APPENDIX
VARIABLE LENGTH AND VARIABLE QUALITY AUDIO STEGANOGRAPHY

Anonymous authors
Paper under double-blind review

A NETWORK ARCHITECTURES

This section details the network architecture design. All convolutional layers here adopt Weight Standardization (Qiao et al., 2019) and Group Normalization (Wu & He, 2018) in order to handle with the small batch size given the limited computational resource. We use ReLU activation function before all convolutional layers.

Layer	Concatenate	Resample	Output Shape
Container	-	-	c ← 224 x 224 x 3
Secret	-	-	s ← 224 x 224 x 2
Conv	(c, s)	-	e0 ← 224 x 224 x 5
Conv	-	MaxPool	e1 ← 112 x 112 x 64
Conv	-	MaxPool	e2 ← 56 x 56 x 128
Conv	-	MaxPool	e3 ← 28 x 28 x 256
Conv	-	MaxPool	e4 ← 14 x 14 x 512
Conv	-	MaxPool	7 x 7 x 1024
Conv	(this, e4)	Upsample	14 x 14 x 512
Conv	(this, e3)	Upsample	28 x 28 x 256
Conv	(this, e2)	Upsample	56 x 56 x 128
Conv	(this, e1)	Upsample	112 x 112 x 64
Conv	(this, e0)	Upsample	224 x 224 x 32
Conv — Tanh	-	-	224 x 224 x 3

Table 1: Encoder Architecture

Layer	Concatenate	Resample	Output Shape
Conv	-	-	e0 ← 224 x 224 x 3
Conv	-	MaxPool	e1 ← 112 x 112 x 64
Conv	-	MaxPool	e2 ← 56 x 56 x 128
Conv	-	MaxPool	e3 ← 28 x 28 x 256
Conv	-	MaxPool	e4 ← 14 x 14 x 512
Conv	-	MaxPool	7 x 7 x 1024
Conv	(this, e4)	Upsample	14 x 14 x 512
Conv	(this, e3)	Upsample	28 x 28 x 256
Conv	(this, e2)	Upsample	56 x 56 x 128
Conv	(this, e1)	Upsample	112 x 112 x 64
Conv	(this, e0)	Upsample	224 x 224 x 32
Conv — Tanh	-	-	224 x 224 x 5
Split	-	-	224 x 224 x (2, 3)

Table 2: Decoder Architecture

Encoder. (Table 1). The encoder architecture follows a simple U-Net structure. The container and the secret data has equal spatial resolution, with three channels in the container for RGB and two channels in the secret for the real and imaginary part of the STFT output. They are channel-wise concatenated from the first layer. The Tanh layer at the end normalizes the output to $(-1, 1)$.

Decoder. (Table 2). The decoder architecture is identical to that of the encoder, except that it does not have to concatenate anything in the input layer, and that the output is split in the channel dimension to represent the updated container and the recovered audio chunk.

Conditioning. (Table 2). We insert the FiLM (Perez et al., 2018) layer before every activation function in both the encoder and the decoder. The input to the FiLM layer is the last layer's feature map and a tensor filled with $\log_{10} \gamma$ extended to the same spatial dimension.

REFERENCES

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Micro-batch training with batch-channel normalization and weight standardization. *arXiv preprint arXiv:1903.10520*, 2019.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.