DEVELOPING A TRANSFORMER-BASED APPROACH FOR FUSING INFRARED AND VISIBLE IMAGES FOR IMPROVED OBJECT DETECTION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

AYTEKIN ERDOGAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

AUGUST 2023

Developing A Transformer-Based Approach for Fusing Infrared and Visible Images for Improved Object Detection

submitted by **AYTEKIN ERDOGAN** in partial fulfillment of the requirements for the degree of **Master of Science in Information Systems Department, Middle East Technical University** by,

Prof. Dr. Director of Institute _____
Dean, **Graduate School of Informatics**

Prof. Dr. Head of Department _____
Head of Department, **Information Systems**

Assoc. Prof. Dr. Supervisor _____
Supervisor, **Department, School**

Assoc. Prof. Dr. Co-supervisor if Exists _____
Co-supervisor, **Department, School**

Examining Committee Members:

Prof. Dr. Committee Member 1
Department, School _____

Assoc. Prof. Dr. Committee Member 2
Department, School _____

Assoc. Prof. Dr. Committee Member 3
Department, School _____

Assist. Prof. Dr. Committee Member 4
Department, School _____

Assist. Prof. Dr. Committee Member 5
Department, School _____

Date: 28.08.2019

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Aytekin Erdogan

Signature :

ABSTRACT

DEVELOPING A TRANSFORMER-BASED APPROACH FOR FUSING INFRARED AND VISIBLE IMAGES FOR IMPROVED OBJECT DETECTION

Erdogan, Aytekin

M.S., Department of Information Systems

Supervisor: Assoc. Prof. Dr. Supervisor

Co-Supervisor: Assoc. Prof. Dr. Co-supervisor if Exists

August 2023, **??** pages

English abstract here

Keywords: A keyword, another keyword, some other keywords

ÖZ

TÜRKÇE BAŞLIK

Erdogan, Aytekin
Yüksek Lisans, Bilişim Sistemleri Bölümü
Tez Yöneticisi: Doç. Dr. Supervisor
Ortak Tez Yöneticisi: Doç. Dr. Co-supervisor if Exists

Ağustos 2023, **??** sayfa

Türkçe öz buraya

Anahtar Kelimeler: Bir anahtar kelime, başka bir anahtar kelime, başka anahtar kelimeler

To the memories of my beloved friends Murat Tekin and Ragip Enes Katran

ACKNOWLEDGMENTS

Acknowledgments here

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

IF	Image Fusion
VIF	Visual and Infrared Image Fusion
AI	Artificial Intelligence
CNN	Convolutional Neural Networks
GAN	General Adveserial Networks
SR	Sparse Representation
MST	Multi-Scale Transformation
LRR	Low-Rank Representation
NAS	Neaural Architecture Search

CHAPTER 1

INTRODUCTION

Image Fusion is a computer vision task that has been taken place for many years. Gathering all the complementary usefull informations into single image is called image fusion, *a.k.a* IF. Visual and Infrared Image Fusion, *henceforth will be referred to as VIF*, is a subfield of iamge fusion. Since the first study [**?**] in 1989, VIF is actively studied. In the era of AI, new methods such as CNN, GAN, auto-encoder, transformers are also applied to the VIF problem.

1.1 Research Questions

1.2 Contributions of the Study

1.3 Organization of the Thesis

CHAPTER 2

RELATED WORK

Image fusion algorithms can be categorized based on several factors, including whether they employ learning methods or hand-crafted steps, based on predefined loss functions, and whether labeled datasets are involved in the process. Learning-based methods involve the use of machine learning techniques, such as CNN, GAN, Transformers, Auto-encoders, to learn the features and relationships between input images. In contrast, hand-crafted approaches involve the manual selection and design of specific features and fusion rules.Whether a methods is end-to-end is defined as do they require one or more handcrafted steps. Loss functions are commonly used in learning-based methods to measure the quality of the fused image and to guide the training process. Based on the loss function, the process can be classified as self-supervised, supervised or unsupervised. Finally, labeled datasets can be used to train learning-based methods, where the some form of label is available and ground-truth labels are known, or for evaluating the performance of fusion algorithms. By considering these factors, researchers can select the most appropriate image fusion algorithm for their specific application requirements.

The available VIF methods can be categorized into two groups: traditional methods, which were widely used before the advent of AI, and learning-based methods. Regardless of the classification, all methods consist of three main components: image feature extraction, fusion of features from multiple images, and reconstruction of the image from the fused features. During the feature extraction stage, features are extracted from multiple images. In the fusion stage, the extracted features are compared, and complementary features are incorporated into a single feature map or set. Finally, in the reconstruction stage, the image is reconstructed from the fused set of features. All related studies aim to improve one or more of these stages in the process.

Learning-based fusion methods are often used to combine information from multiple sources to obtain more accurate and informative results. These methods can be categorized based on whether they require ground truth labels or not. The ground truth annotations don't have to be a form of fused image. If ground truth labels are needed, they are known as supervised methods. On the other hand, if no ground truth labels are required, they are referred to as unsupervised or self-supervised methods. These fusion techniques are commonly employed in convolutional-based methods, which are widely used in image processing and computer vision applications. Regardless of the specific method used, the goal of learning-based fusion methods is to improve the quality and accuracy of the resulting output by leveraging information from multiple sources.

Learning-based image fusion algorithms are commonly categorized based on the type of network used in the algorithm. There are several types of networks that are frequently used, including CNN-based,

auto-encoder based, GAN-based, transformer-based networks and others which include hybrid or handcrafted steps. Convolutional neural network (CNN)-based methods are often used due to their ability to extract features from input images and produce high-quality results. Auto-encoder-based methods are also popular, as they can effectively compress and decompress image information to obtain a fused image. Generative adversarial network (GAN)-based methods use a combination of generator and discriminator networks to produce high-quality fused images. Transformer-based methods have recently gained popularity due to their ability to process long sequences of input data efficiently. Other methods, such as sparse representation-based methods and wavelet-based methods, are also used in learning-based image fusion. Ultimately, the choice of network depends on the specific requirements of the task at hand and the available resources.

In addition to categorizing fusion algorithms based on the type of network used, another way to differentiate them is whether they are end-to-end. An end-to-end algorithm is one that can take raw input data and produce the desired output directly, without any intermediate hand-crafted steps. In the context of image fusion, end-to-end algorithms are those that can take multiple input images and produce a fused image without the need for manual feature extraction or other preprocessing steps. These algorithms are often preferred, as they can be more efficient and less prone to errors compared to non-end-to-end methods that require manual intervention. In contrast, non-end-to-end algorithms may require additional processing steps, such as registration and feature extraction, to produce the final fused image. While these steps can provide additional control over the fusion process, they can also introduce additional complexity and reduce the overall efficiency of the algorithm.

2.1 Traditional Algorithms

The traditional image fusion algorithms have been extensively studied in the literature. However, they are not without shortcomings. One of the major issues with these methods is the presence of handcrafted steps, which can lead to suboptimal results. In addition, the time complexity of some methods can be quite high, making them impractical for real-world applications.

Sparse representation (SR)[**?**,] based methods are a popular choice for image fusion. However, they suffer from several limitations. For example, methods such as [**?**] and [**?**] require dictionary learning, which can significantly increase the time complexity of the algorithm. Furthermore, these methods include handcrafted steps, which can limit their generalizability.

Another commonly used approach for image fusion is multi-scale transformation (MST) based methods. These methods, such as [**?**] and [**?**], can be quite effective at capturing various characteristics of images at different scales. However, they too suffer from limitations. One major issue is their lack of generalizability, which can make them less effective in certain scenarios.

Low-rank representation (LRR) based methods are another popular choice for image fusion. These methods, such as [**?**]liu2012robust, are particularly effective at dealing with noise and other forms of image degradation. However, like the other methods, they too have limitations. For example, they may not be suitable for all types of images, particularly those with complex textures or patterns.

In summary, the success of traditional image fusion algorithms heavily relies on the quality of the feature extraction method used. While there are many different methods available, each with its own

strengths and weaknesses, it is important to carefully consider the specific requirements of the problem at hand before selecting a particular method.

2.2 CNN Based Deep Learning Algorithms

First image fusion method that utilizes CNN is created by Liu et al [**?**]. The paper proposed a method to fuse infrared and visible images using Convolutional Neural Networks (CNNs). The approach consisted of four main steps: preprocessing, feature extraction, fusion strategy, and reconstruction. First, the input images were resized and normalized. Then, a CNN model was trained to extract features from both the visible and infrared images. The extracted features were then fused using a weighted sum fusion strategy, and the weights were learned during training. Finally, the fused feature maps were transformed back into the image domain using a deconvolutional network. The proposed method was evaluated on various datasets and outperformed other state-of-the-art fusion methods in terms of objective quality metrics and subjective visual quality. The proposed method in the paper also utilizes a multiscale approach to fuse the source images. The images are decomposed into different scales using Laplacian and Gaussian pyramids. This allows the model to better preserve details in both the visible and infrared images during the fusion process. To train the model, high-quality images and their blurred versions are used. These blurred versions are generated by applying multiscale Gaussian filtering and random sampling, which helps the model to learn features that are robust to image blurring and noise. It's also important to note that the proposed method is a pioneering deep learning-based approach to VIF, which is a metric used to evaluate the visual quality of the fused images. By introducing CNNs to VIF, the proposed method achieved better fusion results than other state-of-the-art methods available at that time.

CNN-based methods can be categorized into two groups: supervised and unsupervised methods. While both methods use CNNs to extract features from the input images, supervised methods require labeled data during the training process, whereas unsupervised methods do not. The majority of CNN-based image fusion methods are unsupervised, as labeled data is often difficult or expensive to obtain. However, to improve the performance of supervised methods, extra steps, such as data augmentation or transfer learning, are often employed. Despite the success of both supervised and unsupervised CNN-based methods, there is still room for improvement in image fusion techniques, particularly in scenarios where the input images have significant differences in terms of illumination, resolution, or noise.

2.2.1 Supervised CNNs

Image fusion is a challenging task, as obtaining ground truth data is often difficult or impossible. However, there are several workarounds that have been proposed to generate pseudo annotations and ground truth data. One approach is to use the results of another fusion method as ground truth data. This can be done by comparing the output of the method being evaluated to the output of a known or well-established fusion method. Another approach is to generate blurred versions of the high-quality input images and use these blurred images as labels. For example, in the study by Liu et al. [**?**], blurred versions of the high-quality input images were used to train a deep learning-based fusion model. Similarly, in the study by An et al. [**?**], the results of other fusion methods were used as annotations to train a convolutional neural network for image fusion. While these approaches are not

a substitute for true ground truth data, they provide a means to evaluate and improve image fusion methods, especially in scenarios where obtaining actual ground truth is challenging or impossible.

It is also wise to state that many CNN-based image fusion methods rely on transfer learning, where pre-trained models, such as ResNet50 [**?**, **?**], VGG19[**?**, **?**], VGG16 [**?**], and DenseNet-201 [**?**], are used to extract features from source images. The extracted features are then processed before being fused using manually designed rules, such as weighted averaging or combining the fused base and detail parts. Some studies generate weight maps based on the extracted features [**?**, **?**, **?**], while others optimize a loss function computed based on the extracted features using the L-BFGS method [**?**]. These transfer learning-based methods provide a means to leverage pre-existing knowledge in large-scale datasets and improve the performance of image fusion methods.

2.2.2 Unsupervised CNNs

In the field of infrared visual image fusion, most related studies are unsupervised. Due to the lack of a ground truth for this task, loss functions are typically defined as a function of source images and related evaluation metrics. This means that the performance of the fusion method is evaluated based on how well the fused image aligns with the intended objective, rather than how closely it matches a predefined target. While this approach can be challenging due to the subjective nature of image quality evaluation, it remains a popular method for evaluating unsupervised image fusion techniques.

Several studies in the field of image fusion have explored the use of unsupervised convolutional neural networks (CNNs) to improve the process. These CNNs can be applied to either a single part of the process or the entire process. For instance, Liu et al. [**?**] decompose source images into a base and a detail parts using CNNs, while Hou et al. [**?**] uses CNNs in the feature extraction and reconstruction parts of the process. Some studies, such as Xu et al. [**?**] and Mustafa et al. [**?**], use unsupervised CNNs for the entire image fusion process. By utilizing unsupervised CNNs, these studies aim to improve the overall performance of the image fusion process, particularly for tasks such as infrared visual image fusion.

In addition to unsupervised CNNs, there are other methods that can be used to improve the performance of infrared and visual image fusion. Residual connections [**?**] and dense connections [**?**] are examples of techniques that have been used to enhance feature propagation within neural networks. Attention mechanisms [**?**] have also been applied to focus on important features and exclude irrelevant ones, while multi-scale and multilevel features have been used to capture details across different spatial and frequency ranges. Contrastive learning [**?**] and neural architecture search [**?**] are other approaches that have been used to improve the performance of image fusion methods. Image and feature decomposition techniques can also be used to decompose the source images into different components and extract features from them. By combining these various methods, researchers can develop more sophisticated image fusion systems that produce higher-quality results.

2.3 Autoencoder Based Deep Learning Algorithms

Autoencoder is a type of neural network that is widely utilized for unsupervised learning tasks, including but not limited to dimensionality reduction, data compression, and anomaly detection. The

fundamental principle of autoencoder is to compress the input data into a lower-dimensional representation, which is commonly referred to as the latent space. The compressed representation is then utilized to reconstruct the original input data. The objective is to minimize the difference between the input data and the reconstructed output, which requires the autoencoder to learn a compressed representation that captures the most essential features of the input data. The concept of autoencoder was first introduced by Hinton et al.[**?**].

In the context of infrared visual image fusion, the autoencoder technique is leveraged to extract features from source images using the encoder stage, while the decoder stage reconstructs the fused image. The training process typically involves two stages: firstly, the autoencoder is trained using source images, either infrared, visual or both, without any fusion. Subsequently, the fusion step is integrated and the entire model is trained. It is also common practice to employ large datasets for the first training stage. It is worth noting that the autoencoder approach, which is used in this method, differs from transfer learning methods discussed in Section **??** in that an autoencoder is trained from scratch, while pre-trained models are utilized in transfer learning with minimal fine tuning. One of the pioneering works in infrared and visual image fusion using autoencoder is DenseFuse [**?**]. To pre-train the network without the fusion step, the well-known MS-COCO dataset [**?**] is utilized in the first stage training. Other noteworthy studies include Raza et al. [**?**], Fu et al. [**?**], Jian et al. [**?**], Wang et al. [**?**], and finally, Zhao et al. [**?**].

There are still open research questions for this part since rgb and infrared images are different in structure and only rgb image is utilez in autoencoder's training. There are studies that focuses on this difference. IVFENet [**?**], an encoder and two decoders are used in the pretraining stage to alleviate the vital information loss. Liu et al. [**?**] use two encoders and one unified decoder, which utilizes a NAS technique for the model. Measures for performance increase in Section **??** are also used with autoencoder based models such as dense connections [**?**, **?**, **?**], attention mechanism [**?**], multiscale features [**?**, **?**], multi-level features [**?**].

CHAPTER 3

USER EXPERIMENT

In this chapter, the details of the user experiment are presented.

3.1 Research Method and Experiment Design

CHAPTER 4

USER EXPERIMENT

In this chapter, the details of the user experiment are presented.

4.1 Research Method and Experiment Design

CHAPTER 5

CONCLUSION AND FUTURE WORK

REFERENCES

[1] A. Toet, L. J. Van Ruyven, and J. M. Valeton, "Merging thermal and visual images by a contrast pyramid," *Optical engineering*, vol. 28, no. 7, pp. 789–792, 1989.

[2] C. Liu, Y. Qi, and W. Ding, "Infrared and visible image fusion method based on saliency detection in sparse domain," *Infrared Physics & Technology*, vol. 83, pp. 94–102, 2017.

[3] Y. Bin, Y. Chao, and H. Guoyu, "Efficient image fusion with approximate sparse representation," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 14, no. 04, p. 1650024, 2016.

[4] Q. Zhang, Y. Fu, H. Li, and J. Zou, "Dictionary learning method for joint sparse representation-based image fusion," *Optical Engineering*, vol. 52, no. 5, pp. 057006–057006, 2013.

[5] H.-M. Hu, J. Wu, B. Li, Q. Guo, and J. Zheng, "An adaptive fusion algorithm for visible and infrared videos based on entropy and the cumulative distribution of gray levels," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2706–2719, 2017.

[6] K. He, D. Zhou, X. Zhang, R. Nie, Q. Wang, and X. Jin, "Infrared and visible image fusion based on target extraction in the nonsubsampled contourlet transform domain," *Journal of Applied Remote Sensing*, vol. 11, no. 1, pp. 015011–015011, 2017.

[7] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, "Infrared and visible image fusion with convolutional neural networks," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 16, no. 03, p. 1850018, 2018.

[8] W.-B. An and H.-M. Wang, "Infrared and visible image fusion with supervised convolutional neural network," *Optik*, vol. 219, p. 165120, 2020.

[9] H. Li, X.-j. Wu, and T. S. Durrani, "Infrared and visible image fusion with resnet and zero-phase component analysis," *Infrared Physics & Technology*, vol. 102, p. 103039, 2019.

[10] G. Li, Y. Lin, and X. Qu, "An infrared and visible image fusion method based on multi-scale transformation and norm optimization," *Information Fusion*, vol. 71, pp. 109–129, 2021.

[11] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *2018 24th international conference on pattern recognition (ICPR)*, pp. 2705–2710, IEEE, 2018.

[12] X. Ren, F. Meng, T. Hu, Z. Liu, and C. Wang, "Infrared-visible image fusion based on convolutional neural networks (cnn)," in *Intelligence Science and Big Data Engineering: 8th International Conference, IScIDE 2018, Lanzhou, China, August 18–19, 2018, Revised Selected Papers 8*, pp. 301–307, Springer, 2018.

[13] Y. Yang, J.-X. Liu, S.-Y. Huang, H.-Y. Lu, and W.-Y. Wen, "Vmdm-fusion: a saliency feature representation method for infrared and visible image fusion," *Signal, Image and Video Processing*, pp. 1–9, 2021.

[14] Y. Li, J. Wang, Z. Miao, and J. Wang, "Unsupervised densely attention network for infrared and visible image fusion," *Multimedia Tools and Applications*, vol. 79, no. 45-46, pp. 34685–34696, 2020.

[15] Y. Liu, L. Dong, Y. Ji, and W. Xu, "Infrared and visible image fusion through details preservation," *Sensors*, vol. 19, no. 20, p. 4556, 2019.

[16] R. Hou, D. Zhou, R. Nie, D. Liu, L. Xiong, Y. Guo, and C. Yu, "Vif-net: an unsupervised framework for infrared and visible image fusion," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 640–651, 2020.

[17] H. Xu, M. Gong, X. Tian, J. Huang, and J. Ma, "Cufd: An encoder–decoder network for visible and infrared image fusion based on common and unique feature decomposition," *Computer Vision and Image Understanding*, vol. 218, p. 103407, 2022.

[18] H. T. Mustafa, J. Yang, H. Mustafa, and M. Zareapoor, "Infrared and visible image fusion based on dilated residual attention network," *Optik*, vol. 224, p. 165409, 2020.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[21] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Advances in neural information processing systems*, pp. 2204–2212, 2014.

[22] G. E. Hinton and R. R. Salakhutdinov, "Dimensionality reduction by learning an invariant mapping," *arXiv preprint arXiv:0704.2550*, 2006.

[23] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," in *International Conference on Learning Representations*, 2017.

[24] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[25] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755, Springer, 2014.

[26] A. Raza, H. Huo, and T. Fang, "Pfaf-net: Pyramid feature network for multimodal fusion," *IEEE Sensors Letters*, vol. 4, no. 12, pp. 1–4, 2020.

[27] Y. Fu and X.-J. Wu, "A dual-branch network for infrared and visible image fusion," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 10675–10680, IEEE, 2021.

[28] L. Jian, X. Yang, Z. Liu, G. Jeon, M. Gao, and D. Chisholm, "Sedrfuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–15, 2020.

[29] Z. Wang, Y. Wu, J. Wang, J. Xu, and W. Shao, "Res2fusion: Infrared and visible image fusion based on dense res2net and double nonlocal attention models," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[30] Z. Zhao, S. Xu, J. Zhang, C. Liang, C. Zhang, and J. Liu, "Efficient and model-based infrared and visible image fusion via algorithm unrolling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1186–1196, 2021.

[31] F. Zhao, W. Zhao, L. Yao, and Y. Liu, "Self-supervised feature adaption for infrared and visible image fusion," *Information Fusion*, vol. 76, pp. 189–203, 2021.

[32] J. Liu, Y. Wu, Z. Huang, R. Liu, and X. Fan, "Smoa: Searching a modality-oriented architecture for infrared and visible image fusion," *IEEE Signal Processing Letters*, vol. 28, pp. 1818–1822, 2021.

[33] Y. Pan, D. Pi, I. A. Khan, Z. U. Khan, J. Chen, and H. Meng, "Densenetfuse: A study of deep unsupervised densenet to infrared and visual image fusion," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–13, 2021.

[34] Z. Wang, J. Wang, Y. Wu, J. Xu, and X. Zhang, "Unfusion: A unified multi-scale densely connected network for infrared and visible image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3360–3374, 2021.

[35] Z. Li, H. Wu, L. Cheng, S. Luo, and M. Chen, "Infrared and visible fusion imaging via double-layer fusion denoising neural network," *Digital Signal Processing*, vol. 123, p. 103433, 2022.

[36] Y. Peng, G. Liu, X. Xu, D. P. Bavirisetti, X. Gu, and X. Zhang, "Mfdetection: A highly generalized object detection network unified with multilevel heterogeneous image fusion," *Optik*, vol. 266, p. 169599, 2022.

APPENDIX A

TABLES FOR RELATED WORK CHAPTER

A.1 Summary of the Studies

APPENDIX B

EXTRA MATERIAL

APPENDIX C

INSTRUMENTS AND ETHICAL CLEARANCE