VSGAN: Visual Saliency guided Generative Adversarial Network for data augmentation

Zhang Jun, Xian Chuahua, Alexandre Bruckert, Patrick Le Callet, Guiqing Li, Hongmin Cai

VSGAN: Visual Saliency guided Generative Adversarial Network for data augmentation

Jun ZHANG
zhangj9n@gmail.com
South China University of Technology
Guangzhou, China

Chuhua Xian*
chhxian@scut.edu.cn
South China University of Technology
Guangzhou, China

Alexandre Bruckert
alexandre.bruckert@univ-nantes.fr
Nantes Université, École Centrale
Nantes, CNRS, LS2N, UMR 6004
F44000 Nantes, France

Patrick Le Callet
Patrick.Le-Callet@univ-nantes.fr
Nantes Université, École Centrale
Nantes, CNRS, LS2N, UMR 6004
F44000 Nantes, France

Guiqing Li
ligq@scut.edu.cn
South China University of Technology
Guangzhou, China

Hongmin Cai
hmcai@scut.edu.cn
South China University of Technology
Guangzhou, China

ABSTRACT

Deep learning approaches have allowed for a great leap in the performances of visual saliency models. However, the lack of annotated data remains the main challenge for visual saliency prediction. In this paper, we leverage image inpainting methods to synthesize augmented images, which is done by completing the weakly-salient areas, and propose a Visual Saliency guided Generative Adversarial Network (VSGAN) that contains a dual encoder to extract multi-scale features and a generator equipped with visual saliency guided modulation to synthesize high fidelity and diversity results. Extensive experimental results show that our method outperforms state-of-the-art methods for image inpainting on visual saliency datasets, and demonstrate the effectiveness of VSGAN for visual saliency data augmentation both quantitatively and qualitatively.

CCS CONCEPTS

• **Computing methodologies** → *Interest point and salient region detections.*

KEYWORDS

Visual saliency, image inpainting, GAN, data augmentation

ACM Reference Format:
Jun ZHANG, Chuhua Xian, Alexandre Bruckert, Patrick Le Callet, Guiqing Li, and Hongmin Cai. 2023. VSGAN: Visual Saliency guided Generative Adversarial Network for data augmentation. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Visual saliency prediction aims to infer which areas of an image are attractive to human eyes. It is widely used in the fields of image compression, cognition studies and image quality assessment [20]. In the context of multimedia data, understanding visual attention is a key step in understanding user experience, thus highlighting the need for efficient and reliable visual saliency models. In the last decade, deep learning based approaches have shown great advantages over conventional methods. Compared to the extraction of hand-craft features following biological evidence, data-driven methods [1] show superior performance and robustness, especially when dealing with higher-order information.

Nevertheless, such methods need high-volume datasets, and the expensive cost of eye-tracking experiments causes existing public datasets to remain of a limited sizes, and often focus on particular styles [3].

One of the most popular methods proposed to alleviate this burden of gathering eye-tracking data is to pretrain a visual saliency network on a mouse-tracking dataset [8], which is then fine-tuned on smaller eye-tracking datasets. However, a form of bias may be introduced due to the domain shift between the two data modalities, culminating in suboptimal performances with regard to the training datasets size. Another solution, used in numerous domains of computer vision is data augmentation, i.e. the addition of altered or transformed images in the dataset. This includes classic ways, such as shape or color transforms, as well as novel methods relying on Generative Adversarial Networks (GANs) [24, 25].

As for visual saliency prediction, classic ways for data augmentation have been studied in a few works. Kim and Milabfar [12] studied the effect of noise images and proposed a model using noisy images which demonstrated the negative influences. Tilke et al. [9] investigated the use of different resolutions on the images and made a comparison for gaze distribution. Che et al. [3] found that some augmentation methods are not label-preserving, such as crops, by doing a fine-grained analysis of human gaze with different transformations. However, these methods may have potential effects on visual attention. Indeed, images and their corresponding saliency maps are intrinsically related, and altering the image may also alter the associated map. To overcome it, we propose a novel data augmentation method based on image inpainting to preserve the salient regions of an image, and generate diverse backgrounds in weakly-salient regions.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Compared to classic data augmentation techniques, our method avoids modifying the salient areas of the images, allowing for the preservation of the original saliency map. The diversity in the generated weakly-salient areas also allows for large-scale data augments, which further allow for improvements in visual saliency models when trained on this data. More qualitative and quantitative analysis will be further explored in the experimental part. Overall, we propose the following three contributions:

- Visual Saliency guided GAN (VSGAN), a new inpainting network formed with a dual encoder and a generator equipped with visual saliency guided modulation.
- A comparison with exiting image inpainting networks showing that VSGAN achieves better outcomes on visual saliency datasets.
- An evaluation of the significant improvement allowed by VSGAN data augmentation when applied to visual saliency modeling.

2 METHODOLOGY

In order to alleviate the burden of gathering eye-tracking data, we propose a Visual Saliency guided Generative Adversarial Network (VSGAN) as a data augmentation tool, where the overall pipeline is illustrated in Figure 1 (top). VSGAN consists of a dual encoder to generate the content vector $\mathbf{c}$ when given salient regions as input, a mask guided modulation to dynamically modulate the content vector, and a generator to synthesize the inpainted results. Note that we use an additional random vector $\mathbf{z}$ to enable diverse results. The detailed structure of each module is elaborated in the subsections below.

2.1 Dual encoder

We feed the partial image and the corresponding mask into the dual encoder to produce a content vector $\mathbf{c}$. The dual encoder is composed of a local and global branch, as depicted in Figure 1 (left). For the local branch, a series of convolution blocks with a convolutional layer and a 2x downsampling layer are employed, resulting in feature maps with multiple levels $[f_{loc}^1, f_{loc}^2, ..., f_{loc}^{N_{loc}}]$. We propose additional transformer blocks [5] to obtain extra context since convolution lacks the ability to capture long-range information. The transformer blocks are applied on the higher levels of the encoder, producing feature maps $[f_{glb}^1, f_{glb}^2, ..., f_{glb}^{N_{glb}}]$. We fuse the feature maps from each branch via element-wise multiplication. The multiplied features will be sent to the generator by skip connections. In the last layer, the feature maps are flattened into vectors respectively in both branches, which are then concatenated and processed by a linear layer to generate content vector $\mathbf{c}$.

2.2 Mask guided modulation

As shown in Figure 1 (right), we regard the corresponding salient map of the input image as the mask. In order to handle the problem of masks in various sizes, we propose a mask guided modulation to dynamically modulate the content vector $\mathbf{c}$ which is inspired by [19]. In our implementation, the mask is resized to adapt to the spatial size of the last-level feature maps in the dual encoder. By using two convolutional layers, two feature maps are generated,

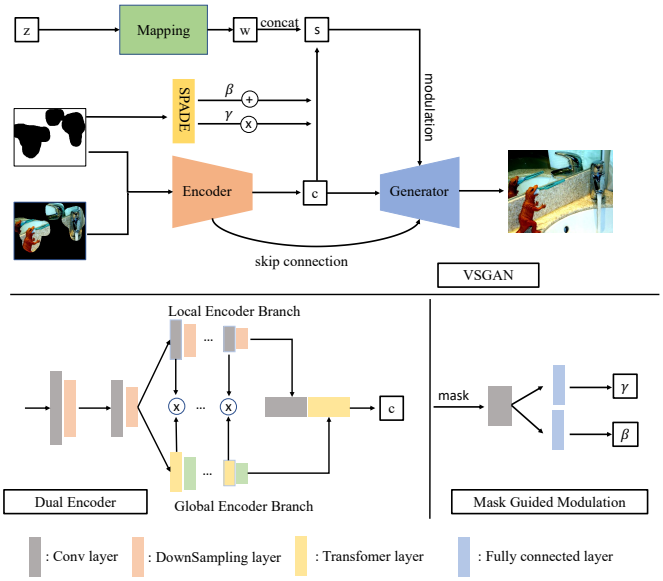


Figure 1: VSGAN is composed of a dual encoder and a generator. Dual encoder: The combination of the local encoder and global encoder branch can be utilized to extract multi-scale features. Mask Guided Modulation: It is implemented by two convolution layers to capture spatial information from mask.

which are then flattened as scale vectors γ and bias vectors β. Also, a linear layer is utilized to transfer γ and β to the proper dimension as $\mathbf{c}$. Finally, the network gains more spatial information based on the input mask via an affine transformation as $\mathbf{c}_{mod} = \gamma\mathbf{c} + \beta$.

2.3 Generator

Recall that our goal is to generate plausible images with each partial image. To this end, we use a random vector $\mathbf{z}$ to enable some degree of diversity in the generated areas. Following StyleGAN [11], a non-linear network is leveraged to map the vector $\mathbf{z}$ to an embedding $\mathbf{w}$. Inspired by CO-MOD-GAN [23], the embedding $\mathbf{w}$ is concatenated with $\mathbf{c_{mod}}$ as global style embedding $\mathbf{s}$ to modulate the generator. Our generator has the symmetric architecture as the local encoder branch with convolution and upsampling layers. Additionally, skip connections, fused from the dual encoder branches, are used to enhance the quality of the synthesized image. Different from [23], our content vector $\mathbf{c}$ is guided by a modulation layer, and the skip connections are derived from the dual encoder as stated before. In this way, more spatial and global information is provided when trading off quality and diversity.

3 EXPERIMENTS

Implementation Details We conduct the data augmentation experiments on SALICON dataset [8] which contains 15000 images and the corresponding saliency maps. All images and saliency maps are resized to (320,256). By default, N_{loc} and N_{glb} are set to be 7 and 5, respectively. We set 0.7 as the threshold to generate mask

Table 1: Ablation study.

Method	FID	LPIPS	PIDS	UIDS
CO-MOD-GAN	10.95	0.2475	4.68	17.73
w/o DE	10.46	0.2426	5.48	19.32
w/o MGM	10.39	**0.2380**	5.06	18.46
VSGAN(full model)	**9.92**	0.2409	**7.24**	**21.30**

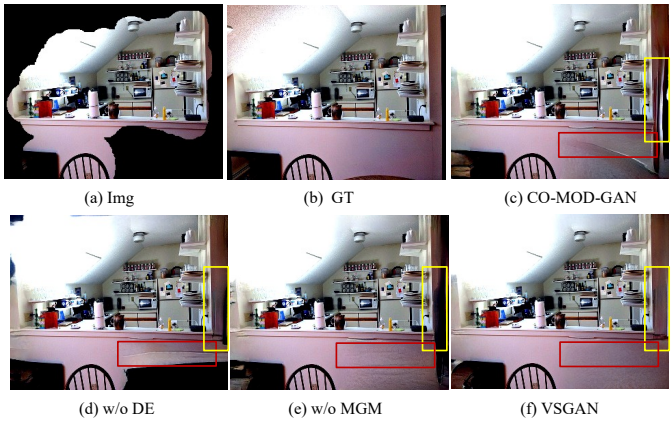


Figure 2: Ablation study. The differences are highlighted red and yellow boxes.

from saliency map. Our VSGAN is trained with the same loss functions as in [11] by using Adam optimizer [13]. All experiments are conducted on one NVIDIA 3090 GPU with learning rate 0.02 and batch size 2.

Evaluation metrics We evaluate the results on SALICON test set by Fréchet Inception Distance (FID) [6], Perceptual Image Patch Similarity Distance (LPIPS) [22] and Paired/Unpaired Inception Discriminative Score (P-IDS/U-IDS) [23]. To evaluate the effectiveness of our methods for data augmentation, we train two visual saliency models [4, 14] where different augmentation strategies are applied. The metrics we use are two variants of Area under ROC curve (AUC-Borji, sAUC), Normalized scanpath saliency (NSS), Similarity (SIM), Pearson's correlation coefficient (CC), Kullback-Lieber divergence (KLDIV) [15].

3.1 Ablation Study

We perform an ablation study to evaluate the efficiency of two key modules in our network: dual encoder(DE) and mask guided modulation(MGM). Four variants are implemented to study the impact of each component. CO-MOD-GAN can be regard as the reference model without our proposed component. All these models in the ablation study are trained and evaluated on the SALICON dataset. Quantitative and qualitative comparisons are shown in Table 1 and Figure 2.

Dual encoder. To study the capability of our dual encoder, we remove the global branch from our network and train on SALICON dataset with 50 epochs (DE in Table 1). As shown in Table 1, we find significantly better results with the full VSGAN on all metrics,

Figure 3: Comparison with Exiting methods. Details processed by our method are visually consistent with input images (marked in red boxes) when comparing to others.

Table 2: Comparison with Existing methods.

Method	FID	LPIPS	PIDS	UIDS
Pix2pix	40.81	0.40	0.0	0.0
Pconv	45.55	0.39	0.0	0.0
PDGAN	32.83	0.38	0.0	0.0
CO-MOD-GAN	10.95	0.25	4.68	17.73
VSGAN	**9.92**	**0.24**	**7.24**	**21.30**

especially on FID, PIDS and UIDS, which suggest that the results on DE lost some fidelity in detail. Moreover, we show some visual examples in Figure 2. Compared to (d) and (f), VSGAN has fewer artifacts in the wall (highlighted in red boxes). By using transformer on global encoder and skip-connection, dual encoder in VSGAN has bigger receptive field than traditional encoder, showing the benefits of synthesizing new images.

Mask guided modulation We then evaluate the mask guided modulation by dropping this layer when concatenating the content vector c and style vector s (MGM in Table 1). Similarly, this model is trained on SALICON dataset with 50 epochs. In Table 1, the performances of VSGAN model proves to be better on all metrics. Considering mask guided modulation can add spatial information on content vector c, VSGAN achieve better balance on the diversity and high fidelity when generating new images. To better comparison, we visually compare the results of MGM (c) and VSGAN (f) in Figure 2. VSGAN generates more realistic wall than MGM (masked as yellow boxes).

Table 3: Model complexity.

Models	# Params of G	# Params of D
Pix2pix	48M	5M
PDGAN	235M	5M
CO-MOD-GAN	80M	29M
VSGAN	130M	29M

3.2 Comparison with existing methods

We compare our method with some leading image inpainting models, including Pix2pix [7], Pconv [16], PDGAN [17], CO-MOD-GAN [23]. The above networks are trained on SALICON dataset with the setting described in their original papers. As shown in Table 2, our method significantly outperforms all other methods in FID, LPIPS, PIDS and UIDS. As shown in Figure 3, the results of Pix2pix, Pconv and PDGAN have blur or droplet-like artifacts in some regions (marked in red boxes). Comparing to CO-MOD-GAN, VSGAN can generate more content-consistent results and create fewer visual artifacts as visualized in (f) and (g).

3.3 Pluralistic Generation

The diversity of our network comes from the StyleGAN architecture. The latent code $\mathbf{z}$ is randomly sampled from a normal distribution and can be seen as a style code which control the inpainting areas. We visualize some results of stochastic image synthesis in Figure 4. In the first row, the roofs in (c), (d) and (e) are different, same as the street. In the second row, VSGAN generates types of groves which have a different form and remain realistic overall.

3.4 Model complexity

We conduct a comparison of parameter numbers with competing methods in Table 3. Our VSGAN has 130M parameters for generator and 29M for discriminator. Although our model has more parameters than some other models, this increased parameter count allows it to better capture the complexity of the underlying data and achieve higher accuracy in the predictions.

3.5 Effectiveness for data augmentation

In this section, we propose an evaluation of data augmentation methods for visual saliency models. Firstly, we studied the effect of classic data augmentation methods. To extend the work of Che et al. [3], we choose 10 wildly used traditional image transformations that contains cropping, flipping, color transform, random rotation, contrast, JPEG, noise, shearing, inversion and mirroring. The implementation detail is introduced in the Table 4. Secondly, to evaluate the performance of our VSGAN as data augmentation method, we mask training images with the saliency maps and generate new images to expand the visual saliency datasets.

In detail, we use the above data augmentation methods to expand one time of the training dataset and then leverage these data to train the visual saliency models. To simplify the representation of each models, we named the model by the data augmentation technique used. Specifically, the baseline models are trained using only the original datasets. Additionally, we create a super-set composed all of the data augmentation techniques that improve the performances

Table 4: Classic methods for data augmentation.

Methods	Generation Details
Cropping	CAT2000 (1720, 980), MIT1003 (800, 600), Random
Flipping	Flipping horizontally and vertically
Color Transform	BGR to HSV
Random Rotation	Rotation degree between 60° and 160°, Random
Contrast	Converting an image range from [0.3, 0.7] to [0, 1]
JPEG	Compression ratio 5
Noise	Gaussian noise (var = 0.1)
Shearing	Shearing matrix [[1,0,0],[0.5,1,0],[0,0,1]]
Inversion	Flipping vertically
Mirroring	Flipping horizontally

Table 5: Results of MSI-Net on CAT2000 test set.

	KLDIV ↓	NSS ↑	sAUC ↑	SIM ↑	CC ↑	AUC_Borji ↑
SALICON	0.4072	2.2438	0.7725	0.6808	0.8295	0.7852
baseline	0.3259	2.3937	0.7890	0.7292	0.8671	0.8007
Cropping	0.2677	2.4719	0.7973	**0.7589**	**0.8946**	0.8087
Color Transform	0.3068	2.4200	0.7914	0.7382	0.8758	0.8030
Random Rotation	0.3230	2.3990	0.7865	0.7316	0.8690	0.7979
Contrast	0.2853	2.4569	0.7912	0.7513	0.8890	0.8022
JPEG	0.2766	2.4722	0.7900	0.7563	0.8939	0.8007
Shearing	0.3150	2.4110	0.7883	0.7352	0.8724	0.7996
Noise	0.2767	**2.4737**	0.7915	0.7567	0.8941	0.8023
Flipping	0.3307	2.3819	0.7898	0.7265	0.8635	0.8016
Mirroring	0.3773	2.3169	0.7781	0.6987	0.8414	0.7896
Inversion	0.3713	2.3355	0.7819	0.7040	0.8483	0.7935
Valid	0.2668	2.4723	0.7983	0.7578	0.8871	0.8091
VSGAN	**0.2656**	2.4734	**0.7988**	0.7563	0.8876	**0.8099**

of the model, and refer this set as Valid set. To evaluate the efficacy of VSGAN as a data augmentation tool, we use a combination of the Valid set and the images generated by VSGAN to train the visual saliency models. In order to get fair and convincing results, we choose two visual saliency models [14] [4] and four visual saliency datasets: SALICON [8], CAT2000 [2], MIT1003 [10] and OSIE [21]. The detail of these models is introduced in the supplementary file. We train them on SALICON dataset and then fine-tune on the visual saliency datasets.

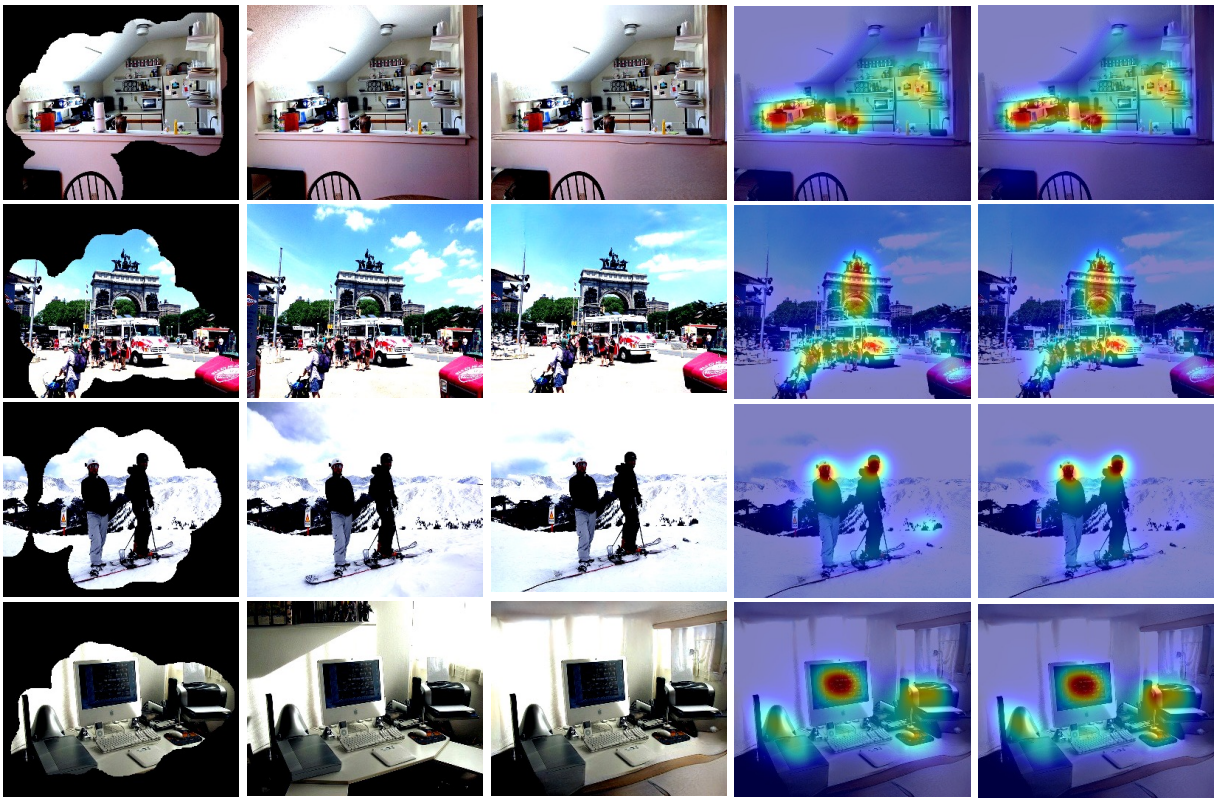
To make a investigation and analysis on whether the generated image preserve the same saliency with GT, we conduct visual saliency experiments using a visual saliency model[18] to generate visual saliency maps for both the origin and our inpainting images. The results of this experiment, as shown in Figure 5, indicate that the saliency maps for the original and inpainting images are nearly identical, demonstrating that our inpainting method is capable of preserving the same saliency as the GT. In the next, we introduce the results of data augmentation for visual saliency models in detail.

3.5.1 CAT2000. For CAT2000 dataset, we divide training set, test set and validation set with the ratio of 8:1:1. Since the CAT2000 dataset has 20 categories, we ensure that our sets are uniformly balanced.

The results of MSI-Net on CAT2000 test set are introduced in Table 5. We find that Cropping, Color Transform, Random Rotation, Contrast, JPEG, Noise methods have positive effect to mitigate overfitting, and thus we consider these transformations as our "Valid set". However, Flipping, Shearing, Inversion, Mirroring methods have negative effect. It seems that these kind of transformations

(a) Img (b) GT (c) VSGAN(1) (d) VSGAN(2) (e) VSGAN(3)

Figure 4: Pluralistic Generation. The diversity is highlighted in red boxes.



a. Mask Image b. Origin Image c. VSGAN d. GT e. VSGAN

Figure 5: Analysis on whether the generated image preserve the same saliency with GT.

may change the ground truth thus providing the model with bad instruction. As we can see from the last two rows in Table 5, VSGAN achieve better performance than Valid, showing the usefulness of this augmentation method in the context of visual saliency prediction.

Similarly to MSI-Net results, the SAM-ResNet model shows significant improvement when trained with augmented data, compared with the original CAT2000 dataset, as shown in Table 6. The performance of models including Flipping, Mirroring, Inversion are ranking last three models on all models. It means that Flipping, Mirroring and Inversion have bad influence on some images

Table 6: Results of SAM-ResNet on CAT2000 test set.

	KLDIV ↓	NSS ↑	sAUC ↑	SIM ↑	CC ↑	AUC_Borji ↑
SALICON	1.2183	0.9243	0.7180	0.4550	0.4017	0.7181
baseline	0.8018	1.6400	0.7179	0.5819	0.7070	0.7179
Cropping	0.5613	1.8021	0.7303	0.6417	0.7784	0.7304
Color Transform	0.5365	1.8144	0.7446	0.6453	0.7849	0.7446
Random Rotation	0.5439	1.8035	0.7476	0.6436	0.7817	0.7476
Contrast	**0.5297**	**1.8095**	0.7477	0.6441	0.7840	0.7476
JPEG	0.5542	1.8095	0.7477	0.6441	0.7840	**0.7511**
Shearing	0.5498	1.7781	0.7491	0.6340	0.7683	0.7492
Noise	0.5363	1.8088	0.7444	0.6449	0.7830	0.7444
Flipping	0.5685	1.7775	0.7420	0.6332	0.7686	0.7420
Mirroring	0.5682	1.7983	0.7332	0.6423	0.7791	0.7332
Inversion	0.5712	1.7762	**0.7499**	0.6337	0.7693	0.7500
Valid	0.5372	1.8106	0.7345	0.6452	0.7871	0.7463
VSGAN	0.5345	1.8153	0.7348	**0.6461**	**0.7889**	0.7506

Table 7: Results of MSI-Net on OSIE dataset.

	KLDIV ↓	NSS ↑	sAUC ↑	SIM ↑	CC ↑	AUC_Borji ↑
SALICON	1.2781	1.0368	0.6973	0.4126	0.3789	0.6973
baseline	0.6695	1.9517	0.7517	0.6078	0.7200	0.7517
Cropping	**0.6126**	1.9655	0.7677	0.6183	0.7350	**0.7677**
Flipping	0.6582	1.9837	0.7619	0.6134	0.7250	0.7620
Color Transform	0.6511	1.9665	0.7585	0.6126	0.7251	0.7585
Random Rotation	0.6278	1.9686	0.7625	0.6157	0.7304	0.7625
Contrast	0.6753	1.9608	0.7522	0.6093	0.7233	0.7522
JPEG	0.6510	1.9781	0.7587	0.6149	0.7281	0.7587
Noise	0.6684	1.9790	0.7536	0.6130	0.7276	0.7536
Shearing	0.7039	1.9112	0.7504	0.5984	0.7068	0.7504
Inversion	0.7975	1.7584	0.7346	0.5507	0.6505	0.7345
Mirror	0.7444	1.7974	0.7655	0.5584	0.6652	0.7654
Valid	0.6588	**1.9891**	0.7689	0.6201	0.7368	0.7623
VSGAN	0.6595	1.9876	**0.7732**	**0.6221**	**0.7385**	0.7647

of CAT2000 dataset, more or less. VSGAN, however, allows for a slight improvement of performances when coupled with the Valid set. The different of performance on MSI-Net and SAM-ResNet may come from the different size of the parameters. In fact, MSI-Net is much lighter than SAM-ResNet. When trained on CAT2000 training set, it almost gets the best performance. On the contrary, SAM-ResNet has more parameters to be fit which needs more data to get convergence.

3.5.2 MIT1003 and OSIE. MIT1003 dataset is split into training and validation set with a ratio of 800:203. Since MIT1003 dataset only contains 1003 pairs of data, we used MIT1003 dataset to fine-tune the best model training by SALICON dataset and tested on OSIE dataset to get the baseline.

We conduct similar experiments with CAT2000 experiments. In the Table 7, we find that the results are very similar to whose on CAT2000 test dataset. The only difference is that Flipping has positive effect to mitigate overfitting. When using Flipping method to change the images and visual saliency maps on CAT2000 dataset, it may generate wrong ground truth since the CAT2000 dataset has 20 different sets data which contains Flipping. All other kinds of methods showed the same results with those in CAT2000 dataset. For the last two lines of Table 7, VSGAN ranks 1st in most metrics.

On SAM-ResNet, the results are very different. As is shown on Table 8, the models training on the MIT1003 training set coupled with augmented datasets perform worse than the model training

Table 8: Results of SAM-ResNet on OSIE dataset.

	KLDIV ↓	NSS ↑	sAUC ↑	SIM ↑	CC ↑	AUC_Borji ↑
SALICON	1.5282	0.5048	0.6472	0.3210	0.1870	0.6472
baseline	**0.7160**	**2.0233**	**0.7489**	**0.6180**	**0.7270**	**0.7489**
Cropping	1.4832	1.0273	0.6968	0.4140	0.3783	0.6968
Color Transform	1.6598	0.8524	0.7046	0.3784	0.3134	0.7046
Random Rotation	1.6109	0.7851	0.6989	0.3691	0.2862	0.6989
Contrast	1.4600	0.9103	0.7181	0.3906	0.3364	0.7181
JPEG	1.6179	0.8930	0.7002	0.3827	0.3279	0.7002
Shearing	1.5869	0.9787	0.7046	0.3983	0.3639	0.7046
Noise	1.5869	0.9787	0.7182	0.4132	0.3934	0.7181
Flipping	1.4565	1.0365	0.7224	0.4104	0.3886	0.7224
Mirroring	1.5969	0.9448	0.7109	0.4132	0.3934	0.7181
Inversion	1.5567	0.9166	0.7090	0.3878	0.3360	0.7002
Valid	-	-	-	-	-	-
VSGAN	1.3572	1.3201	0.7228	0.4536	0.4872	0.7235

on MIT1003 training set. It may be that SAM-ResNet is over-fitting on the MIT1003 dataset, and thus doubling the size of the dataset might be detrimental to the performances of the model on the OSIE dataset. Another explanation for this gap in performances for the augmentations between MSI-Net and SAM-ResNet may cause by their architecture and loss functions. As we mentioned before, MSI-Net is a lighter network and used KL-divergence loss. SAM-ResNet used a linear combination of three three different loss, thus creating more constraints, making the network more prone to overfit on the training set. From the last row of Table 8, we also notice the overfitting problem of VSGAN but less than others.

4 CONCLUSION

In this paper, we transfer the problem of data augmentation for visual saliency models into image inpainting task by masking the weakly-salient area of the images and synthesizing new images to extend the training data. We have proposed VSGAN to achieve good performance for image inpainting when work on visual saliency datasets. Experiments are conducted on the several datasets to show the effectiveness of our proposed method on visual saliency modeling.

REFERENCES

[1] Ali Borji. 2019. Saliency prediction in the deep learning era: Successes and limitations. *IEEE transactions on pattern analysis and machine intelligence* 43, 2 (2019), 679–700.

[2] Ali Borji and Laurent Itti. 2015. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581* (2015).

[3] Zhaohui Che, Ali Borji, Guangtao Zhai, Xiongkuo Min, Guodong Guo, and Patrick Le Callet. 2019. How is gaze influenced by image transformations? dataset and model. *IEEE Transactions on Image Processing* 29 (2019), 2287–2300.

[4] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2018. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing* 27, 10 (2018), 5142–5154.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

[8] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1072–1080.

[9] Tilke Judd, Fredo Durand, and Antonio Torralba. 2011. Fixations on low-resolution images. *Journal of Vision* 11, 4 (2011), 14–14.

[10] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*. IEEE, 2106–2113.

[11] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

[12] Chelhwon Kim and Peyman Milanfar. 2013. Visual saliency in noisy images. *Journal of vision* 13, 4 (2013), 5–5.

[13] A KingaD. 2015. A methodforstochasticoptimization. *Anon. InternationalConferenceon Learning Representations. SanDego: ICLR* (2015).

[14] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. 2020. Contextual encoder–decoder network for visual saliency prediction. *Neural Networks* 129 (2020), 261–270.

[15] Matthias Kummerer, Thomas SA Wallis, and Matthias Bethge. 2018. Saliency benchmarking made easy: Separating models, maps and metrics. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 770–787.

[16] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*. 85–100.

[17] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. 2021. Pd-gan: Probabilistic diverse gan for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9371–9381.

[18] Jianxun Lou, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu. 2021. TranSalNet: Visual saliency prediction using transformers. *CoRR* abs/2110.03593 (2021). arXiv:2110.03593 https://arxiv.org/abs/2110.03593

[19] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2337–2346.

[20] Xu Wang, Lin Ma, Sam Kwong, and Yu Zhou. 2018. Quaternion representation based visual saliency for stereoscopic image quality assessment. *Signal Processing* 145 (2018), 202–213.

[21] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. 2014. Predicting human gaze beyond pixels. *Journal of vision* 14, 1 (2014), 28–28.

[22] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 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.

[23] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. 2021. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428* (2021).

[24] Xu Zheng, Tejo Chalasani, Koustav Ghosal, Sebastian Lutz, and Aljosa Smolic. 2019. Stada: Style transfer as data augmentation. *arXiv preprint arXiv:1909.01056* (2019).

[25] Xinyue Zhu, Yifan Liu, Jiahong Li, Tao Wan, and Zengchang Qin. 2018. Emotion classification with data augmentation using generative adversarial networks. In *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 349–360.