IMAGE STEGANALYSIS WITH CONVOLUTIONAL VISION TRANSFORMER

Ge Luo, Ping Wei, Shuwen Zhu, Xinpeng Zhang, Zhenxing Qian*, Sheng Li*

School of Computer Science, Fudan University, Shanghai, China

ABSTRACT

Recent research has shown that deep learning based methods offer more accurate detection for image steganalysis than the traditional detection paradigm based on rich media models. Existing network architectures based on deep learning, however, stack more and more convolutional layers to increase local receptive fields for image steganalysis. Limited by hardware, the detector with several convolutional layers may not extract features of steganography images from a global perspective effectively. In this paper, we propose a Convolutional Vision Transformer for image steganalysis, which can capture both local and global dependencies among noise features. In image processing phase, our network preserves CNN frame for its capacity of producing image noise residuals. Different from previous methods, we utilize the attention mechanism of vision transformer for feature extraction and classification. The proposed network is validated on two public image datasets (BOSSbase 1.01 and ALASKA #2). Experimental results demonstrate that our network performs well over fixed-size dataset and arbitrary-size dataset.

***Index Terms*—** Steganalysis, deep leraning, convolutional vision transformer

1. INTRODUCTION

Image steganography is a way of private communication by hiding secret information in selected images or generated images [1, 2]. Contrary to steganography, approaches to image steganalysis techniques are developed for detecting the hidden information. They can be divided into two categories: hand-crafted features based methods and deep learning based methods. Focusing on hand-crafted features, methods usually utilize diverse manually defined features to improve the detection performance [3, 4, 5], especially statistical features of the correlation between neighboring pixels [6, 7] and features based on selection channel [8, 9]. However, methods based on hand-crafted features are limited by the need for a great deal of human expertise. Recently, deep learning based methods, have achieved state-of-the-art performance for image steganalysis [10, 11, 12, 13]. Compared to traditional methods, CNN-based architectures can extract statistical features from images automatically, for instance, SRNet [14] made a definite improvement on detection accuracy without fixed high-pass filters. To steganalyze arbitrary-size images, statistical moments [15] and a siamese backbone [16] have been proposed to handle heterogeneous datasets effectively.

Although existing CNN-based methods for steganalysis show the most promising performance, global relations among steganographic signal features are far from being utilized adequately by stacking multiple convolutions. Generally, image steganography methods embed secret signal into comparatively complex parts of images to reduce statistic changes. From a global perspective, we argue that different parts of image provide available correlation information to distinguish steganography images from normal images. When detecting large images, aggregation capacity becomes inefficient, in spite of stacking extra convolutional layers.

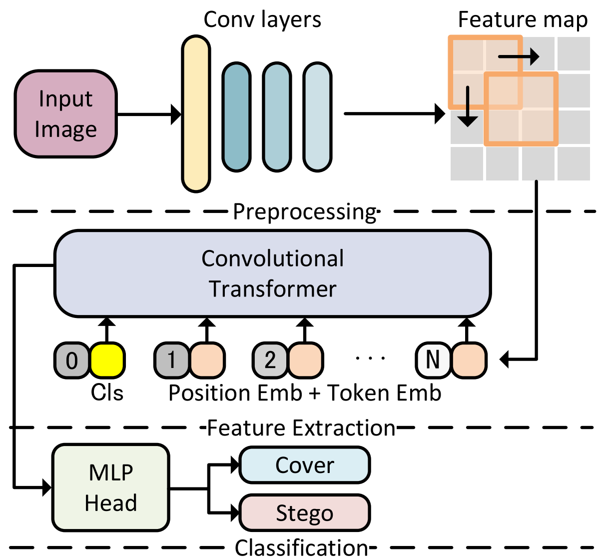


Fig. 1: Overview of proposed network for image steganalysis.

In this paper, as shown in Fig. 1, we propose a Convolutional Vision Transformer for image steganalysis. Vision Transformer (ViT) [17] have achieved considerable results in various vision classification tasks. To extract the feature vector of noise residual further effectively from a global perspective, especially for large images, additional convolutional layers are replaced with modified convolutional transformer blocks [18]. The original image and steganographic image are denoted by "cover" and "stego" respectively.

Our main contributions are as follows:

- We propose a Convolutional Vision Transformer for image steganalysis, which extracts the noise residuals efficiently from both local and global perspectives.
- Our proposed network applies convolutional layers with channel attention module to utilize global information in preprocessing phase.
- Introducing convolutional transformer, global self-attention enables the steganalysis network to learn the relationships among noise residuals in feature extraction phase.
- Our network provides a satisfying performance on fixed-size dataset, which achieves a boost of detection accuracy compared with existing methods on the heterogeneous dataset.

Supported by the National Science Foundation of China (Grant U1936214, Grant U20B2051, Grant 62072114, Grant U20A20178), and Project of Shanghai Science and Technology Commission (Grant 21010500200). *Corresponding Authors are: zhangxinpeng@fudan.edu.cn and zxqian@fudan.edu.cn

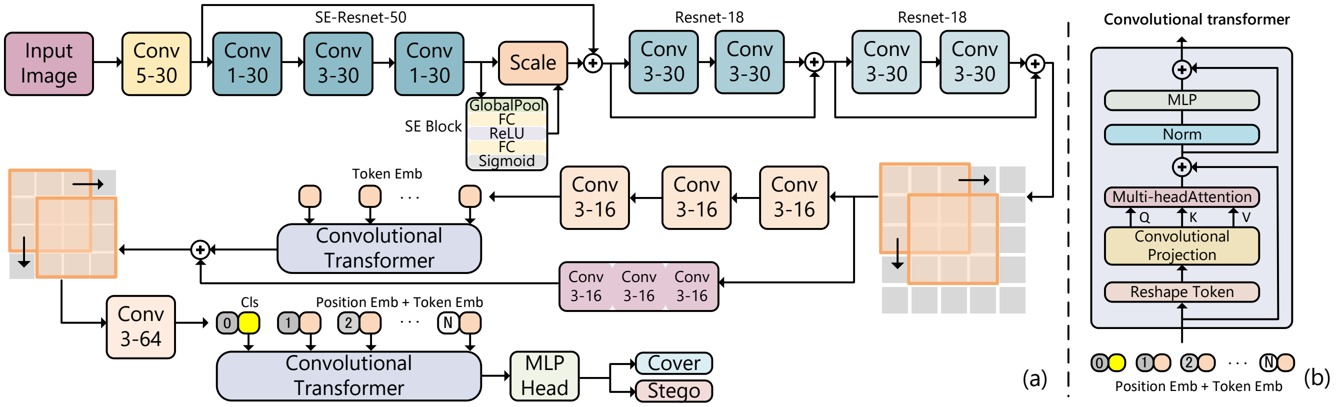


Fig. 2: (a) The detailed architecture of the proposed network. (b) Details of the Convolutional Transformer used in feature extraction phase.

2. OUR APPROACH

2.1. Network Architecture

Fig. 2(a) illustrates the detailed architecture of our proposed network, which mainly comprises three phases: preprocessing, feature extraction, classification. It should be noted that we use the "Conv[kernel size]-[the number of channels]" to denote the convolutional layers with "kernel size $\times$ kernel size $\times$ channel numbers" in Fig. 2(a). The detailed internal structure of convolutional transformer has been shown in Fig. 2(b).

In preprocessing phase, the weights of the first convolutional layer are initialized with SRM filter kernels (30 "square 5×5") proposed in [7, 12]. In Fig. 2, it is represented by the first yellow box (Conv5-30). As discussed in SRNet [14], pooling may decrease the energy of image noise residuals, and shortcut connection is beneficial to image steganalysis network. Inspired by the same strategy in [14, 16], we use a bottleneck building block (Resnet50 layer) and two basic building blocks (Resnet18 layers) [19] without pooling. In a bottleneck building block represented by three dark cyan boxes, we add an extra channel attention module named "Squeeze-and-Excitation" [20], which is simple but effective. Two light cyan boxes mark the subsequent stages (Resnet18 layers) respectively. Removing pooling from Resnet simply, each stage consists of 'Conv3-30 (or Conv1-30)' layers, batch normalization (BN) layers, rectified linear units (ReLU) in this phase.

In feature extraction phase, two modified convolutional transformer based on CVT's version [18] are applied to extract the feature vector of noise residuals from a global perspective orderly. Different from classical vision transformer used in ViT[17], output feature maps from the previous stage are mapped into new feature maps with convolution operation firstly. In consideration of the peculiarity of steganographic signal distribution, we use the large receptive field to extract features further by a convolutional layer with 7×7 kernel. In practice, tree continuous mere convolutional layers (without any BN layer, RelU, pooling layer), denoted by salmon boxes (Conv3-16) in Fig. 2, replace such a 7×7 convolutional layer in the same way as VGGNet [21]. Then output feature map is flattened into token embeddings as input into the subsequent convolutional transformer(represented by mediumpurple box). Fusing layer-wise local features output by preprocessing phase and global noise features output by convolutional transformer, three extra convolutional layers (denoted by pink boxes) are designed as a shortcut connection. After being reshaped into 2D feature map, transformer output could be processed with similar operations in the next stage (module consisting of a single convolutional layer (Conv3-64) and a convolutional transformer). Of special note is that positional embedding is added to token embedding just in the second stage in this phase.

In classification phase, regarding image steganalysis as a binary classification task, a class token is used in the second convolutional transformer for steganalysis task via an MLP (i.e. fully connected) head similar to ViT [17] and CVT [18]. When dealing with arbitrarily sized input images, it can be appended to a Size-Independent Detector(SID) [15], which is used to extract statistical moments (the maximum, minimum, variance, and average) of last output feature maps. After concatenation, it can be fed directly to fully connected layer similar to SID used in SiaStegNet [16].

2.2. Channel Attention Module

Similar to previous steganalysis networks [12, 15, 16], the first layer is initialized by 30 SRM filters (5×5). We hold that CNN-based architecture can make use of relationships among 30 channels to improve the quality of noise residuals representations with channel attention mechanism. With channel attention, strong noise features can be emphasized while the useful ones suppressed. It's worth noting and distinguishing that channel attention is far different from Knowledge of Selection Channel proposed in Yenet [12] which introduced the selection channel via a parallel branch in the first layer.

To fuse local spatial correlation learned by convolutional layer and global information from different channels, we use a Resnet-50 [19] with channel attention module named "Squeeze and Excitation" block [20]. Denoted by stacked layers (Globalpooling layer, FC layer, ReLU, FC layer, Sigmoid function) one by one in Fig. 2, 'SE' block can be used effectively with residual shortcut connection. To be specific, "Squeeze" operation compresses the input feature maps $U \in \mathbb{R}^{H \times W \times C}$ into the $1 \times 1 \times C$ maps ($1 \times 1 \times 30$), and "Excitation" operation maps subsequent output to 30 channel weights. Scaling output from three convolutional layers with these channel weights in a 'Scale' block, the final output of first stage (SE-Resnet-50) can be obtained with residual shortcut.

2.3. Convolutional transformer

Vision transformers have been successfully applied to various vision classification tasks over the past two years. Different from ViT [17], convolutional transformer retains part of convolution operations in transformer to model local spatial relationships when capturing global dependencies [18]. For image steganalysis, we argue that network which combines CNN-based architecture with convolutional transformer can model both local and global dependencies from image noise residuals.

For input into the following transformer block, feature maps can achieve spatial downsampling and enrich spatial representation via convolution operations, and then be flattened into the compatible size for transformer. In Fig. 2(b), detailed internal architecture of Convolutional Transformer has been depicted. Instead of a linear projection, convolutional transformer uses a convolutional projection before self-attention. Therefore, input tokens should be reshaped into 2D token map for subsequent separable convolution operations. Next, Projected tokens are flattened into 1D as $Q/K/V$ vectors used for multi-head attention computation. Besides convolutional projection, it just consists of a Multi-head Attention and MLP Head block with a Layernorm (LN) before each of them.

2.4. Positional Embedding

As discussed in other vision tasks, positional embedding used in transformer can provide positional information to capture global image features effectively. However, the version of CVT [18] drops the positional embedding from the whole network. They hold that convolution operations in token embedding and projection can utilize spatial information from global and local perspectives sufficiently.

In our proposed network, we still use the standard learnable 1D position embeddings, since we find that convolutional layers of limited depth may not learn global relations among noise residuals features completely, especially for detecting large steganographic images. But the remarkable thing is that we just add positional embeddings to token embeddings in the second convolutional transformer, and we remove the same operation in the first transformer in feature extraction phase. The distribution of secret messages embedded into complex areas is sparse, and signal energy is fairly weak. In the first transformer, the network may wrongly regard positional embeddings as steganographic signals, which may hurt detection performance.

2.5. Convolutional Projection

Instead of linear projection before self-attention block, convolutional projection enables transformer to emphasize local spatial information. When detecting fixed-size images (256×256), we use a unsqueezed convolutional projection with stride $= 1$ for computing $Q/K/V$ metrics. To improve efficiency for processing arbitrary-sized images (like 1024×1024), squeezed convolutional projection (the undersampling of K and V matrices) with stride $= 2$ can reduce computation cost effectively.

3. EXPERIMENTS

3.1. Experiment Settings

Two content-adaptive steganography methods, HILL [22], WOW [23], have been employed to generate stego images respectively. We use the Matlab implementations online[1] with random embedding key. Our proposed network is compared with SRnet [14] and SiaStegNet [16]. All the experimental results were obtained using Nvidia GTX 1080Ti GPU cards.

[1] http://dde.binghamton.edu/download/

Table 1: Steganalysis accuracy comparison of SRNet, SiaStegNet, and ours for two embedding algorithms at 0.2 and 0.4 bpps on fixed-size dataset.

Network \ Algorithm	HILL		WOW	
	0.2	0.4	0.2	0.4
SRNet [14]	**77.25**	85.43	**85.61**	91.73
SiaStegNet [16]	77.12	**85.86**	85.58	91.67
Ours	76.83	85.61	85.25	**92.10**

3.2. Datasets

The first dataset we used for fixed-size experiments comes from the BOSSbase 1.01 [24] consisting of 10000 native resolution images. We use the same operations as [16] to process these images. After being cropped into squares and resized to 256×256 by imresize function with the bilinear interpolation in Matlab, we get a new dataset of size 256×256, which is used for fixed-size images steganalysis. We select 12000/2000/6000 images of BOSSbase 1.01 as cover and stego images for the training/validation/test sets (without overlap among all 20000 images).

The second one for arbitrary-sized images was obtained from ALASKA #2 [25], which includes ALASKA_v2_TIFF_512 and ALASKA_TIFF_VariousSize, generally shortened to ALASKA_512 and ALASKA_VAR. ALSAKA_512 includes about 80000 512×512 images, then we select 24000/4000 images as covers and stegos for the training/validation sets randomly. ALASKA_VAR contains 16 sets of images of different sizes. We select 750 images of each set as test set (totaled 12000), and there is no overlap among all 40000 images.

3.3. Hyper-parameters

The initial learning rate is set to 0.0001, and it would be reduced to 0.00001 after 300 epochs (500 epochs in total for training). The batch size is set to 32, due to GPU memory limitation. The depths of convolutional transformers in feature extraction phase are set to 1 and 2 respectively, and parameter r in SE-Resnet-50 is set to 5.

3.4. Results on fixed-size dataset

In Table 1, we report results for the detection accuracy when steganalyzing HILL and WOW algorithms at payload 0.2 and 0.4 on BOSS_256. Grounded on a novel architecture rather different from those CNN-based methods, our network is well-matched with SRNet [14] and SiaStegNet [16] in performance for detecting fixed-size image. Our network has an accuracy of 92.10%, which is 0.43% and 0.37% higher than SiaStegNet and SRNet respectively for WOW at 0.4 bpp. Certainly, they show further excellent performances for HILL than our network. However, the capacity of fusing global and local information supplied by convolutional transformer has not been realized fully for images steganalysis with 256×256 size, which is propitious to large images relatively.

3.5. Results on arbitrary-size dataset

In deep learning, Siamese Network has been a common and efficient architecture for various vision tasks since proposed in [26]. Regarded as a backbone proposed in [16] for image steganalysis primarily, an input image is partitioned into two areas (left and right), and both of them are fed into parallel subnets (arbitrary CNN-based

Table 2: Steganalysis accuracy comparison of SiaStegNet and ours for WOW embedding algorithms on arbitrary-size dataset. 512×512 images are used for training, and images of all sizes are tested directly without any retraining.

	SiaStegNet	Our Network (Siamese)
512×512	**77.23**	73.04
512×640	**76.84**	73.95
640×512	**77.05**	76.17
512×720	**76.37**	72.39
720×512	75.91	**75.95**
640×640	**76.46**	73.82
640×720	**76.25**	72.99
720×640	**75.88**	75.39
720×720	**74.96**	73.40
512×1024	75.12	**77.60**
1024×512	75.27	**77.04**
640×1024	74.90	**76.03**
1024×640	74.03	**76.82**
720×1024	72.36	**75.91**
1024×720	73.19	**78.23**
1024×1024	71.65	**77.10**

Table 3: Steganalysis accuracy comparison of our network and compositions with different architectures.

Architecture	Image Size	
	512×512	1024×1024
Ours +Sia (Use Version)	73.71	77.24
Ours +Sia+SID	74.15	78.06
Ours +Sia−PE	72.67	75.41
Ours +Sia+PE (first stage)	72.43	75.02
Ours +Sia−Transformer	68.82	69.64
Ours +Sia−CA	72.08	75.30

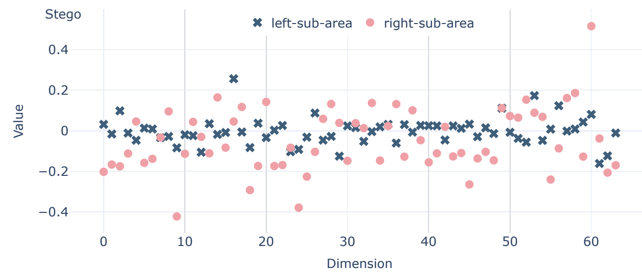
network). Capturing relationships between two areas, SiaStegNet [16] can distinguish stegos from covers efficiently.

We replace the subnet with our convolutional vision transfomer based on Siamese backbone. WOW is used at payload of 0.4 bpp to generate correspondingly 512×512 stego images. Furthermore, the payloads for algorithm on images of different sizes are adjusted for constant statistical detectability, which is processed similarly as [16] according to square root law [27].

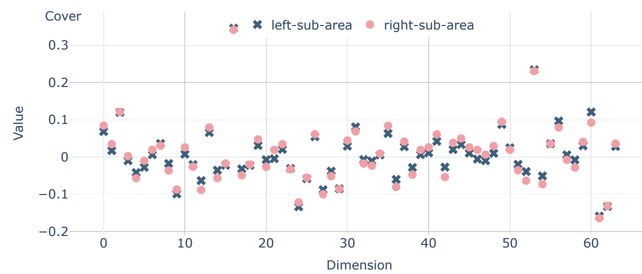
As reported in Table 2, benefiting from SID detector, the performance of SiaStegNet is superior to ours slightly when detecting images of smaller sizes (512×512 to 720×720). Obviously, when tested on large images, our network exceeds SiaStegNet, which is 2% to 5% better compared to their performances. It is notable that SID detector has not been added to our network in these experiments.

3.6. Ablation Study

Table 3 shows performance comparisons of our network and our network with different mentioned architectures when testing images of different sizes, including Siamese backbone(Sia), SID detector, Positional Embedding (PE) and Channel Attention (CA). One of the most notable results is that the detection accuracy drops sharply by 4.9% and 7.6% when we remove convolutional transformer from our network. When detecting larger images, SID detector and the



(a) Output vectors extracted from a stego



(b) Output vectors extracted from a cover

Fig. 3: 64-D vectors of two sub-nets output from last convolutional transformer block when inputting a stego and a cover respectively.

Siamese backbone can provide noticeable increases to detection performances respectively. Different from removing positional embeddings in CVT[18], we hold that position information of noise features are further underutilized, especially for detecting large images. Curiously, detection performance is decreased by 1.28% when we add positional embedding in the first convolutional transformer similarly as in the second one. A reasonable guess is that extra position information may be confused with noise residuals here. Focusing on the efficiency of transformer to process global relations for steganalysis, as depicted in Fig. 3, there is a quite difference between values from different areas at the same dimension when testing a stego image, of which values are almost equal for a cover image. For a cover, the mean value of differences on the 64 dimensions equals to 0.013, while the average for a stego reaches as high as 0.148. Therefore, convolutional transformer can extract efficient relationships for distinguishing stegos from covers.

4. CONCLUSION

In this paper, we propose a Convolutional Vision Transformer network for spatial steganalysis. Combined with previous CNN-based architecture effectively, we make further full use of self-attention provided by vision transformer for handling heterogeneous datasets.

We conclude the advantages of our network as follows: (1) Fusing channel attention into preprocessing phase, information can be utilized to produce globalized image residuals. (2) We use convoluntional transformer to extract features of noise residuals from both local and global perspectives in feature extraction phase. (3) We add positional embeddings to tokens embedding at a reasonable step to enhance global attention for extra improvement on detection accuracy. Future work will pivot on novel architectures based on vision transformer for image steganalysis.

5. REFERENCES

[1] Vahid Sedighi, Rémi Cogranne, and Jessica Fridrich, "Content-adaptive steganography by minimizing statistical detectability," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 2, pp. 221–234, 2015.

[2] Tomáš Pevnỳ, Tomáš Filler, and Patrick Bas, "Using high-dimensional image models to perform highly undetectable steganography," in *International Workshop on Information Hiding*. Springer, 2010, pp. 161–177.

[3] Siwei Lyu and Hany Farid, "Detecting hidden messages using higher-order statistics and support vector machines," in *International Workshop on information hiding*. Springer, 2002, pp. 340–354.

[4] Ismail Avcibas, Nasir Memon, and Bülent Sankur, "Steganalysis using image quality metrics," *IEEE transactions on Image Processing*, vol. 12, no. 2, pp. 221–229, 2003.

[5] İsmail Avcıbaş, Mehdi Kharrazi, Nasir Memon, and Bülent Sankur, "Image steganalysis with binary similarity measures," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 17, pp. 1–9, 2005.

[6] Tomáš Pevny, Patrick Bas, and Jessica Fridrich, "Steganalysis by subtractive pixel adjacency matrix," *IEEE Transactions on information Forensics and Security*, vol. 5, no. 2, pp. 215–224, 2010.

[7] Jessica Fridrich and Jan Kodovsky, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012.

[8] Tomas Denemark, Vahid Sedighi, Vojtech Holub, Rémi Cogranne, and Jessica Fridrich, "Selection-channel-aware rich model for steganalysis of digital images," in *2014 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2014, pp. 48–53.

[9] Weixuan Tang, Haodong Li, Weiqi Luo, and Jiwu Huang, "Adaptive steganalysis against wow embedding algorithm," in *Proceedings of the 2nd ACM workshop on Information hiding and multimedia security*, 2014, pp. 91–96.

[10] Yinlong Qian, Jing Dong, Wei Wang, and Tieniu Tan, "Deep learning for steganalysis via convolutional neural networks," in *Media Watermarking, Security, and Forensics 2015*. International Society for Optics and Photonics, 2015, vol. 9409, p. 94090J.

[11] Guanshuo Xu, Han-Zhou Wu, and Yun-Qing Shi, "Structural design of convolutional neural networks for steganalysis," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 708–712, 2016.

[12] Jian Ye, Jiangqun Ni, and Yang Yi, "Deep learning hierarchical representations for image steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 11, pp. 2545–2557, 2017.

[13] Mehdi Yedroudj, Frédéric Comby, and Marc Chaumont, "Yedroudj-net: An efficient cnn for spatial steganalysis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2092–2096.

[14] Mehdi Boroumand, Mo Chen, and Jessica Fridrich, "Deep residual network for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1181–1193, 2018.

[15] Clement Fuji Tsang and Jessica Fridrich, "Steganalyzing images of arbitrary size with cnns," *Electronic Imaging*, vol. 2018, no. 7, pp. 121–1, 2018.

[16] Weike You, Hong Zhang, and Xianfeng Zhao, "A siamese cnn for image steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 291–306, 2020.

[17] 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.

[18] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang, "Cvt: Introducing convolutions to vision transformers," *arXiv preprint arXiv:2103.15808*, 2021.

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

[20] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[21] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] Bin Li, Ming Wang, Jiwu Huang, and Xiaolong Li, "A new cost function for spatial image steganography," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 4206–4210.

[23] Vojtěch Holub and Jessica Fridrich, "Designing steganographic distortion using directional filters," in *2012 IEEE International workshop on information forensics and security (WIFS)*. IEEE, 2012, pp. 234–239.

[24] Patrick Bas, Tomáš Filler, and Tomáš Pevnỳ, "Break our steganographic system: the ins and outs of organizing boss," in *International workshop on information hiding*. Springer, 2011, pp. 59–70.

[25] R. Cogranne, Q. Giboulot, and P. Bas., "ocumentation of alaskav2 dataset scripts: A hint movingtowards steganography and steganalysis into the wild," `https://alaska.utt.fr/`, 2019, Accessed: 2021-05-02.

[26] Sumit Chopra, Raia Hadsell, and Yann LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE, 2005, vol. 1, pp. 539–546.

[27] Andrew D Ker, Tomáš Pevnỳ, Jan Kodovskỳ, and Jessica Fridrich, "The square root law of steganographic capacity," in *Proceedings of the 10th ACM workshop on Multimedia and security*, 2008, pp. 107–116.