MGQFormer: Mask-Guided Query-Based Transformer for Image Manipulation Localization

Kunlun Zeng*, Ri Cheng*, Weimin Tan†, Bo Yan†

School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
klceng22@m.fudan.edu.cn, rcheng22@m.fudan.edu.cn, wmtan@fudan.edu.cn, byan@fudan.edu.cn

Abstract

Deep learning-based models have made great progress in image tampering localization, which aims to distinguish between manipulated and authentic regions. However, these models suffer from inefficient training. This is because they use ground-truth mask labels mainly through the cross-entropy loss, which prioritizes per-pixel precision but disregards the spatial location and shape details of manipulated regions. To address this problem, we propose a Mask-Guided Query-based Transformer Framework (MGQFormer), which uses ground-truth masks to guide the learnable query token (LQT) in identifying the forged regions. Specifically, we extract feature embeddings of ground-truth masks as the guiding query token (GQT) and feed GQT and LQT into MGQFormer to estimate fake regions, respectively. Then we make MGQFormer learn the position and shape information in ground-truth mask labels by proposing a mask-guided loss to reduce the feature distance between GQT and LQT. We also observe that such mask-guided training strategy has a significant impact on the convergence speed of MGQFormer training. Extensive experiments on multiple benchmarks show that our method significantly improves over state-of-the-art methods.

Introduction

Digital image manipulation risk has become more serious in recent years due to advances in deep generative models and editing techniques. An increasing number of image processing applications are emerging and easily accessible to produce tampered images in a visually imperceptible way. Traditional local image editing method including copy-move, splicing, and removal, as a currently common forgery category, requires meticulous and skillful processing. Recent deep generative models, such as GAN (Goodfellow et al. 2020) and diffusion (Ho, Jain, and Abbeel 2020) models, can generate realistic false contents in designated areas or use language prompts to modify image semantics and style. As a result, these manipulated images cause various social security issues and can mislead the public. Consequently, it is a realistic requirement to develop a reliable model to accurately locate the manipulated region.

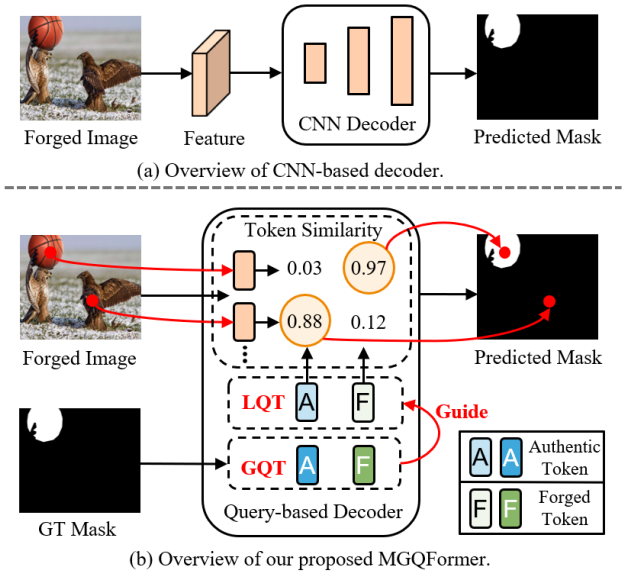

Figure 1: The difference between previous methods and ours. Our method uses a query-based transformer which is efficient and explainable. Token similarity means the softmax result of the scalar product between the image feature and query tokens. In addition, we use ground-truth masks to guide the learnable query token (LQT) in identifying the authentic and forged regions.

Despite significant advancements that have been made, existing image manipulation localization networks suffer from two shortcomings that lead to poor performance. First, these methods employ the convolutional neural network (CNN) in the final decoder process to classify the per-pixel feature (Lin et al. 2023; Zhang et al. 2021), as shown in Figure 1 (a). However, the local receptive field property of convolution filters restricts access to global information in the image. To address this problem, we propose using a query-based transformer for the image manipulation localization task. Figure 1 (b) shows that the query-based single-stage method utilizes learnable query tokens (LQT) to select pixel embeddings that are highly similar to itself, which makes network processes more explainable and effectively exploits

*These authors contributed equally.
†Corresponding authors: Bo Yan, Weimin Tan.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

the attention mechanism of the transformer.

The second shortcoming is that these image manipulation localization networks mainly utilize ground-truth mask labels through cross-entropy loss. However, cross-entropy loss does not exploit the spatial location and shape details of manipulated areas. This is because cross-entropy loss operates at the pixel level to evaluate whether each position estimation is correct, stressing per-pixel precision. As a result, the network training is inefficient. To address this problem, we feed ground-truth masks into the MGQFormer to guide the network focusing on forged regions, leading to an efficient training process.

In this paper, we propose a Mask-Guided Query-based Transformer framework (MGQFormer), which uses ground-truth masks to guide the learnable query token (LQT) in identifying forged regions. During training, we first use a multi-branch feature extractor to extract space-channel-aware features from the RGB input image. It uses two distinct transformer encoders to extract features from an RGB input image and its noise map, respectively. Then, it uses spatial and channel attention to fuse RGB image and noise map features with different distributions and domains. Finally, the fused feature is fed into our proposed query-based transformer decoder to output the location of the forged region in the image. As shown in Figure 1(b), we utilize authentic and forged LQT to distinguish manipulated regions from authentic ones. The closer a token is to the authentic query token, the more authentic it is, whereas the closer it is to a forged query token, the more fraudulent it is.

In order to force the LQT to concentrate on the forged regions, we extract the ground-truth mask feature as authentic and forged guiding query tokens (GQT) and input them into the decoder to also estimate the location of forged regions. Since GQT comes from the ground-truth mask, which is the target of the predicted mask, the GQT will contain the spatial location and shape details of forged regions. Hence, we propose a mask-guided loss to reduce the feature distance between GQT and LQT. After the model is trained, the LQT also makes the network focus on the position and shape of the forge region. As a result, we only use LQT during inference to locate manipulated regions in our query-based transformer decoder.

In summary, our main contributions are summarized as follows:

- We introduce the Mask-Guided Query-based Transformer, which contains a query-based transformer decoder utilizing the learnable query token (LQT) to locate the manipulated regions.
- We propose a mask-guided training approach, which applies the guiding query token (GQT) extracted from the GT mask as the guidance to refine LQT. In addition, we design mask-guided loss to force the GQT to guide LQT concentrating on the spatial location and shape details of manipulated regions.
- We conduct extensive experiments on multiple benchmarks and demonstrate that our method achieves state-of-the-art performance on several datasets.

Related Work

Image Manipulation Localization

Although early methods achieve excellent performance on a specific type of manipulation, including splicing (Cozzolino, Poggi, and Verdoliva 2015b; Huh et al. 2018; Kniaz, Knyaz, and Remondino 2019; Lyu, Pan, and Zhang 2014; Salloum, Ren, and Kuo 2018; Wu, Abd-Almageed, and Natarajan 2017), copy-move (Cozzolino, Poggi, and Verdoliva 2015a; D'Amiano et al. 2018; Islam et al. 2020; Wu, Abd-Almageed, and Natarajan 2018b), and removal (Wu and Zhou 2021; Wu, Abd-Almageed, and Natarajan 2018a; Yang et al. 2020; Zhu et al. 2018), they cannot generalize well to other unknown and diverse forgery, restricting their practical application. Recent studies (Zhou et al. 2018; Wu, AbdAlmageed, and Natarajan 2019; Hu et al. 2020; Liu et al. 2022; Wang et al. 2022; Chen et al. 2021; Cozzolino and Verdoliva 2019) attempt to build a unified model to tackle multiple forgery types. RGB-N (Zhou et al. 2018) adopts the steganalysis-rich model and Faster R-CNN (Ren et al. 2015), but it can only output bounding boxes instead of segmenting masks. SPAN (Hu et al. 2020) models spatial correlation at multiple scales through the pyramid structure of local self-attention blocks. PSCC-Net (Liu et al. 2022) utilizes a progressive mechanism and spatial and channel-wise correlations to enhance feature representation. ObjectFormer (Wang et al. 2022) combines RGB features and frequency features to identify the tampering artifacts, and ERMPC (Li et al. 2023) exploits the edge information to model the inconsistency between the forged and authentic regions. In this work, we exploit the novel query-based model and fulfill the task by introducing ground-truth masks that serve as guidance.

Efficient Training for Query-based Transformers

Query-based transformers employ learnable query embeddings to generate predictions (Strudel et al. 2021; Li et al. 2022b; Cheng, Schwing, and Kirillov 2021), and benefit from global attention that they can capture information from the whole image, achieving a better result than convolutional networks. However, it causes the problem that the process of training becomes difficult due to the global computation. For example, DETR (Zhu et al. 2020) suffers from low efficiency in training, which requires 500 epochs. Therefore, methods aiming to ease training for Transformers are proposed. DN-DETR (Li et al. 2022a) comes up with the idea of adding noised ground-truth boxes as positional queries for denoising training, and this approach is proved effective to speed up detection. Except for the target of the detection, Mask2Former (Cheng et al. 2022) proposes mask attention in segmentation which adds predicted masks as attention masks and speeds up query refinement compared with other query-based models. FastInst (He et al. 2023) uses instance activation-guided queries, which selects pixels with high semantics from feature maps and holds rich information about potential objects at the initial to improve the efficiency of query iterations in the Transformer decoder. MP-Former (Zhang et al. 2023) aims to address the inconsistent prediction problem in Mask2Former that leads to the

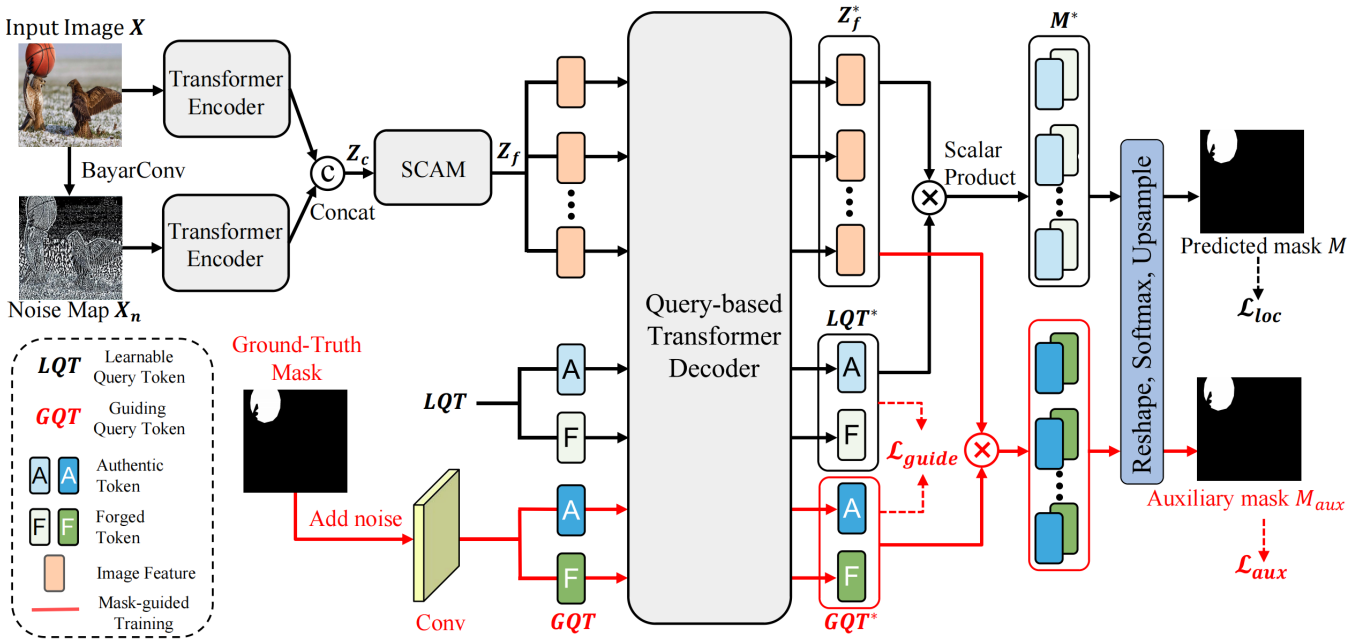

Figure 2: An overview of the proposed framework MGQFormer consisting of two-branch transformer encoder, a fusing module, and a mask-guided transformer decoder. During training, the input is a suspicious image ($H \times W \times 3$) and a ground-truth mask, the output includes a predicted mask and an auxiliary mask ($H \times W \times 1$) which are both involved in loss computation ($\mathcal{L}_{loc}$ and $\mathcal{L}_{aux}$). Note that the red-line part is mask-guided training and is not required during inference.

low utilization of queries and uses noised GT masks as the attention masks to further stable training at the early stage. These methods apply the guidance to Transformer decoder and intend to refine class queries efficiently. Our approach differs from previous query-based segmentation methods as we use extra tokens and encode the ground truth mask into GQT. We further propose auxiliary loss and mask-guided loss to guide LQT refinement.

Method

Our approach aims to identify the manipulated area in a suspicious image using a mask-guided query-based transformer (MGQFormer). Figure 2 is an overview of our framework. We denote the input image as $X \in \mathbb{R}^{H \times W \times 3}$, where H and W are the height and width of the image, respectively. We first extract the RGB and noise feature from the input image with BayarConv and Transformer Encoder. Then, the multi-modal features are fused by a spatial and channel attention module (SCAM). We design two learnable query tokens (LQT) to represent authentic and forged features, which are used to search manipulated regions in our proposed query-based transformer decoder. To make the query token refine effectively and our query-based decoder converge rapidly, we propose a mask-guided training strategy, which exploits the ground-truth mask's spatial location and shape details. Specifically, we input the noised GT masks into MGQFormer to obtain guiding query token (GQT) and auxiliary masks M_{aux}. Then, an auxiliary loss $\mathcal{L}_{aux}$ is used to make GQT contain forged regions' spatial and shape information. Furthermore, we propose a mask-guided loss $\mathcal{L}_{guide}$ to reduce the distance between LQT and GQT.

Multi-Branch Feature Extractor

The image manipulation localization usually contains elaborate post-processing, making detecting minute differences and forgery traces challenging for the RGB domain. Therefore, we employ a two-branch transformer encoder to entirely exploit the information from two domains. A BayarConv first processes the input image X to extract the noise feature $X_n \in \mathbb{R}^{H \times W \times 3}$. Then the input image and noise map are sent to Transformer Encoder. Specifically, we divide X and X_n into patches with size P, and the patch is reshaped to embeddings $X_p \in \mathbb{R}^{N \times D}$, where $N = HW/P^2$ is the number of patches, and D is the dimension of the embedding. Learnable position embeddings $\mathbf{pos} \in \mathbb{R}^{N \times D}$ are added to image embeddings to produce the sequenced tokens $Z = X_p + \mathbf{pos}$, then these tokens are processed through L Transformer layers. The same settlement mentioned above is also performed on the noise branch. After the Transformer Encoder, the output of two branches is concatenated and we get $Z_c \in \mathbb{R}^{N \times 2D}$, which is used for subsequent fusing.

The contextualized tokens from the two-branch Transformer Encoder have distinct domains and separate distributions. Therefore, we use the Spatial and Channel Attention Module (SCAM) to fulfill this task. We first reshape the tokens Z_c and use a convolution layer to get $Z_m \in \mathbb{R}^{h \times w \times c}$, where $h = H/P$, $w = W/P$ and $c = D$. Next we project

and transpose Z_m as $V = proj(Z_m) \in \mathbb{R}^{hw \times c}$, $K = proj(Z_m) \in \mathbb{R}^{hw \times c}$ and $Q = transpose(proj(Z_m)) \in \mathbb{R}^{c \times hw}$, where each $proj$ is a distinct projection layer including a 1×1 convolution and a reshape operation. Then we perform the channel attention module as follows:

$$CAM(Z_m) = proj(V(softmax(QK))). \quad (1)$$

In the meanwhile, we proceed to compute spatial attention, nearly the same as channel attention, except for transposed Q and K. Subsequently, we can obtain the contextualized tokens as follows:

$$SAM(Z_m) = proj(softmax(Q^T K^T)V), \quad (2)$$

$$Z_f = CAM(Z_m) + SAM(Z_m) + Z_m. \quad (3)$$

Then the image feature tokens $Z_f \in \mathbb{R}^{N \times D}$ are sent to query-based transformer decoder.

Mask Transformer Decoder

We first introduce the decoder at the stage of inference. For the proposed query-based transformer decoder, we employ authentic and forged learnable query tokens $\mathbf{LQT} \in \mathbb{R}^{2 \times D}$. The queries are randomly initialized and represent the forged and authentic features. Specifically, the image feature tokens Z_f and $\mathbf{LQT}$ are processed simultaneously by the decoder consisting of n transformer-based layers. During the attention mechanism, $\mathbf{LQT}$ interacts with feature tokens Z_f and extracts the rich forgery information. After that, we obtain the contextualized image feature Z_f^* and the $\mathbf{LQT}^*$. Then, the mask is computed as follows:

$$M^* = norm(proj(Z_f^*)) * (norm(proj(\mathbf{LQT}^*))^T, \quad (4)$$

where $proj$ is a linear layer, $norm$ represents L_2 normalization, and we get the $M^* \in \mathbb{R}^{N \times 2}$ by performing a scalar product between refined image features and learnable query tokens. To get the final mask, we reshape the sequence to the mask $M^{**} \in \mathbb{R}^{h \times w \times 2}$ and apply a softmax on the class dimension:

$$M = upsample(norm(softmax(M^{**}))), \quad (5)$$

where $M \in \mathbb{R}^{H \times W}$ is the predicted mask, and $upsample$ is a bilinear upsampling operation to resize the mask to the same size as the input image. In conclusion, our query-based method utilizes authentic and forged $\mathbf{LQT}$ to select regions that are highly similar to itself, which makes the process of predicting the forged regions more explainable and effective. Next, we will describe the training stage for the transformer decoder in the following section.

Mask-Guided Training

The query-based model has achieved great success in the corresponding tasks. However, these models have been proven to suffer from the low efficiency of query refinement. Previous methods have proposed approaches like denoising (Li et al. 2022a) and masked attention (Cheng et al. 2022). We point out that previous methods lack direct supervision of LQT by the location and shape details of forged regions, leading to inefficient training. These methods mainly utilize ground-truth masks through cross-entropy loss, prioritizing per-pixel precision. To address this issue, we propose a mask-guided training strategy, which uses guiding query tokens (GQT) to force the LQT to focus on the location and shape of forged regions. GQT is obtained by extracting the feature of the noised ground-truth mask, and we use the auxiliary loss to make GQT contain the spatial and shape information of forged regions. As a result, the convergence speed of MGQFormer training will be improved.

Specifically, we first add noise to the ground-truth mask. This step is because predicting the auxiliary mask from the pristine ground-truth mask may be too simple for transformer decoder and retard training. We apply point noises to the mask, analogous to DN-DETR (Li et al. 2022a) for box denoising training, to obtain more robust models. We randomly select the points within the mask and invert the original value to represent the distinct region. In addition, we use a tuned parameter μ to denote the noised percentage of area, so the number of noised points is $\mu \cdot HW$.

Given the noised mask, we further convert the mask to GQT with a convolution network to maintain the spatial information in the mask, and the ground-truth mask $G \in \mathbb{R}^{H \times W}$ is transformed into the $\mathbf{GQT} \in \mathbb{R}^{2 \times N}$. After that, the $\mathbf{GQT}$ together with image features Z_f and $\mathbf{LQT}$ are sent to the transformer decoder. In the decoder, the ground-truth information $\mathbf{GQT}$ serves as guidance to interact with other queries and assists the decoder in refining the $\mathbf{LQT}$.

After the transformer decoder, we obtain the image feature Z_f^* and query tokens $\mathbf{LQT}^*$ and $\mathbf{GQT}^*$, which are already guided by the ground-truth token $\mathbf{GQT}$. An auxiliary mask $M_{aux} \in \mathbb{R}^{H \times W}$ is further calculated by performing scalar product on Z_f^* and $\mathbf{GQT}^*$ with the same process described in the mask transformer decoder section. We then make the M_{aux} get involved in the loss computation.

Auxiliary Loss. Since we employ a convolution network to convert the ground-truth mask to queries, and the mask is noised to keep robustness, supervision for the convolution network is needed to make the auxiliary mask more accurate. Therefore, we use the pixel-level cross-entropy loss as follows to make GQT contain the spatial and shape information of forged regions:

$$\mathcal{L}_{aux} = -\sum_{i=1}^{HW} G_i \cdot \log(M_{aux,i}) \quad (6)$$

where $G \in \mathbb{R}^{H \times W}$ is the ground-truth mask. Note that we calculate the auxiliary loss using the pristine GT masks G without applying noises to make the model predict the desired accurate mask.

Mask-Gudied Loss. The purpose of $\mathbf{GQT}$ is to guide the $\mathbf{LQT}$, and both are processed the same to generate the predicted mask M and auxiliary mask M_{aux}. Thus, we expect the $\mathbf{LQT}$ to become similar to $\mathbf{GQT}$ to make the prediction more precise. A cosine similarity loss is applied to reduce the distance of both queries, which can be formulated as:

$$\mathcal{L}_{guide} = 1 - cos(\mathbf{LQT}^*, \mathbf{GQT}^*) \quad (7)$$

where cos denotes computing the cosine similarity.

Method	Col.	CASIA	NIST16	IMD20	Avg.
ManTra.	82.4	81.7	79.5	74.8	79.6
SPAN	93.6	79.7	84.0	75.0	83.1
Object.	95.5	84.3	87.2	82.1	87.3
ERMPC	96.8	87.6	**89.5**	85.6	89.9
Ours	**97.1**	**88.6**	86.2	**88.3**	**90.1**

Table 1: Comparison of manipulation localization AUC(%) scores of different pre-trained models.

Loss Function

The total loss function $\mathcal{L}$ includes three parts: the auxiliary loss to make M_{aux} accurate, the mask-guided loss to make $\mathbf{LQT}^*$ and $\mathbf{GQT}^*$ closer, and the localization loss $\mathcal{L}_{loc}$ for the predicted mask M, which employs the same cross-entropy loss as auxiliary loss:

$$\mathcal{L} = \mathcal{L}_{loc} + \mathcal{L}_{aux} + \lambda \mathcal{L}_{guide} \tag{8}$$

where λ is a weight parameter and set to 0.5 during training.

Experiment

Experiment Setup

Testing Datasets. We first pre-train our model with the dataset synthesized by PSCC-Net (Liu et al. 2022). Then we evaluate our model on CASIA dataset (Dong, Wang, and Tan 2013), Columbia dataset (Hsu and Chang 2006), NIST16 dataset (Guan et al. 2019) and IMD20 dataset (Novozamsky, Mahdian, and Saic 2020). Specifically, CASIA provides splicing and copy-move images, which widely appear in the image forgery field. Columbia consists of 180 splicing images, which are uncompressed and lack post-process. NIST16 is a challenging dataset with 564 high-resolution images that are hard for the eyes to recognize. IMD20 collects 35,000 real images captured by different camera models and is comprised of different types of manipulation generated by various inpainting methods.

Evaluation Metrics. To evaluate the localization performance of the proposed MGQFormer, following PSCC-Net (Liu et al. 2022), we report the image-level F1 score and Area Under Curve (AUC) as the evaluation metric. We adopt the fixed threshold to binarize the predicted masks, which are necessary to calculate F1 scores.

Implementation Details. The MGQFormer is implemented on the Pytorch with an NVIDIA GTX 1080 Ti GPU. All input images are resized to 384 × 384. We use Adam as the optimizer, and the learning rate decays from 2.5e-7 to 1.5e-8 with a batch size of 2. The feature extractor is initialized using the ImageNet pre-trained ViT model weights (Steiner et al. 2021) with 12 layers and a patch size of 16, while the decoder is initialized using random weights from a truncated normal distribution with 6 layers.

Comparison with State-of-the-Art Methods

We compare our model with other state-of-the-art methods under two settings: 1) training on the synthetic dataset and evaluating on the full test datasets. 2) fine-tuning the pre-trained model on the training split of test datasets and evaluating on their test split. For the pre-trained model, we evaluate the performance with ManTraNet (Wu, AbdAlmageed, and Natarajan 2019), SPAN (Hu et al. 2020), ObjectFormer (Wang et al. 2022), and ERMPC (Li et al. 2023), while further comparing with RGB-N (Zhou et al. 2018) and PSCCNet (Liu et al. 2022) for the fine-tuned model.

Method	CASIA	
	AUC	F1
RGB-N	79.5	40.8
SPAN	83.8	38.2
PSCCNet	87.5	55.4
ObjectFormer	88.2	57.9
ERMPC	90.4	58.6
Ours	**91.5**	**58.8**

Table 2: Comparison of manipulation localization results using fine-tuned models.

Pre-trained Model. Table 1 reports the best localization AUC(%) scores with pre-trained models. We can observe that MGQFormer achieves the highest performance on Columbia, CASIA, IMD20 and the average AUC(%) of all datasets, and gets competitive performance on NIST16. In particular, MGQFormer achieves 88.3 % on the real-world IMD20 dataset and outperforms ERMPC by 2.7%. This validates our method has the outstanding ability of capturing tampering traces and the generalization to high-quality datasets. On NIST16 dataset, we fail to achieve the best performance. We believe that the performance of Transformer networks is influenced by the training resolution. High performance can be fully achieved if the resolution at test time is close to training. However, NIST16 is a high-resolution dataset that greatly exceeds our training dataset.

Fine-tuned Model. To compensate for the difference in visual quality between the synthesized datasets and standard datasets, the network weights of the pre-trained model are used to initiate the fine-tuned models, which will be trained on the training split of CASIA dataset. As shown in Table 2, we compare the AUC and F1 results (%) with other methods, and our model achieves the best performance, which demonstrates that MGQFormer can capture subtle tampering artifacts effectively by query.

Robustness Evaluation

We apply different image distortion methods on raw images from Columbia dataset and evaluate the robustness of our MGQFormer. The distortion types include: 1) Resize the image with different scales, 2) Gaussian blurring with a kernel size k, and 3) JPEG compression with a quality factor q. We compare the manipulation localization performance (AUC scores) of our pre-trained models on the pristine dataset and the corrupted data, and report the results in Table 3. Compared to previous methods, MGQFormer has the best robustness against all distortions. Especially, when facing the resizing and JPEG Compress, the performance of our method

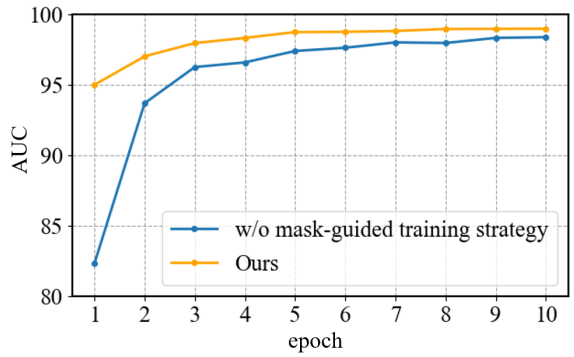

Figure 3: AUC scores (%) of MGQFormer without and with mask-guided training on the validation split of the synthesized dataset in different training epochs.

Distortion	SPAN	PSCC-Net	Ours	
no distortion	93.6	98.19	97.10	
Resize (0.78×)	89.99	93.40	**96.69**	↓0.41
Resize (0.25×)	69.08	78.41	**96.85**	↓0.25
Blur (k=3)	78.97	84.18	**92.75**	↓4.35
Blur (k=15)	67.7	73.24	**77.84**	↓19.26
Compress (q=100)	93.32	**97.97**	97.05	↓0.06
Compress (q=50)	74.62	89.11	**95.62**	↓1.48

Table 3: Localization performance on Columbia dataset under various distortions. AUC scores are reported (in %), (Blur: Gaussian Blur, Compress: JPEG Compress).

drops a little, denoting that the patch-wise MGQFormer has robustness against low-quality images.

Ablation Analysis

The design of MGQFormer contains the multi-branch feature extractor and mask-guided training. The multi-branch feature extractor employs an additional BayarConv branch to exploit the noise information and fuse both domains using SCAM. The mask-guided training is utilized to add ground-truth information, which guides the LQT to focus on the target area and improve the efficiency of the query refinement.

Ablation Study of Noise Branch. The quantitative results are listed in Table 4. The baseline denotes that we just use a single encoder and the query-based transformer decoder. To evaluate the the effectiveness of noise branch, we use a single RGB branch and remove SCAM. We can observe that without the noise branch, the AUC scores drop by 1.1% on Columbia and 2.3% on CASIA. The performance promotion validates that the use of multi-branch feature extractor effectively improves the performance of our model.

Ablation Study of Mask-guided Training. To prove the impression of Mask-guided training, we leave only LQT in Transformer decoder with the image feature and take out the input of ground-truth mask during training. As shown in Table 4, without mask-guided training, the AUC scores decrease by 2.8% on Columbia and 3.6% on CASIA.

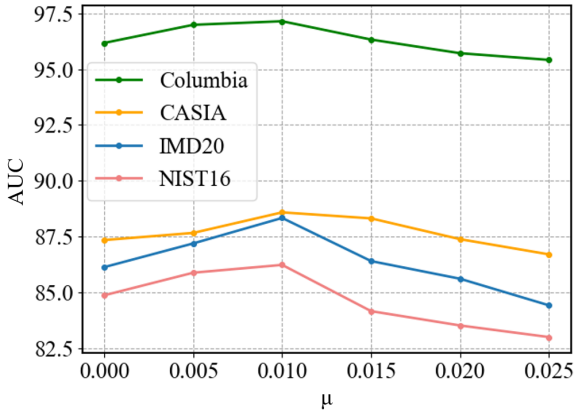

Figure 4: The effect of parameter μ in mask-guided training.

Variants	Columbia	CASIA
baseline	94.2	81.4
w/o noise branch	96.0	86.3
w/o MG training	94.3	85.0
w/o noised mask	96.2	87.4
Ours	**97.1**	**88.6**

Table 4: Ablation study of noise branch, mask-guided training (MG training), and applying noises to GT guiding masks on CASIA and Columbia dataset with pre-trained models.

Except for the promotion of localization, mask-guided training further boosts the speed of the convergence. To evaluate this effect, we compare the result of the presence and absence of the training strategy in different epochs. As shown in Figure 3, we display the AUC (%) scores on the validation split of the synthesized dataset during training. Proof by facts, MGQFormer significantly boosts training at the beginning, which surpasses the model without mask-guided training by 12.7% in the first epoch, and prominently reaches the convergence faster. This reveals that GQT certainly helps the Transformer decoder to improve the efficiency of refining LQT.

Ablation Study of applying noises to GT guiding masks. In Figure 4, we show the different values of parameter μ denoting the percentage of noised points to verify its effect over Columbia and IMD20. With its increasing, the ground truth mask has more noised points to get a more robust and generalized model; however, a large value may cause damage to the spatial information and mislead the network. In contrast, a smaller value of μ provides a more accurate ground-truth mask but may be too easy for the model to predict the auxiliary mask and retard training. It can be seen from the comparison that the setting of 0.01 is the optimal solution. The usage of point noises achieves 0.9%/1.2% AUC gain as shown in Table 4.

Visualization Results

Qualitative Results. As shown in Figure 5, we provide predicted forgery masks of various methods. We can ob-

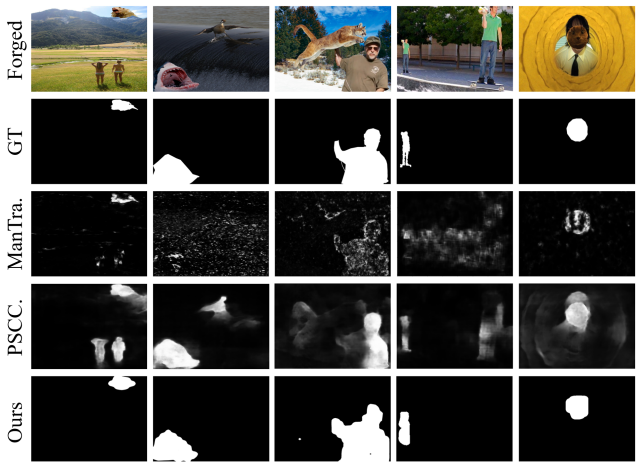

Figure 5: Visualization of the predicted manipulation mask by different methods. From top to bottom, we show forged images, GT masks, predictions of ManTraNet, PSCC-Net, and ours.

serve that PSCC-Net and ManTraNet either output the false region or make unclear predictions. The comparison of visualization results demonstrates that our method can not only locate the tampering regions more accurately but also output clear regions. It benefits from the multi-modal information and the query-based transformer decoder that employs global attention to generate masks.

Visualization of Mask-guided Training. To verify the effectiveness of the mask-guided training, we show the masks predicted by MGQFormer, the masks generated without mask-guided training, and further the auxiliary masks in Figure 6. It is clear that MGQFormer makes use of ground-truth masks to focus on the forged region, which can be seen from the similarity between the predicted mask and the auxiliary mask. Specifically, the network without mask-guided training will make false judgments about objects that are relatively small.

In Figure 7, we further show the differences between attention maps for the LQT representing the forgery in Transformer decoder from MGQFormer and attention maps without mask-guided training. It is obvious that with mask-guided training, the LQT can focus on the target area accurately due to the guidance of the GQT. In contrast, the LQT without mask-guided training does not detect forgery well and is even assigned to the totally opposite region representing the authentic place. This comparison demonstrates that the proposed GQT containing the spatial and shape information from GT mask can force the LQT to concentrate on the correct type of area that we assign to the LQT.

Conclusion

In this paper, we propose a novel mask-guided query-based transformer framework (MGQFormer). In detail, the first step is to extract RGB and noise features with a two-branch transformer encoder and further fuse them. In the second step, we convert noised ground truth masks to the guiding query token (GQT) and feed GQT and LQT into MGQFormer to estimate fake regions, respectively. We further propose auxiliary loss and mask-guided loss to guide LQT refinement. Visualization results show that the proposed mask-guided training strategy has a significant impact on the convergence speed of MGQFormer training and the performance of localization. Extensive experimental results on several benchmarks demonstrate the effectiveness of our algorithm.

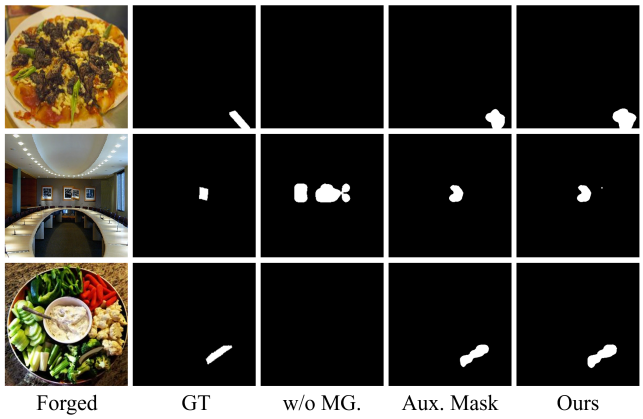

Figure 6: Visualization of results' comparison for the proposed mask-guided training. From left to right, we show the forged images, GT masks, predicted masks generated w/o MG-training, the auxiliary masks, and the predicted masks from MGQFormer.

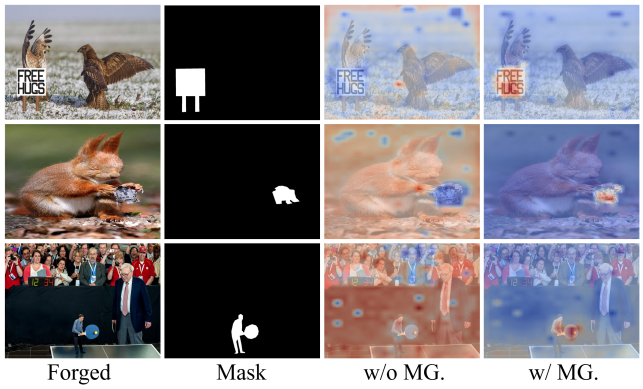

Figure 7: Visualization of attention maps for the proposed mask-guided training. From left to right, we display the forged images, GT masks, and attention maps for LQT representing the forgery in Transformer decoder without (w/o) and with (w/) mask-guided training, respectively.

Acknowledgments

This work is supported by NSFC (GrantNo.: U2001209 and 62372117) and Natural Science Foundation of Shanghai (21ZR1406600).

References

Chen, X.; Dong, C.; Ji, J.; Cao, J.; and Li, X. 2021. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14185–14193.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Cheng, B.; Schwing, A.; and Kirillov, A. 2021. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34: 17864–17875.

Cozzolino, D.; Poggi, G.; and Verdoliva, L. 2015a. Efficient dense-field copy–move forgery detection. *IEEE Transactions on Information Forensics and Security*, 10(11): 2284–2297.

Cozzolino, D.; Poggi, G.; and Verdoliva, L. 2015b. Splicebuster: A new blind image splicing detector. In *2015 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–6. IEEE.

Cozzolino, D.; and Verdoliva, L. 2019. Noiseprint: A CNN-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15: 144–159.

Dong, J.; Wang, W.; and Tan, T. 2013. Casia image tampering detection evaluation database. In *2013 IEEE China summit and international conference on signal and information processing*, 422–426. IEEE.

D'Amiano, L.; Cozzolino, D.; Poggi, G.; and Verdoliva, L. 2018. A patchmatch-based dense-field algorithm for video copy–move detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(3): 669–682.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

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

He, J.; Li, P.; Geng, Y.; and Xie, X. 2023. FastInst: A Simple Query-Based Model for Real-Time Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23663–23672.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, 6840–6851. Curran Associates, Inc.

Hsu, J.; and Chang, S. 2006. Columbia uncompressed image splicing detection evaluation dataset. *Columbia DVMM Research Lab*, 6.

Hu, X.; Zhang, Z.; Jiang, Z.; Chaudhuri, S.; Yang, Z.; and Nevatia, R. 2020. 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*, 312–328. Springer.

Huh, M.; Liu, A.; Owens, A.; and Efros, A. A. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European conference on computer vision (ECCV)*, 101–117.

Islam, A.; Long, C.; Basharat, A.; and Hoogs, A. 2020. Doa-gan: Dual-order attentive generative adversarial network for image copy-move forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4676–4685.

Kniaz, V. V.; Knyaz, V.; and Remondino, F. 2019. The point where reality meets fantasy: Mixed adversarial generators for image splice detection. *Advances in neural information processing systems*, 32.

Li, D.; Zhu, J.; Wang, M.; Liu, J.; Fu, X.; and Zha, Z.-J. 2023. Edge-Aware Regional Message Passing Controller for Image Forgery Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8222–8232.

Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L. M.; and Zhang, L. 2022a. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13619–13627.

Li, Z.; Wang, W.; Xie, E.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; Luo, P.; and Lu, T. 2022b. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1280–1289.

Lin, X.; Wang, S.; Deng, J.; Fu, Y.; Bai, X.; Chen, X.; Qu, X.; and Tang, W. 2023. Image manipulation detection by multiple tampering traces and edge artifact enhancement. *Pattern Recognition*, 133: 109026.

Liu, X.; Liu, Y.; Chen, J.; and Liu, X. 2022. 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.

Lyu, S.; Pan, X.; and Zhang, X. 2014. Exposing region splicing forgeries with blind local noise estimation. *International journal of computer vision*, 110: 202–221.

Novozamsky, A.; Mahdian, B.; and Saic, S. 2020. IMD2020: A large-scale annotated dataset tailored for detecting manipulated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 71–80.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Salloum, R.; Ren, Y.; and Kuo, C.-C. J. 2018. Image splicing localization using a multi-task fully convolutional network (MFCN). *Journal of Visual Communication and Image Representation*, 51: 201–209.

Steiner, A.; Kolesnikov, A.; Zhai, X.; Wightman, R.; Uszkoreit, J.; and Beyer, L. 2021. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7262–7272.

Wang, J.; Wu, Z.; Chen, J.; Han, X.; Shrivastava, A.; Lim, S.-N.; and Jiang, Y.-G. 2022. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2364–2373.

Wu, H.; and Zhou, J. 2021. IID-Net: Image inpainting detection network via neural architecture search and attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3): 1172–1185.

Wu, Y.; Abd-Almageed, W.; and Natarajan, P. 2017. Deep matching and validation network: An end-to-end solution to constrained image splicing localization and detection. In *Proceedings of the 25th ACM international conference on Multimedia*, 1480–1502.

Wu, Y.; Abd-Almageed, W.; and Natarajan, P. 2018a. Busternet: Detecting copy-move image forgery with source/target localization. In *Proceedings of the European conference on computer vision (ECCV)*, 168–184.

Wu, Y.; Abd-Almageed, W.; and Natarajan, P. 2018b. Image copy-move forgery detection via an end-to-end deep neural network. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1907–1915. IEEE.

Wu, Y.; AbdAlmageed, W.; and Natarajan, P. 2019. Mantranet: 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*, 9543–9552.

Yang, Q.; Yu, D.; Zhang, Z.; Yao, Y.; and Chen, L. 2020. Spatiotemporal trident networks: Detection and localization of object removal tampering in video passive forensics. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10): 4131–4144.

Zhang, H.; Li, F.; Xu, H.; Huang, S.; Liu, S.; Ni, L. M.; and Zhang, L. 2023. MP-Former: Mask-piloted transformer for image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18074–18083.

Zhang, Y.; Zhu, G.; Wu, L.; Kwong, S.; Zhang, H.; and Zhou, Y. 2021. Multi-task SE-network for image splicing localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7): 4828–4840.

Zhou, P.; Han, X.; Morariu, V. I.; and Davis, L. S. 2018. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1053–1061.

Zhu, X.; Qian, Y.; Zhao, X.; Sun, B.; and Sun, Y. 2018. A deep learning approach to patch-based image inpainting forensics. *Signal Processing: Image Communication*, 67: 90–99.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.