Learning Interpretable Forensic Representations via Local Window Modulation

Sowmen Das
University of Cambridge, UK
sd973@cam.ac.uk

Md. Ruhul Amin
Fordham University, USA
mamin17@fordham.edu

Abstract

The majority of existing image forgeries involve augmenting a specific region of the source image which leaves detectable artifacts and forensic traces. These distinguishing features are mostly found in and around the local neighborhood of the manipulated pixels. However, patch-based detection approaches quickly become intractable due to inefficient computation and low robustness. In this work, we investigate how to effectively learn these forensic representations using local window-based attention techniques. We propose Forensic Modulation Network (ForMoNet) that uses focal modulation and gated attention layers to automatically identify the long and short-range context for any query pixel. Furthermore, the network is more interpretable and computationally efficient than standard self-attention, which is critical for real-world applications. Our evaluation of various benchmarks shows that ForMoNet outperforms existing transformer-based forensic networks by 6% to 11% on different forgeries.

1. Introduction

Manipulation or forgery is the act of altering the original content of data and presenting it in a distorted form which can be detrimental to society and personal lives. Recent advancements in generative modeling has brought this problem into more focus. However, manually performed manipulations or "*photoshopping*" is still one of the easiest and most widely used way that images are manipulated in the wild. Since generative models like Diffusion and GAN generally create an image from scratch, these fall more under the category of *fake* images, rather than *manipulated*. In this work we focus on detecting image manipulations and understanding how well neural network based detectors can represent the forensic traces in images.

Image forgery localization is a type of semantic segmentation problem where we are only concerned with the tampered regions than the semantic objects. Manipulated regions can vary widely in scale, shape and content. This makes it a challenging task for existing FCN architectures since they are optimized towards summarizing the semantic context of an image rather than identifying the forensic traces. Typical solutions to this problem include training with huge amounts of data on deeper networks [65, 40], using multiple networks to learn separate modalities [24, 35, 78, 12, 57], or relying solely on noise and discriminatory signals [76, 16, 5, 56]. However, such deep networks are generally difficult to interpret and even harder to understand the cases where they fail [17]. Self-attention is capable of effectively modeling the pairwise forensic relations of an image through their dense correction architectures [27, 28, 46, 58]. However, vision-transformer mod-

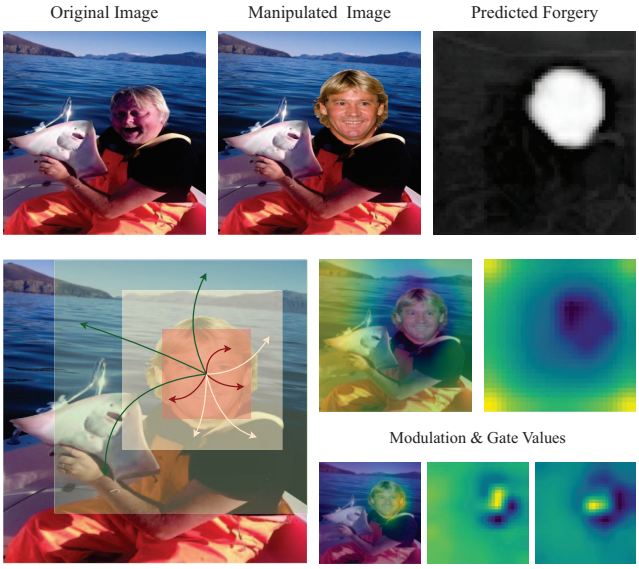

Figure 1: The top row shows an example of an original and manipulated image, followed by our network's predicted localization mask. The bottom left image illustrates how the forged region's local and global neighbours differ in their interaction and feature granularity with a specific query pixel. The modulation and gated layers of the local window attention learn to recognise these differences and can extract forensic features within the image.

els have the drawback of requiring large computation overhead in order to generate all pair attentions. Additionally, by replacing convolutions with self-attention, they require large amounts of data and parameters to learn the spatial dependencies of images. Since these models are heavily over-parameterized [2], it is difficult to interpret their feature space. As a result, their applications for forensics are still lacking in terms of generalisation and explainability for widespread application.

In this work we re-investigate the fundamental property of manipulated images, which is to learn the boundary between modified and authentic pixel regions. Instead of computing entire image representations, we focus on learning local neighbourhood differences. In the case of forgeries such as splicing, copy-move or inpainting, the tampered region is always localized to a certain portion of the image. As seen in Fig. 1, pixels within the tampered section are more similar to one another and, are greatly distinct from the remaining portion of the image. This creates a clear discrepancy between the overall feature space which can be identified using frequency or compression analysis [38, 61, 16]. To locate the forged pixels from these boundaries a number of patch and window based analysis frameworks have been proposed over time [75, 48]. Although a straightforward concept, these manual patch based approaches are greatly limited by the size of their windows, influence to post processing attributions, and unshared feature granularity. As a result, these methods never gained widespread traction.

To tackle the challenges of utilizing local patch features, we propose our Forensic Modulation Network (ForMoNet) that uses *focal* modulation [70] and a dynamic window based attention decoder to automatically learn local feature disparities and neighborhood interactions. Forensic features within an image are already quite subtle and easily gets diminished when using deep or complex neural networks. Transformers using self-attention modules compute a $\mathcal{O}(n^2)$ map at each layer of the network, and they need deep layers to effectively capture the receptive field of the entire image. This dilutes the subtle forensic details and regional variations. Instead of focusing on the entire image, focal modulation dynamically identifies the appropriate query windows and attends to their local context. Focal attention adapts the granularity of short and long range interactions based on their importance to a particular region. This is similar to the local receptive field of CNN's, but is better at identifying token correlations. Moreover, to further model the multi-scale dependencies of forged regions, we use a multi window decoder to combine the intermediate feature maps and generate a hierarchical representation. This improves the decoder's ability in localizing forged regions of arbitrary scale through intermediate position-mixing [53] and pooling operations. ForMoNet achieves 93.1% AUC(%) on CASIA, and 85.0% AUC(%) on IMD2020 benchmarks, outperforming existing transformer based forensic networks by a margin of 6% - 11%, while also ensuring model explainability.

2. Related Works

Existing research on image manipulation have generally divided forgery operations into three distinct categories – splicing, copy-move and inpainting. Each operation confines the region of the image being manipulated to a local group of pixels. Although post-processing like blur, or compression is applied to the final result, these operations still leave behind some distinct artifacts due to the feature misalignment of the local and global regions. These artifacts include, PRNU sensor information [14], compression and noise features[76, 56], camera model features [9, 48], etc. Various filters and kernels have been developed overtime [34, 44, 22] to learn these artifacts, but they are not robust.

Earlier neural network based detectors largely revolved around using deep FCNs and large amounts of data to learn the underlying forensic patterns [65, 33, 40]. Many works have shown that the extent on data requirement can be somewhat circumvented by utilizing additional noise and steganalytic features [3, 4] through either pre or post fusion before the final classification [79, 37, 24, 36]. Contrastive learning [46, 71, 73] has also shown promise in identifying difference structures between local windows by learning the comparative representations. Additionally, various works have also explored the use of classification [49, 11, 45, 23], object proposal [6, 62], segmentation [8, 7, 74], and self-supervised frameworks [80, 1] for manipulation detection and localization.

Self-Attention is the core component of Transformer architectures which was primarily developed for long-range modeling of language tokens. Self-attention involves computing a correspondence map between all pairs of query tokens i.e. pixels of an image to determine correlations over features. In image forensics this can be used to model the global dependency and pixel relationships, which is one of the key requirements of forgery detection. However, computing all-pair attention is an expensive operation requiring $\mathcal{O}(n^2)$ complexity over the number of pixels. Furthermore, since vision transformers do not use convolution layers for spatial modeling, they require much more data to learn the same semantic structure. Different variations of transformers have been proposed to reduce the complexity of operations [10, 55, 60, 15, 64]. These methods have also been extended to the task of image segmentation via the transformers' encoder-decoder framework [52, 66, 77, 13]. They either use a hierarchical pyramid or transformer encoder backbone to generate intermediate image features and consequently merge them through selective attention to generate the segmentation mask.

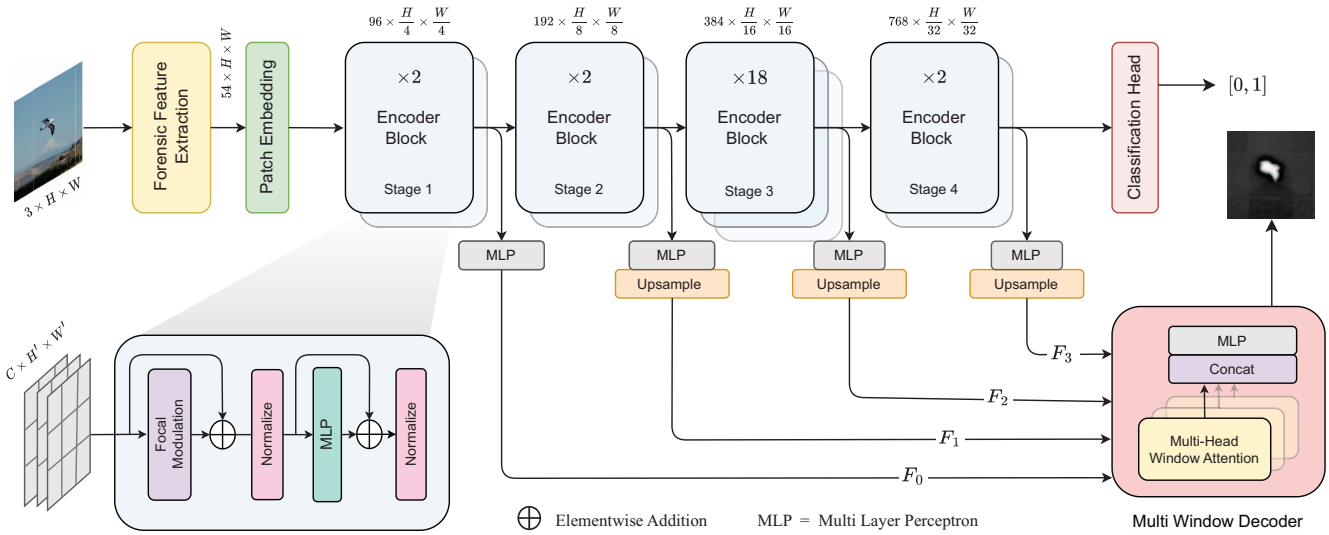

Figure 2: The overall architecture of our proposed Forensic Modulation Network (ForMoNet). The network is composed of two sections; (i) A four stage transformer encoder that uses focal modulation and locally gated attention. In between each stage, the features are downsampled and normalized. The output of the final stage goes through a classification head to generate binary labels. (ii) A multi-window decoder that processes upsampled features from each encoder stage to generate the localization mask. To upscale the forensic embeddings to their precise spatial positions, the decoder employs multi-head window attention followed by MLP layers.

For our work we utilize the comparative relationship modeling capabilities of attention for a local neighborhood span instead of looking at the whole image. Although generating a global relationship mapping would be considered ideal in order to identify which pixels are out of distribution i.e forged, this operation is both exhaustive and detrimental for forensic analysis due to their abstruse nature. Using local attention we can achieve the same or even better representations by focusing on a smaller window and prevent feature dilution. Existing attention transformers implement this operation either via shifted windows [42, 41, 20], dilated windows [29, 30], window based clustering [59, 67], and dynamic kernel or patch operations [50, 72, 39]. But these methods use heavy query interactions and aggregation for visual tokens. They also lack the semantic advantages of convolution layers which is necessary to understand the spatial dependency for localization. Additionally, earlier research [17] has shown that gated networks are able to filter the necessary forensic traces from observable image features for better forgery detection. Thus, we make use of focal modulation networks [69, 70] which performs attention over adaptive neighborhood contexts in a *focal* manner through a sequence of hierarchical gating. This results in faster computation over each window and better interpretation of the gated values. Gating over selective windows reinforces the underlying features and propagates the forensic modulations during the upscaling and decoding phase. This leads to better information sharing through the network and improvements in detection of multi-scale forgeries.

3. Proposed Method

3.1. Overview

The proposed architecture as shown in Fig. 2 is composed of two modules; an attention encoder that generates hierarchical feature maps through focal modulation, and a multi window attention decoder which combines the multi granularity features to generate the segmentation map. The input image $I \in \mathbb{R}^{3\times H\times W}$ is first passed through parallel feature suppression modules which include SRM filter [78], ELA conversion [17], and a constrained Bayer-conv layer [5] to generate additional noise and steganalytic features. The resulting concatenated input $X \in \mathbb{R}^{54\times H\times W} = \{\text{Conv2D}(I), \text{SRM(I)}, \text{Bayer}(I), \text{ELA}(I)\}$ is passed through a a 4×4 patch-embedding layer which projects the patches into hidden features with dimension $d = 96$. This spatial feature map then passes through four stages of focal transformer blocks. Each stage $i \in \{1, 2, 3, 4\}$ consists of N_i focal attention layers followed by a downsampling layer. The features of the four intermediate stages are each passed through an MLP before upsampling and being passed to the decoder. The localization map is generated by the decoder through hierarchical window attention and mixing of these intermediate representations. Additionally, the final feature of the encoder is also passed through a dense layer to generate a classification label.

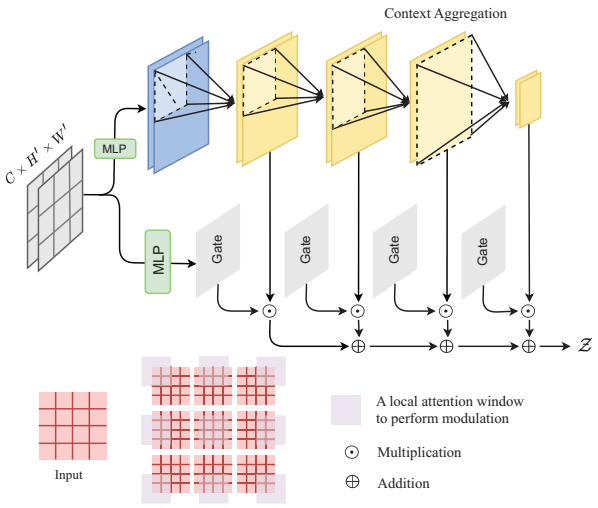

Figure 3: The structure of the focal modulation block [69]. The input feature map is processed in parallel by the context aggregation and gating branches to produce the modulation vector $\mathcal{Z}$. The bottom row shows how the neighborhood features are captured by the local window as the input passes through the layers.

3.2. Focal Modulation

Traditional self-attention is a generic encoding process that produces a feature representation $y_i \in \mathbb{R}^C$ for each token $x_i \in \mathbb{R}^C$ in an input feature map $X \in \mathbb{R}^{C \times H \times W}$, by computing an interaction with its neighbors in X and aggregating over the contexts. This interaction becomes more expensive as the neighborhood's size increases. In contrast, focal modulation requires *first* producing the context of the entire input through an aggregation and *then* computing the modulated interaction with this aggregated vector. This enables the interactions to be *focused* on the actual context of the input, as opposed to being influenced by specific values. The two stages of the procedure are depicted in Fig. 3. This method allows both channel and space specific modulation. In addition, gating results in tokens that are decoupled from one another, enabling control over fine and coarse contexts. The modulation values of each query x_i are determined as:

$$y_i = q(x_i) \odot h(\{G^l\}_{l=1}^L \cdot \{Z^l\}_{l=1}^L)$$

where $q(\cdot), h(\cdot)$ are linear transforms, G^l and Z^l are the respective sets of gating and context values, and $\odot$ is an element-wise multiplication operation.

Local Context Aggregation: The input feature map X is consecutively passed through a series of L windowed transformations that generate the local context for each focal level. For forensic localization, all relevant windows for multi-scale hierarchical contexts must be examined. A single pooling would aggregate the visual semantics as opposed to the forensic traces. Sequential aggregation is performed by inspecting each patch within the feature map and compressing them using learnable and structure-dependent depth-wise convolutions (DWConv2D). Each successive focal level has a receptive field $r_l = 1 + \sum_{i=1}^{l}(k^l - 1)$ that is larger than its kernel size k^l. This is because the current level can use the aggregated values from all preceding levels to capture both short and long range contexts at varying granularities. This allows us to summarise the semantics of an image and locate forensic traces. Output features Z^l of each focal level $l \in \{1, ..., L\}$ are computed as follows:

$$Z^l = f_\psi^l(Z^{l-1})$$
$$f_\psi \triangleq \text{ReLU}(\text{DWConv2D}(Z^{l-1})$$

The weights learned by the last modulation level for each encoder stage are depicted in Fig. 4. Because earlier stages have smaller windows, they can only highlight the distinctive features of the local region. However, as the features propagate, the receptive field of the modulation window expands and the network is able to differentiate between the image and forensic space. Even in the presence of multiple forged patches, the network can locate both minima.

Attention Gating: The set of feature maps $\{Z^l\}_{l=1}^L$ obtained through the context aggregation process is transformed into the final modulation vector via sequential gating. In most cases, the relationship between a query pixel and its surrounding pixels is determined by the semantic information of the image. But we want to extract the forensic features that were aggregated in the previous step. The gating process allows control over how much we sample from coarse or fine-grained features. By applying this operation over the L feature maps, we can encode both local fine-grained and global coarse-grained features from different levels for each query. The set of gating vectors $\mathcal{G} = \{G^l\}_{l=1}^L = f_g(X)$, where $G^l \in \mathbb{R}^{1 \times H \times W}$, is first generated by applying a linear layer to the input X. These are then used to generate the final modulation vector as,

$$\mathcal{Z} = \sum_{l=1}^{L} G^l \odot Z^l \in \mathbb{R}^{C \times H \times W}$$

The final step in focal modulation is multiplying the modulation vector with the input using element-wise multiplication. By using multiple context levels and gates, the attention is able to adapt its receptive window for any particular query. This allows controlled communication across neighboring regions, visualized in Fig. 3. Fig. 4 depicts the attention gates of the final encoder stage. We can see that these gates are crucial in filtering forensic traces from the global image space. Each gate is activated by a unique set of characteristics such as boundary edges or inner distributions. By combing their activation through the hierarchical

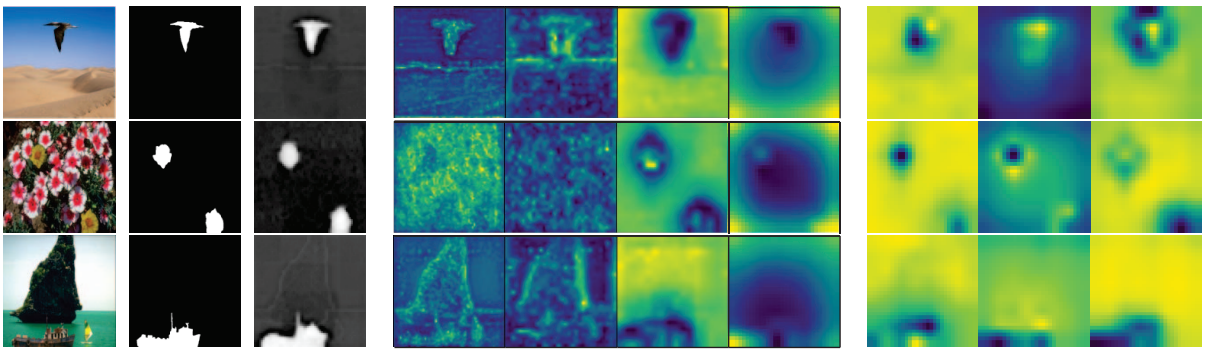

Figure 4: Visualization of the modulation weights of each stage, and attention gates of the last stage learnt by the network.

addition process, we are able to utilize the combined feature space for segregation.

3.3. Multi Window Decoder

The decoder is responsible for integrating over the learned representations provided by the encoder and generate the spatial localization map of the forged regions. It is important that we are able to correlate the spatial positions with the identified forensic features. Instead of iteratively combining and transforming the intermediate features $\{F_0, F_1, F_2, F_3\}$, we pool the values together and perform a multi-window attention over multiple spatial sizes. This is similar to how multi-head attention in ViT [21] works. As shown in Fig. 5, branches of the attention heads provide multiple hierarchies of receptive fields for the local windows. The short path realigns the feature output with the low level image context. The resulting values are linearly transformed through a series of MLP's and upsampled to generate the final segmentation.

Each head of large window attention [68] performs a series of spatial and channel mixing operations over the specific window size. Through iterative downsampling in the encoder, the spatial positions become abstract. To realign these activations with the image space, the attention uses a sequence of parallel token and channel-mixing MLPs. This is the core of MLP-Mixer [54] that allows better communication between the spatial positions. Given a 2D feature map $x \in \mathbb{R}^{C\times H\times W}$ and a query patch $x_p \in \mathbb{R}^{C\times P\times P}$, the operation of a single attention head can be formulated as,

$$\begin{aligned}\tilde{x_p} &= \text{Reshape}(C, P^2)(x_p)\\ z_h &= \text{Reshape}(C, P, P)(\varphi(\{\text{MLP}_h(\tilde{x_p})\}))\\ \text{MLP}_h(\tilde{x_p}) &= \mathbf{W}_2\sigma(\mathbf{W}_1^{\mathrm{T}}\tilde{x_p} + \mathbf{b})\end{aligned}$$

where $\mathbf{W}_1 \in \mathbb{R}^{HW\times d}$ and $\mathbf{W}_2 \in \mathbb{R}^{d\times HW}$ are both learned linear transformations along with a sigmoid non-linearity, and φ is an average pooling operation performed over the multiple outputs of the mlp-mixer. z_h is the output of the h-th attention head. The outputs of each head for the specific query patch x_p are concatenated to create the position-mixed content $\mathbf{z} = \text{concat}(\{z_1, z_2, ..., z_h\})$. The final learned attention after combining the outputs of each head is formulated as,

$$\begin{aligned}A &= \text{softmax}\left(\frac{(\mathbf{W}_q x)\,(\mathbf{W}_k \mathbf{z})^{\mathrm{T}}}{\sqrt{d}}\right)(\mathbf{W}_v \mathbf{z})\\ \hat{\mathbf{A}} &= \text{concat}(\{A_1, A_2, ..., A_h\})\mathbf{W}_m\end{aligned}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{C\times d}$ are the learned linear transforms and $\mathbf{W}_m$ is the learned weights that aggregates multiple attention values.

The use of multiple window attention heads is essential in retrieving the relevant patch activations for a particular position. The high level encoder features summarize the forged context with narrower windows, while the lower ones have a larger window that correlates the position of that region to its surroundings. Instead of arbitrarily upscaling the forged pixel's position, the attention heads collect important response values across several windows and finetune the exact location.

3.4. Complexity Analysis

Vanilla self-attention in ViT takes $\mathcal{O}((HW)^2C + HW \times (3C^2))$ complexity [69]. As the size of an image increases, or the number of tokens increase, $(HW)^2$ increases quadratically. Focal modulation uses three linear projections $q(\cdot), h(\cdot)$, and $f_\psi(\cdot)$. The final complexity of the modulation operation is $\mathcal{O}\left(HW \times \left(3C^2 + C(2L+3) + C\sum_\ell \left(k^\ell\right)^2\right)\right)$ [69], which is devoid of the $(HW)^2$ constraining term.

4. Experiments

4.1. Data & Setup

We follow the evaluation protocols in [65, 17] for training and validation. The model was trained on four types

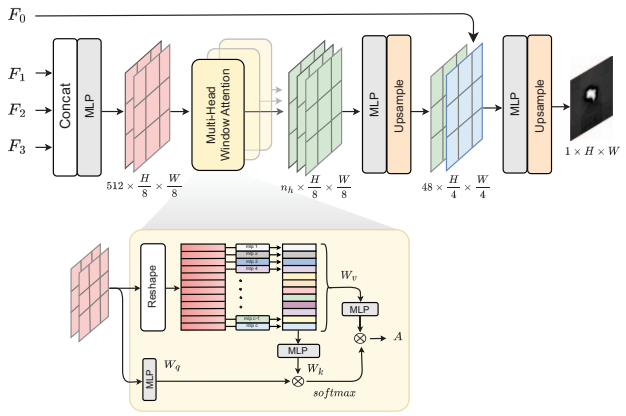

Figure 5: The decoder consisting of multi-head window attention [68] which uses the channel and position mixing framework of MLP-mixer to generate the output. The decoder takes in $\{F_0, F_1, F_2, F_3\}$ the encoder features and through sequential upscaling generates the final localization mask.

of data samples – copy-move, splicing, inpainting, and authentic images taken from Dresden [25], MS COCO synthetic [4], Defacto [43], and IMD-Real [47] datasets. We use the standard datasets: CASIA [19], NIST16 [26], COVERAGE [63] and IMD-2020 [47] for pre-trained and fine-tuned evaluation. For finetuning we perform train/test split of the datasets independently, following the process outlined in [78, 40].

4.2. Loss Function

The train data consists of input images $I \in \mathbb{R}^{3\times H\times W}$ and binary ground-truth masks $M \in [0,1]^{1\times H\times W}$. The network was trained end-to-end using a multi-task loss function following [17] which combines both detection and localization losses as follows,

$$\mathcal{L} = w_c \cdot \mathcal{L}_{BCE} + w_d \cdot \mathcal{L}_{DSC} + w_f \cdot \mathcal{L}_{FL}$$

$$\text{where, } \mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot \log(\hat{y}_i)$$

$$\mathcal{L}_{DSC}(P,G) = -\log\left(\frac{2\cdot|P\cap G|+\epsilon}{|P|+|G|+\epsilon}\right)$$

$$\mathcal{L}_{FL}(P) = -\alpha_t (1-P)^{\gamma}\log(P)$$

here $\mathcal{L}$ is the final loss which is a weighted sum of the classification BCE loss, and Dice and Focal losses for the segmentation map. By combining both losses, we are able to reinforce the segmentation by the classifier predictions. Additionally, dice and focal loss significantly improves out of bound errors and reduces false positives. The hyper-parameters were set as $\epsilon = 10^{-7}, \gamma = 2, w_c = 1, w_d = 1.10$, and $w_f = 1.15$.

Method	CASIA	NIST16	COVERAGE	IMD2020
MantraNet [65]	81.7	79.5	81.9	74.8
SPAN [33]	79.7	84.0	92.2	75.0
PSCC-Net [40]	82.9	85.5	84.7	80.6
ObjectFormer [58]	84.3	**87.2**	**92.8**	82.1
ForMoNet (Ours)	**86.4**	84.6	85.2	**83.9**

Table 1: Pixel-level AUC comparison with pre-trained models.

4.3. Benchmark Evaluation

We compare the network's performance for both detection and segmentation against existing transformer based and SOTA forgery detection models. We report the existing values of other models as mentioned in their papers or from existing re-implementations [27, 57, 18].

Pre-trained: For pre-trained model evaluation, we compare our model's localization performance to that of another transformer-based network, ObjectFormer [58], as well as existing SOTA models such as MantraNet [65], SPAN [33], and PSCC-Net [40]. Table 1 shows that different models perform very differently on independent benchmarks. CASIA is a balanced dataset that contains a variety of forgeries. On CASIA, ForMoNet obtains an AUC of 86.4, surpassing ObjectFormer by 2.5% and SPAN by 8.4%. On IMD2020, our network got an AUC of 83.9, outperforming PSCC-Net by 4% and MantraNet by 12%. This indicates that ForMoNet can handle a wide range of real-world manipulations. For NIST16 and COVERAGE, which are primarily concerned with splicing and copy-move, we lag by approximately 2.9% and 7.5%, respectively. The explanation for this could be that the synthesised training data used by other models more closely mimics the distribution of specific forgeries.

Fine-tuned: We fine-tune the pre-trained model on each individual dataset and compare it to a few other attention-based models in Table 2. On CASIA and IMD2020, we outperform contemporary transformer attention-based networks as well as SOTA architectures, as we have in the past. On CASIA, we surpass TransForensics by 11% with an AUC of 93.1. On NIST16, where we had previously failed, our model improved its own performance by 13%, outperforming MVSS-Net by 12% and MSMG by 10%. Performance on COVERAGE has also seen similar gains. This demonstrates that the preceding performance issues can be solved by increasing the number of target-specific manipulations and training ForMoNet on additional synthetic data.

Manipulation Detection: To assess the performance of image level forgery classification, we compare the F1 score of pre-trained ForMoNet against various architectures using the evaluation policies in [17, 40]. Table 3 shows that

	Methods	CASIA	NIST16	COVERAGE	IMD2020
	RGB-N [78]	79.5 / 40.8	93.7 / 72.2	81.7 / 43.7	-
	PSCC-Net [40]	87.5 / 55.4	**99.6** / 81.9	**94.1** / 72.3	80.6 / -
	CAT-Net [34]	70.4 / 20.3	75.1 / 17.3	75.3 / 28.8	78.6 / 23.9
	GCA-Net [17]	92.2 / 71.2	95.3 / 84.5	87.4 / 69.5	82.4 / 42.6
attention-based	TransForensics [28]	85.0 / 62.7	-	88.3 / 67.4*	84.8 / -
attention-based	MVSS-Net [18]	74.8 / 39.0	82.1 / 44.1	81.1 / 41.8	81.7 / 41.1
attention-based	SPAN [33]	83.8 / 38.2	96.1 / 58.2	93.7 / 55.8	75.0 / -
attention-based	MSMG [57]	72.6 / 42.5	83.1 / 49.2	85.3 / 48.0	-
attention-based	ForMoNet (Ours)	**93.1 / 73.4**	95.1 / **84.7**	86.2 / 65.1	**85.0 / 43.8**

Table 2: Pixel-level AUC/F1 performance of image forgery localization using fine-tuned models on unseen test splits. * designates pre-trained models.

Method	Image-Level F1 Score
MantraNet [65]	56.69
SPAN [33]	63.48
PSCC-Net [40]	66.88
GCA-Net [17]	85.51
MVSS-Net [18]	75.80
ForMoNet (Ours)	**87.95**

Table 3: Comparison of image-level detection performance on CASIA detection set.

our network outperforms the other approaches by a significant margin achieving a detection score of 87.95. This is also evident from the intermediate modulation and gate visualisations shown in the previous sections. This is because existing approaches classify the image as a threshold over the number of detected pixels from the localization mask, while we utilize a separate classification head.

The presented benchmark evaluations support our claim that using windowed attention and focal modulation can considerably improve the model in learning forensic representations. Although ForMoNet lags behind some SOTA approaches, it is still superior to the majority of existing attention models. Furthermore, the network is interpretable, which makes it more conducive to future research into the usefulness of this approach.

4.4. Ablation Study

We evaluate how different hyper parameters like window sizes, model configuration, or attention types effect the performance of the network. We report the pixel-level AUC and F1 score for the following experiments which were done on the CASIA validation set.

Focal Parameters: For the focal attention in the encoder we need to specify number of focal levels (L) and the window size (k^l) at each level. Increasing the window size to a large extent can defeat the purpose of our local attention learning, while decreasing it too much makes it harder for the model to learn the local context (Table 4(a)). Additionally, increasing the number of focal levels control the extent of the receptive field for that modulation window. A large receptive field can dilute the attention focus and increase complexity (Table 4(b)). From experiments we found that a window size of 4, and $L = 3$ focal levels gives the optimal results for an input image of size 256×256.

(a) Window Size	AUC(%)	F1
k = 2	89.2	71.1
k = 4	89.5	71.2
k = 6	89.1	71.0
k = 9	88.2	70.8
(b) Focal Levels		
L = 2	89.1	71.0
L = 3	89.6	71.2
L = 4	89.4	71.2
(c) Decoder Heads		
h = 4	89.6	71.2
h = 6	89.8	71.3
h = 8	89.8	71.2

(d) Aggregation Type	#Params	AUC(%)	F1
Conv2D	84M	76.1	56.6
DWConv2D	50M	89.6	71.2
SE	50M	74.3	52.8
scSE	57M	76.4	57.0
(e) Model Depth			
2-2-12-2	39M	87.4	68.3
2-2-18-2	50M	89.6	71.2
2-2-22-2	58M	89.9	71.3

Table 4: Ablation study of various network parameters on the CASIA validation set.

Decoder Heads: The number of heads used in the decoder during localization is a hyper-parameter that control which specific feature is pooled from the encoder vectors. The original intention of multi-head attention was that by distributing the input across windows, each is able to learn a different set of features. But this increases number of trainable parameters, and some heads might not even learn anything useful. From Table 4(c) we see that increasing heads from 4 to 6 does not have much impact to the final result.

Model Depth: The depth or number of layers in a model directly correlates to the number of trainable parameters as well as the capacity of the model. However, for forensic tasks, increasing model depth can cause the subtle features to get diluted and difficult to identify. We can see from Table 4(e) that increasing depth to higher degrees does not provide the equal amount of performance.

Spatial Aggregation: During focal modulation, instead of pooling, we aggregate the structure and semantics of the input features using depth-wise convolutions. We compare how other similar operations including standard convolution (Conv2D), squeeze-and-excitation (SE) [32], and spatial-and-channel excitation (scSE) [51] would perform. We see from Table 4(d) that the baseline with depth-wise convolution is better than others. This might be because the excitation modules collapse the input channels and perform a linear redistribution over the feature space. Although the output retains the same 2D structure, it is not as effective for patch based attention aggregation.

4.5. Robustness Analysis

To verify the robustness of ForMoNet, we examine the change in the network's performance due to various post processing operations on the input. For this purpose, we use images from the NIST16 test set and degrade the images using different distortion settings as detailed in [40]. These include Gaussian Blur with kernel size k, JPEG Compression with a quality factor q, and Additive Gaussian Noise using standard deviation σ. From Fig. 6 we see that the network can adapt to various degrees of attributions similar to other existing methods. Increasing the Gaussian blur from $k = 0$ to $k = 15$ reduced the localization AUC by only 3.7% compared to SPAN which fell by 5.7%. Similarly, for Gaussian noise addition the performance of ForMoNet degraded by 5.3%, and for JPEG compression by 1.7%.

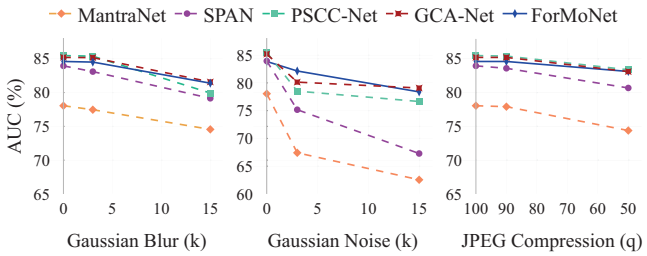

Figure 6: Localization AUC performance on NIST16 dataset under different post-processing distortions.

5. Limitations & Future Works

Decoder Optimization: The primary objective of our research was to learn the local forensic representations of a manipulated image using the focal modulation technique and window attention. The encoder is primarily responsible for extracting the forensic traces. The decoder simply maps and upscales the modulated encoder features based on position estimations. However, there are some instances where the encoder correctly classifies the image as forged, but the decoder is unable to localize the position accurately. Fig. 7 shows that, while there are some peaks in the output localization map, they are being overtaken by the surrounding noise. Nonetheless, the gate and modulation activation maps tell us that the encoder was successfully able to identify the forensic traces. One possible solution to this problem could be specific decoder optimization by freezing the encoder and only tuning the decoder layers. We could also use different decoder methods, such as masked-attention [13], which uses the groundtruth mask to improve position accuracy, or feature alignment methods [31] , which eliminates the upscaling operation with learnable alignment functions. It is worth pointing out that we were able to identify this problem because of the networks' interpretability.

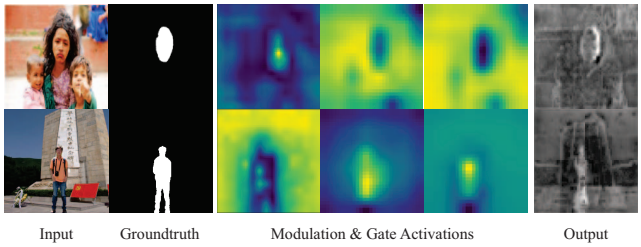

Figure 7: Illustration of the decoders' limitations. Although the encoder activations could capture the forensic traces, the decoder was unable to generate a noise-free localization

Extensions to GNN: Our research confirmed that neighborhood-based searching is more effective than global feature processing in identifying forensic differences within an image. This concept can be expanded further using graph learning techniques. If we can encode the image as a graph, with each node containing the features for a specific subset or group of pixels, the problem becomes identifying the most activating local clusters. We can encode multi-domain features such as noise response, compression kernels, DCT responses, and steganalysis outputs within each node. By processing this image graph with Graph Neural Networks, we can ideally identify the local subgroup based on these features. Furthermore, SOTA GNNs can easily learn from a graph with millions of nodes. As a result, this would also solve the issue of image scale and resolution. However, we would still need to figure out how to convert the image into a graph representation.

6. Conclusion

In this work we investigated how to effectively learn the local neighborhood representations of a manipulated image in order to identify forged regions. To that end, we proposed the Forensic Modulation Network (ForMoNet), a new architecture that uses focal modulation and window attentions to automatically identify and learn these local neighborhoods in an image. The network can better identify forensic features at a lower computational cost by utilizing context aggregation and gated forwarding. Furthermore, this process is highly interpretable, allowing us to determine whether or not the network is identifying the correct regions. We can provide explainable interpretations of the models' findings by moving away from black box networks, which is critical for real-world applications. We evaluated ForMoNet against existing transformer-based forensic models as well as other SOTA architectures and found that the network outperforms some of these models by a factor of 6% to 11% in specific tasks. This demonstrates that the model is explainable as well as generalizable to multi forgery detection and localization.

References

[1] Susmit Agrawal, Prabhat Kumar, Siddharth Seth, Toufiq Parag, Maneesh Singh, and R Venkatesh Babu. Sisl: self-supervised image signature learning for splicing detection & localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22–32, 2022. 2

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. arxiv e-prints, art. *arXiv preprint arXiv:1811.04918*, 2018. 2

[3] Quentin Bammey, Rafael Grompone von Gioi, and Jean-Michel Morel. An adaptive neural network for unsupervised mosaic consistency analysis in image forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14204, 2020. 2

[4] Jawadul H Bappy, Cody Simons, Lakshmanan Nataraj, BS Manjunath, and Amit K Roy-Chowdhury. Hybrid lstm and encoder-decoder architecture for detection of image forgeries. *IEEE Transactions on Image Processing*, 2019. 2, 6

[5] B. Bayar and M. C. Stamm. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security*, 2018. 1, 3

[6] T. Aaron Gulliver & Saif alZahir Belal Ahmed. Image splicing detection using mask-rcnn. *Signal, Image and Video Processing*, 2020. 2

[7] Xiuli Bi, Yanbin Liu, Bin Xiao, Weisheng Li, Chi-Man Pun, Guoyin Wang, and Xinbo Gao. D-unet: A dual-encoder u-net for image splicing forgery detection and localization, 2020. 2

[8] X. Bi, Y. Wei, B. Xiao, and W. Li. Rru-net: The ringed residual u-net for image splicing forgery detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 30–39, 2019. 2

[9] L. Bondi, S. Lameri, D. Güera, P. Bestagini, E. J. Delp, and S. Tubaro. Tampering detection and localization through clustering of camera-based cnn features. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1855–1864, 2017. 2

[10] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 2

[11] Ivan Castillo Camacho and Kai Wang. A simple and effective initialization of cnn for forensics of image processing operations. In *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, IH&MMSec'19, page 107–112, New York, NY, USA, 2019. Association for Computing Machinery. 2

[12] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14185–14193, 2021. 1

[13] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. 2021. 2, 8

[14] Giovanni Chierchia, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. A bayesian-mrf approach for prnu-based image forgery detection. *IEEE Transactions on Information Forensics and Security*, 9, 2014. 2

[15] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34:9355–9366, 2021. 2

[16] Davide Cozzolino and Luisa Verdoliva. Noiseprint: a cnn-based camera model fingerprint, 2018. 1, 2

[17] Sowmen Das, Md Islam, Md Amin, et al. Gca-net: utilizing gated context attention for improving image forgery localization and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 81–90, 2022. 1, 3, 5, 6, 7

[18] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 6, 7

[19] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. pages 422–426, 07 2013. 6

[20] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022. 3

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[22] Gokhan Egri and Todd Zickler. Stegapos: Preventing crops and splices with imperceptible positional encodings. *arXiv preprint arXiv:2104.12290*, 2021. 2

[23] Moawad I. Dessowky Ghada M. El Banby Ashraf A. M. Khalaf Ahmed S. Elkorany & Fathi E. Abd. El-Samie Faten Maher Al Azrak, Ahmed Sedik. An efficient method for image forgery detection based on trigonometric transforms and deep learning. *Multimedia Tools and Applications*, 2020. 2

[24] Zan Gao, Chao Sun, Zhiyong Cheng, Weili Guan, Anan Liu, and Meng Wang. Tbnet: A two-stream boundary-aware network for generic image manipulation localization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 1, 2

[25] Thomas Gloe and Rainer Böhme. The 'dresden image database' for benchmarking digital image forensics. Association for Computing Machinery, 2010. 6

[26] H. Guan, M. Kozak, E. Robertson, Y. Lee, A. N. Yates, A. Delgado, D. Zhou, T. Kheyrkhah, J. Smith, and J. Fiscus. Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 63–72, 2019. 6

[27] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. *arXiv preprint arXiv:2212.10957*, 2022. 1, 6

[28] Jing Hao, Zhixin Zhang, Shicai Yang, Di Xie, and Shiliang Pu. Transforensics: image forgery localization with dense self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15055–15064, 2021. 1, 7

[29] Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer. *arXiv preprint arXiv:2209.15001*, 2022. 3

[30] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. *arXiv preprint arXiv:2204.07143*, 2022. 3

[31] Hanzhe Hu, Yinbo Chen, Jiarui Xu, Shubhankar Borse, Hong Cai, Fatih Porikli, and Xiaolong Wang. Learning implicit feature alignment function for semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 487–505. Springer, 2022. 8

[32] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 7

[33] Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. Span: Spatial pyramid attention network for image manipulation localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 312–328. Springer, 2020. 2, 6, 7

[34] Myung-Joon Kwon, In-Jae Yu, Seung-Hun Nam, and Heung-Kyu Lee. Cat-net: Compression artifact tracing network for detection and localization of image splicing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 375–384, 2021. 2, 7

[35] Fengyong Li, Zhenjia Pei, Weimin Wei, Jing Li, Chuan Qin, et al. Image forgery detection using tamper-guided dual self-attention network with multiresolution hybrid feature. *Security and Communication Networks*, 2022, 2022. 1

[36] Fengyong Li, Zhenjia Pei, Weimin Wei, Jing Li, Chuan Qin, et al. Image forgery detection using tamper-guided dual self-attention network with multiresolution hybrid feature. *Security and Communication Networks*, 2022, 2022. 2

[37] Wei-Yun Liang, Jing Xu, and Xiao Jin. Tripinet: Tripartite progressive integration network for image manipulation localization. *arXiv preprint arXiv:2212.12841*, 2022. 2

[38] Zhouchen Lin, Junfeng He, Xiaoou Tang, and Chi-Keung Tang. Fast, automatic and fine-grained tampered jpeg image detection via dct coefficient analysis. *Pattern Recognition*, 42(11):2492 – 2501, 2009. 2

[39] Kai Liu, Tianyi Wu, Cong Liu, and Guodong Guo. Dynamic group transformer: A general vision transformer backbone with dynamic group attention. *arXiv preprint arXiv:2203.03937*, 2022. 3

[40] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7517, 2022. 1, 2, 6, 7, 8

[41] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 3

[42] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[43] Gaël MAHFOUDI, Badr TAJINI, Florent RETRAINT, Frédéric MORAIN-NICOLIER, Jean Luc DUGELAY, and Marc PIC. Defacto: Image and face manipulation dataset. In *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019. 6

[44] Hannes Mareen, Dante Vanden Bussche, Fabrizio Guillaro, Davide Cozzolino, Glenn Van Wallendael, Peter Lambert, and Luisa Verdoliva. Comprint: Image forgery detection and localization using compression fingerprints. *arXiv preprint arXiv:2210.02227*, 2022. 2

[45] Aniruddha Mazumdar, Jaya Singh, Yosha Singh Tomar, and Prabin Kumar Bora. Universal image manipulation detection using deep siamese convolutional neural network, 2018. 2

[46] Fahim Faisal Niloy, Kishor Kumar Bhaumik, and Simon S Woo. Cfl-net: Image forgery localization using contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4642–4651, 2023. 1, 2

[47] A. Novozámský, B. Mahdian, and S. Saic. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 71–80, 2020. 6

[48] Abdul Muntakim Rafi, Uday Kamal, Rakibul Hoque, Abid Abrar, Sowmitra Das, Robert Laganière, and Md. Kamrul Hasan. Application of densenet in camera model identification and post-processing detection, 2019. 2

[49] Yuan Rao and Jiangqun Ni. A deep learning approach to detection of splicing and copy-move forgeries. 2016. 2

[50] Pengzhen Ren, Changlin Li, Guangrun Wang, Yun Xiao, Qing Du, Xiaodan Liang, and Xiaojun Chang. Beyond fixation: Dynamic window visual transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11987–11997, 2022. 3

[51] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 421–429. Springer, 2018. 7

[52] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 2

[53] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 2

[54] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 5

[55] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 2

[56] Vinay Verma, Deepak Singh, and Nitin Khanna. Block-level double jpeg compression detection for image forgery localization, 2020. 1, 2

[57] Fengsheng Wang and Leyi Wei. Msmg-net: Multi-scale multi-grained supervised metworks for multi-task image manipulation detection and localization. *arXiv preprint arXiv:2211.03140*, 2022. 1, 6, 7

[58] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2022. 1, 6

[59] Pichao Wang, Xue Wang, Fan Wang, Ming Lin, Shuning Chang, Hao Li, and Rong Jin. Kvt: k-nn attention for boosting vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 285–302. Springer, 2022. 3

[60] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 2

[61] N. B. A. Warif, M. Y. I. Idris, A. W. A. Wahab, and R. Salleh. An evaluation of error level analysis in image forensics. In *2015 5th IEEE International Conference on System Engineering and Technology (ICSET)*, pages 23–28, 2015. 2

[62] Liqing Wei, Wu, Dong, Zhang, and Sun. Developing an image manipulation detection algorithm based on edge detection and faster r-cnn. *Symmetry*, 11:1223, 10 2019. 2

[63] B. Wen, Y. Zhu, R. Subramanian, T. Ng, X. Shen, and S. Winkler. Coverage — a novel database for copy-move forgery detection. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 161–165, 2016. 6

[64] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 2

[65] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9543–9552, 2019. 1, 2, 5, 6, 7

[66] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2

[67] Yutong Xie, Jianpeng Zhang, Yong Xia, Anton van den Hengel, and Qi Wu. Clustr: Exploring efficient self-attention via clustering for vision transformers. *arXiv preprint arXiv:2208.13138*, 2022. 3

[68] Haotian Yan, Chuang Zhang, and Ming Wu. Lawin transformer: Improving semantic segmentation transformer with multi-scale representations via large window attention. *arXiv preprint arXiv:2201.01615*, 2022. 5, 6

[69] Jianwei Yang, Chunyuan Li, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Focal modulation networks, 2022. 3, 4, 5

[70] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers, 2021. 2, 3

[71] Qilin Yin, Jinwei Wang, Wei Lu, and Xiangyang Luo. Contrastive learning based multi-task network for image manipulation detection. *Signal Processing*, 201:108709, 2022. 2

[72] Sukmin Yun, Hankook Lee, Jaehyung Kim, and Jinwoo Shin. Patch-level representation learning for self-supervised vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8354–8363, 2022. 3

[73] Yuyuan Zeng, Bowen Zhao, Shanzhao Qiu, Tao Dai, and Shu-Tao Xia. Towards effective image manipulation detection with proposal contrastive learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[74] Rongyu Zhang and Jiangqun Ni. A dense u-net with cross-layer intersection for detection and localization of image forgery. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020. 2

[75] Xueqi Zhang, Shuo Wang, Chenyu Liu, Min Zhang, Xiaohan Liu, and Haiyong Xie. Thinking in patch: Towards generalizable forgery detection with patch transformation. In *PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part III 18*, pages 337–352. Springer, 2021. 2

[76] Zhongyuan Zhang, Yi Qian, Yanxiang Zhao, Lin Zhu, and Jinjin Wang. Noise and edge based dual branch image manipulation detection. *arXiv preprint arXiv:2207.00724*, 2022. 1, 2

[77] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 2

[78] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Learning rich features for image manipulation detection, 2018. 1, 3, 6, 7

[79] Haochen Zhu, Gang Cao, and Mo Zhao. Effective image tampering localization with multi-scale convnext feature fusion, 2022. 2

[80] Long Zhuo, Shunquan Tan, Bin Li, and Jiwu Huang. Self-adversarial training incorporating forgery attention for image forgery localization. *IEEE Transactions on Information Forensics and Security*, 17:819–834, 2022. 2