DENet: Disentangled Embedding Network for Visible Watermark Removal

Ruizhou Sun[1,2*], **Yukun Su**[1,3*], **Qingyao Wu**[1,4†]

[1]School of Software Engineering, South China University of Technology
[2]Key Laboratory of Big Data and Intelligent Robot, Ministry of Education
[3]Pazhou Lab, Guangzhou, China
[4]Peng Cheng Laboratory, China
ruizhousun@foxmail.com

Abstract

Adding visible watermark into image is a common copyright protection method of medias. Meanwhile, public research on watermark removal can be utilized as an adversarial technology to help the further development of watermarking. Existing watermark removal methods mainly adopt multi-task learning networks, which locate the watermark and restore the background simultaneously. However, these approaches view the task as an image-to-image reconstruction problem, where they only impose supervision after the final output, making the high-level semantic features shared between different tasks. To this end, inspired by the two-stage coarse-refinement network, we propose a novel contrastive learning mechanism to disentangle the high-level embedding semantic information of the images and watermarks, driving the respective network branch more oriented. Specifically, the proposed mechanism is leveraged for watermark image decomposition, which aims to decouple the clean image and watermark hints in the high-level embedding space. This can guarantee the learning representation of the restored image enjoy more task-specific cues. In addition, we introduce a self-attention-based enhancement module, which promotes the network's ability to capture semantic information among different regions, leading to further improvement on the contrastive learning mechanism. To validate the effectiveness of our proposed method, extensive experiments are conducted on different challenging benchmarks. Experimental evaluations show that our approach can achieve state-of-the-art performance and yield high-quality images. The code is available at: https://github.com/lianchengmingjue/DENet.

Introduction

As an important carrier, social media provides a platform for us to deliver and share different images and video content, whose security and robustness have become an important trend in recent research. Among them, watermarking (Cox et al. 2002; Katzenbeisser and Petitcolas 2000) is a common copyright protection approach of digital media. Meanwhile, many researchers pay attention to the watermark removal methods as an adversarial technology to boost the further development of digital watermarks.

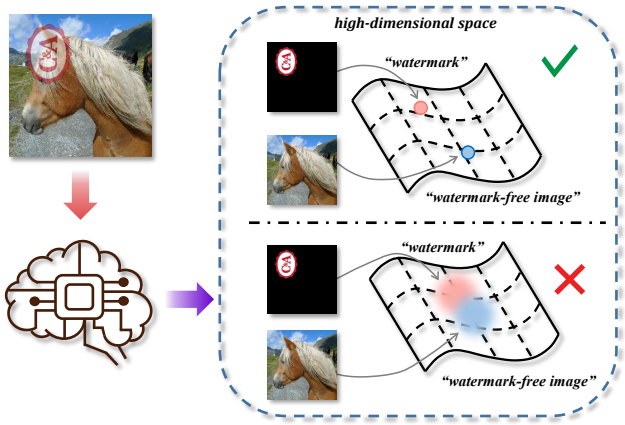


Figure 1: The main idea of our proposed DENet. Our goal is to disentangle the watermark-free image and watermark embeddings in the high-level space, which can provide more explicit cues for the decoder to reconstruct the images. The decoupled embedding of the watermark images has a more separate distribution of the potential content.

Some early watermark image removal methods (van Renesse 1996; Pei and Zeng 2006) adopt traditional techniques like nonlinear pixel domain technology and Independent Component Analysis (ICA) algorithm to decompose the watermark-free images and watermarks. With the flourish of deep learning, recent methods view the watermark image removal task as the image-to-image translation issue. Benefiting from the end-to-end reconstruction network (Ronneberger, Fischer, and Brox 2015) and generative adversarial networks (Goodfellow et al. 2014), the existing methods can achieve satisfactory watermark removal performance. More recently, Cun *et al.* (Cun and Pun 2021) and Liang *et al.* (Liang et al. 2021) propose to utilize multi-branch networks to locate and separate the watermark and image and finally refine the reconstructed images.

However, the above mentioned methods ignore the high-level semantic embedding of the watermark and watermark-free image feature. To be specific, the previous methods capture the watermark image's semantic information within a single encoder, which will make their learning representation less discriminative and separate, as shown in Fig 1

*These authors contributed equally.
†Corresponding author.

bottom. These uncertain cues are then fed into the decoder for further image reconstruction, which may harm the network to model the useful information, explicitly. With this in mind, our goal is to disentangle their semantic embeddings in the high-dimensional space, driving the network focusing on different parts of the watermarks and clean image, as shown in Fig 1 top. This can provide valuable cues to the subsequent networks to map watermark images to watermark-free ones maintaining high-quality patterns.

To this end, we propose a disentangled embedding network for visible watermark removal, termed as DENet. Specifically, we design a contrastive learning mechanism to decouple the semantic information of both the watermarks and watermark-free images. In this paper, we organize two sets of contrast loss constraints and construct a Siamese network to obtain positive and negative pairs. In addition, we introduce a self-attention-based enhancement module, which aims to strengthen the perception of features in different regions. In this paper, we make an early attempt to explore contrastive learning in the watermark image removal, and we experimentally investigate the features manifolds of both the disentangled embeddings. Extensive experimental evaluations on different challenging benchmarks, including LOGO-H, LOGO-L and LOGO-Gray (Cun and Pun 2021) and the qualitative intermediate visualization all validate the effectiveness of our proposed method. We expect this work will provide a new perspective for considering the relationship between supervised and contrastive learning.

Our contributions can be summarized as follows:

- We make the first endeavour to explore the impact of different embedding of clean images and watermarks in high-dimensional space and experimentally pull the positive pairs while pushing the negative pairs away, which, to our best knowledge, has not been well explored.
- We propose a disentangled embedding network for watermark removal, which aims to decouple the image and watermark representations in the high-level embedding space by contrastive learning mechanism to obtain more oriented features for reconstruction. Besides, we introduce a self-attention-based enhancement module for network to capture information from different regions.
- Extensive experimental evaluations on different datasets validate the superiority and effectiveness of our proposed method, which can achieve new state-of-the-art performance and yield high-quality output.

Related Work

Watermark Removal

Watermark removal share the similar scheme with image dehazing (He, Sun, and Tang 2010; Zhang and Patel 2018; Liu et al. 2022), deraining(Qian et al. 2018; Ren et al. 2019) and shadow removal(Cun, Pun, and Shi 2020; Liu et al. 2021). In essence, they are all the task of recovering the source image from a damaged image, but there are still differences that cannot be ignored in specific applications. For dehazing and deraining, the interference factors, such as haze and raindrop, permeate the whole image. Moreover, there are plenty of repeated patterns and semantics between different regions within the same image and even among different images. However, watermarks are usually concentrated in a local area of the image, and each watermark exists independently with a unique information representation. For shadow removal, the shadow usually appears as a meaningless grey area, while the watermark, as a symbol of media copyright, is usually colorful and meaningful. Therefore, these differences make watermark removal a unique and challenging research topic.

In the early works (Huang and Wu 2004; Pei and Zeng 2006; Park, Tai, and Kweon 2012), researchers mainly relied on hand-crafted features. Huang *et al.* proposed a visual-watermark attack scheme based on traditional image inpainting. Pei *et al.* used Independent Component Analysis (ICA) to separate source images from watermarked and referenced images. Park *et al.* formulated the problem using Bayes' rule via cross-channel correlation assumption. These methods not only rely on manual features but also require users to locate the watermark, which has a detrimental effect on usability. To avoid manual intervention, Dekel *et al.* (Dekel et al. 2017) assumed that different images have the same watermark, but it's unrealistic in real-world applications

With the development of deep learning, a number of data-driven neural network methods have emerged. Some of them (Li et al. 2019; Cao et al. 2019) only treated the watermark removal as an image-to-image translation task. Other alternative methods (Hertz et al. 2019; Liu, Zhu, and Bai 2021; Cun and Pun 2021; Liang et al. 2021) with better performance considered both watermark localization and watermark removal within a multi-task learning framework. Hertz *et al.* first removed visual motifs from images blindly and pioneered the single encoder with multi-decoder architecture for multi-task watermark removal. Pun *et al.* proposed a two-stage network for prediction and refinement, respectively. Inspired by multi-level feature fusion, Liang *et al.* designed several complex and elaborate modules to enhance the quality of generated images. Nevertheless, none of the above methods realize the importance of disentangling high-level semantics embedding between different tasks and still only progressively approach different goals.

Contrastive Learning

Contrastive learning is an effective method widely used in the field of unsupervised visual representation learning (He et al. 2020; Chen et al. 2020; Su et al. 2021; Su, Lin, and Wu 2021; Su et al. 2022). The core idea of contrastive learning is to construct positive and negative pairs, and adopt loss functions such as Info-NCE (Wu, Wu, and Huang 2021) to narrow the distance between positive pairs and widen the distance between negative pairs. Moreover, This idea have achieved impressive performance in image translation (Park et al. 2020) and image harmonization (Liang and Pun 2022) by properly constructing positive and negative samples, combined with patch-wised contrastive loss. To the best of our knowledge, this mechanism has not been explored in the field of watermark removal.

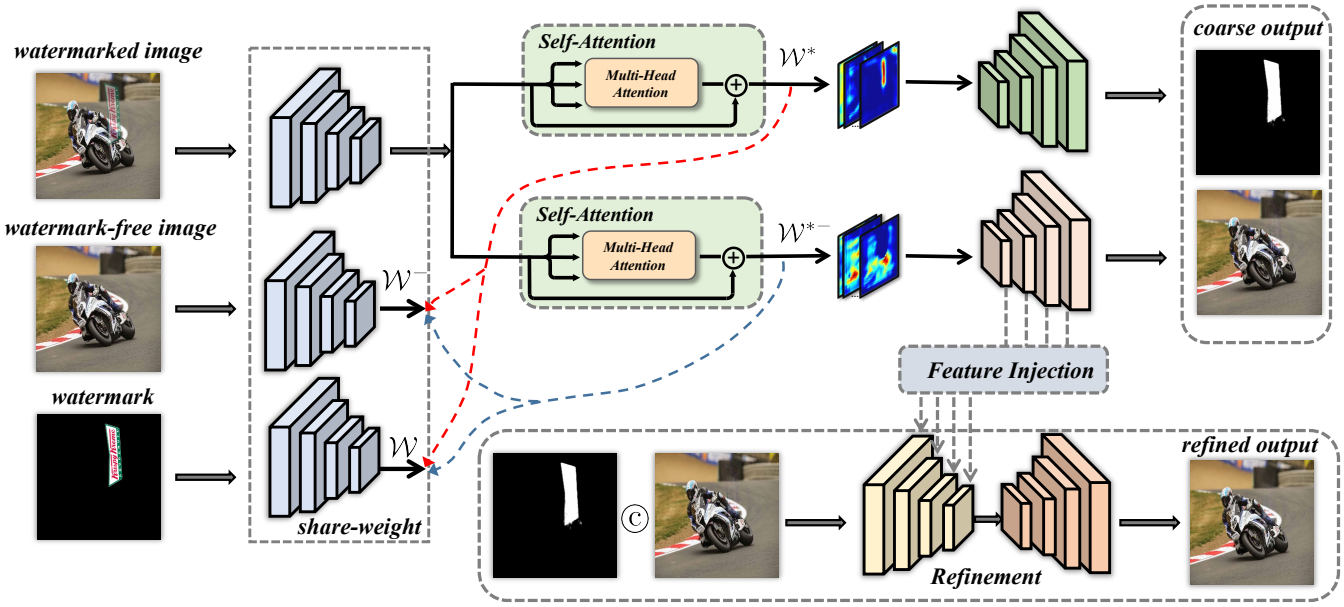


Figure 2: The overview of our proposed method. In the training phase, given the triple input including watermarked image, watermark-free image and watermark, they are first encoded by a share-weight convolutional backbone. Then the output feature of the watermarked image is passed through the self-attention enhancement block, which yields two different embeddings $\mathcal{W}^*$ and $\mathcal{W}^{*-}$. For the query $\mathcal{W}^{*-}$, we then construct the positive pairs as $\{\mathcal{W}^{*-}, \mathcal{W}^-\}$ and the negative pair as $\{\mathcal{W}^{*-}, \mathcal{W}\}$. For the query $\mathcal{W}^*$, vice versa. We aim to minimize the positive pairs distances while maximizing the negative pairs distances in the high-dimensional space to decompose the distributions between the watermark and watermark-free image. Then, these features are fed into the decoder and yield coarse results. Later, the coarse results are combined together and viewed as the input for the second stage refinement network, which finally produces the refined output. During the testing phase, given only the watermarked image, it will undergo the pre-trained network and finally reconstruct to the watermark-free image.

Methodology

In this paper, we address the watermark image removal issue from a new perspective. Our goal is to disentangle the embedding of the watermark and watermark-free image in the high-level space, thereby explicitly providing oriented valuable decoupling information to the network. In particular, we propose an effective framework, termed as DENet. In the following sections, we will show the overall network architecture and the proposed disentangled embedding module in detail.

Overall Architecture

The overview of our DENet is shown in Fig 2. Inspired by the former multi-stage refinement networks (Cun and Pun 2021; Liu, Zhu, and Bai 2021; Liang et al. 2021), we also adopt the cascade coarse-to-refine network to perform watermark image removal. However, our proposed network is totally different from the previous works. To be specific, given the triple input that can be easily accessed in the datasets, including the watermarked image, watermark-free image and watermark, they are first fed into a share-weight encoder to capture the semantic information. Then, we can obtain the semantic embedding $\mathcal{W}^-$ and $\mathcal{W}$ corresponding to the watermark-free image feature and watermark feature. Later, the watermarked image feature is further sent to the self-attention block to model semantic information among different pixel regions, which will yield two different embeddings $\mathcal{W}^*$ and $\mathcal{W}^{*-}$.

Afterward, we construct two triplets $\{\mathcal{W}^{*-}, \mathcal{W}^-, \mathcal{W}\}$ and $\{\mathcal{W}^*, \mathcal{W}^-, \mathcal{W}\}$, as illustrated by the blue and red arrows in Fig 2. For each triplet, our goal is to minimize the positive pairs distance while maximizing the negative pairs distances. In this way, we can decouple their semantic embedding in the high-level space and obtain more oriented semantic features, which is beneficial for the subsequent decoder network branch. We will give a detailed elaboration of the disentangled embedding in the following section. Later, the two different disentangled embedding cues are fed into the corresponding decoder for the watermark removal and watermark localization, which yields to coarse watermark-free image and the watermark mask, respectively. And then, the coarse results are concatenated together and passed through the refinement network in Unet (Ronneberger, Fischer, and Brox 2015) architecture. Note that we inject the decoder information from the coarse stage into the refinement stage to aggregate more potential hints, which is similar to (Liang et al. 2021). Concretely, we adopt tensor element-wise addition operation for two feature maps within the same resolution. Finally, the refinement network will produce the refined reconstructed watermark-free image.

Disentangled Embedding Mechanism

In order to disentangle the embeddings of the different learning features, we introduce contrastive learning mechanism (He et al. 2020) to perform this task. As mentioned above, we can yield two different watermark-free image embedding $\mathcal{W}^{*-}$ and watermark embedding $\mathcal{W}^{*}$ as illustrated in Fig 2, for simplicity, we take $\mathcal{W}^{*-}$ as an example to elaborate our disentangled embedding mechanism.

Specifically, given the watermark-free image embedding $\mathcal{W}^{*-} \in \mathbb{R}^{N\times C\times H\times W}$ and the encoder feature $\mathcal{W}, \mathcal{W}^{-} \in \mathbb{R}^{N\times C\times H\times W}$, we first downsample the watermark mask given in the dataset to $M \in \mathbb{R}^{N\times 1\times H\times W}$, and reformulate the embedding as follows:

$$n = \varphi(\mathcal{W} \odot M), \tag{1}$$

$$p = \varphi(\mathcal{W}^{-} \odot M), \tag{2}$$

$$q = \varphi(\mathcal{W}^{*-} \odot M), \tag{3}$$

where $\odot$ denotes element-wise multiplication, φ indicates global average pooling. Query q is supposed to be similar to its positive key p and dissimilar to negative key n.

In practice, we do not directly feed the watermark image into the share-weight encoder. Instead, we fill the black region of the watermark with the same content as the watermark-free image, otherwise the useless information of 0 will cause damage to the encoder. And here we use Eqn. (3) to remove the content that was filled before.

Naïve Distance: One might only consider the positive pairs and try to make their embedding distance closer with L_2 loss, which can be performed as follow:

$$\mathcal{L}_{\text{naïve}} = \frac{1}{N}\sum_{i=1}^{N} ||q_i - p_i||_2. \tag{4}$$

However, this naïve idea does not achieve the best performance, as shown in the experimental section.

Triplet Distance: For a more comprehensive consideration of both the positive and negative pairs, we simultaneously punish the triplet as follows:

$$\mathcal{L}_{CL}^{-} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(q_i \cdot p_i/\tau)}{\exp(q_i \cdot p_i/\tau) + \exp(q_i \cdot n_i/\tau)}, \tag{5}$$

where τ is a temperature parameter that controls the weight. $\mathcal{L}_{CL}^{+}$ is in reverse, viewing W^{*} as query, $\{W, W^{-}\}$ as positive and negative keys, respectively. Therefore, the total contrastive learning loss is as follows:

$$\mathcal{L}_{CL} = \mathcal{L}_{CL}^{+} + \mathcal{L}_{CL}^{-}. \tag{6}$$

Moreover, we present intuitive visualization for the effects of contrastive learning, which can be found in the experimental section.

Self-Attention Block

Although the above mechanism provides additional supervision for disentangling network learning, it is hard to achieve ideal embedding with only conventional convolution layers. Inspired by (Vaswani et al. 2017; Wang et al. 2020), we adopt multi-head attention to capture information from different regions. Given the input X, for i-th head, we then formulate the function as follow:

$$\begin{aligned} Q_i, K_i, V_i &= XW_i^{Q}, XW_i^{K}, XW_i^{V}, \\ \text{head}_i &= \text{softmax}(\frac{Q_i K_i^{T}}{\sqrt{d_k}})V_i, \end{aligned} \tag{7}$$

where W_i^{Q}, W_i^{K} and W_i^{V} are three projection matrices. d_k is the feature dimension. We finally concatenate all the features from different heads as output.

Likewise, we have visualized the attention map, which will be explained in detail in the experimental section.

End-to-end Training

Our proposed framework is an end-to-end network, and all modules are updated in one backpropagation. The loss function used in the training phase is as follows.

Following (Hertz et al. 2019), binary cross-entropy loss is applied to supervise watermark mask $\hat{M}$ with its ground truth M

$$\mathcal{L}_{mask} = -\sum_{i,j}(M_{i,j}\log\hat{M}_{i,j} + (1-M_{i,j})\log(1-\hat{M}_{i,j})). \tag{8}$$

Given watermark-free ground-truth image I, coarse output image $\hat{I}_{coarse}$ and refined output $\hat{I}_{refine}$, we employ L_1 loss to squeeze the gap between the ground truth and prediction output.

$$\mathcal{L}_{coarse} = ||I - \hat{I}_{\text{coarse}}||_1, \tag{9}$$

$$\mathcal{L}_{refine} = ||I - \hat{I}_{\text{refine}}||_1. \tag{10}$$

Similar to (Liang et al. 2021; Cun and Pun 2021), we add additional deep perceptual loss (Johnson, Alahi, and Fei-Fei 2016) for higher quality output

$$\mathcal{L}_{perc} = \sum_{k\in 1,2,3} ||\Phi_{vgg}^{k}(\hat{I}) - \Phi_{vgg}^{k}(I)||_1, \tag{11}$$

where $\Phi_{vgg}^{k}(\cdot)$ denotes the activation map of k-th layer in the pre-trained VGG16 (Simonyan and Zisserman 2014).

Finally, all the above loss functions are combined to get the final loss function with controllable hyper-parameters:

$$\begin{aligned} \mathcal{L}_{all} = \quad & \mathcal{L}_{\text{coarse}} + \mathcal{L}_{\text{refine}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \\ & \lambda_{\text{contrast}}\mathcal{L}_{\text{CL}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}}. \end{aligned} \tag{12}$$

τ	0.02	0.05	0.07	0.1	0.2
PSNR	44.04	43.9	**44.24**	44.1	44.07

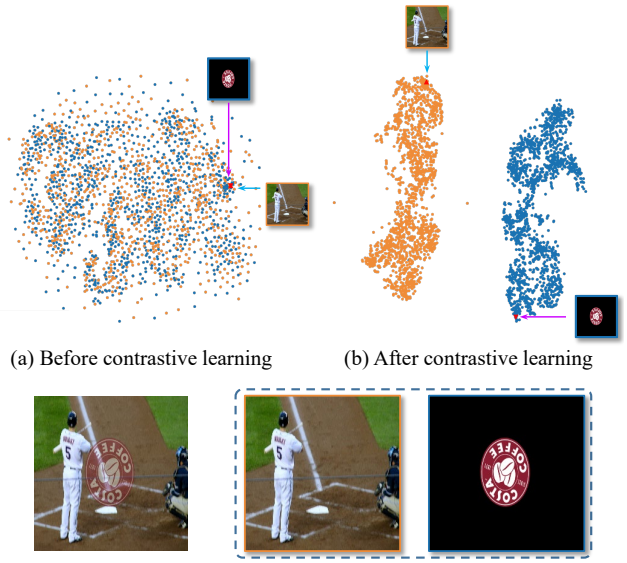
Table 1: Effects of the temperature parameter τ in contrastive loss on LOGO-L dataset.

h	1	4	8	16
PSNR	43.98	**44.24**	44.16	44.19

Table 2: Effects of the multi-head number h in self-attention block on LOGO-L dataset.

Baseline	Refinement	attention	contrast	PSNR
✓				42.89
✓	✓			43.58
✓	✓	✓		43.88
✓	✓	✓	naïve	44.01
✓	✓	✓	✓	**44.24**

Table 3: Analysis of different modules on LOGO-L dataset.



(a) Before contrastive learning (b) After contrastive learning

(c) From left to right: Watermarked Image, Watermark-free Image, Watermark. Their features are embedded into t-SNE as sample.

Figure 3: The t-SNE (Van der Maaten and Hinton 2008) visualization of the embedding distributions of the watermark (blue) and watermark-free image (orange) before and after the proposed contrastive learning mechanism.

Experiment

In this section, we start by introducing the datasets and implementation details. Then, we provide extensive ablation experiments with enriched visualization results to investigate the benefit of each module. Finally, we compare our DENet with state-of-the-art methods on various datasets, including LOGO-L, LOGO-H and LOGO-Gray (Cun and Pun 2021). The experimental results demonstrate our effectiveness both qualitatively and quantitatively.

Datasets and Implementation Details

Similar to the existing watermark removal method (Cun and Pun 2021), all the experiments are conducted on the LOGO series dataset. **LOGO-L**: LOGO-L contains 12151 watermarked images for training and 2025 images for testing. In this dataset, the watermark transparency range is between 35% and 60%, and the watermark size is also resized to 35% to 60% of the original image. **LOGO-H**: This dataset contains the same number of images as LOGO-L, but the watermark size in this dataset accounts for 60% to 80%, and the transparency is set from 60% to 85%. Thus, this is a harder dataset compared to LOGO-L due to the missing texture and the larger watermark area. **LOGO-Gray**: This dataset also includes 12151 images for training and 2025 images for testing. Different from the above two datasets, the embedded watermarks only contain grey-scale images.

Our method is implemented with Pytorch (Paszke et al. 2019). The models are trained for 200 epochs, where the input image resolution is 256×256. We choose Adam (Kingma and Ba 2014) as optimizer with learning rate of 1e-3, batch size 16. The hyper-parameters in (7) are $\lambda_{\text{mask}} = 1, \lambda_{\text{vgg}} = 0.25, \lambda_{\text{contrast}} = 0.25$, respectively. Following previous work(Cun and Pun 2021; Liang et al. 2021), we evaluate our method on several popular metrics, such as Peak Signal-to-Noise Radio (PSNR), Structural Similarity (SSIM) (Wang et al. 2004) and the deep perceptual similarity (LPIPS) (Zhang et al. 2018).

Ablation Study

In this section, we perform extensive ablation experiments to demonstrate the necessity and significance of each module of our network. Specifically, we first conduct sensitivity analysis experiments of the parameters. Then, we remove all modules and add them back incrementally to explore their effectiveness. Most notably, we present impressive visualization results, which provide a glimpse into the interpretability of our proposed modules.

Sensitivity analysis: Following (Chen et al. 2020; Vaswani et al. 2017), we search for optimal temperature parameter τ and head number h w.r.t. PSNR on LOGO-L dataset. Tab 1 and Tab 2 show the effect of using different τ and h. For τ, we find that when the value is set to 0.07, it can achieve the best performance. Moreover, we find that when $h = 1$, that is, when multi-head attention degenerates into single-head attention, the performance drops significantly. However, when h continues to increase, the performance improvement is not obvious and even decreases. Therefore, we set temperature parameter τ to 0.07 and multi-head numbers h to 4 in the following experiments.

Methods	LOGO-H			LOGO-L			LOGO-Gray		
	PSNR ↑	SSIM ↑	LPIPS↓	PSNR ↑	SSIM ↑	LPIPS↓	PSNR ↑	SSIM ↑	LPIPS↓
UNet	30.51	0.9612	5.44	34.87	0.9814	2.97	32.15	0.9728	3.53
SIRF	32.35	0.9673	8.01	36.25	0.9825	6.55	34.33	0.9782	6.72
BS^{2}AM	31.93	0.9677	4.45	36.11	0.9839	2.23	32.91	0.9754	3.05
DHAN	35.68	0.9809	6.61	38.54	0.9887	5.91	36.39	0.9836	5.94
BVMR	36.51	0.9799	2.37	40.24	0.9895	1.26	38.90	0.9873	1.15
SplitNet	40.05	0.9897	1.15	42.53	0.9924	0.87	42.01	0.9928	0.73
SLBR	40.56	0.9913	1.06	44.10	0.9947	0.70	42.21	0.9936	0.69
DENet (*Ours*)	**40.83**	**0.9919**	**0.89**	**44.24**	**0.9954**	**0.54**	**42.60**	**0.9944**	**0.53**

Table 4: Quantitative comparisons of our DENet with the other state-of-the-art methods on LOGO-H, LOGO-L and LOGO-Gray datasets. We choose PSNR, SSIM and LPIPS (in percentage) as metrics. The best results are marked in bold.

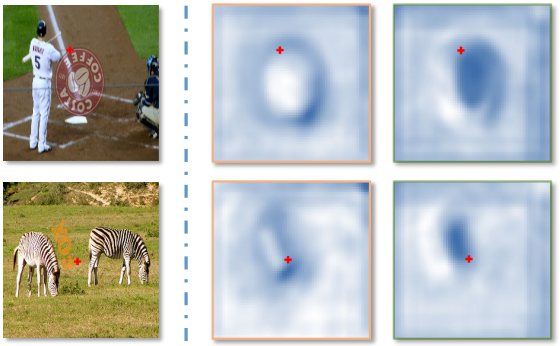


Figure 4: Visualization of the attention maps. The red cross denotes the selected pixels, with similar feature representation in blue color and dissimilar features in white.

Individual module analysis: As shown in Tab 3, we investigate the effectiveness of each module in our framework by removing all and adding them back incrementally. We started with a basic structure of ResUnet as (Hertz et al. 2019), except that it has two branches for predicting watermark location and watermark-free image, respectively. The performance of this baseline method is reported in the first row of Tab 3. Then, consistent with SLBR (Liang et al. 2021), we add Unet with feature injection as refinement net, which is shown in the second row. It can be seen that the performance will improve. After that, we add self-attention module in both branches to focus on different regions, leading to the third row in Tab 3. By comparing the second row with the third row, it demonstrates the effectiveness of the self-attention module. In addition, we introduce the contrastive learning mechanism and try to add contrastive loss function. In the fourth row, we first add the naïve $l2$ loss described in Eqn. (4). Then, we choose the proposed contrastive loss function instead, and the performance is reported in the fifth row. By comparing the second row to the fifth row, we can observe that the loss function in Eqn. (5) is more efficient, resulting in the best performance.

Visualization of contrastive learning mechanism: To further study the effectiveness of our proposed contrastive learning method, we obtain the feature W^*, W^{*-} described in Fig 2 and perform t-SNE algorithm (Van der Maaten and Hinton 2008; Poličar, Stražar, and Zupan 2019). As illustrated in Fig 3(a), without our contrastive learning mechanism, the watermark feature W^*(blue points) and the watermark-free image feature W^{*-}(orange points) are extremely confusing, and no effective distinction boundary can be seen. That is, watermark and watermark-free images share highly entangled in high-level semantics. This is easy to understand because they come from the same encoder, and the loss function acts on the output image, which is far from deep layers. On the contrary, as shown in Fig 3(b), after applying contrastive learning, the features of both watermark and watermark-free images are clearly clustered and separated from each other. It indicates that the deep features of different branches have their own orientation, thus the subsequent learning can be more focused on their respective tasks, such as watermark localization and removal. In other words, this experiment strongly demonstrates that our proposed disentangled embedding mechanism can efficiently distinguish the watermark image feature and decompose the intrinsic components, leading to better performance.

Visualization of self-attention module: To further verify the effectiveness of our proposed self-attention module, we present the attention map from Eqn. (7) according to (Wang et al. 2020). We first calculate AttnMap $= \frac{1}{h}\sum_{i=1}^{h} \text{softmax}(\frac{Q_i K_i^T}{\sqrt{d_k}})$ and then visualize it as shown in Fig 4. In each row, the selected pixels are the same, except that the second and third attention map come from the watermark removal and watermark localization branch, respectively. Although the location is the same, the area of concern is so different. For the watermark removal branch, the pixels within the watermark are more similar to other pixels around the watermark, which can help the network perceive the original watermark-free image information and restore the texture of the area covered by the watermark. Mean-

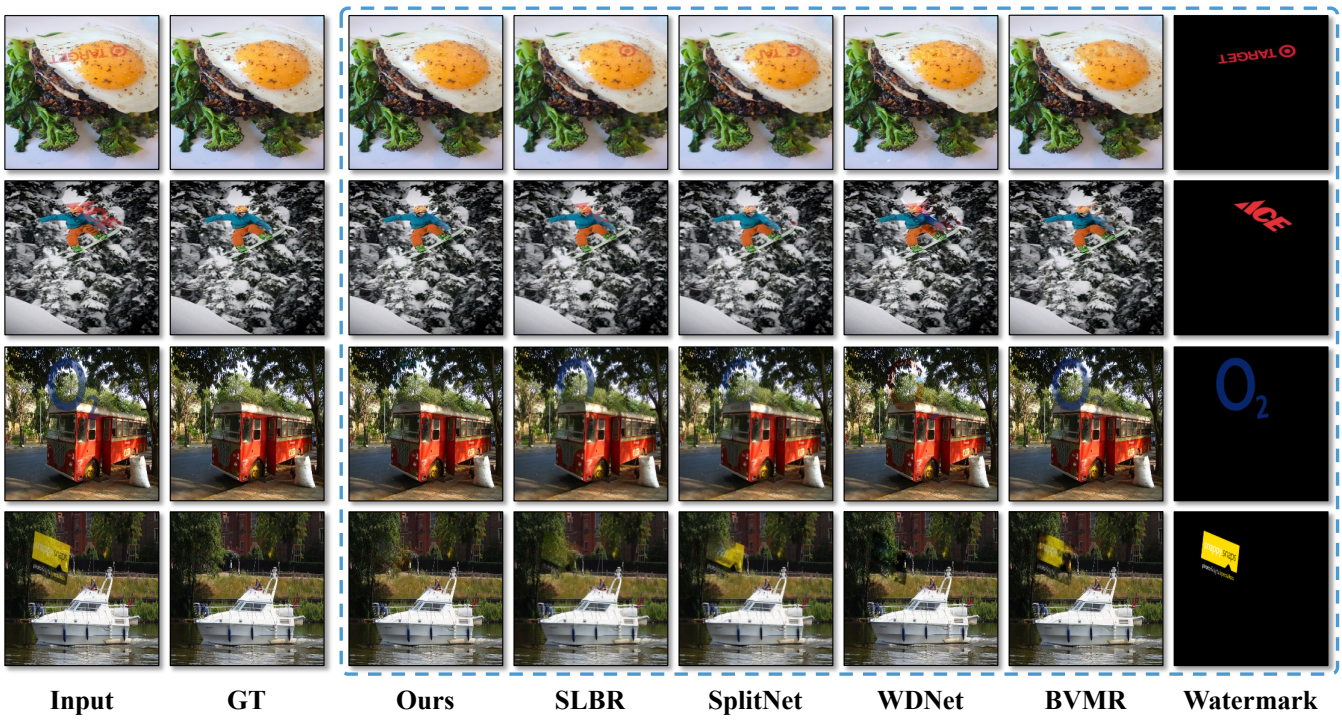


Figure 5: Qualitative comparisons with other state-of-the-art watermark removal methods. Our approach can yield more high-quality reconstructed output without watermark artifacts.

while, for the localization branch, the pixels in the watermark pay more attention to other pixels inside the watermark because the localization branch only cares about completely segmenting the watermark. Therefore, it proves that our self-attention module can capture information in different regions according to the corresponding task.

Comparisons with State-of-the-art Methods

The quantitative comparison of our proposed DENet with other existing watermark removal methods is summarized in Tab 4. Among them, Unet (Ronneberger, Fischer, and Brox 2015), SIRF (Zhang, Ng, and Chen 2018), BS^{2}AM (Cun and Pun 2020), DHAN (Cun, Pun, and Shi 2020) are migrated from related tasks such as blind image harmonization, shadow removal, etc. BVMR (Hertz et al. 2019), SplitNet (Cun and Pun 2021), and SLBR (Liang et al. 2021) are the latest technologies dedicated to watermark removal.

Our framework outperforms all the other methods on the three datasets and achieves new state-of-the-art performance. The various experimental results are sufficient to prove the effectiveness of our approach, which starts from a disentangling embedding perspective. It should be pointed out that we have not designed any complex modules tailored for watermark removal, such as SMR,MBE (Liang et al. 2021) and S^{2}AM (Cun and Pun 2021). We only use two general approaches (contrast learning and self-attention) to build framework based on our understanding of watermark removal. Furthermore, extensive ablation experiments and visualizations support the interpretability of our method.

Moreover, we present qualitative results of our DENet approach compared with other methods in Fig 5. From left to right, we display the input watermarked image, the ground-truth of the watermark-free image, the watermark-free image generated by different methods, and the ground-truth of the watermark. We can observe that DENet produces the most satisfactory results, which shows better watermark localization and repaired texture in some complicated areas. For example, in the first row, SLBR and SplitNet fail to identify the complete region of the watermark. In the third row, WDnet mistakenly paints the watermark area with incongruent colors, while BVMR does not remove the watermark at all. In these cases, our DENet can accurately identify the watermark region and restore the original texture properly.

Conclusion

In this paper, we propose a disentangled embedding network (DENet) for visible watermark removal. Specifically, through a contrastive learning mechanism, we decouple the image and watermark representation in the high-level embedding space, leading to obtaining more oriented features for localization and reconstruction, respectively. To further improve the network‘s ability to capture information from different regions, a self-attention module is introduced. Comparisons on various datasets show that DENet can achieve state-of-the-art performance qualitatively and quantitatively. In addition, extensive ablation experiments and visualizations have demonstrated the effectiveness and interpretability of our proposed method.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) 62272172 and 61876208, Tiptop Scientific and Technical Innovative Youth Talents of Guangdong Special Support Program 2019TQ05X200 and 2022 Tencent Wechat Rhino-Bird Focused Research Program (Tencent WeChat RBFR2022008), and the Major Key Project of PCL under Grant PCL2021A09.

References

Cao, Z.; Niu, S.; Zhang, J.; and Wang, X. 2019. Generative adversarial networks model for visible watermark removal. *IET Image Processing*, 13(10): 1783–1789.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Cox, I. J.; Miller, M. L.; Bloom, J. A.; and Honsinger, C. 2002. *Digital watermarking*, volume 53. Springer.

Cun, X.; and Pun, C.-M. 2020. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing*, 29: 4759–4771.

Cun, X.; and Pun, C.-M. 2021. Split then refine: stacked attention-guided ResUNets for blind single image visible watermark removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1184–1192.

Cun, X.; Pun, C.-M.; and Shi, C. 2020. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10680–10687.

Dekel, T.; Rubinstein, M.; Liu, C.; and Freeman, W. T. 2017. On the effectiveness of visible watermarks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2146–2154.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Sun, J.; and Tang, X. 2010. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12): 2341–2353.

Hertz, A.; Fogel, S.; Hanocka, R.; Giryes, R.; and Cohen-Or, D. 2019. Blind visual motif removal from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6858–6867.

Huang, C.-H.; and Wu, J.-L. 2004. Attacking visible watermarking schemes. *IEEE transactions on multimedia*, 6(1): 16–30.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, 694–711. Springer.

Katzenbeisser, S.; and Petitcolas, F. 2000. Digital watermarking. *Artech House, London*, 2: 2.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, X.; Lu, C.; Cheng, D.; Li, W.-H.; Cao, M.; Liu, B.; Ma, J.; and Zheng, W.-S. 2019. Towards photo-realistic visible watermark removal with conditional generative adversarial networks. In *International Conference on Image and Graphics*, 345–356. Springer.

Liang, J.; Niu, L.; Guo, F.; Long, T.; and Zhang, L. 2021. Visible Watermark Removal via Self-calibrated Localization and Background Refinement. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4426–4434.

Liang, J.; and Pun, C.-M. 2022. Image Harmonization with Region-wise Contrastive Learning. *arXiv preprint arXiv:2205.14058*.

Liu, H.; Wu, Z.; Li, L.; Salehkalaibar, S.; Chen, J.; and Wang, K. 2022. Towards Multi-Domain Single Image Dehazing via Test-Time Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5831–5840.

Liu, Y.; Zhu, Z.; and Bai, X. 2021. Wdnet: Watermark-decomposition network for visible watermark removal. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3685–3693.

Liu, Z.; Yin, H.; Wu, X.; Wu, Z.; Mi, Y.; and Wang, S. 2021. From shadow generation to shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4927–4936.

Park, J.; Tai, Y.-W.; and Kweon, I. S. 2012. Identigram/watermark removal using cross-channel correlation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 446–453. IEEE.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, 319–345. Springer.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pei, S.-C.; and Zeng, Y.-C. 2006. A novel image recovery algorithm for visible watermarked images. *IEEE Transactions on information forensics and security*, 1(4): 543–550.

Poličar, P. G.; Stražar, M.; and Zupan, B. 2019. openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. *BioRxiv*, 731877.

Qian, R.; Tan, R. T.; Yang, W.; Su, J.; and Liu, J. 2018. Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2482–2491.

Ren, D.; Zuo, W.; Hu, Q.; Zhu, P.; and Meng, D. 2019. Progressive image deraining networks: A better and simpler baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3937–3946.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Su, Y.; Lin, G.; Hao, Y.; Cao, Y.; Wang, W.; and Wu, Q. 2022. Self-supervised object localization with joint graph partition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2289–2297.

Su, Y.; Lin, G.; Sun, R.; Hao, Y.; and Wu, Q. 2021. Modeling the uncertainty for self-supervised 3d skeleton action representation learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, 769–778.

Su, Y.; Lin, G.; and Wu, Q. 2021. Self-supervised 3d skeleton action representation learning with motion consistency and continuity. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13328–13338.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

van Renesse, R. 1996. Protecting publicly-available images with a visible image watermark. *Optical Security and Counterfeit Deterrence Techniques*, 2659: 126–133.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y.; Zhang, J.; Kan, M.; Shan, S.; and Chen, X. 2020. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12275–12284.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wu, C.; Wu, F.; and Huang, Y. 2021. Rethinking InfoNCE: How Many Negative Samples Do You Need? *arXiv preprint arXiv:2105.13003*.

Zhang, H.; and Patel, V. M. 2018. Densely connected pyramid dehazing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3194–3203.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, X.; Ng, R.; and Chen, Q. 2018. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4786–4794.