects in the foreground of the original image and obtain obscure distortion images. Then, we employ our SDM method to compute the scores between the original image and

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFC0801001, in part by the NSFC under Grant 61571413, Grant 61632001, and Grant 61390514, and in part by the Intel ICRI MNC.

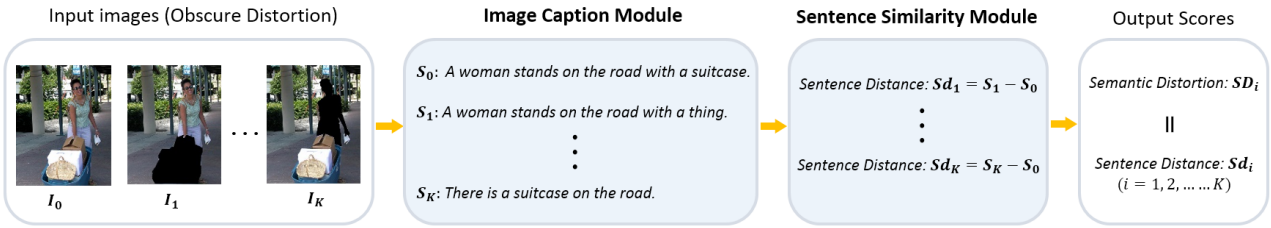


Fig. 1: The framework of our proposed SDM method.

distorted images. Finally, a dataset is chosento to evaluate the semantic accuracy of our method. Two important factors are needed to be considered as follows:

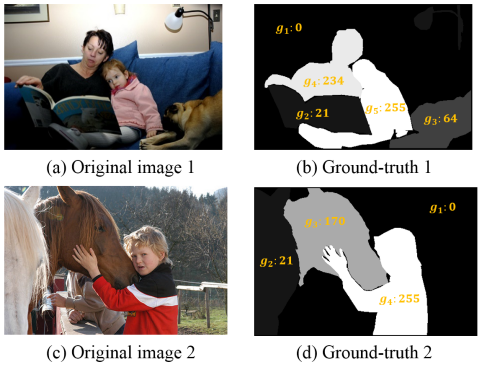
1) Obscure Object Distortion: We produce distorted images by obscuring, which is to simulate the effect of encryption. Therefore, the object in the foreground of evaluation dataset needs to be clearly segmented.

2) Semantic Ground-Truth: There should be clear value in the corresponding region of the segmentation object, which can be taken as the ground-truth of semantic information.

In conclusion, we choose a saliency dataset called SalObj [21] as the evaluation dataset. The saliency value in this dataset is adopted to represent the semantic information ground-truth of the object in the foreground, which we name the "semantic saliency ground-truth".

In this paper, we first propose a Semantic Distortion Measurement (SDM) method based on the image caption module and the sentence similarity module. Second, we construct a ROSS encryption system by improving the existing ROI video encryption system [22]. Finally, two experiments are conducted to show the effectiveness of our SDM method compared with other metrics. The first experiment is to evaluate the semantic accuracy of our method based on a saliency dataset. The second demonstrates that the scores obtained by our method can reflect the encrpted content more effectively. To the best of our knowledge, this is the first work that proposes a novel image-to-text-based IQA method to measure the image semantic distortion and applies it to the encrypted scene with good results [14], [15], [16].

The remainder of the paper is organized as follows. Section II presents the proposed SDM method. Next, we introduce our ROSS encryption system in Section III. Then, the experimental results are presented in Section IV. Finally, the conclusion is given in Section V.



(a) Original image 1 (b) Ground-truth 1

(c) Original image 2 (d) Ground-truth 2

Fig. 2: Examples of the SalObj dataset.

II. Semantic Distortion Measurement Method

In this section, we first introduce the evaluation dataset called SalObj [21] used in our paper. Then we present the details of our proposed semantic-level FR-IQA method called Semantic Distortion Measurement (SDM).

A. Evaluation Dataset

To the best of our knowledge, there is no semantic dataset for video encryption yet. Meanwhile, based on the two factors and the key assumption mentioned above, i.e. the obscure object distortion and the semantic ground-truth, we use a saliency dataset as a semantic saliency dataset to evaluate the semantic accuracy of our SDM method, which can provide the semantic information of objects in the foreground. Saliency is usually divided into two categories: fixation prediction and salient object segmentation [23]. The fixation prediction deals with predicting the locations that the human observer focus on, while the salient object segmentation aims to segment the most salient object. The SalObj dataset is a high-quality saliency dataset that offers both fixation and salient object segmentation ground-truth by augmenting 850 images from the PASCAL 2010 [24] dataset.

Specifically, we utilize the salient object segmentation subset of the SalObj dataset. As illustrated in Fig. 2, figures (a) and (c) are the original images. Figures (b) and (d) are the salient object ground-truth that we use. The salient object segmentation experiment in the SalObj dataset is conducted by 12 subjects labeling the salient objects. The grayscale value (0∼255) of each segmented object indicates the degree of saliency (the higher the saliency, the higher the value).

The SalObj saliency dataset has clear segmentation,which can be utilized to obscure different semantic targets. Most importantly, the different levels of grayscale value in these segmentation regions can be regarded as the semantic importance of the corresponding target, which we define as semantic saliency ground-truth. To our best knowledge, it is the only publically available dataset with different levels of saliency objects, while other salient object datasets are just divided into two levels, i.e., "saliency" (binary 1) and "not saliency" (binary 0). Therefore, we choose the SalObj dataset as the semantic saliency benchmark for our SDM method.

B. SDM Algorithm

Inspired by the fact that children learn new semantic concepts by observing the visual words and listening to the descriptions from their parents [25], we propose a novel Semantic Distortion Measurement (SDM) method. The basic

idea of the proposed SDM method is the domain transformation described as follows.

The framework of our proposed SDM method is shown in Fig. 1. Let I_0 be the original image. First, we introduce distortion to the original image and then obtain distorted images. Specifically, based on the assumption that semantic information mainly exists in foreground objects and the relationship between them, we obscure each semantic object region in the foreground, which can obtain the corresponding obscure distortion images $I_1 \sim I_K$.

Second, we convert the original image and distorted images to the corresponding sentences with the state-of-the-art image caption model named neural image caption (NIC) [26]. In other words, the original image I_0 and distorted images $I_1 \sim I_K$ are transformed to the original sentence S_0 and distorted sentence $S_1 \sim S_K$, respectively.

Third, we compute the sentence distance Sd_i between the S_0 and $S_i (i = 1, 2, \ldots, K)$ with the word mover's distance (WMD) [27] and the semantic propositional image caption evaluation (SPICE) [28] algorithms, where K denotes to the number of distorted images. Also, we name them SDM-WMD and SDM-SPICE algorithms.

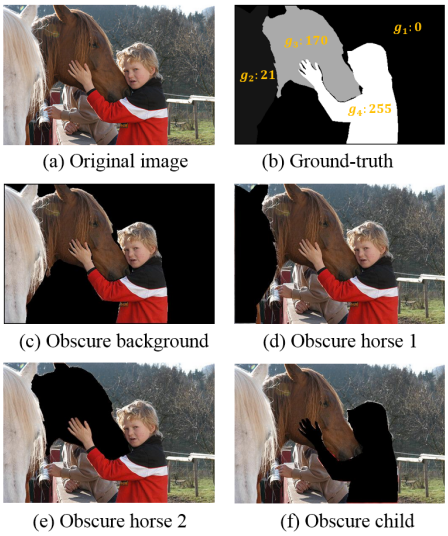
Finally, through this transformation, we can use the visible image caption to represent the abstract semantic information in the image. The degree of distortion of the image SD_i is equal to the corresponding distance between sentences Sd_i.

III. Ross Encryption System

In this section, we first present our constructed Region Of Semantic Saliency (ROSS) system in details. Then, we introduce the concept of 'Sensitivity' in our experiment in order to measure the changing degree of different IQA scores for different encryption regions.

A. System Overview

For most specific scenarios, we can assume that the semantic information mainly exists in foreground objects. Meanwhile, obvious semantic information cannot be accessed because the background is fixed and unchanging. Based on this assumption, here is the problem of traditional ROI encryption system: The ROI is usually chosen by the encryptor manually [29], [22], [30] and the ROI selected by people may not be the most meaningful and semantic goals for others, which is not conducive to the realization of fully automated semantic system design. Moreover, it is time-consuming and not impractical to select the encryption area artificially before each encryption. Therefore, when actually building the system, we need to use automatic semantic target detection module instead of manually selecting tracking module. Particularly, we re-implement the code and replace the Open-Tracking-Learning-Detection (Open-TLD) algorithm of [22] with the automatic detection model Mask R-CNN in [31]. Thus, we realize a Region Of Semantic Saliency (ROSS) video encryption to verify the effectiveness of our SDM algorithm.



(a) Original image (b) Ground-truth

(c) Obscure background (d) Obscure horse 1

(e) Obscure horse 2 (f) Obscure child

Fig. 3: Illustration of experimental ground truth and test images.

B. Sensitivity Measurement

To measure the "semantic contents descriptive ability" of IQA metrics/methods for different encrypted regions, we introduce the 'Sensitivity' index. In machine learning, we can use the relative standard deviation (RSD) to describe the stability of a classifier as follows [32]:

$$RSD = \frac{\sigma_c}{\mu_c} \times 100\%, \tag{1}$$

where σ_c and μ_c are the standard deviation and mean of the classification values, respectively. In particular, we consider only two values C_1 and C_2 in video encryption scene. C_1 and C_2 are the two IQA scores corresponding to two different encryption regions R_1 and R_2. We then define the 'Sensitivity' as follows:

$$Sensitivity = RSD_2 = \frac{\sqrt{\sum_{j=1}^{2}(C_j - \mu_c)^2}}{\frac{1}{2}(C_1 + C_2)} = \frac{|C_1 - C_2|}{C_1 + C_2}. \tag{2}$$

Formula (2) shows that under the same kind of distortion, the greater value of 'Sensitivity', the more sensitive the current metric is for different encrypted regions, which is able to be seen as a better ability to distinguish the image content.

IV. Experiment Results

In this section, we first evaluate the performance of our SDM algorithm on the salient object segmentation subset of the SalObj dataset. Then, we apply the proposed SDM method to assess the semantic distortion caused by ROSS encryption in order to explore the sensitivity of different methods.

A. Performance Evaluation

Our experimental ground truth and test images are given in Fig. 3. Figures (a) and (b) are the reference/original image and the semantic saliency ground truth respectively. It is worth noting that there are four separate regions (background, horse 1, horse 2, child) of different semantic saliency levels in figure (b) and the semantic saliency ground truth is $\boldsymbol{G} = \{g_1, g_2, g_3, g_4\} = \{0, 21, 170, 255\}$.

TABLE I: Performance Comparison on SalObj Dataset

	SDM-WMD	*SDM-SPICE*	*PSNR*	*SSIM*	*MSE*	*MSSIM*	*VSNR*
SROCC	**0.7319**	**0.7189**	0.7032	0.6938	0.7042	0.6953	0.7010
KROCC	**0.6742**	**0.6674**	0.6356	0.6265	0.6373	0.6315	0.6362
	VIFP	*UOI*	*IFC*	*NQM*	*WSNR*	*SNR*	*VIF*
SROCC	0.6854	0.6905	0.6842	0.7081	0.7039	0.7042	0.6763
KROCC	0.6200	0.6232	0.6179	0.6426	0.6379	0.6373	0.6021

(a) Automatic Detection of ROSS (b) ROSS encryption effect

(c) Manual Selection of ROI (d) ROI encryption effect

Fig. 4: The comparison of the effect of ROSS encryption and ROI encryption ($QP = 32$).

First, we obscure these four semantic segmentation regions in sequence to simulate the influence of encryption (making it impossible for the human eye to identify objects in the image), resulting in figure (c-f).

Second, we adopt fourteen kinds of typical FR-IQA metrics such as PSNR, SSIM as comparison to our algorithms to compute the score between the obscure distortion images (figure (c-f)) and the original image (figure (a)), such as the SD-WMD score is $S_{SD-WMD} = \{w_1, w_2, w_3, w_4\}$ and the PSNR score is $S_{PSNR} = \{p_1, p_2, p_3, p_4\}$.

Finally, we adopt the Spearman rank order correlation coefficient (SROCC) and Kendall rank-order correlation coefficient (KROCC) to evaluate the performance. The results in Table I demonstrate that our proposed SDM outperforms other state-of-the-art algorithms. The higher correlation of our method indicates that our method can indeed reflect higher-level semantic information than the lower-level signal and structure information.

B. Sensitivity of Video Encryption

To compare the 'Sensitivity' of our SDM with other metrics, we test a video sequence called Kimono1 which is coded in the randomaccess configuration with the quantization parameter $QP = 32$. Then, as shown in Fig. 4, we select two different encryption regions. From Fig. 4, we can see that figures (a) and (b) are the semantic encrypted region (woman in the foreground) automatically detected by our ROSS system and the corresponding encryption effect. Meanwhile, figures (c) and (d) show the defect of non-semantic encrypted region (tree in the background) caused by selecting ROI manually. As a result, some mistakes may be made on account of the subjective selection. Therefore, our system is beneficial to exclude some interference from the background in advance and define the scope of semantic goals objectively. Note that the detected regions in Fig. 4 are rectangular due to the 'Tile' mechanism used in the encryption system, i.e. each Tile region of High Efficiency Video Coding (HEVC) is rectangular [22].

Table II shows the test results of the video sequence in Fig. 4. The two scores of SSIM are almost the same for two different encryption regions. Similarly, the sensitivity of PSNR is also weak. Nevertheless, the values of 'Sensitivity' of our SDM-WMD and SDM-SPICE are 60.21% and 86.28% respectively, which verifies that our approach has the better ability to identify semantic targets and background relative to other metrics in our ROSS encryption system.

Furthermore, we change the video quantization parameters $QP = \{22, 27, 32, 37, 42, 47\}$. Fig. 5 shows that the 'Sensitivity' performance with different QPs. It can be seen that our method is robust for different video QPs, which is still higher than the PSNR and SSIM even for low quality.

TABLE II: The Test Results of Video Encryption ($QP = 32$)

	PSNR	*SSIM*	*SDM-WMD*	*SDM-SPICE*
Woman	15.5087	0.6987	3.1672	0.7865
Tree	13.1816	0.6542	0.7865	0.6796
Sensitivity	8.09%	3.29%	**60.21%**	**86.28%**

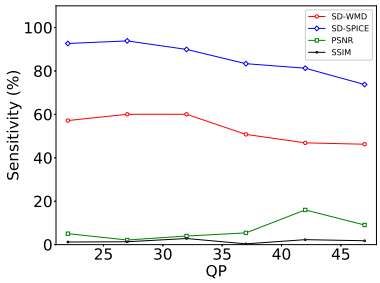


Fig. 5: The 'Sensitivity' performance with different QPs in Kimono1.

V. Conclusion

In this paper, we propose a novel algorithm named SDM to measure image semantic distortion in video encryption. This method is based on domain transformation and verified to have better accuracy in semantic distortion measurement compared with other well-known metrics. Furthermore, we construct a ROSS encryption system to show the high sensitivity and good robustness of our proposed method.

Since the profile of obscuring objects can still provide some semantic information for people to recognize, we may design a subjective experiment aiming to replace the object saliency values of SalObj. Besides, the accuracy of detection module in the ROSS system can be improved in future study.

REFERENCES

[1] Z. Shahid and W. Puech, "Visual protection of hevc video by selective encryption of cabac binstrings," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 24–36, 2014.

[2] L. Dubois, W. Puech, and J. Blanc-Talon, "Smart selective encryption of H.264/AVC videos using confidentiality metrics," *Annals of telecommunications - annales des télécommunications*, vol. 69, no. 11-12, pp. 569–583, 2014.

[3] M. Farajallah, "Chaos-based crypto and joint crypto-compression systems for images and videos," Ph.D. dissertation, UNIVERSITE DE NANTES, 2015.

[4] Q. Cao, Y. Ying, and P. Li, "Similarity metric learning for face recognition," in *Computer Vision (ICCV), IEEE International Conference on*. IEEE, 2013, pp. 2408–2415.

[5] L. Tong, F. Dai, Y. Zhang, and J. Li, "Visual security evaluation for video encryption," in *International Conference on Multimedia 2010, October*, 2010, pp. 835–838.

[6] N. Joshi and A. Kapoor, "Learning a blind measure of perceptual image quality," in *Computer Vision and Pattern Recognition*, 2011, pp. 305–312.

[7] Y. Fang, K. Ma, Z. Wang, W. Lin, Z. Fang, and G. Zhai, "No-reference quality assessment of contrast-distorted images based on natural scene statistics," *IEEE Signal Processing Letters*, vol. 22, no. 7, pp. 838–842, 2015.

[8] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012.

[9] R. Soundararajan and A. C. Bovik, "RRED indices: reduced reference entropic differencing for image quality assessment," *IEEE Transactions on Image Processing*, vol. 21, pp. 517–26, 2012.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[11] D. M. Chandler and S. S. Hemami, "VSNR: a wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284–2298, 2007.

[12] J. Preiss, F. Fernandes, and P. Urban, "Color-image quality assessment: from prediction to optimization." *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1366–1378, 2014.

[13] S. Winkler, "Issues in vision modeling for perceptual video quality assessment," *Signal Processing*, vol. 78, no. 2, pp. 231–252, 1999.

[14] P. Zhang, W. Zhou, L. Wu, and H. Li, "SOM: Semantic obviousness metric for image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 2394–2402.

[15] D. Li, T. Jiang, and M. Jiang, "Exploiting High-Level Semantics for No-Reference Image Quality Assessment of Realistic Blur Images." ACM Press, 2017, pp. 378–386.

[16] D. Liu, D. Wang, and H. Li, "Recognizable or not: Towards image semantic quality assessment for compression," *Sensing and Imaging*, vol. 18, no. 1, p. 1, 2017.

[17] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision*, 2010, pp. 2106–2113.

[18] K. Y. Chang, T. L. Liu, H. T. Chen, and S. H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *International Conference on Computer Vision*, 2011, pp. 914–921.

[19] L. Tong, F. Dai, Y. Zhang, and J. Li, "Prediction restricted h. 264/avc video scrambling for privacy protection," *Electronics letters*, vol. 46, no. 1, pp. 47–49, 2010.

[20] X. Ma, W. K. Zeng, L. T. Yang, D. Zou, and H. Jin, "Lossless ROI Privacy Protection of H.264/AVC Compressed Surveillance Videos," *IEEE Transactions on Emerging Topics in Computing*, vol. 4, no. 3, pp. 349–362, 2016.

[21] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The Secrets of Salient Object Segmentation," in *Computer Vision and Pattern Recognition*, 2014, pp. 280–287.

[22] X. Cheng and H. Li, "Encryption system integrated with roi and tracking for high efficiency video coding," *International Proceedings of Computer Science and Information Technology*, vol. 59, p. 177, 2014.

[23] J. Wang, H. R. Tavakoli, and J. Laaksonen, "Fixation Prediction in Videos Using Unsupervised Hierarchical Features," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jul. 2017, pp. 2225–2232.

[24] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[25] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille, "Learning like a child: Fast novel visual concept learning from sentence descriptions of images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2533–2541.

[26] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[27] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, "From word embeddings to document distances," in *International Conference on International Conference on Machine Learning*, 2015, pp. 957–966.

[28] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European Conference on Computer Vision*. Springer, 2016, pp. 382–398.

[29] M. Farajallah, W. Hamidouche, O. Déforges, and S. El Assad, "Roi encryption for the hevc coded video contents," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 3096–3100.

[30] C. Bergeron, N. Sidaty, W. Hamidouche, B. Boyadjis, J. Le Feuvre, and Y. Lim, "Real-time selective encryption solution based on roi for mpeg-a visual identity management af," in *International Conference on Digital Signal Processing (DSP)*. IEEE, 2017, pp. 1–5.

[31] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2980–2988.

[32] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, and H. J. W. L. Aerts, "Machine Learning methods for Quantitative Radiomic Biomarkers," *Scientific Reports*, vol. 5, p. 13087, Aug. 2015.