En-VStegNET: Video Steganography using spatio-temporal feature enhancement with 3D-CNN and Hourglass

Aman Jaiswal
Dept. of Computer Science
IIT Dharwad
Dharwad, India
160010004@iitdh.ac.in

Suraj Kumar
CEA Department
GLA University
Mathura, India
csksuraj17@gmail.com

Aditya Nigam
SCEE
IIT Mandi
Mandi, India
aditya@iitmandi.ac.in

***Abstract*—Learning Spatio-temporal features has shown improved performance on tasks involving video analysis using deep learning, and the deep learning community has used these features to solve a varied variety of problems. Video steganography is one such problem where learning these features for a video can help improve the performance of steganography. Steganography is the practice of concealing confidential information, to protect the information from an adversary, into an ordinary cover message in a way that the cover message does not seem suspicious to the adversary. Recent deep-learning-based steganography methods have proven to improve the secrecy and capacity of steganography over traditional techniques. In this paper, we propose a novel state-of-the-art deep 3D-CNN architecture with enhancement feature learning for full video steganography. The proposed model outperforms the current state-of-the-art methods for full video steganography both qualitatively and quantitatively. We have validated our model by comparing it with new as well as traditional steganography techniques, on quality and different statistical metrics, namely, PSNR, SSIM, APD, VIF at the frame, and video level. Moreover, to check the undetectability of our model, we have subjected our model to detection by steganalysis tools like SRNet. Results of fine-tuning classifiers, like ResNet and Inception-v3, to detect steganographic messages from ordinary messages maintains our model's undetectability and accuracy.**

I. Introduction

In everyday life, we come across a lot of data whose confidentiality, secrecy, and ownership must be ensured. For instance, with the advent of cloud-storage [13], many individuals and organizations prefer to store their data on the cloud, as it provides a mechanism to conveniently and easily access and share data over the network. Preventing such information from being disclosed is of utmost importance, as these data may contain crucial confidential information.

Cryptography and Steganography [5]–[7], [14] both provide methods to ensure the security of confidential data on public channels. Fig. 1 highlights the central difference between cryptography and steganography techniques. In cryptography, Alice sends a secret message to Bob on a public channel by encoding the plain-text into a cipher-text using some encryption algorithm, on receiving the message, Bob uses the corresponding decryption algorithm to reveal the secret message. However, in steganography, Alice hides the secret message into a different cover message using a hiding algorithm such that the cover message appears undefiled and sends it to Bob on the public channel. On the receiving end, Bob reveals the secret message using the corresponding revealing algorithm. In both situations, there is Eve, who is eavesdropping the conversation between Alice and Bob. In the case of cryptography, looking at the cipher-text, Eve can guess the presence of secret communication between Alice and Bob and in some cases, even decipher the secret message. In the case of steganography, it is difficult for Eve to identify secret communication because the steganographic message is indistinguishable from the cover message.

The significant advantage of Steganography over Cryptography is that cryptography techniques only focus on hiding the secret message and not the existence of secret communication, whereas steganography techniques provide a method for communication that the adversary does not deem suspicious. Other than covert communication, steganography can be used for digital watermarking [6], [15] without compromising the integrity of the cover message.

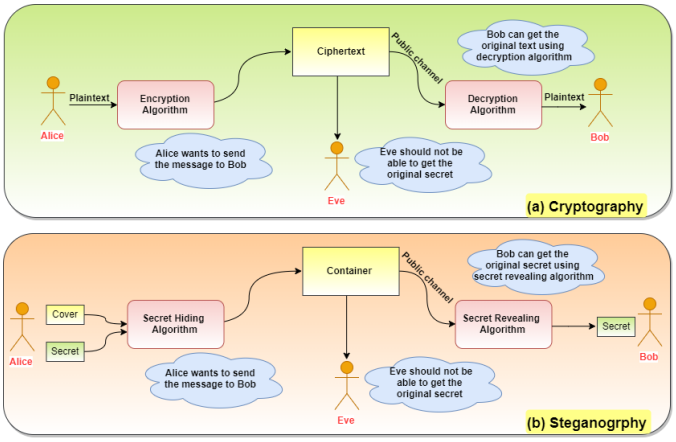

Fig. 1: Cryptography vs Steganography

Over the years, steganography has become a popular technique for covert communication as well as digital water-

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

marking, as it conceals not only the message but also hides the mere existence of a secret message, which is essential for privacy-sensitive communication. The main feature of steganography is also a challenge qualifying the performance of steganography. Designing a steganography algorithm that not only hides the information within a cover message but also preserves the integrity of the cover message as well as allows for recovery of the secret message is a challenging task. Embedding information within a cover message can alter the visual appearance as well as the underlying statistics of the cover, making it prone to discovery by visual or statistical analysis. The performance of a steganography algorithm is judged on its *Undetectability*: similarity between cover message and container message (steganographic message), *Capacity*: the amount of information embedded in the cover message and *Reproducibility*: how accurately the secret message can be recovered from the container message. Capacity and Undetectability are highly correlated; longer the secret message, higher will be the capacity, and more the cover will be altered, making it susceptible to discovery.

Video is continuing to grow both in popularity and importance all across the internet, therefore, video steganography [7], [9], [17] has recently started to gain traction in the research community. Using image steganography techniques for frame-wise video steganography can be thought of as a possible solution but it is not necessarily optimal, as it does not take into consideration the temporal coherence between successive video frames. Recently, Kumar et al. proposed VStegNET [7] in BMVC'19, for full-video steganography and VStegNET is the current state-of-the-art model to solve the problem of full video steganography. For this paper, we further explored the task of hiding a full-sized video into another video of the same size and proposed a novel deep 3D CNN architecture inspired by traditional auto-encoders [18] for full video steganography. The proposed model outperforms the current state-of-the-art VStegNET [7]. Our key contributions are as follows:

- A novel deep 3D CNN architecture, outperforming the current state-of-the-art VStegNET (BMVC'19) [7] for full video steganography.
- Qualitative as well as Quantitative analysis of the model at both frame and video level with metrics like APD [5], [9], SSIM [24], PSNR [25], and VIF [26] to show the effectiveness of our proposed model.
- Rigorously tested the 'undetectability' of the model by testing it against traditional as well as deep steganalysis tools.
- Experimental analysis along with ablation study, using payload capacity, failure cases, drawbacks, etc. of the model is done to test the generalizability and to validate the performance of the proposed model.
- Superiority of the model is maintained by comparing the model with other state-of-the-art models like NIPS'17 [5], HCCVS [9], and VStegNET [7].

The rest of the paper is organized as follows: Section II describes some of the well known state-of-the-art techniques for digital steganography. Section III explains our proposed methodology, subsequently, the results and experiments are explained in section IV. Finally, we conclude our proposal in section V with future remarks.

II. Related work

One of the provincial methods of steganography is the LSB (Least Significant Bit) steganography [1] which, as the name suggests, hides the secret message in the least significant bits of the cover image. To ensure that the variation in the cover image is minimal, manipulating only the least significant bits is a good strategy; however, LSB steganography loses the information from the cover image. Since LSB uses a hand-crafted technique for hiding messages, by design, the steganographic images produced are not visually different, but it alters the underlying statistics of the cover image, making it prone to reliable detection by steganalysis [19].

More sophisticated methods have been designed that preserve the underlying image statistics and work on designing distortion functions that force the embedding process to localize to more noisy and challenging to model parts of the image. Advanced steganographic techniques focus on minimizing the designed distortion functions between the cover and the steganographic image.

All distortion based steganographic techniques have the same end goal: to localize the information to more noisy and complex regions of the image by minimizing the distortion function; they differ only in their approach of defining the distortion function. Highly Undetectable steGO (HUGO) [2] is one of the most secure and content-adaptive steganography technique that hides the secret payload spatially in the image. The distortion function is based on Subtractive Pixel Adjacency Matrix (SPAM) [20] feature vectors to adaptively identify noisy regions or complex textures in the image to hide the payload. Likewise, Wavelets Obtained Weights (WOW) [3] is an additive steganography technique having the same capacity as HUGO. S-Uniward [4] uses, "distortion function based on the sum of relative changes of coefficients in a directional filter bank decomposition of the cover". The only problem with these techniques is their low capacity of only 0.2 bits per pixel (bpp).

With advancements in deep learning research, recent steganography techniques pose steganography as an unsupervised learning task and train deep neural networks that have shown to outperform traditional steganography algorithms to achieve a capacity as high as 8 bpp. Recently, Baluja et al. [5] proposed an image steganography technique that uses deep convolutional neural networks to hide a complete image within another image. They use an autoencoder based deep learning model for the task and train it with the weighted sum of reconstruction loss between the secret and revealed secret image and the cover and container image. They have achieved an embedding capacity of 100%. HiDDeN [6] is another deep learning model for not only image steganography but also watermarking. Their system hides an n-bit message within

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

an image such that the steganographic image is robust to perturbations like blurring, cropping, and lossy compression.

Deep learning techniques used for full video steganography, where the goal is to hide one complete video into another video of the same size, are most relevant to our work. Video steganography is becoming popular with the research community as videos have a temporal dimension that provides more redundancy and proves to be useful in hiding the secret message and increasing the capacity of steganography. Recently, Weng et al. proposed HCCVS (ICMR'19) [9], which uses 2D-convolutions with temporal residual modeling to hide and reveal the secret message. They use two convolutional neural networks: one to hide the reference frames and the other to hide residual frames and train the networks to solve the problem of full video steganography. In contrast to the HCCVS approach, Kumar et al. proposed VStegNET (BMVC'19) [7], which uses 3D-convolutions [16] with hourglass network [21] to inherently do the temporal modeling and achieved a payload capacity of 24 bpp. VStegNET [7] is the current state-of-the-art model for full video steganography.

III. Proposed Methodology

Our proposed architecture consists of two connected deep 3D convolutional neural networks: a hiding network (HN), which is used by Alice to hide the secret video frames into cover video frames thereby generating container video frames, and a revealing network (RN), which is used by Bob to extract the secret video frames from the container video frames as shown in Fig. 2. Both hiding network (HN) and revealing network (RN) are based on traditional encoder-decoder architecture, where the frames go through a series of down-sampling steps followed by a series of up-sampling steps. Our network resembles the down-sampling step of the encoder, but we have designed a gradual up-sampling step for the decoder, as compared to successive up-sampling. The network specifications are described in detail in section III-A.

In this paper, we have used 3D-convolutions to exploit the Spatio-temporal [16], [22], [23] relationship between consecutive video frames and appropriately designed skip-connections, max-pooling, and up-sampling layers to make the model most suitable for full video steganography.

A. Network specifications

The overall architecture consists of two connected deep 3D convolutional neural networks:

- **Hiding Network (HN)** : HN takes 8 cover video frames and 8 secret video frames, concatenated along the temporal dimension, as input. The network comprises of 5 levels of feature extraction and reconstruction each. In the feature extraction phase, the feature maps are convoluted with 3 3D-convolutions of shape $3 \times 3 \times 3$, each applied with 32, 64, 128, 256, and 512 filters at consecutive levels, respectively. After convolution, at every level except the last, the feature maps are down-sampled using max-pooling as shown in Fig. 3. In the reconstruction phase, before convolution operation, all the feature maps from

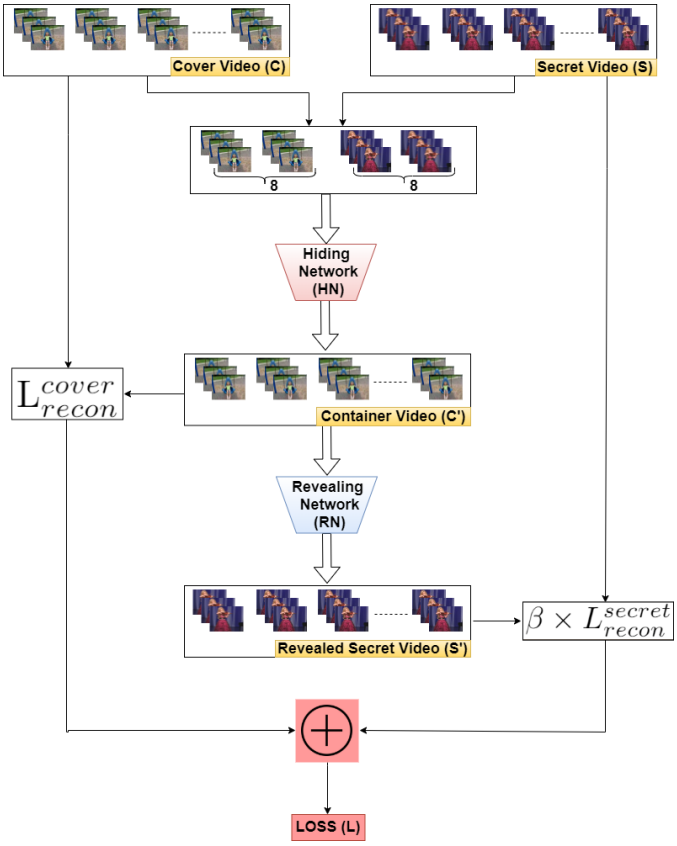

Fig. 2: Proposed Overall Architecture

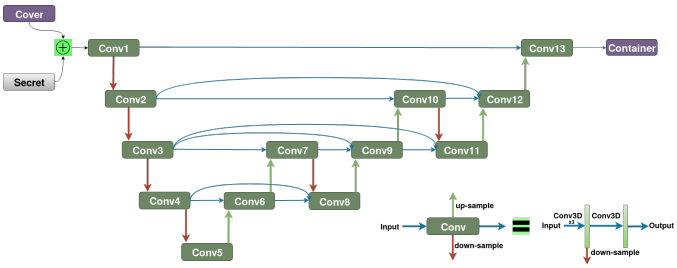

Fig. 3: Overview of the Hiding Network

the corresponding level are collected and concatenated with the current feature map, except for the first level. As the output of 'Conv1' is $16 \times 320 \times 240 \times 32$ and the output 'Conv12' followed by up-sampling is $8 \times 320 \times 240 \times 64$, a temporal pooling of 'Conv1' is required to be able to concatenate 'Conv1' with up-sampled 'Conv12'. The skip-connections at each level vouch for any feature loss that might happen during the extraction or reconstruction phase as demonstrated by [28]. In contrary to traditional decoders, which continuously up-sample the feature maps for reconstruction, our proposed architecture undergoes a down-sampling step every two up-sampling steps, which eventually helps in feature enhancement at every level of reconstruction. The output of HN is the 8 container video frames of dimension $320 \times 240 \times 3$ each, which

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

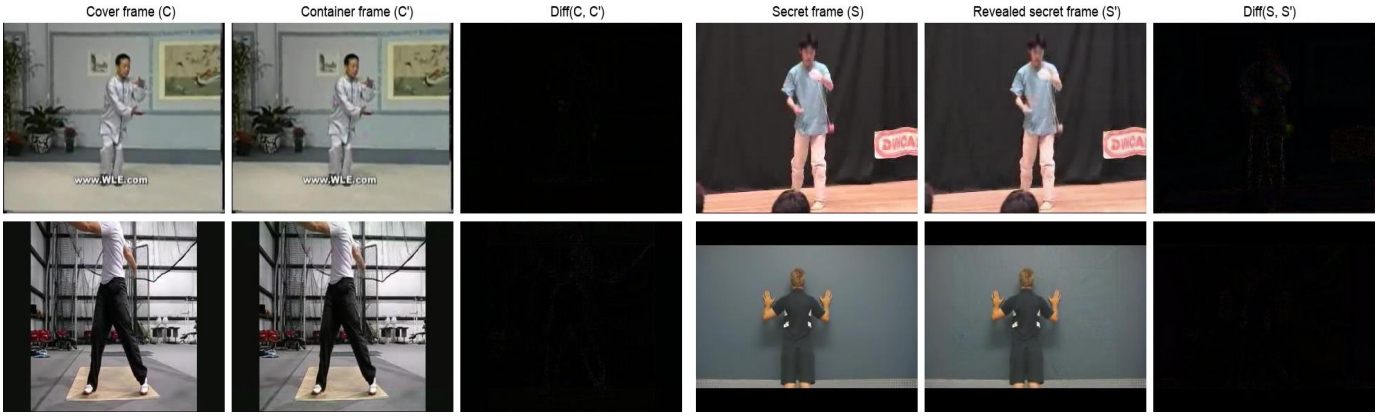

Fig. 4: Results of our model showcasing the container (Column 2) and revealed secret (Column 5) frames generated by hiding and revealing networks, respectively, along with the difference maps between the cover-container (Column 3) and secret-revealed secret (Column 6) video frames.

are visually similar to the cover video frames.

- **Revealing Network (RN)** : The architecture of RN is exactly similar to the HN, except for two design requirements: First, the input to RN is 8 container video frames of dimension $320 \times 240 \times 3$ each. Second, because the input is only 8 container video frames, there is no need for an additional temporal pooling for skip connection between 'Conv1' and up-sampled 'Conv12'. The output of RN is 8 revealed secret frames of dimension $320 \times 240 \times 3$ each.

B. Loss function

The loss function to regularize the training of the proposed model should be designed in the way that optimization of the loss function should:

- Ensure that the container frames are visually indistinguishable from the cover frames.
- Ensure clear reconstruction of the hidden secret frames

The loss function of our proposed model is described below:

$$\begin{aligned} Loss(C, C', S, S') &= L^{cover}_{container} + \beta * L^{secret}_{revealed\_secret} \\ &= ||C - C'||_2 + \beta * ||S - S'||_2 \end{aligned}$$

The term β is introduced as to weight the reconstruction errors between cover-container video frames and secret-revealed secret video frames.

IV. Experiments and Results

As a proof of concept, we trained our model on an action recognition dataset, UCF101 [8]. Fig. 4 showcases some of the results of our model. It is evident from Fig. 4 that our model produces container frames which are visually indistinguishable from the cover frames, also, the revealed secret frames are very similar in appearance to the original secret frames. Difference maps between the cover and container video frames as well as the secret and revealed secret video frames shown in column 3 and column 6 of Fig. 4, respectively, justifies our claim.

For Quantitative analysis, the performance of our model is evaluated on three different axes, namely: *Undetectablity*, *Capacity* and *Reproducibility*.

- Undetectablity: difficulty in detecting the hidden message. It is measured as the robustness of the model against various steganalysis tools. As a proxy to this, similarity between the cover and container message can also be measured using metrics like, Peak Signal-to-Noise Ratio **(PSNR)**, Absolute Pixel Discrepancy **(APD)**, Visual Information Fidelity [26] **(VIF)** and Structural Similarity Index **(SSIM)**.
- Capacity: number of message bits that are hidden per video bit. Primarily measured in Bits Per Pixel **(BPP)**.
- Reproducibility: similarity between the original secret frames and revealed secret frames. It can be measured with the same metrics as undetectability namely, **PSNR, SSIM, APD** & **VIF**.

The goal is to achieve higher PSNR, VIF, and SSIM and low APD.

All the experiments in this paper were performed on a workstation with Nvidia Geforce 1080Ti GPU Card.

A. Dataset Specifications

UCF101 dataset [8] is an action recognition dataset of 13,320 videos. This dataset was chosen because it comprises of videos with variation in the camera motion, different illumination conditions, different object scale, pose and viewpoint, cluttered background, etc. The size of each frame of the video is $320 \times 240 \times 3$, with a mean clip size of 7.21 seconds with a rate of 25 frames per second.

B. Video and Frame sampling

The input to the model is 8 frames of both cover and secret video, so frames are extracted from both cover and secret videos. These two videos can be of different length or different frame rates, therefore, the number of frames in both the videos can be different. As the number of frames is different, we do

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

a temporal equalization before feeding them to the model, as suggested by BMVC'19 VStegNET [7].

Temporal equalization is done as follows: Assume N_1 and N_2 are the number of frames in the two chosen videos since the model processes 8 frames of both the videos at a time, the highest multiple of 8 that is smaller than or equal to the minimum of N_1 or N_2 is chosen as N, the number of frames fed to the model.

$$N = min(N_1, N_2) - min(N_1, N_2) mod 8$$

C. Training and Testing methodology

The complete dataset is partitioned into training and test set with 10,000 videos in the training set and the remaining 3,320 videos in the testing set. The model was trained end-to-end using Adam optimizer with Loss(C, C', S, S') as an error signal and learning rate $\alpha = 1e-4$.

For training, any two videos are randomly sampled from the set of 10,000 videos (Total possible combination = ${}^{10000}C_2$), without loss of generality one is taken to be cover video and the other as secret video. The training was done for two values of β, 0.75, and 1.00. For $\beta = 0.75$ our proposed model converged after training on 6,250 video pairs and for $\beta = 1.00$, it converged after 8,500 video pair. The converged model was observed to generalize well over unseen test pairs as well.

For testing, we generated a list of 500 video pairs from the 3,320 videos in the testing set (Total possible combination = ${}^{3320}C_2$). The testing was done on these fixed 500 video pairs so that comparison with other state-of-the-art techniques can be reported. The results reported in Table I and II for different video steganography models are averaged over all these fixed 500 test samples.

D. Comprehensive Results

To show the superiority of our model, we have compared the results of our model with other well-know and state-of-the-art models for steganography.

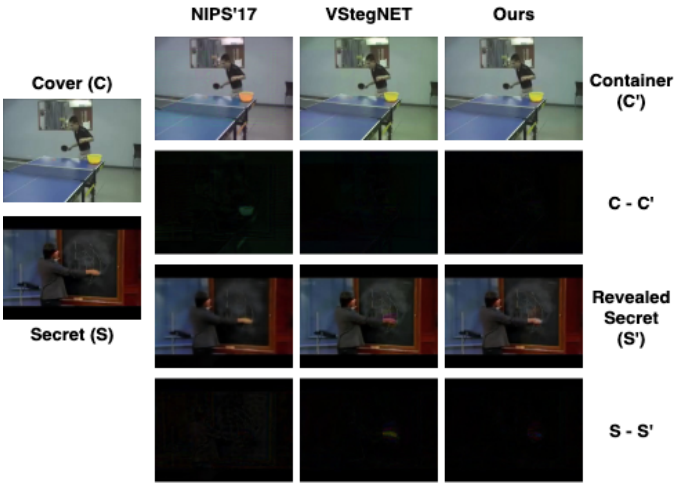

Fig. 5: Qualitative comparative analysis between NIPS'17 Image model [5], BMVC'19 VStegNET [7], and Our model.

In table I and table II, we have reported the comparative results of our proposed model against LSB steganography [1], NIPS'17 Image model [5], ICMR'19 HCCVS [9] and BMVC'19 VStegNET [7], on different performance metrics. The code for LSB [1] and HCCVS [9] was not available so we have borrowed the results from VStegNET [7], whereas, we have trained NIPS'17 Image model and BMVC'19 VStegNET on our dataset and the results for the same are reported. The results of all the performance metrics show that our model outperforms other well-known state-of-the-art steganography techniques.

Model	$\|\|C - C'\|\|_2$	$\|\|S - S'\|\|_2$
LSB [1]	6.64	8.64
Baluja et. al. [5]	6.31	4.97
HCCVS [9]	3.80	5.84
VStegNET [7]	3.23	4.70
Ours	**2.87**	**4.11**

TABLE I: Comparative Results of our proposed method on APD metric

Model		**VStegNET**		**Ours**	
		β **= 0.75**	β **= 1.0**	β **= 0.75**	β **= 1.0**
APD	**C - C'**	3.23	3.51	**2.87**	2.98
	S - S'	4.70	4.68	4.11	**3.97**
PSNR	**C - C'**	35.57	34.97	**36.62**	36.27
	S - S'	31.60	31.88	32.88	**33.24**
SSIM	**C - C'**	0.94	0.94	**0.95**	**0.95**
	S - S'	0.92	0.93	0.93	**0.94**
VIF	**C - C'**	0.71	0.69	**0.74**	0.72
	S - S'	0.60	0.60	**0.65**	**0.65**

TABLE II: Comparison of our model on different statistical metrics with VStegNET [7].

The qualitative results of Our model, NIPS'17 Image model [5] and BMVC'19 VStegNET [7] are shown in Fig. 5 and 7. The comparison in Fig. 5 show that our model generates a container frame that is color accurate and closer in visual appearance to the cover frame. The container frame generated using the NIPS'17 Image model has visible traces of the secret frame also evident in the difference-map, whereas the one generated using VStegNET has a yellow hue all across the image, making it suspicious to an adversary. Not only the container frame, but the secret revealed by our model is also sharper as compared to blurry results of NIPS'17 Image model and BMVC'19 VStegNET. The complete revealed secret frame is smooth in case of NIPS'17 Image model, whereas noticeable blurriness is present near the hand of the person in case of BMVC'19 VStegNET.

Fig. 7 helps in justifying our claims, wherein we zoom-in a 135×135 patch from both the container and revealed secret frames. The zoomed-in patches of the container frame show traces of the man's face from the secret frame in case of both NIPS'17 Image model and BMVC'19 VStegNET, which is

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

(a) Hiding Network activation maps

(b) Revealing Network activation maps

Fig. 6: Activation maps generated by Hiding and Revealing networks at the same level because of sequence of up-sampling and down-sampling steps.

easily noticeable to an adversary, whereas no such trace can be identified in our case depicting the better quality of our container frames. The zoomed-in patch of the secret frames is smooth both in the case of NIPS'17 Image model and BMVC'19 VStegNET, whereas our model reveals a much sharper secret.

Fig. 7: Visual comparison of our model with NIPS'17 image model [5] and BMVC'19 VStegNET [7]. First and third column shows the container and revealed secret frames generated by corresponding methods and the second column shows 135×135 patch which is zoomed to show the superiority of Our model.

E. Performance analysis

To investigate the performance gain, we analyzed the architecture and feature maps generated by both hiding and revealing networks. Our approach is similar to BMVC'19 VStegNET [7] in using an encoder-decoder architecture for both the hiding and revealing networks. However, our decoder is different in the way described in section III-A which helps to gradually learn and improve upon the features required to hide the secret video frames in case of hiding network and to reconstruct the secret video frames from the container video frame in case of revealing network.

Fig. 6a & 6b show activation maps generated at level 2 by hiding and revealing network, respectively, for both the feature extraction and reconstruction phases. As we can see in Fig. 6a the second feature map during the reconstruction phase is more similar to its correspondent in the extraction phase as compared to the first one at the same level, showing that gradual up-sampling helps in improving upon the features required to hide the secret frames. Whereas in Fig. 6b we can see that the second feature map during reconstruction reveals more information about the secret hidden as compared to the first, thereby proving that gradual up-sampling helps enhance the activation maps at the time of revealing as well.

Model	Accuracy
Inception-V3	0.52
ResNet50	0.49

TABLE III: Classification Accuracy: Both Inception [11] and ResNet [10] were fine tuned for the task of detection starting with weights of ImageNet [12]

F. Steganalysis

Steganalysis [19], [27], the study of detecting messages hidden using steganography, is a parallel research field to steganography. We have tested the undetectability of our model by subjecting it to state-of-the-art steganalysis techniques.

In general, the cover message is destroyed after hiding the secret message. However, we analyzed the scenario where the adversary somehow has access to labeled cover and container video frames by training ResNet [10] and Inception [11] classifiers on these labeled video frames, the results for which are reported in table III. The classification accuracy of the two models shown in Table III shows nothing but random guessing.

Recently, steganalysis tools based on deep learning have proven to be better than traditional machine learning techniques which relied on hand-crafted SRM [27] features. We have tested our model against SRNet [19], a deep-learning-based technique for steganalysis. Fig. 8 tells us the number of training examples required for the adversarial networks to achieve almost perfect accuracy of detection. It is evident from the figure that our proposed model requires almost double the number of samples (approximately 30000) than state-of-the-art BMVC'19 VStegNET [7].

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

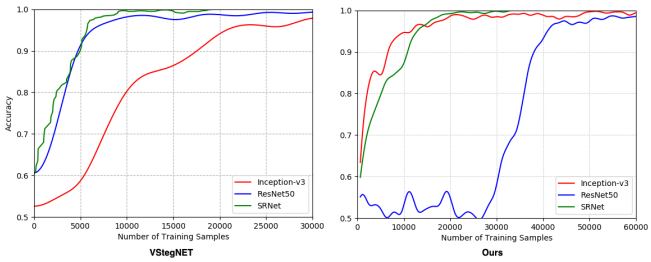

Fig. 8: Robustness analysis against adversarial attacks

We have also subjected the container frames generated by our model to revelation by revealing networks of NIPS'17 Image model [5] and BMVC'19 VStegNET [7], the results of which are shown in Fig. 9. Meaningless revealed secrets by [5], [7] demonstrate that the container frames generated by our model are secure to revelation by other steganography models.

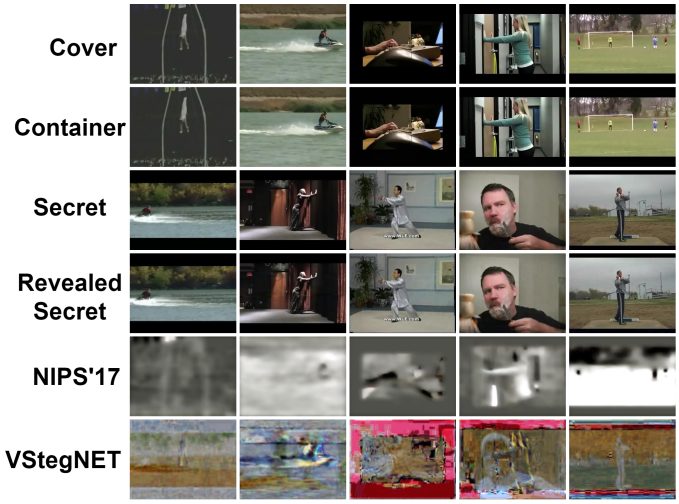

Fig. 9: Revealing network of NIPS'17 [5] and VStegNET [7] produces meaningless results when given the container images generated by our model

V. Conclusion and Future Work

In this paper, we have proposed a novel deep 3D convolutional neural network for full video steganography, inspired by traditional encoder-decoder architecture and Spatio-temporal modeling of videos, which outperforms the current state-of-the-art methods. We have demonstrated the superiority of the network by comparing, both qualitatively and quantitatively, with well-know state-of-the-art methods for digital steganography.

A minor extension will be to make β trainable and allow the network to learn the weights of the reconstruction losses. A major extension over the present model will be to analyze the effect of perturbations like lossy compression, blurring, cropping, trimming, etc. to the container video and make the model robust to such kind of alterations. Appropriately modeling the redundancy between the secret video frames can help increase capacity. These are the two major improvements that can be done to the present technique and we plan to incorporate these changes in our current state-of-the-art method for video steganography.

References

[1] D. Neeta, K. Snehal, and D. Jacobs. Implementation of lsb steganography and its evaluation for various bits. In 2006 1st International Conference on Digital Information Management, pages 173–178, Dec 2007. doi: 10.1109/ICDIM.2007.369349

[2] Tomáš Pevný, Tomáš Filler, and Patrick Bas. 2010. Using high-dimensional image models to perform highly undetectable steganography. In Proceedings of the 12th international conference on Information hiding (IH'10), Rainer Böhme, Philip W. L. Fong, and Reihaneh Safavi-Naini (Eds.). Springer-Verlag, Berlin, Heidelberg, 161-177.

[3] Holub, Vojtech and Jessica J. Fridrich. "Designing steganographic distortion using directional filters." 2012 IEEE International Workshop on Information Forensics and Security (WIFS) (2012): 234-239.

[4] Vojtˇech Holub, Jessica Fridrich, and Tomáš Denemark. Universal distortion function for steganography in an arbitrary domain.EURASIP Journal on Information Security,2014(1):1, Jan 2014. ISSN 1687-417X. doi: 10.1186/1687-417X-2014-1.

[5] Baluja, Shumeet. "Hiding Images in Plain Sight: Deep Steganography." NIPS (2017).

[6] Zhu, J., Kaplan, R., Johnson, J., & Fei-Fei, L. (2018). HiDDeN: Hiding Data With Deep Networks. ArXiv, abs/1807.09937.

[7] Mishra Aayush, Kumar Suraj, Nigam Aditya, Islam Saiful, "VStegNET: Video Steganography Network using Spatio-Temporal features and Micro-Bottleneck", BMVC 2019, https://bmvc2019.org/wp-content/uploads/papers/0966-paper.pdf.

[8] Soomro, Khurram and Zamir, Amir Roshan and Shah, Mubarak (2012), UCF101: A dataset of 101 human actions classes from videos in the wild, arXiv preprint arXiv:1212.0402

[9] Xinyu Weng, Yongzhi Li, Lu Chi, Yadong Mu. 2019. High-Capacity Convolutional Video Steganography with Temporal Residual Modeling. In 2019 International Conference on Multimedia Retrieval (ICMR'19), June 10–13, 2019, Ottawa, ON, Canada. ACM, New York, NY, USA. 9 pages. DOI: https://doi.org/10.1145/3323873.3325011.

[10] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778.doi: 10.1109/CVPR.2016.90

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 2818-2826. doi: 10.1109/CVPR.2016.308

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 6 (May 2017), 84-90. DOI: https://doi.org/10.1145/3065386

[13] Sun, Y., Zhang, J., Xiong, Y., & Zhu, G. (2014). Data Security and Privacy in Cloud Computing. International Journal of Distributed Sensor Networks. https://doi.org/10.1155/2014/190903

[14] Zenon Hrytskiv, Sviatoslav Voloshynovskiy, and Yuriy B. Rytsar. Cryptography and steganography of video information in modern communications. 1998.

[15] Nagai, Yuki et al. "Digital watermarking for deep neural networks." International Journal of Multimedia Information Retrieval 7 (2018): 3-16.

[16] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15, pages 4489–4497, Washington, DC, USA, 2015. IEEE Computer Society. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.510.

[17] Sadek, M. M., Khalifa, A. S., & Mostafa, M. G. M. (2014). Video steganography: a comprehensive review. Multimedia Tools and Applications, 74(17), 7063–7094. doi:10.1007/s11042-014-1952-z

[18] Pierre Baldi. 2011. Autoencoders, unsupervised learning and deep architectures. In Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning workshop - Volume 27 (UTLW'11), Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham Taylor, and Daniel Silver (Eds.), Vol. 27. JMLR.org 37-50.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

[19] M. Boroumand, M. Chen and J. Fridrich, "Deep Residual Network for Steganalysis of Digital Images," in IEEE Transactions on Information Forensics and Security, vol. 14, no. 5, pp. 1181-1193, May 2019. doi: 10.1109/TIFS.2018.2871749
[20] T. Pevny, P. Bas and J. Fridrich, "Steganalysis by Subtractive Pixel Adjacency Matrix," in IEEE Transactions on Information Forensics and Security, vol. 5, no. 2, pp. 215-224, June 2010. doi: 10.1109/TIFS.2010.2045842
[21] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In ECCV, 2016.
[22] Taylor G.W., Fergus R., LeCun Y., Bregler C. (2010) Convolutional Learning of Spatio-temporal Features. In: Daniilidis K., Maragos P., Paragios N. (eds) Computer Vision – ECCV 2010. ECCV 2010. Lecture Notes in Computer Science, vol 6316. Springer, Berlin, Heidelberg
[23] Xie S., Sun C., Huang J., Tu Z., Murphy K. (2018) Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. In: Ferrari V., Hebert M., Sminchisescu C., Weiss Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol 11219. Springer, Cham
[24] Zhou Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," in IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, April 2004. doi: 10.1109/TIP.2003.819861
[25] A. Horé and D. Ziou, "Image Quality Metrics: PSNR vs. SSIM," 2010 20th International Conference on Pattern Recognition, Istanbul, 2010, pp. 2366-2369. doi: 10.1109/ICPR.2010.579
[26] H. R. Sheikh and A. C. Bovik. 2006. Image information and visual quality. IEEE Transactions on Image Processing 15, 2 (2006), 430–444
[27] J. Fridrich and J. Kodovsky, "Rich Models for Steganalysis of Digital Images," in IEEE Transactions on Information Forensics and Security, vol. 7, no. 3, pp. 868-882, June 2012. doi: 10.1109/TIFS.2012.2190402.
[28] Ronneberger, Olaf et al. "U-Net: Convolutional Networks for Biomedical Image Segmentation." ArXiv abs/1505.04597 (2015): n. pag.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE