A Comparison of steganography, watermarker, and LAACA

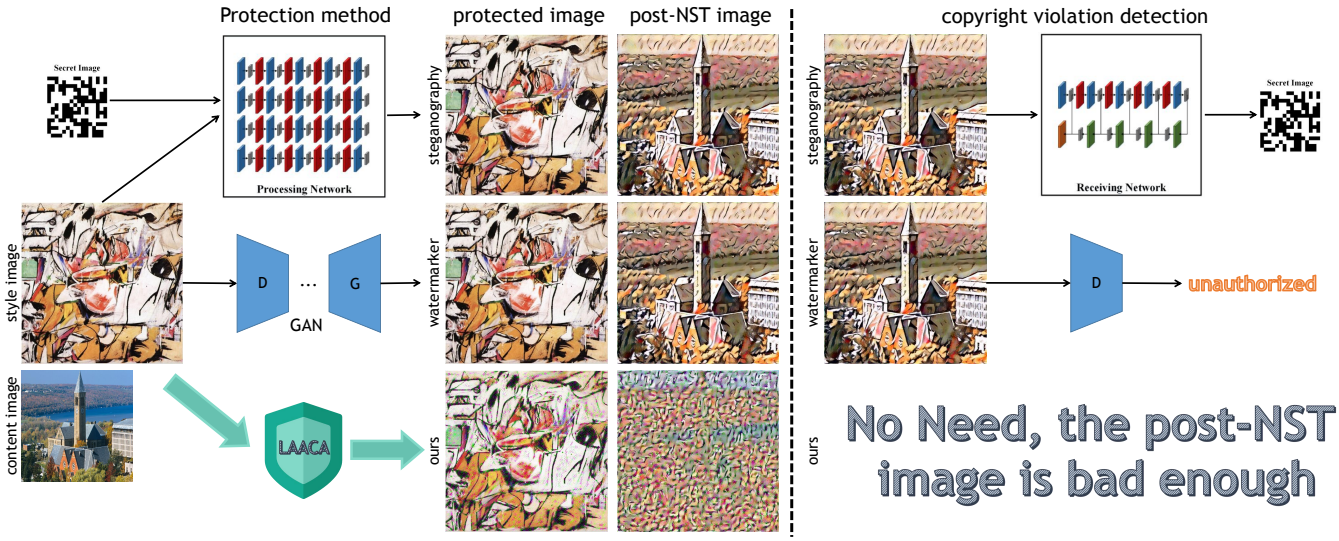

Figure 1. The comparison of steganography, watermarker, and our proposed method.

Figure 1 highlights the fundamental difference between our proposed method and previous copyright protection techniques related to style transfer. Existing methods, such as neural steganography and watermarking, rely on a pair of networks for encoding and decoding to embed either fixed (watermarking) or customizable (steganography) information within images. After the style transfer process, these methods employ specialized components to detect the embedded watermark or steganographic information. While these approaches can help artists protect their intellectual property to a certain extent, they have limitations. Not all instances of unauthorized artwork usage will be discovered by the artists, and pursuing legal action involves considerable time and financial costs.

In contrast, our method takes a proactive approach to prevent copyright infringement at its source. By protecting the images in such a way that the quality of style transfer generated from them significantly deteriorates, potential infringers are discouraged from using these low-quality images altogether. This eliminates the need for artists to actively monitor for unauthorized use and reduces the reliance on costly legal proceedings.

Our method focuses on preventing copyright infringement before it occurs, offering a more efficient and effective solution compared to traditional watermarking and steganography techniques.

B Extra experimental results

In the section, we supplement experiments which are not shown in the main paper, and illustrate exact values for those ablation figures in the main paper. Experiments were performed on an NVIDIA A100-SXM4-80GB, except where noted. All code are implemented with PyTorch [**?**], the random seed is fixed as 3407 [**?**].

B.1 Runtime

From Table 1, we can observe that our proposed LAACA method demonstrates a high efficiency in terms of runtime. Specifically, with other hyperparameters fixed as $k = 4$, $\alpha = 8$, and $\epsilon = 80$, the average runtime of LAACA is only 0.24s, 1.14s, and 2.27s for attack steps $T = 10$, $T = 50$, and $T = 100$, respectively. This rapid execution speed highlights the practicality of LAACA in real-world scenarios, where quick generation of adversarial examples is often crucial. Moreover, the runtime scales linearly with the number of attack steps, indicating the stability and predictability of our method's computational complexity. It is worth noting that LAACA achieves such efficiency without the need for model training, as it directly utilizes openly available pre-trained models. This feature further enhances the accessibility and usability of our method, making it highly convenient for researchers and practitioners to apply LAACA in their respective fields.

Table 1. Runtime of LAACA for difference choice of T.

attack steps T	10	50	100
average runtime	0.24s	1.14s	2.27s

In addition, we also test the performance of our proposed method on a consumer-level GPU. We found that the Nvidia RTX 3060 is the most-used GPU in the latest Steam Hardware Survey.[1] Table 2 shows the runtime and VRAM our method performed with different resolutions on this GPU, evidencing our method can process high-resolution images with acceptable computational cost, showcasing its potential practical applicability.

Table 2. Runtime and VRAM occupation of LAACA for difference resolutions.

resolution	VRAM	average time
256	632MB	3.6s
512	2050MB	11.8s
1024	7470MB	45.7s

B.2 Raw data for ablations

To provide a more comprehensive view of the experimental results presented in the main text, we include the raw data in tabular form in the following subsections. This table contains the original values used to generate the plots and figures, allowing for more comprehensive reference and verification of our findings.

B.3 Ablation raw data for ℓ_p norm

k	style images	ℓ_2 Neural Style Transfer Methods					
		Gatys	AdaIN	OST	SANet	EFDM	Average
3	136.1223	236.6225	202.4433	231.2812	206.3182	218.0069	218.9344
4	135.2614	233.1434	204.9743	233.5398	206.9464	217.6859	219.2580
5	134.8922	237.9601	207.1167	232.4886	204.7050	215.1189	219.4779
6	134.6510	224.2473	206.5098	232.5409	202.6859	215.5023	216.2972

α	style images	ℓ_2 Neural Style Transfer Methods					
		Gatys	AdaIN	OST	SANet	EFDM	Average
3	119.7790	222.9151	192.9840	225.0621	198.0758	210.2352	209.8544
4	130.8325	242.6859	204.7829	230.6003	201.2357	217.8570	219.4324
5	134.1788	237.2644	206.1372	232.7555	205.5454	223.4854	221.0376
6	135.2614	233.1434	204.9743	233.5398	206.9464	217.6859	219.2580

ϵ	style images	ℓ_2 Neural Style Transfer Methods					
		Gatys	AdaIN	OST	SANet	EFDM	Average
10	19.4547	101.1880	82.7188	146.6981	123.8358	91.4558	109.1793
20	38.1657	115.4231	117.9879	171.3367	148.5377	131.5788	136.9728
40	73.5585	164.5961	157.0720	199.0368	177.0844	170.8019	173.7182
80	135.2614	233.1434	204.9743	233.5398	206.9464	217.6859	219.2580

k	style images	ℓ_∞ Neural Style Transfer Methods					
		Gatys	AdaIN	OST	SANet	EFDM	Average
3	0.3137	0.9857	0.9773	0.9878	0.9839	0.9694	0.9808
4	0.3137	0.9877	0.9808	0.9897	0.9856	0.9736	0.9835
5	0.3137	0.9899	0.9810	0.9892	0.9851	0.9709	0.9832
6	0.3137	0.9875	0.9817	0.9910	0.9808	0.9707	0.9823

α	style images	ℓ_∞ Neural Style Transfer Methods					
		Gatys	AdaIN	OST	SANet	EFDM	Average
1	0.3137	0.9855	0.9791	0.9912	0.9813	0.9647	0.9804
2	0.3137	0.9915	0.9835	0.9921	0.9829	0.9686	0.9837
4	0.3137	0.9888	0.9814	0.9893	0.9829	0.9724	0.9830
8	0.3137	0.9877	0.9808	0.9897	0.9856	0.9736	0.9835

ϵ	style images	ℓ_∞ Neural Style Transfer Methods					
		Gatys	AdaIN	OST	SANet	EFDM	Average
1	0.3137	0.9855	0.9791	0.9912	0.9813	0.9647	0.9804
2	0.3137	0.9915	0.9835	0.9921	0.9829	0.9686	0.9837
4	0.3137	0.9888	0.9814	0.9893	0.9829	0.9724	0.9830
8	0.3137	0.9877	0.9808	0.9897	0.9856	0.9736	0.9835

[1] https://store.steampowered.com/hwsurvey/videocard/

B.4 Ablation for ACDM

k	style images	ACDM Neural Style Transfer Methods					
		Gatys	AdaIN	OST	SANet	EFDM	Average
3	0.0485	0.1385	0.1809	0.0672	0.1336	0.1700	0.1380
4	0.0495	0.1322	0.1913	0.0711	0.1300	0.1798	0.1409
5	0.0499	0.1390	0.2023	0.0730	0.1322	0.1879	0.1469
6	0.0505	0.1276	0.2047	0.0744	0.1345	0.1937	0.1470

α	style images	ACDM Neural Style Transfer Methods					
		Gatys	AdaIN	OST	SANet	EFDM	Average
1	0.0403	0.1235	0.1458	0.0505	0.1138	0.1509	0.1169
2	0.0441	0.1409	0.1667	0.0572	0.1260	0.1620	0.1306
4	0.0480	0.1391	0.1820	0.0653	0.1373	0.1844	0.1416
8	0.0495	0.1322	0.1913	0.0711	0.1300	0.1798	0.1409

ϵ	style images	ACDM Neural Style Transfer Methods					
		Gatys	AdaIN	OST	SANet	EFDM	Average
10	0.0049	0.0482	0.0409	0.0141	0.0347	0.0387	0.0353
20	0.0097	0.0522	0.0659	0.0203	0.0521	0.0660	0.0513
40	0.0228	0.0810	0.1197	0.0376	0.0935	0.1116	0.0887
80	0.0495	0.1322	0.1913	0.0711	0.1300	0.1798	0.1409

B.5 Ablation raw data for SSIMc

k	style images	SSIMc Neural Style Transfer Methods					
		Gatys	AdaIN	OST	SANet	EFDM	Average
3	0.6162	0.2330	0.3761	0.3677	0.3108	0.3492	0.3273
4	0.6130	0.2392	0.3891	0.3671	0.3150	0.3674	0.3356
5	0.6145	0.2293	0.3935	0.3675	0.3181	0.3733	0.3363
6	0.6171	0.2522	0.4009	0.3678	0.3259	0.3778	0.3449

α	style images	SSIMc Neural Style Transfer Methods					
		Gatys	AdaIN	OST	SANet	EFDM	Average
1	0.6841	0.2575	0.4014	0.3798	0.3253	0.3448	0.3417
2	0.6531	0.2245	0.3806	0.3683	0.3233	0.3397	0.3273
4	0.6326	0.2328	0.3797	0.3664	0.3145	0.3473	0.3281
8	0.6130	0.2392	0.3891	0.3671	0.3150	0.3674	0.3356

ϵ	style images	SSIMc Neural Style Transfer Methods					
		Gatys	AdaIN	OST	SANet	EFDM	Average
10	0.9500	0.7332	0.7824	0.6507	0.5905	0.7524	0.7018
20	0.9051	0.6590	0.6386	0.5662	0.5013	0.6028	0.5936
40	0.7945	0.4593	0.4983	0.4715	0.4102	0.4715	0.4622
80	0.6130	0.2392	0.3891	0.3671	0.3150	0.3674	0.3356

C User Study

To validate the effectiveness of the leveraged metrics (including our proposed metric ACDM), and prevent scenarios where the method performs well on all metrics but fails in practical applications, we conduct a user study. This study serves as a complementary evaluation to corroborate the validity and reliability of ACDM and other leveraged metrics in the context of our task.

This multi-faceted evaluation strategy helps to mitigate the risk of relying solely on metric-based performance, which may not always translate well to practical usability. By considering both quantitative measures and qualitative user feedback, we can gain a more holistic understanding of the method's effectiveness and robustness in preserving image quality under adversarial attacks.

We conducted a completely anonymous online survey by randomly inviting 37 individuals with no potential interest conflict in our work. Participants were required not to be color-blind. Due to the limited scope of our advertising, the subjects of this user study were recruited from university students.

They were asked to rate around 200 images on a scale of 1 to 5 in each set, assessing both pre-attack and post-attack style images, as well as 5 sets of pre-attack and post-attack NST images. The rating was based on the degree of difference between the images; a higher score indicates greater dissimilarity. During the survey, participants were unaware of the NST method corresponding to each column. No hints were given in terms of how we design and implement the attack, i.e., we did not mention in the survey that our proposed method targeted aspects like color, texture, or strokes. Moreover, we did not set any specific criteria for judging the two images, relying entirely on the subjective feelings of the participants.

Table 3 shows the high perceived dissimilarity between the pre- and post-attack NST images (4.5) is consistent with the changes captured by ACDM and other metrics. Conversely, the relatively low perceived dissimilarity between the original and attacked style images (1.8)

corroborates the metrics' indication. These findings demonstrate that with the collaboration of ACDM and other metrics, we effectively evaluate perceptually relevant changes in the context of attack on NST, providing a reliable assessment of LAACA's performance.

Table 3. User study evaluation results, lower score means better image quality.

	style image	Neural Style Transfer methods					
		Gatys	AdaIN	OST	SANet	EFDM	NST average
score	1.8345	4.5456	4.7753	3.7736	4.3024	4.8784	4.4551

The following images are some samples used in our user study, additionally serving to demonstrate the effectiveness of our proposed method. Like the images in the full paper, they are organized in sets, with each set spanning two lines. In every set, the first column displays the pre-attack and post-attack style images. The second column features the content image. The remaining columns each represent a pair of pre-attack and post-attack Neural Style Transfer (NST) images, showcasing different methods: Gatys (the first work on NST 2015), AdaIN (ICCV 2017), OST (ICCV 2019), SANet (CVPR 2019), and EFDM (CVPR 2022).