ORAL PRESENTATION

A Comparison of Image Fusion Quality Metrics

Emre Bendeş1* (ORCID: https://orcid.org/0000-0002-9432-3278)

*1 Nevşehir Hacı Bektaş Veli University, Engineering-Architecture Faculty, Computer Engineering Department, Nevşehir, Turkey.

*Corresponding author e-mail: emrebendes@nevsehir.edu.tr

Abstract
Measuring the fitness of the fused image plays a key role in image fusion applications. For a learning process performed in a machine learning algorithm, the result of the fusion should be evaluated numerically. In the literature, there are well-known quality metrics developed for this purpose. Each metric evaluates the quality of the image using a different method. However, to be used in the learning process, the quality metrics must be able to provide results compatible with the change in the image's visual quality. In this study, synthetic images with known quality levels were created for this purpose. The scoring accuracy of six quality metrics commonly used in the literature was compared with these test images and the results were evaluated.
Keywords: image fusion, image quality metrics, image processing.

INTRODUCTION

Merging the information contained in two or more images to obtain a single image with more meaningful information is called image fusion. Image fusion is used in a variety of fields. We can divide image fusion applications into two: single-mode and multi-mode image fusion (Huang et al. 2020). While single-mode image fusion captures images of the scene with different parameters from the same sensor, multi-mode image fusion captures the same scene with different sensors.

The best-known application of single-mode image fusion is multi-focus image fusion. In multi-focus image fusion, images captured at different focal lengths are combined (Xiaole Ma et al. 2021). Due to the structure of the camera hardware, it is known that the images can only be focused to a certain distance. Therefore, objects near the focused distance will appear sharp in the image, while objects closer or farther away will be perceived as blurred depending on the distance. If it is desired to produce an image where every part of the scene is clearly visible, it is necessary to transfer the clear areas in the source images captured with different focal lengths into a fused image. The most important step in this process is to determine the sharpness of the source images.

On the other hand, multi-mode image fusion also has a wide range of applications (Aslantas et al. 2014). Medical applications combining CT images sensitive to hard tissue and MR images sensitive to soft tissue are common in the literature (Li et al. 2021). Military applications are also well-known examples of multi-mode image fusion. Battlefield observation (Wanyi Zhang et al. 2020) and concealed weapon detection are examples of this application area. Another multi-modal image fusion application is enhanced night vision (Yanling Chen et al. 2022; Nashwan Jasim Hussein et al. 2017), which aims to improve perception in dark environments. Another common application is remote sensing where multispectral images are combined (Kurban 2021).

The importance of the information obtained through the fusion process depends on the application area. Regardless of the purpose of image fusion, the resulting fused image should be evaluated by objective or subjective tests. Subjective evaluation is a process of interpreting the quality of the image based on human perception. Human perception can be influenced depending on environmental conditions (Vladimir Petrović 2007). Moreover, it is not suitable for use in an intelligent software algorithm since it is a time-consuming process (Aslantas et al. 2014). For this reason, a function that objectively provides a numerical value about the quality of the image is needed. This function is called the objective quality metric. Many quality metrics have been proposed in the literature (Aslantas and Bendes 2015; Xydeas and Petrovic 2000; Guihong Qu et al. 2002; Eskicioglu and Fisher 1995; V. Aslantas and R. Kurban 2009).

Metrics can be used to compare fusion methods as well as a measure of fitness value in processes performed with deep learning or optimization algorithms (Liu et al. 2018; Zhang 2021; Zhou et al. 2019; Sen Qiu et al. 2021; Li et al. 2018; Aslantas et al. 2014). It is important to compare quality metrics for such applications.

Quality metrics are expected to give better scores the better the fused image is. An image fusion process should transfer the meaningful information in the source images into the fused image. Therefore, defining the quality of a metric in terms of image fusion is not only about what the fused image contains, but also about how much meaningful information the source images provide. Meaningful information is edge information that can be perceived in an image (Zhiyun Xue ,Zhong-Ying Zhang ,Rick S. Blum 2005).

In this context, synthetic test image sets, whose quality can be determined independently of human perception, were created for single-mode and multi-mode image fusion in this study to compare quality metrics. These sets consist of input images and some fused images whose visual quality order is known in advance. In this way, it was possible to observe how the quality metrics perform as the quality of the image fusion increases.

The rest of the paper is organized as follows: Section 2 describes the image quality metrics used in experiments. In Section 3 test images sets are introduced and experimental results are discussed. Finally, the conclusion is presented.

IMAGE QUALITY METRICS

Quality metrics used in experiments are defined in this section. First and second source images are represented as I_1 and I_2 respectively and fused image is represented as I_f in the equations.

The structural similarity (*SSIM*) is a metric that yields similarity between images (Wang et al. 2004). the mean SSIM index value between the two images computed as follows:

$$MSSIM(x,y) = \frac{1}{M}\sum_{j=1}^{M} SSIMI(x_j, y_j) \tag{1}$$

Where *x* is fused image and *y* is source image. The *SSIM* is calculated as:

$$SSIMI(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2}$$

Where μ_x and μ_y are averages and, σ_x and σ_y are variance. σ_{xy} is the covariance of inputs. $C_1 = 6.5025$ and $C_2 = 58.5225$. We used SSIM metric in our experiments as follows:

$$SSIM = MSSIM(I_1, I_f) + MSSIM(I_2, I_f) \tag{3}$$

Quality of edge (*QoE*) is a measure of how much edge from the input images transferred to fused image (Xydeas and Petrovic 2000). *QoE* is use the sobel operators for this purpose.

$$QoE = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} k^a(i,j)w^a(i,j) + k^b(i,j)w^b(i,j)}{\sum_{i=1}^{n}\sum_{j=1}^{m} w^a(i,j) + w^b(i,j)} \tag{4}$$

where w^a and w^b are weighting coefficients based on sobel edge strength of the input images, k^a ve k^b edge preservation coefficients.

Mutual information (MI) make use of Kullback-Leibler measure to determine shared information between two images as follows (Guihong Qu et al. 2002):

$$M(x,y) = \sum_{i,j} P_{xy}(i,j)\log\frac{P_{xy}(i,j)}{P_x(i)P_y(j)} \tag{5}$$

Where P_{xy} is the normalized joint gray level histogram of the input images. P_x and P_y are normalize histogram od input images respectively. Maximum contribution of input images is desired in image fusion applications. Consequently, sum of the shared information between input images and fused image are used as metric value in the *MI* as follows:

$$MI = M(I_1, I_f) + M(I_2, I_f) \tag{6}$$

Standard deviation (STD) is another quality metric used in the experiments. Diversity of intensity value is an indicator of image quality. This metric make evaluation just by using fused image as follows:

$$STD = \sqrt{\frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} \left[I(i,j) - \bar{I} \right]^2} \tag{7}$$

Where *i* and *j* are pixel coordinates and $\bar{I}$ is the mean value of the pixel values of the fused image.
Spatial Frequency (SF) measures the general activity that expresses the changes in the image (Eskicioglu and Fisher 1995). It is built on first-order differences between neighboring pixels. First, the gradients of the pixel differences for the rows (*R*) and columns (*C*) in the I image are calculated as follows:

$$C = \left[\frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} \left[I(i,j) - I(i-1,j) \right]^2 \right]^{1/2} \tag{8}$$

$$R = \left[\frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} \left[I(i,j) - I(i,j-1) \right]^2 \right]^{1/2} \tag{9}$$

The final *SF* value is expressed as:

$$SF = \sqrt{C^2 + R^2} \tag{10}$$

Sum of the correlation of differences (SCD) is a measure of how much the input image make contribution to the fused image (Aslantas and Bendes 2015). The main idea in the *SCD* metric is the fact that the fused image is built from the source images. Therefore, the difference between any source image and the fused image will give information about the contribution of the other source image. These differences expressed as:

$$F_1 = I_f - I_1 \tag{8}$$

$$F_2 = I_f - I_2 \tag{9}$$

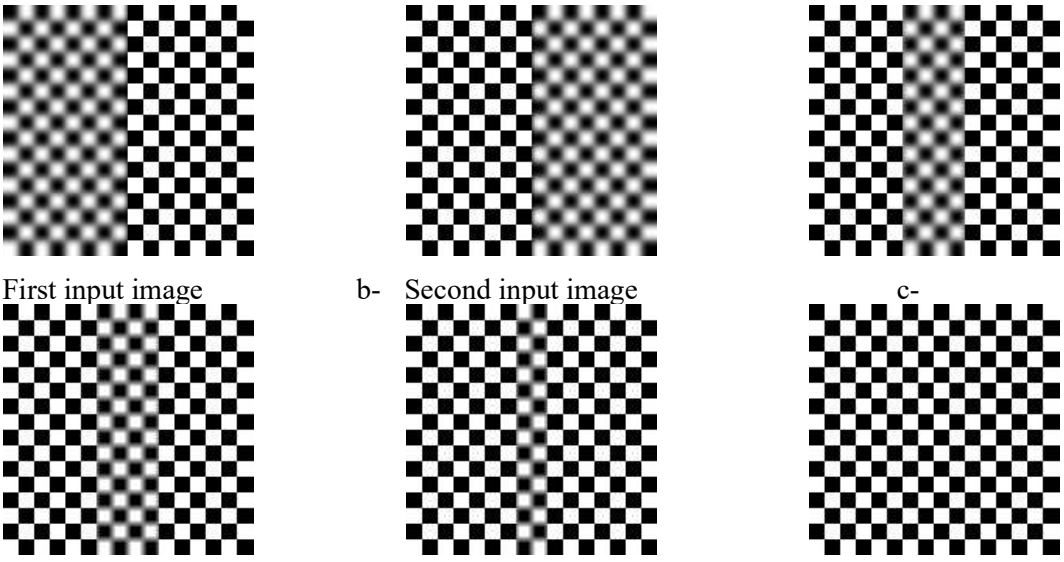
The SCD value is calculated as follows:

$$SCD = r(F_1, I_1) + r(F_2, I_2) \tag{10}$$

Where *r(.)* is the Pearson correlation between two images.

TEST IMAGES

Some sets of test images are produced synthetically. In this way, it is known in advance what the fused image should contain. Consequently, a visual evaluation is possible that is not influenced by environmental conditions and personal perception.

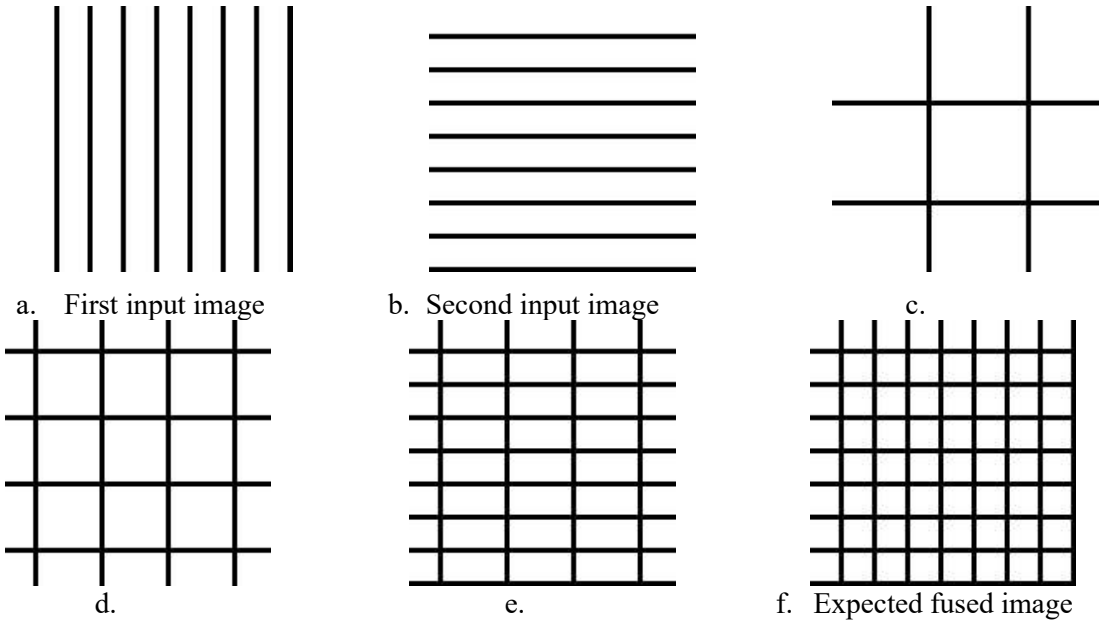


a- First input image b- Second input image c-

d- e- f- Expected fused image

Figure 1. The first test image set

Figure 1 shows the first synthetic test image set. This set was created for multi-focus image fusion. There are some blurred areas in the first and second input images. It is expected that there should be no blurred areas in a fused image. Therefore, in this test image set, we know which areas of the input images should be moved to the fused image. The sharp image has no blurred areas, as shown in Figure 1 f.



a. First input image b. Second input image c.

d. e. f. Expected fused image

Figure 2. The second test image set

In multi-sensor image fusion, the input images of the same scene are acquired from different sensors such as the thermal, the visible, or the MR. Therefore, the intensity values in all the input images have different meanings. Regardless of what the intensity values stand for, something that can be detected by one sensor in the scene will cause intensity variations in the image. That is, the edges would show up on the image. Accordingly, in Figure 2, some edges were added to the input images and all detectable information was represented in the expected fused image.

EVOLUTIONAL RESULTS

To show the performance of the metric, we first establish the matches between the images in the test sets to define the inputs and outputs for image fusion. Each image set contains two predefined source images. These images became the input images for each matching. In the matching process, each image sequence is assumed to be the output of the fusion process and the codes starting with "F" (fusion) are specified for matching. The mappings between F1 and F6 were obtained from the first image set, and the mappings between F7 and F12 were obtained from the second image set. These mappings are listed in Table 1.

Table 1. Image fusion tests

Test	First Input	Second Input	Fused Image
F1	Figure 1 a-	Figure 1 b-	Figure 1 a-
F2	Figure 1 a-	Figure 1 b-	Figure 1 b-
F3	Figure 1 a-	Figure 1 b-	Figure 1 f-
F4	Figure 1 a-	Figure 1 b-	Figure 1 d-
F5	Figure 1 a-	Figure 1 b-	Figure 1 e
F6	Figure 1 a-	Figure 1 b-	Figure 1 f
F7	Figure 2 a	Figure 2 b	Figure 2 a
F8	Figure 2 a	Figure 2 b	Figure 2 b
F9	Figure 2 a	Figure 2 b	Figure 2 c
F10	Figure 2 a	Figure 2 b	Figure 2 d
F11	Figure 2 a	Figure 2 b	Figure 2 e
F12	Figure 2 a	Figure 2 b	Figure 2 f

Table 2 shows the test quality scores for the first set of images. The desired result for this set is that it receives the best value for the F6 test, where Figure 1 f is the output. The SSIM QoE STD and SCD metrics give the best values for the desired image when examining the values. However, not every metric can adhere to the order mentioned in the previous section.

Table 2. Metric values for multi-focus tests.

Test	*SSIM*	*MI*	*QoE*	*SF*	*STD*	*SCD*
F1	1,4481	**3,1927**	0,5245	9,2724	11,1597	NaN
F2	1,4481	**3,1927**	0,5245	**9,2770**	11,1597	NaN
F3	1,4327	2,1887	0,5632	7,4659	11,2214	0,7766
F4	1,4350	2,0008	0,5457	6,7196	11,2460	0,8959
F5	1,4356	1,9195	0,5801	5,8796	11,2688	1,0341
F6	**1,4490**	1,8407	**0,6013**	4,8894	**11,2916**	**1,1699**

While the various metrics provide the best value for the right image, it turns out that they cannot rank properly for the intentionally produced images. To better illustrate this situation, linear normalized metric values are given in Figure 3. Each test value for each metric is presented in the form of a graph. When looking at the graphs, it becomes clear that not every metric can provide the desired ranking. While the metrics MI and SF provide an opposite score to the desired results, it is observed that a parallel score is obtained for metrics other than SSIM. For the SSIM metric, the result in the first two tests gives a high value as it is one of the image inputs. The SCD metric calculates the difference between the input image and the output image. The result of this calculation is a zero matrix in the first two tests. Due to the zero matrix in the correlation calculation used by its mathematical structure, it could not produce a numerical value in F1 and F2 tests, but in the other tests, it could produce the desired graph

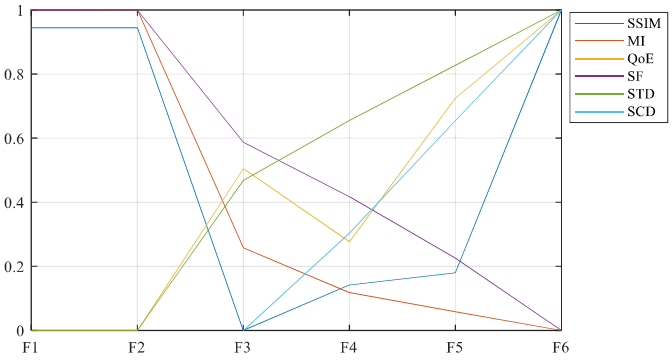


Figure 3. Normalized metric values for multi-focus results.

Table 3 shows the evaluation results of the quality metrics for the second image set. As with the first image set, the final predefined test for the second image set includes the desired fusion result, and the metrics are expected to provide the best result for this test. The best values are highlighted in bold in the table. Looking at the values in the table, each metric has achieved the best value in the F12 test.

Test	SSIM	MI	QoE	SF	STD	SCD
F7	1,1909	1,2977	0,5105	78,0775	91,0532	NaN
F8	1,1909	1,2977	0,5103	78,0775	91,0532	NaN
F9	1,0231	1,1026	0,2995	55,9404	66,5827	0,4689
F10	1,1339	1,1529	0,4958	77,5552	89,6247s	0,9054
F11	1,2045	1,2566	0,6417	91,6649	104,5379	1,3846
F12	**1,2160**	**1,3202**	**0,7413**	**101,8006**	**114,1801**	**1,8200**

Looking at the variation in metric scores between tests for the second set of images in Figure 4, after the two high scores in tests F7 and F8, there is an increase from test F9 to test F12. In the first two tests, the result image is one of the input images. Considering that the meaningful information in this set of images is the horizontal and vertical lines in the source images, we can observe that the total number of edges and the number of data transmitted from the source images in F7 and F8 is higher than that of the result images in tests F9, F10 and F11. All metrics provided the expected results except that the first two tests did not provide results due to the structure of the SCD.

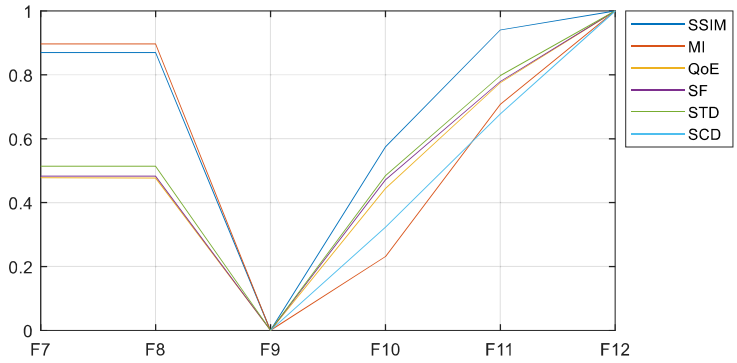


Figure 1. Normalized metric values for multi-sensor results.

The importance of metrics that provide numerical results compatible with visual quality in Deep Learning or optimization processes cannot be ignored. Metrics are functions that provide numerical values about the fitness of the fused image and are used in objective evaluation. If machine learning is to be performed for image fusion, the metric values would play a key role in the training process of the system because the result images that the system would produce should be evaluated objectively and the development should be observed numerically. At this point, the values produced by the metrics should be compatible with the subjective evaluation. The experimental tests conducted show how the metrics produce results in different applications of image fusion.

CONCLUSION

In this paper, six image fusion quality metrics from the literature are compared in terms of their compatibility with visual quality changes. To compare the results of the quality metrics, two image sets are synthetically created. Each image set consists of two predefined input images and four fused images. The input images of the first image set contain blurred images for use in multi-focus image fusion. The second image set contains input images consisting of synthetically replaced edges.

The experimental result shows that quality metrics that use the input image for evaluation are more consistent with visual quality. Since sharpness increases the standard deviation, *STD* gives better results in the multi-focus application as the overall sharpness increases. In both sets of tests, *QoE*, *SCD* and *SSIM* provide more acceptable results than the others for all tests.

REFERENCES

Aslantas, Veysel; Bendes, Emre (2015): A new image quality metric for image fusion: The sum of the correlations of differences. In AEU - International Journal of Electronics and Communications 69 (12), pp. 1890–1896. DOI: 10.1016/j.aeue.2015.09.004.

Aslantas, Veysel; Bendes, Emre; Kurban, Rifat; Toprak, Ahmet (2014): New optimised region-based multi-scale image fusion method for thermal and visible images. In Image Processing, IET 8, pp. 289–299. DOI: 10.1049/iet-ipr.2012.0667.

Eskicioglu, A. M.; Fisher, P. S. (1995): Image quality measures and their performance (43). In IEEE Transactions on Communications (12).

Guihong Qu; Dali Zhang; Pingfan Yan (2002): Information measure for performance of image fusion. In Electronics Letters 38 (7), pp. 313–315. Available online at https://digital-library.theiet.org/content/journals/10.1049/el_20020212.

Huang, Bing; Yang, Feng; Yin, Mengxiao; Mo, Xiaoying; Zhong, Cheng; Fantacci, Maria E. (2020): A Review of Multimodal Medical Image Fusion Techniques. In Computational and Mathematical Methods in Medicine 2020, p. 8279342. DOI: 10.1155/2020/8279342.

Kurban, Tuba (2021): Fusion of remotely sensed infrared and visible images using Shearlet transform and backtracking search algorithm. In International Journal of Remote Sensing 42 (13), pp. 5087–5104. DOI: 10.1080/01431161.2021.1910370.

Li, Hui; Wu, Xiao-Jun; Kittler, Josef (2018): Infrared and Visible Image Fusion using a Deep Learning Framework, pp. 2705–2710. DOI: 10.1109/ICPR.2018.8546006.

Li, Yi; Zhao, Junli; Lv, Zhihan; Li, Jinhua (2021): Medical image fusion method by deep learning. In International Journal of Cognitive Computing in Engineering 2, pp. 21–29. DOI: 10.1016/j.ijcce.2020.12.004.

Liu, Yu; Chen, Xun; Wang, Zengfu; Wang, Z. Jane; Ward, Rabab K.; Wang, Xuesong (2018): Deep learning for pixel-level image fusion: Recent advances and future prospects. In Information Fusion 42, pp. 158–173. DOI: 10.1016/j.inffus.2017.10.007.

Nashwan Jasim Hussein; Fei Hu; Feng He (2017): Multisensor of thermal and visual images to detect concealed weapon using harmony search image fusion approach. In Pattern Recognition Letters 94, pp. 219–227. DOI: 10.1016/j.patrec.2016.12.011.

Sen Qiu; Hongkai Zhao; Nan Jiang; Zhelong Wang; Long Liu; Yi An et al. (2021): Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. In Information Fusion. DOI: 10.1016/j.inffus.2021.11.006.

V. Aslantas; R. Kurban (2009): A comparison of criterion functions for fusion of multi-focus noisy images. In Optics Communications 282 (16), pp. 3231–3242. DOI: 10.1016/j.optcom.2009.05.021.

Vladimir Petrović (2007): Subjective tests for image fusion evaluation and objective metric validation. In Information Fusion 8 (2), pp. 208–216. DOI: 10.1016/j.inffus.2005.05.001.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; Simoncelli, E. P. (2004): Image quality assessment: From error visibility to structural similarity. In IEEE Transactions on Image Processing 13 (4), pp. 600–612. DOI: 10.1109/TIP.2003.819861.

Wanyi Zhang; Xiuhua Fu; Wei Li (2020): The intelligent vehicle target recognition algorithm based on target infrared features combined with lidar. In Computer Communications 155, pp. 158–165. DOI: 10.1016/j.comcom.2020.03.013.

Xiaole Ma; Zhihai Wang; Shaohai Hu (2021): Multi-focus image fusion based on multi-scale sparse representation. In Journal of Visual Communication and Image Representation 81, p. 103328. DOI: 10.1016/j.jvcir.2021.103328.

Xydeas, Costas; Petrovic, Vladimir (2000): Objective pixel-level image fusion performance measure: SPIE (4051).

Yanling Chen; Lianglun Cheng; Heng Wu; Fei Mo; Ziyang Chen (2022): Infrared and visible image fusion based on iterative differential thermal information filter. In Optics and Lasers in Engineering 148, p. 106776. DOI: 10.1016/j.optlaseng.2021.106776.

Zhang, Chengfang (2021): Convolution dictionary learning for visible-infrared image fusion via local processing. In Procedia Computer Science 183, pp. 609–615. DOI: 10.1016/j.procs.2021.02.104.

Zhiyun Xue ,Zhong-Ying Zhang ,Rick S. Blum (2005): An Overview of Image Fusion.

Zhou, Tongxue; Ruan, Su; Canu, Stéphane (2019): A review: Deep learning for medical image segmentation using multi-modality fusion. In Array 3-4, p. 100004. DOI: 10.1016/j.array.2019.100004.