Gabor Feature Network for Transformer-based Building Change Detection Model in Remote Sensing

Priscilla Indira Osa, Josiane Zerubia, Zoltan Kato

GABOR FEATURE NETWORK FOR TRANSFORMER-BASED BUILDING CHANGE DETECTION MODEL IN REMOTE SENSING

Priscilla Indira Osa[*1,2] *Josiane Zerubia*[2] *Zoltan Kato*[3,4]

[1] University of Genoa, DITEN dept., Italy, [2] Inria, Université Côte d'Azur, France
[3] University of Szeged, Institute of Informatics, Hungary, [4] J. Selye University, Komarno, Slovakia

ABSTRACT

Detecting building change in bitemporal remote sensing (RS) imagery requires a model to highlight the changes in buildings and ignore the irrelevant changes of other objects and sensing conditions. Buildings have comparatively less diverse textures than other objects and appear as repetitive visual patterns on RS images. In this paper, we propose Gabor Feature Network (GFN) to extract the distinctive repetitive texture features of buildings. Furthermore, we also design Feature Fusion Module (FFM) to fuse the extracted multiscale features from GFN with the features from a Transformer-based encoder to pass on the texture features to different parts of the model. Using GFN and FFM, we design a Transformer-based model, called GabFormer for building change detection. Experimental results on the LEVIR-CD and WHU-CD datasets indicate that GabFormer outperforms other SOTA models and in particular show significant improvement in the generalization capability. Our code is available on https://github.com/Ayana-Inria/GabFormer.

Index Terms— Transformer, Gabor feature, building change detection, remote sensing, image analysis

1. INTRODUCTION

Building Change Detection (BCD) in remote sensing (RS) is a process to identify changes of buildings on two or more images of a specific geographical location taken at different times [1]. The typical BCD task aims to create a change map that highlights appearance and disappearance of buildings i.e., newly-built and destroyed/dismantled buildings. BCD has been the core procedure behind a broad range of applications, such as urban growth analysis [2], and disaster assessment and recovery [3].

The rapid evolution of computer algorithms such as machine-learning-based methods has facilitated automatic BCD methods, replacing the manual tasks that are time-consuming and need more expensive labor costs. Moreover, the availability of open BCD datasets [4, 5] pushes the BCD model's development to the domain of deep learning (DL). In order to design a robust BCD model, it is necessary for a model not only to differentiate between changed and unchanged pixels on the image, but also to distinguish the changes in the object of interest (buildings) from the changes of other objects (e.g., vegetation and roads), as well as to ignore the insignificant changes on buildings (e.g., change of buildings' color).

Deep Convolutional Neural Networks (CNN) have demonstrated promising performance in addressing the aforemen-

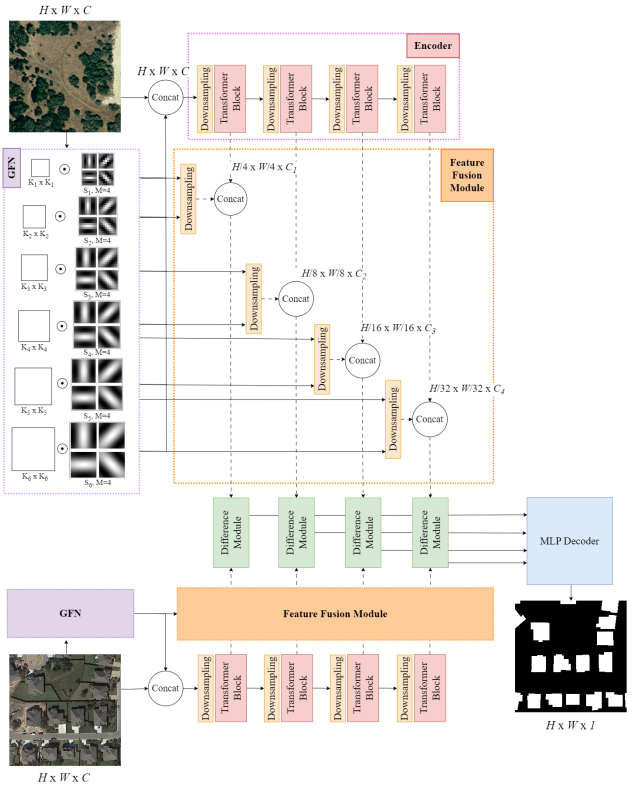

Fig. 1. The overall architecture of GabFormer

*The first author performed the work while at Inria, Université Côte d'Azur, France. University of Genoa and Université Côte d'Azur (UCA) are part of the Ulysseus Alliance (European University). https://ulysseus.eu/.
The authors acknowledge the internship and conference attendance funding support by Inria, France (BMI-NF), the grant of University of Szeged (NKFIH K135728), and the scholarship by French government.

tioned complexities of the BCD task [4, 6, 7, 8]. CNN-based methods utilize several strategies to tackle the difficulties of the BCD problems, such as using metric-based learning technique [4, 7], as well as integrating attention mechanisms [4, 8, 9]. Metric-based learning transforms the features to the embedding space and trains the network to minimize the distance of no change pixels and maximize the distance for change pixels [7]. Attention-based approaches put weights on the important features in the dimension intended to be highlighted e.g., temporal attention relates the features of the bitemporal images that accentuate the change [4]. Recently, Transformer which has the self-attention as its building block, has been applied to the BCD task because of its efficiency in capturing the global context of the features. Transformer is incorporated in combination with the CNN [10], or is utilized fully without the feature extraction by the CNN [11]. However, none of the above-mentioned methods explicitly perform feature extractions that are characteristic to the particular texture properties of building inspite of its importance in differentiating buildings from other objects in the BCD task. Unlike textures of other land covers typically found on RS images, building's textures do not have much variations and buildings in bird's eye view can be characterized visibly by its repeating visual patterns. Recent publication takes into account the pattern of the object to determine the shape of neighboring area based on the geometry of the object [12]. In our case that focuses on building, we believe that the building's pattern can be extracted easily without changing the geometrical variation. Theoretically, both CNN and Transformer can learn texture features from the training image data [13, 14]. However, learning such features usually need a large amount of data, and open BCD datasets generally contain relatively limited quantity of images (thousands to tens of thousands) compared to popular large datasets used to train CNN and Transformer networks such as ImageNet [15] (around 14 million images), and JFT-3B [16] (around 3 billion images). Thus, it is questionable that the networks can learn the texture features effectively with a limited number of images, especially Transformer which fails to learn some specific features if not being trained with sufficient amount of data [14].

Based on the above observation, we propose Gabor Feature Network (GFN) to extract relevant features for a Transformer network. GFN is based on Gabor filter [17], a well known image processing filter to extract repeating texture information of an image. Herein, we propose GFN to maximize the capability of the network to capture textures belonging to buildings. We believe that the texture of buildings is discriminative enough to highlight buildings from other objects present on the RS images. Furthermore, by emphasizing on texture features, we expect to reduce the noises caused by the insignificant changes e.g., those by color changes or weather condition variation. The GFN is constructed by Gabor filters modulated via CNN filters [13], which makes it more robust than merely using Gabor filters alone as the network will learn the weight of the convolution kernels combined with the Gabor filters. In addition to the GFN, we also introduce Feature Fusion Module (FFM) to merge the multiscale Gabor feature maps from the GFN with the features extracted by the encoder of the Transformer network to ensure the information of building's textures is being preserved at matching scales in the deep intermediate layers of the network.

2. METHODOLOGY

2.1. GabFormer

Our model consists in a novel feature extraction and fusion technique which makes use of multiscale Gabor filters. These modules are then used in a Transformer-based bitemporal change detection model, which is built based on ChangeFormer [11]. Fig. 1 shows the proposed architecture of GabFormer : Feature extraction and fusion are based on a new concept by adding Gabor Feature Network (GFN) and introducing Feature Fusion Module (FFM) which are integrated into a Siamese-style network consisting of 4 multi-level pairs of Downsampling-Transformer Block in the hierarchical encoder part, a Difference Module to calculate the difference of multi-scale features coming from pre- and post- change images, and an MLP Decoder to do upsampling operations and produce the final change map.

Pre-change and post-change input images with size $C \times H \times W$ (C, H, W refer to channel, height, width respectively) are fed into the GFN which will extract relevant texture features at multiple scales and orientations. These new features along with the original RGB image are the input for the Downsampling-Transformer Blocks. These encoder blocks provide outputs at multiple resolution, which are fused with corresponding GFN feature maps by the FFM. Essentially, these combined features are the input features of the Difference Module responsible to produce the difference maps of pre- and post- change images which are passed to the MLP Decoder where features are forwarded to the MLP and upsampling operations before being classified to produce the final change map.

2.2. Gabor Feature Network

2D Gabor wavelets [18] are the generalization of the 1D function originally proposed by Gabor [17]. With a capacity to imitate the receptive fields of the mammalian visual cortex [19], Gabor filters are widely used for image analysis and texture feature extraction, especially to capture repetitive visual patterns in an image. We believe that this capability is suitable for our building change detection task as we need to sense the repetitive texture of the clusters of buildings in an area.

Gabor wavelets (filters or kernels) are defined as follows

[13, 19, 20]:

$$G(u,v) = \frac{||\vec{k_{u,v}}||^2}{\sigma^2} e^{-(||\vec{k_{u,v}}||^2||\vec{z}||^2/2\sigma^2)}[e^{i\vec{k_{u,v}}\vec{z}} - e^{-\sigma^2/2}], \tag{1}$$

where $\vec{z} = (x, y)$, $\vec{k_{u,v}} = \begin{pmatrix} k_{jx} \\ k_{jy} \end{pmatrix} = \begin{pmatrix} k_v \cos k_u \\ k_v \sin k_u \end{pmatrix}$, frequency $k_v = (\pi/2)/\sqrt{2}^{(v-1)}$, and orientation $k_u = u\frac{\pi}{U}$ with the scale parameter $v = 1, ..., V$ which determines the frequency of the filter in inverse proportion, the parameter $u = 0, ..., U-1$ that controls the orientation of the filter, and $\sigma = 2\pi$. As indicated in Eq. 1, by giving a set of orientation and scale as the parameters, one Gabor kernel will filter visual repetitive patterns in an image according to those particular orientation and frequency.

The building block of our GFN is based on the Gabor orientation Filter (GoF) originally proposed in [13]. A GoF [13] consists of a group of Gabor filters at a scale v in a set of orientations U

$$\hat{C}_i^v = (C_{i,0}^v, ..., C_{i,U-1}^v), \tag{2}$$

modulated by learnable convolutional filters C_i

$$C_{i,u}^v = C_i \odot G(u, v), \tag{3}$$

where $N \times K \times K$ is the size of the learned filters, $K \times K$ is the spatial size of the Gabor filter, N is the number of channels, and $i = 1, ..., N$. $\odot$ indicates the element-wise product operation. $G(u, v)$ is a Gabor filter with size $K \times K$, orientation u, and scale v. In GFN, we only utilize the real parts of the Gabor filters. $\hat{C}_i^v$ is a GoF with scale v, and a set of orientations U. We construct the GFN with a number of GoFs that contain Gabor filters with different parameters to capture several possible orientations and frequencies.

The proposed GFN produces a set of Gabor feature maps in several frequencies and orientations by convolving the input image with several GoFs. Let I be the input image, and $\hat{C}_i^v$ be a GoF of a scale v and orientations U, the output Gabor features in scale v is obtained by convolution:

$$F_{u,v} = I * \hat{C}_i^v, \tag{4}$$

where $F_{u,v}$ denotes features maps with size $U \times H \times W$. Hence, a collection of Gabor feature maps at different scales V can be defined as:

$$\hat{F} = (F_{u,1}, ..., F_{u,V}) \tag{5}$$

In our GFN, we have to determine the set of orientations and scales such that repetitive textures produced by various buildings in the input image are well captured. Note that we do not need a very high resolution in these parameters, what is critical, however, is to have a well-defined filter-response from our GFN. For that purpose, we simply set $U = 4$ to cover horizontal, vertical, and diagonal orientations. As for the scale, $V = 6$ was set based on our preliminary qualitative evaluation on the building features extracted by Gabor filters. Fig. 2 shows 2 images containing different sizes of buildings which are convolved with different frequencies of Gabor filters. We can observe that $v = 1$ is enough to capture comparatively small buildings shown in the upper row. As we increase the scale (lower the frequency), the filter is responding to bigger objects (the textures of the big buildings on the image in the second row are observable but the small buildings in the first row are getting blurred out). We stop at $v = 6$ where the big buildings begin to be blurry but still have clear texture features. Clearly $v > 6$ will be too large to extract any meaningful textures of the buildings. Of course, the learned filter size must match with the corresponding Gabor filter kernel size K, which is adjusted according to the frequency of the Gabor filter: as the scale increases, the filter needs bigger kernel size to capture one cycle of the filter impulse. Thus, the pairs of scale v and kernel size K utilized in our network are $(v, K) = \{(1, 7), (2, 9), (3, 11), (4, 15), (5, 19), (6, 23)\}$, as shown inside the GFN in Fig. 1.

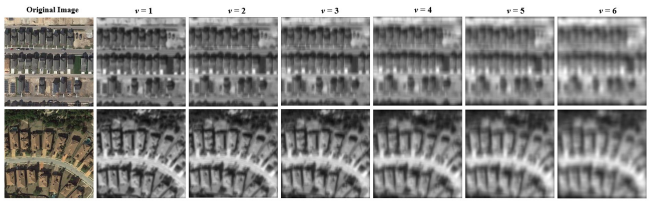

Fig. 2. A sample of Gabor filter outputs with several scales. Note that these are the output of Gabor filters only, not GoFs.

2.3. Feature Fusion Module

Feature Fusion Module aims to combine downsampled Gabor feature maps extracted by GFN and features extracted in the Transformer blocks such that the Difference Module receives the fusion of features at corresponding scales from both parts of the network. Transformer feature maps F_i have a resolution of $C_i \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$, where $i = \{1, 2, 3, 4\}$, and $C_{i+1} > C_i$. We choose a pair of Gabor feature maps to be concatenated with each of these encoder feature maps according to the resolution of the Transformer feature maps (see Fig. 1), i.e. Gabor features from lower frequency (i.e., larger v) in GoFs which capture bigger patterns in the image, are concatenated with the smaller resolution of the Transformer features so the meaningful information after the downsampling operation is kept. For example, referring to Fig. 2, Gabor filter with $v = 6$ only captures big buildings while blurring out the small ones which makes the features keeping the relevant information even if it is downsampled to the smallest spatial resolution $H/32 \times W/32$ of Transformer features, while Gabor features with higher resolution information such as the one from $v = 1$ will only be resized to $H/4 \times W/4$

	LEVIR-CD				WHU-CD*			
Model	Precision	Recall	F1-score	IoU	Precision	Recall	F1-score	IoU
BIT [10]	89.24%	89.37%	89.31%	80.68%	87.65%	90.91%	89.25%	80.59%
ChangeFormer [11]	92.05%	88.80%	90.40%	82.48%	**94.15%**	85.52%	89.63%	81.20%
STANet-PAM [4]	83.81%	**91.00%**	87.26%	77.40%	70.65%	**93.54%**	80.50%	67.37%
GabFormer	**92.87%**	88.54%	**90.66%**	**82.91%**	94.12%	89.45%	**91.73%**	**84.72%**

Table 1. The comparison of quantitative results between GabFormer and State-of-The-Art models on the LEVIR-CD dataset and WHU-CD dataset. The best result is highlighted in bold. * denotes the reimplemented training results.

to preserve the detailed texture information. Downsampling of the i-th Gabor feature map is achieved by 2D MaxPooling operation with a kernel size K_i and stride S_i. The kernel size and stride are set to be the same in each block $K_i = S_i = 2^{i+1}$. We choose the 2D MaxPooling over other more sophisticated options such as the 2D Convolution, because the feature maps do not contain higher resolution information (due to the Gabor filter properties) thus a simple MaxPooling is capable enough to downsample the features. Let F_i be the i-th Transformer feature map, $F_{u,v}$ be a Gabor feature map where $v = \{1, 2, 3, 4, 5, 6\}$, and $Down_i$ be the downsampling operation for the i-th block features. A pair of Gabor features to be combined to the i-th Transformer block features is as follows:

$$F_{u,i} = Concat(F_{u,j}, F_{u,k}), \tag{6}$$

where $j = \{1, 3, 4, 5\}$, and $k = \{2, 4, 5, 6\}$. The output fused features are then defined as

$$\tilde{F_i} = Concat(Down_i(F_{u,i}), F_i) \tag{7}$$

Hence, the size of the output fused feature maps that is passed to the Difference Module are $(C_i + 2U) \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$, where $i = \{1, 2, 3, 4\}$.

3. EXPERIMENTS

3.1. Datasets and Implementation Details

Two building change detection datasets were used to conduct the experiments. **LEVIR-CD** [4] is a building change detection dataset consisting of 637 pairs of very high-resolution (VHR) Google Earth RGB images with a spatial resolution of 0.5 m and a size of 1024×1024 pixels. This dataset highlights the bitemporal change of building development and building decline in 20 different areas in Texas, USA with time span ranging from 5 to 14 years. The images and labels were cropped to 256×256 patches without overlap. The default split of training/validation/test was used, hence the total pairs of images are 7120/1024/2048. **WHU-CD** [5] contains a pair of 0.2 m RGB aerial imagery split into two pairs of images with size of 21243×15354 pixels and 11256×15354 pixels for train and test respectively. This open dataset covers an area in Christchurch, New Zealand and focuses on building changes between 2012 and 2016. The dataset comes with binary labels of change and no change. Cropped images with a size of 256×256 and a random split of train/val/test = 6096/762/762, were utilized in the experiment.

The model was implemented in PyTorch. Our experiments were run on two different GPUs: NVIDIA Quadro GV100, and NVIDIA Quadro RTX 8000. During the training phase, data augmentation such as random flip, random scaling $(0.8 - 1.2)$, random crop, random color jittering, and Gaussian blur was applied. Model's weights were initialized randomly, and models were trained using Cross-Entropy Loss and AdamW optimizer with weight decay of 0.01 and beta values equal to $(0.9, 0.999)$. Initial learning rate was set to 0.0001 which linearly decays to 0. We utilized a batch size of 8 and trained the model for 300 epochs.

We utilize Precision, Recall, F1-score, and Intersection over Union (IoU) as the metrics to evaluate the performance of our model.

3.2. Comparison with SOTA models

We compare the performance of GabFormer with State of the Art (SOTA) methods: **BIT** [10] was chosen as the representative of networks that utilize both CNN and Transformer in its architecture; **ChangeFormer** [11] introduces a pure Transformer-based encoder combined with an MLP decoder without any use of CNNs in the network; while **STANet-PAM** [4] is a CD model that was proposed together with the LEVIR-CD dataset and we chose this model to represent fully CNN network architecture.

Table 1 reports the comparative results of GabFormer with SOTA models on the LEVIR-CD and WHU-CD datasets. As what can be seen in this table, GabFormer outperforms the other methods in terms of both F1-score and IoU. For instance, GabFormer exceeds the pure CNN-based model, STANet-PAM by 3.40% and 5.51%, and improves the Transformer-based ChangeFormer by 0.26% and 0.43% on the LEVIR-CD dataset. We can observe a more significant difference on the WHU-CD dataset where there are 2.10% and 3.53% increases in F1-score and IoU from the second-best-performing model (ChangeFormer). This improvement can be seen visually for example in Fig. 3 patch (a) of both datasets. It can be observed that GabFormer predicts less FP and FN in these sample images. The visualization also indi-

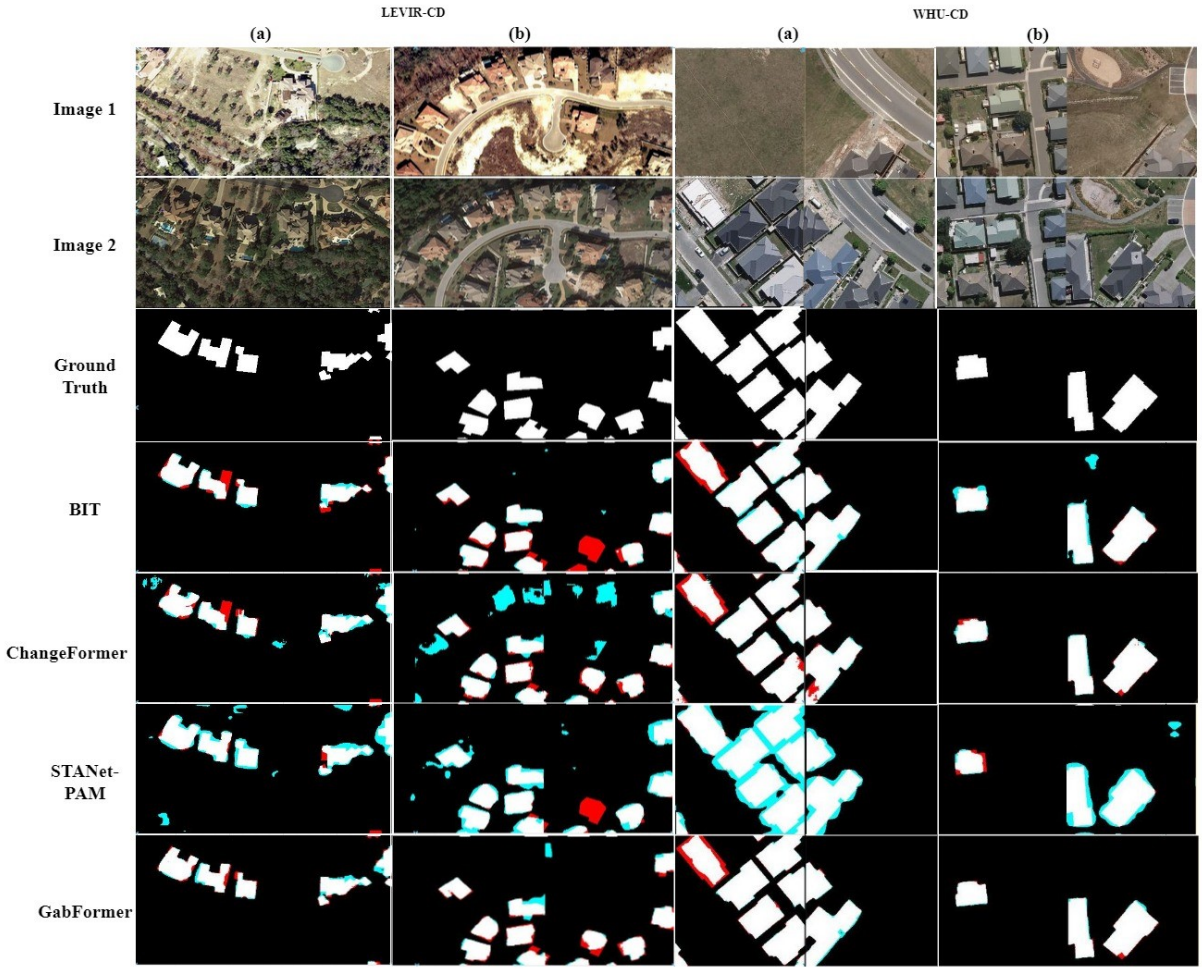

Fig. 3. The visualization comparison of all models. Color representation: TP (white), FP (light blue), TN (black), FN (red).

	Train on LEVIR-CD, test on WHU-CD				Train on WHU-CD, test on LEVIR-CD			
Model	Precision	Recall	F1-score	IoU	Precision	Recall	F1-score	IoU
BIT [10]	58.36%	**79.52%**	67.32%	50.74%	43.79%	8.39%	14.08%	7.58%
ChangeFormer [11]	76.87%	70.10%	73.33%	57.89%	30.90%	5.46%	9.28%	4.86%
GabFormer	**77.90%**	73.08%	**75.41%**	**60.53%**	**47.92%**	**16.96%**	**25.06%**	**14.32%**

Table 2. The cross-dataset performance of the models.

cates that GabFormer is more robust at ignoring the insignificant change such as what is shown in Fig. 3 LEVIR-CD patch (b) where there is a significant difference in color between pre-change and post-change images, as well as WHU-CD patch (b) with changes happened in other land covers. This supports our hypothesis that texture features extraction by the GFN contributes to the reduction of errors caused by unimportant changes such as change of color.

In order to evaluate the generalization capability of the models, we perform a cross-dataset evaluation by measuring the performance of a model that is previously trained using one dataset, on the test split of another dataset. It is interesting to note that the two datasets were taken from different platforms i.e., LEVIR-CD contains satellite images while WHU-CD comprises of aerial imagery. Moreover, both datasets captured images from different areas of the world which have different characteristics of buildings and surrounding environment. The results are displayed in Table 2. We compare GabFormer with the other two SOTA methods that have relatively close quantitative performance, as previously shown in Table 1. The low numbers reported on the (train WHU-CD, test LEVIR-CD) may come from 2 possible reasons: (1) the LEVIR-CD dataset contains more difficult cases such as smaller building footprints (LEVIR-CD has

Model	Precision	Recall	F1-score	IoU
GabFormer	**92.87%**	88.54%	**90.66%**	**82.91%**
GabFormer - FFM (GFN only)	92.67%	88.35%	90.46%	82.58%
GabFormer - FFM - GFN	92.05%	**88.80%**	90.40%	82.48%

Table 3. Ablation study on the effect of removing GFN and FFM from the model.

987 change pixels/instance while WHU-CD has 9296 change pixels/instance on average), (2) WHU-CD has less data for the networks to learn (both the number of training data and the number of change pixels in the dataset i.e., the WHU-CD has approximately 9 million change pixels less than the LEVIR-CD). The results indicate that our proposed GabFormer performs significantly better than the other models both in F1-score and IoU. Being compared to ChangeFormer, GabFormer improves 2.08% F1-score as well as 2.64% IoU when being tested on WHU-CD, and increases F1-score and IoU by 15.78% and 9.46% on the LEVIR-CD evaluation. Since training data for building change detection can be difficult to obtain in a sufficiently large quantity, methods that can learn efficiently from a limited dataset are necessary. However, Transformer-based networks typically require a larger training dataset than traditional architectures. These results thus imply that making use of our Gabor Feature Network and Feature Fusion Module in GabFormer can significantly improve the generalization capability of the network, due to its reduced number of free parameters and the well defined Gabor filter characteristics in terms of capturing repetitive texture features.

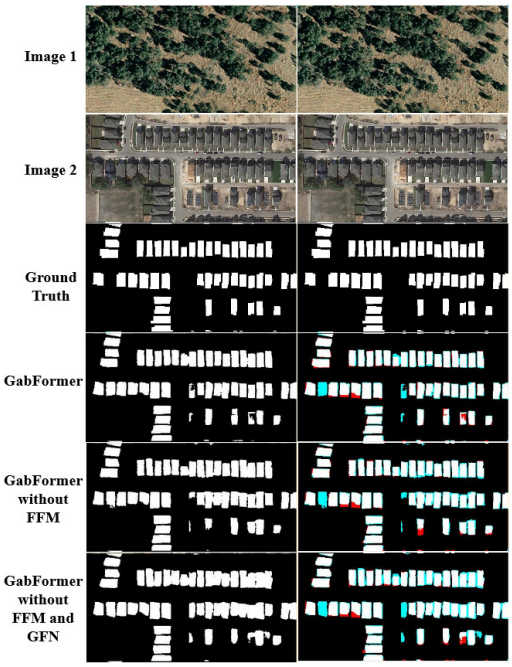

Fig. 4. The visualization of the effect of removing GFN and FFM from the model. Left column shows the prediction without color to have a better visual on the edges. Right column illustrates the results with color representation: TP (white), FP (light blue), TN (black), FN (red).

3.3. Ablation Study

We conduct an ablation study to quantitatively evaluate the effectiveness of the proposed Gabor Feature Network (GFN) and Feature Fusion Module (FFM). As indicated in Table 3, when we remove the FFM and GFN one by one, we observe that the performance decreases in terms F1-score and IoU. This change in performance can also be observed in the visualization in Fig. 4. The left column shows the change map predicted by each model and the results with color indicators of TP, FP, TN, and FN are illustrated on the right side. We can observe that as we remove the FFM, the edges of the predicted pixels located close to each other start to combine together, and when GFN is removed together with FFM, the boundaries between buildings are not clear anymore. This indicates the vital roles of the GFN to extract distinct buildings' textures as well as the FFM in fusing the extracted multiresolution texture features with the features from the encoder, in such a way that adding GFN and FFM makes the model predicts clearer shape and boundary of buildings.

4. CONCLUSION

In this paper, we propose GabFormer, a Transformer-based model which uses Gabor Feature Network (GFN) to extract distinctive building's texture features using a reduced number of free convolution weights. In addition, Feature Fusion Module (FFM) which merges the extracted Gabor features and hierarchical Transformer features at corresponding resolution, is also proposed in such a way that the extracted texture features are passed on to the deep intermediate layers of the network. Based on our experimental evaluation, the proposed GabFormer outperforms SOTA models and we can also see a significant improvement in the generalization ability of the proposed model. Moreover, the ablation study confirms that adding GFN and FFM provides a more precise shape and boundary of the buildings predicted in the change map.

5. REFERENCES

[1] Ying Sun, Xinchang Zhang, Jianfeng Huang, Haiying Wang, and Qinchuan Xin, "Fine-grained building change detection from very high-spatial-resolution remote sensing images based on deep multitask learning," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[2] Peijun Du, Sicong Liu, Paolo Gamba, Kun Tan, and Junshi Xia, "Fusion of difference images for change detection over urban areas," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 4, pp. 1076–1086, 2012.

[3] Seyd Teymoor Seydi, Mahdi Hasanlou, Jocelyn Chanussot, and Pedram Ghamisi, "BDD-Net+: A building damage detection framework based on modified Coat-Net," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 4232–4247, 2023.

[4] Hao Chen and Zhenwei Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, pp. 1662, 2020.

[5] Shunping Ji, Shiqing Wei, and Meng Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 2019.

[6] Rodrigo Caye Daudt, Bertrand Le Saux, and Alexandre Boulch, "Fully convolutional Siamese networks for change detection," in *25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 4063–4067.

[7] Jie Chen, Ziyang Yuan, Jian Peng, Li Chen, Haozhe Huang, Jiawei Zhu, Yu Liu, and Haifeng Li, "DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1194–1206, 2021.

[8] Qingle Guo, Junping Zhang, Shengyu Zhu, Chongxiao Zhong, and Ye Zhang, "Deep multiscale Siamese network with parallel convolutional structure and self-attention for change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[9] Deepanshi, Rahasya Barkur, Devishi Suresh, Shyam Lal, C. Sudhakar Reddy, and P. G. Diwakar, "RSCD-Net: A robust deep learning architecture for change detection from bi-temporal high resolution remote sensing images," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 2, pp. 537–551, 2023.

[10] Hao Chen, Zipeng Qi, and Zhenwei Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[11] Wele Gedara Chaminda Bandara and Vishal M. Patel, "A transformer-based Siamese network for change detection," in *IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2022, pp. 207–210.

[12] Shuwei Huo, Yuan Zhou, Lei Zhang, Yanjie Feng, Wei Xiang, and Sun-Yuan Kung, "Geometric variation adaptive network for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[13] Shangzhen Luan, Chen Chen, Baochang Zhang, Jungong Han, and Jianzhuang Liu, "Gabor convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4357–4366, 2018.

[14] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy, "Do vision transformers see like convolutional neural networks?," in *Neural Information Processing Systems*, 2021.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[16] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer, "Scaling vision transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 12104–12113.

[17] D. Gabor, "Theory of communication," *Journal of Institution of Electrical Engineers*, vol. 93, no. 3, pp. 429–457, 1946.

[18] J.G. Daugman, "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 7, pp. 1169–1179, 1988.

[19] Chengjun Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 467–476, 2002.

[20] Laurenz Wiskott, Jean-Marc Fellous, Norbert Kruger, and Christoph von der Malsburg, "Face recognition by elastic bunch graph matching," in *Intelligent Biometric Techniques in Fingerprint and Face Recognition*. CRC Press, 1999.