Inharmonious Region Localization by Magnifying Domain Discrepancy

Jing Liang[1], Li Niu[1*], Penghao Wu[1], Fengjun Guo[2], Teng Long[2]

[1]MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
[2]INTSIG
{leungjing,ustcnewly,wupenghaoCraig}@sjtu.edu.cn, {fengjun_guo, mike_long}@intsig.net

Abstract

Inharmonious region localization aims to localize the region in a synthetic image which is incompatible with surrounding background. The inharmony issue is mainly attributed to the color and illumination inconsistency produced by image editing techniques. In this work, we tend to transform the input image to another color space to magnify the domain discrepancy between inharmonious region and background, so that the model can identify the inharmonious region more easily. To this end, we present a novel framework consisting of a color mapping module and an inharmonious region localization network, in which the former is equipped with a novel domain discrepancy magnification loss and the latter could be an arbitrary localization network. Extensive experiments on image harmonization dataset show the superiority of our designed framework.

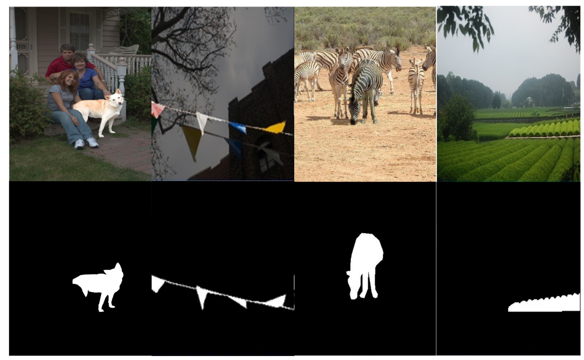

Figure 1: We show the examples of inharmonious synthetic images in the top row and their inharmonious region masks in the bottom row.

Introduction

With the rapid development of image editing techniques and tools (*e.g.*, appearance adjustment, copy-paste), users can blend and edit existing source images to create fantastic images that are only limited by an artist's imagination. However, some manipulated regions in the created synthetic images may have inconsistent color and lighting statistics with the background, which could be attributed to careless editing or the difference among source images (*e.g.*, capture condition, camera setting, artistic style). We refer to such regions as inharmonious regions (Liang, Niu, and Zhang 2021), which will remarkably downgrade the quality and fidelity of synthetic images.

Recently, the task of inharmonious region localization (Liang, Niu, and Zhang 2021) has been proposed to identify the inharmonious regions. When the inharmonious regions are identified, users can manually adjust the inharmonious regions or employ image harmonization methods (Tsai et al. 2017; Cong et al. 2020; Cun and Pun 2020; Cong et al. 2021) to harmonize the inharmonious regions, yielding the images with higher quality and fidelity.

To the best of our knowledge, the only existing inharmonious region localization method is DIRL (Liang, Niu, and Zhang 2021), which attempted to fuse multi-scale features and avoid redundant information. However, DIRL is a rather general model without exploiting the uniqueness of this task, that is, the discrepancy between inharmonious region and background. Besides, the performance of DIRL is still far from satisfactory when the inharmonious region is surrounded by cluttered background or objects that have similar shapes to the inharmonious region.

Considering the uniqueness of inharmonious region localization task, we refer to each suite of color and illumination statistics as one domain following (Cong et al. 2020, 2021). Thus, the inharmonious region and the background belong to two different domains. In this work, we propose a novel method based on a simple intuition: *can we transform the input image to another color space to magnify the domain discrepancy between inharmonious region and background, so that the model can identify the inharmonious region more easily?*

To achieve this goal, we propose a framework composed of two components: one color mapping module and one inharmonious region localization network. First, the color mapping module transforms the input image to another color space. Then, the inharmonious region localization network detects the inharmonious region based on the transformed image. For color mapping module, we extend

*Corresponding Author.

HDRNet (Gharbi et al. 2017) to improved HDRNet (iHDR-Net). HDRNet is popular and has achieved great success in previous works (Zhou et al. 2021; Xia et al. 2020; Wang et al. 2019). Similar to HDRNet, iHDRNet learns region-specific and intensity-specific color transformation parameters, which are applied to transform each input image adaptively. After color transformation, we expect that the domain discrepancy between inharmonious region and background could be magnified, so that the region localization network can identify the inharmonious region more easily. With this purpose, we leverage encoder to extract the domain-aware codes from inharmonious region and background before and after color transformation, in which the domain-aware codes are expected to contain the color and illumination statistics. Then, we design a Domain Discrepancy Magnification (DDM) loss to ensure that the distance of domain-aware codes between inharmonious region and background becomes larger after color transformation. Furthermore, we employ a Direction Invariance (DI) loss to regularize the domain-aware codes. For inharmonious region localization network, we can choose any existing network for region localization and place it under our framework. We refer to our framework as MadisNet (**Ma**gnifying **d**omain d**is**crepancy).

We conduct experiments on the benchmark dataset iHarmony4 (Cong et al. 2020), which shows that our proposed framework outperforms DIRL (Liang, Niu, and Zhang 2021) and the state-of-the-art methods from other related fields. Our contributions can be summarized as follows:

- We devise a simple yet effective inharmonious region localization framework which can accommodate any region localization method.
- We are the first to introduce adaptive color transformation to inharmonious region localization, in which improve HDRNet is used as the color mapping module.
- We propose a novel domain discrepancy magnification loss to magnify the domain discrepancy between inharmonious region and background.
- Extensive experiments demonstrate that our framework outperforms existing methods by a large margin (*e.g.*, IoU is improved from 67.85% to 74.44%).

Related Works

Image Harmonization

Image harmonization, which aims to adjust the appearance of foreground to match background, is a long-standing research topic in computer vision. Prior works (Cohen-Or et al. 2006; Sunkavalli et al. 2010; Jia et al. 2006; Pérez, Gangnet, and Blake 2003; Tao, Johnson, and Paris 2010) focused on transferring low-level appearance statistics from background to foreground. Recently, plenty of end-to-end solutions (Tsai et al. 2017; Cong et al. 2020; Ling et al. 2021; Guo et al. 2021; Sofiiuk, Popenova, and Konushin 2021) have been developed for image harmonization, including the first deep learning method (Tsai et al. 2017), domain translation based methods (Cong et al. 2020, 2021), attention based module (Cun and Pun 2020; Hao, Iizuka, and Fukui 2020). Unfortunately, most of them require inharmonious region mask as input, otherwise the quality of harmonized image will be remarkably degraded. S2AM (Cun and Pun 2020) took blind image harmonization into account and predicted inharmonious region mask. However, mask prediction is not the focus of (Cun and Pun 2020) and the quality of predicted masks is very low.

Inharmonious Region Localization

Inharmonious region localization aims to spot the suspicious regions incompatible with background, from the perspective of color and illumination inconsistency. DIRL (Liang, Niu, and Zhang 2021) was the first work on inharmonious region localization, which utilized bi-directional feature integration, mask-guided dual attention, and global-context guided decoder to dig out inharmonious regions. Nevertheless, DIRL did not consider the uniqueness of this task and its performance awaits further improvement. In this work, we propose a novel framework to magnify the discrepancy between inharmonious region and background, which can help the downstream detector distinguish the inharmonious region from background.

Image Manipulation Localization

Another related topic is image manipulation localization, which targets at distinguishing the tampered region from the pristine background. Copy-move, image splicing, removal, and enhancement are the four well-studied types in image manipulation localization, in which image splicing is the most related topic to our task.

Traditional image manipulation localization methods heavily relied on the prior knowledge or strong assumptions on the inconsistency between tampered region and background, such as noise patterns (Pun, Liu, and Yuan 2016), Color Filter Array interpolation patterns (Ferrara et al. 2012), and JPEG-related compression artifacts (Amerini et al. 2014). Recently, deep learning based methods (Wu, AbdAlmageed, and Natarajan 2019; Bappy et al. 2019; Kniaz, Knyaz, and Remondino 2019; Yang et al. 2020) attempted to tackle the image forgery problem by leveraging local patch comparison (Bayar and Stamm 2016; Rao and Ni 2016; Huh et al. 2018; Bappy et al. 2019), forgery feature extraction (Yang et al. 2020; Wu, AbdAlmageed, and Natarajan 2019; Zhou et al. 2020), adversarial learning (Kniaz, Knyaz, and Remondino 2019), and so on. Different from the above image manipulation localization methods, color and illumination inconsistency is the main focus in inharmonious region localization task.

Learnable Color Transformation

In previous low-level computer vision tasks such as image enhancement, many color mapping techniques have been well explored, which meet our demand for color space manipulation. To name a few, HDRNet (Gharbi et al. 2017) learned a guidance map and a bilateral grid to perform instance-aware linear color transformation. Zeng *et al.* (Zeng et al. 2020) exploited 3D Look Up Table (LUT) for color transformation. DCENet (Guo et al. 2020) iteratively estimated color curve parameters to correct color. In this work, we adopt the improved version of HDRNet (Gharbi et al. 2017) as color mapping module to magnify the domain discrepancy between inharmonious region and background.

Our Approach

Given an input synthetic image $\boldsymbol{I}$, inharmonious region localization targets at predicting a mask $\hat{\boldsymbol{M}}$ that distinguishes the inharmonious region from the background region. Since the perception of inharmonious region is attributed to color and illumination inconsistency, we expect to find a color mapping $\mathcal{F} : \boldsymbol{I} \mapsto \boldsymbol{I}'$ so that the downstream localization network G can capture the discrepancy between inharmonious region and background more easily. As shown in Figure 2, the whole framework consists of two stages: color mapping stage and inharmonious region localization stage. In the color mapping stage, we derive color transformation coefficients $\boldsymbol{A}$ from the color mapping module and perform color transformation to synthetic image $\boldsymbol{I}$ to produce the retouched image $\boldsymbol{I}'$. We assume that the retouched image $\boldsymbol{I}'$ will be exposed larger discrepancy between the inharmonious region and the background. To impose this constraint, we propose a domain discrepancy magnification loss and a direction invariance loss based on the extracted domain-aware codes of inharmonious regions and background regions in $\boldsymbol{I}$ and $\boldsymbol{I}'$. In the inharmonious region localization stage, the retouched image $\boldsymbol{I}'$ is delivered to the localization network G to spot the inharmonious region, yielding the inharmonious mask $\hat{\boldsymbol{M}}$. We will detail two stages in Section and Section respectively.

Color Mapping Stage

Color Manipulation: In some localization tasks (Panzade, Prakash, and Maheshkar 2016; Roy and Bandyophadyay 2013; Cho, Sung, and Jun 2016; Beniak, Pavlovicova, and Oravec 2008), input images are first converted from RGB color space to other color spaces (*e.g.*, HSV (Panzade, Prakash, and Maheshkar 2016; Roy and Bandyophadyay 2013), YCrCb (Cho, Sung, and Jun 2016; Beniak, Pavlovicova, and Oravec 2008)), in which the chroma and illumination distribution are more easily characterized. However, these color mappings are prefixed and cannot satisfy the requirement of inharmonious region localization task. Therefore, we seek to learn an instance-aware color mapping $\mathcal{F} : \boldsymbol{I} \mapsto \boldsymbol{I}'$, to promote the learning of downstream localization network. Considering the popularity of HDRNet (Gharbi et al. 2017) and its remarkable success in color manipulation task (Zhou et al. 2021; Xia et al. 2020; Wang et al. 2019), we build our color mapping module inheriting the spirits of HDRNet. HDRNet (Gharbi et al. 2017) implements local and global feature integration to keep texture details, producing a bilateral grid. To preserve edge information, they also learn an intensity map named guidance map and perform data-dependent lookups in the bilateral grid to generate region-specific and intensity-specific color transformation coefficients. For more technique details, please refer to (Gharbi et al. 2017).

We make two revisions for HDRNet. First, we first use central difference convolution layers (Yu et al. 2020) to extract local features, in which a hyperparameter θ tradeoffs the contribution between vanilla convolution and central difference convolution. As claimed in (Yu et al. 2020), introducing central difference convolution into vanilla convolution can enhance the generalization ability and modeling capacity. Then, we apply a self-attention layer (Zhang et al. 2019) to aggregate global information, which is adept at capturing long-range dependencies between distant pixels. We use the processed features to produce the bilateral grid and the remaining steps are the same as HDRNet. We refer to the improved HDRNet as iHDRNet. The detailed comparison between HDRNet and iHDRNet can be found in Supplementary.

Analogous to HDRNet, iHDRNet learns region-specific and intensity-specific color transformation coefficients $\boldsymbol{A} = [\boldsymbol{K}, \boldsymbol{b}] \in \mathbb{R}^{H\times W\times 3\times 4}$ with $\boldsymbol{K} \in \mathbb{R}^{H\times W\times 3\times 3}$ and $\boldsymbol{b} \in \mathbb{R}^{H\times W\times 3\times 1}$, where H and W are the height and width of input image $\boldsymbol{I}$ respectively. With color transformation coefficients $\boldsymbol{A}$, the inharmonious image $\boldsymbol{I}$ could be mapped to the retouched image $\boldsymbol{I}'$. Formally, for each pixel at location p, $\boldsymbol{I}'(p) = \boldsymbol{A}(p) \cdot [\boldsymbol{I}(p), \mathbf{1}]^T = \boldsymbol{K}(p)\boldsymbol{I}(p) + \boldsymbol{b}(p)$, where $\boldsymbol{K}(p) \in \mathbb{R}^{3\times 3}, \boldsymbol{b}(p) \in \mathbb{R}^{3\times 1}$ are the transform coefficients at location p.

Domain Discrepancy Magnification: We expect that the color and illumination discrepancy between the inharmonious region and the background is enlarged after color transformation. Following (Cong et al. 2020, 2021), we refer to each suite of color and illumination statistics as one domain. Then, we employ a domain encoder E_{dom} to extract the domain-aware codes of inharmonious region and background separately from $\boldsymbol{I}$ and $\boldsymbol{I}'$. Note that we name the extracted code as domain-aware code instead of domain code, because the extracted code is expected to contain the color/illumination statistics but may also contain the content information (*e.g.*, semantic layout). For the latent feature space, we select the commonly used intermediate features from the fixed pre-trained VGG-19 (Simonyan and Zisserman 2014) and pack them into the partial convolution layer (Liu et al. 2018) to derive region-aware features. The domain encoder takes an image and a mask as input. Each partial convolutional layer performs convolution operation only within the masked area, where the mask is updated by rule and the information leakage from the unmasked area is avoided. At the end of E_{dom}, features are averaged along spatial dimensions and projected into a shape-independent domain-aware code. We denote the domain-aware codes of inharmonious region (*resp.*, background) of $\boldsymbol{I}$ as $\boldsymbol{z}_f$ (*resp.*, $\boldsymbol{z}_b$). Similarly, we denote the domain-aware code of inharmonious region (*resp.*, background) of $\boldsymbol{I}'$ as $\boldsymbol{z}'_f$ (*resp.*, $\boldsymbol{z}'_b$). Note that the domain encoder E_{dom} is only used in the training phase, and only the projector is trainable while other components are frozen.

Domain Discrepancy Magnification Loss: To ensure that the color/illumination discrepancy between inharmonious region and background is enlarged, we enforce the distance between the domain-aware codes of inharmonious re-

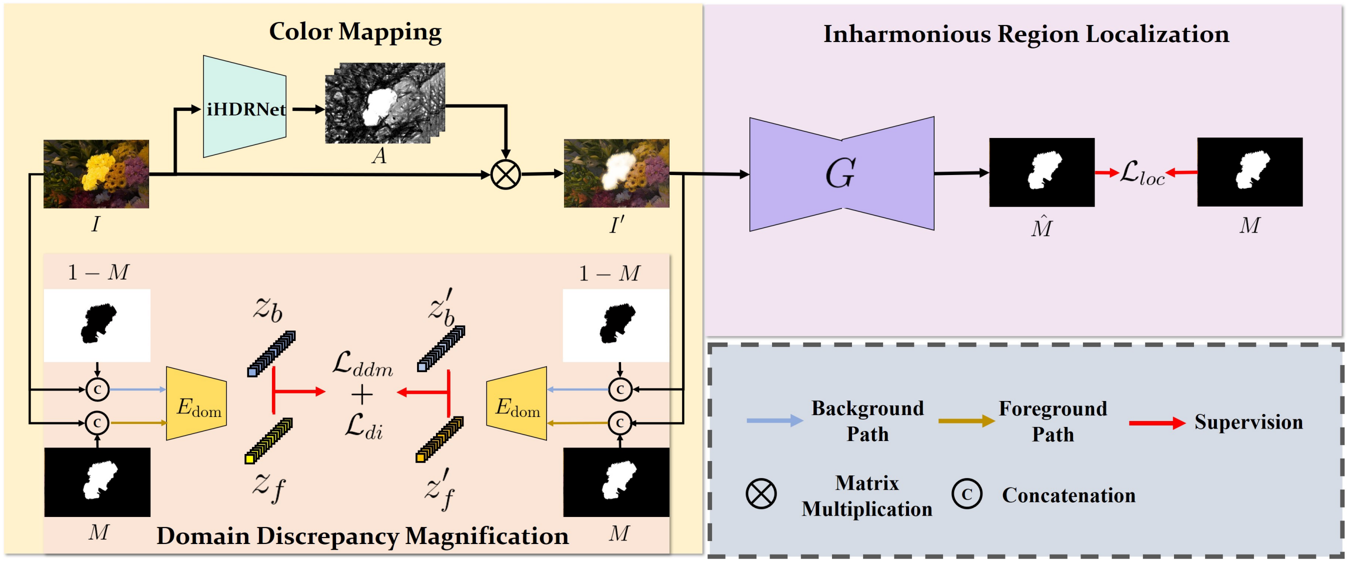

Figure 2: The illustration of our proposed framework which consists of color mapping stage and inharmonious region localization stage. Our color mapping module iHDRNet predicts the color transformation coefficients $\boldsymbol{A}$ for the input image $\boldsymbol{I}$, and the transformed image $\boldsymbol{I}'$ is fed into G to produce the inharmonious region mask $\hat{\boldsymbol{M}}$.

gion and background of retouched image $\boldsymbol{I}'$ to be larger than that of original image $\boldsymbol{I}$. To this end, we propose a novel Domain Discrepancy Magnification (DDM) loss as follows,

$$\mathcal{L}_{ddm} = \max\left(d(\boldsymbol{z}_f, \boldsymbol{z}_b) - d(\boldsymbol{z}'_f, \boldsymbol{z}'_b) + m, 0\right), \tag{1}$$

where $d(\cdot, \cdot)$ measures the Euclidean distance between two domain-aware codes, and the margin m is set as 0.01 via cross-validation. In this way, the distance between $\boldsymbol{z}'_f$ and $\boldsymbol{z}'_b$ is enforced to be larger than the distance between $\boldsymbol{z}_f$ and $\boldsymbol{z}_b$ by a margin m. One issue is that the domain-aware codes may also contain content information (*e.g.*, semantic layout). However, the content difference between inharmonious region and background remains unchanged after color transformation, so we can deem $d(\boldsymbol{z}_f, \boldsymbol{z}_b) - d(\boldsymbol{z}'_f, \boldsymbol{z}'_b)$ as the change in domain difference after color transformation.

Direction Invariance Loss: In practice, we find that solely using (1) might lead to the corruption of domain-aware code space without necessary regularization. Inspired by StyleGAN-NADA (Gal et al. 2021), we calculate the domain discrepancy vector $\Delta\boldsymbol{z} = \boldsymbol{z}_f - \boldsymbol{z}_b$ (*resp.*, $\Delta\boldsymbol{z}' = \boldsymbol{z}'_f - \boldsymbol{z}'_b$) between inharmonious region and background in the input (*resp.*, retouched) image. Then, we align the direction of domain discrepancy vector of input image with that of retouched image, using the following Direction Invariance (DI) loss:

$$\mathcal{L}_{di} = 1 - \langle \Delta\boldsymbol{z}, \Delta\boldsymbol{z}' \rangle, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ means the cosine similarity. Intuitively, we expect that the direction of domain discrepancy roughly stays the same after color transformation. There could be some other possible regularizers for domain-aware codes, but we observe that Direction Invariance (DI) loss in (2) empirically works well.

Inharmonious Region Localization Stage

In the inharmonious region localization stage, the retouched image $\boldsymbol{I}'$ is delivered to the localization network G, which can dig out the inharmonious region from $\boldsymbol{I}'$ and produce the inharmonious mask $\hat{\boldsymbol{M}}$.

The focus of this paper is a novel inharmonious region localization framework by magnifying the domain discrepancy. This framework can accommodate an arbitrary localization network G, such as inharmonious region localization method DIRL (Liang, Niu, and Zhang 2021), segmentation methods (Ronneberger, Fischer, and Brox 2015; Chen et al. 2017), and so on. In our experiments, we try using DIRL (Liang, Niu, and Zhang 2021) and UNet (Ronneberger, Fischer, and Brox 2015) as the localization network.

After determining the region localization network, we wrap up its original loss terms (*e.g.*, binary-cross entropy loss, intersection over union loss) as a localization loss $\mathcal{L}_{loc}$. Together with our proposed domain discrepancy magnification (DDM) loss in (1) and direction invariance (DI) loss in (2), the total loss of our framework could be written as

$$\mathcal{L}_{total} = \lambda_{ddm}\mathcal{L}_{ddm} + \lambda_{di}\mathcal{L}_{di} + \mathcal{L}_{loc}, \tag{3}$$

where the trade-off parameter λ_{ddm} and λ_{di} depend on the downstream localization network.

Experiments

Datasets and Implementation Details

Following (Liang, Niu, and Zhang 2021), we conduct experiments on the image harmonization dataset iHarmony4 (Cong et al. 2020), which provides inharmonious images with their corresponding inharmonious region masks. iHarmony4 is composed of four sub-datasets: HCOCO, HFlickr, HAdobe5K, HDay2Night. For HCOCO and HFlickr datasets, the inharmonious images are obtained by adjusting the color and lighting statistics of foreground. For HAdobe5K and HDay2Night datasets, the inharmonious images are obtained by overlaying the foreground with the counterpart of the same scene retouched with a different style or captured in a different condition. Therefore, the inharmonious images of the four sub-datasets will give people inharmonious perception mainly due to color and lighting inconsistency, which conforms to our definition of the inharmonious region. Moreover, suggested by DIRL (Liang, Niu, and Zhang 2021), we simply discard the images with foreground occupying larger than 50% area, which avoids the ambiguity that background can also be deemed as inharmonious region. Following (Liang, Niu, and Zhang 2021), the training set and test set are tailored to 64255 images and 7237 images respectively.

All experiments are conducted on a workstation with an Intel Xeon 12-core CPU(2.1 GHz), 128GB RAM, and a single Titan RTX GPU. We implement our method using Pytorch (Paszke et al. 2019) with CUDA v10.2 on Ubuntu 18.04 and set the input image size as 256×256. We choose Adam optimizer (Kingma and Ba 2014) with the initial learning rate 0.0001, batch size 8, and momentum parameters $\beta_1 = 0.5, \beta_2 = 0.999$. The hyper-parameter λ_{ddm} and λ_{di} in Eqn. (3) are set as 0.01 for DIRL(Liang, Niu, and Zhang 2021) and 0.001 for UNet (Ronneberger, Fischer, and Brox 2015) respectively. *The detailed network architecture of domain encoder and iHDRNet can be found in Supplementary.*

For quantitative evaluation, we calculate Average Precision (AP), F_1 score, and Intersection over Union (IoU) based on the predicted mask $\hat{M}$ and the ground-truth mask M following (Liang, Niu, and Zhang 2021).

Baselines

To the best of our knowledge, DIRL (Liang, Niu, and Zhang 2021) is the only existing method designed for inharmonious region localization method. Therefore, we also consider other works from related fields. 1) blind image harmonization method S2AM (Cun and Pun 2020); 2) image manipulation detection methods: MantraNet (Wu, AbdAlmageed, and Natarajan 2019), MFCN (Salloum, Ren, and Kuo 2018), MAGritte (Kniaz, Knyaz, and Remondino 2019), H-LSTM (Bappy et al. 2019), SPAN (Hu et al. 2020); 3) salient object detection methods: F3Net (Wei, Wang, and Huang 2020), GATENet (Zhao et al. 2020), MINet (Pang et al. 2020); 4) semantic segmentation methods: UNet (Ronneberger, Fischer, and Brox 2015), DeepLabv3 (Chen et al. 2017), HRNet-OCR (Sun et al. 2019).

Experimental Results

Quantitative Comparison The quantitative results are summarized in Table 1. All of the baseline results are directly copied from (Liang, Niu, and Zhang 2021) except SPAN, GATENet, F3Net, and MINet. For fair comparison, we trained the baselines from scratch. One observation is that image manipulation localization methods (Wu, AbdAlmageed, and Natarajan 2019; Kniaz, Knyaz, and Remondino 2019; Bappy et al. 2019; Hu et al. 2020) are weak in localizing the inharmonious region. One possible explanation is that they focus on the noise pattern and forgery feature extraction while paying less attention to the low-level statistics of color and illumination. We also notice that salient object detection methods (Wei, Wang, and Huang 2020; Zhao et al. 2020) also achieve worse performance than the semantic segmentation methods (Ronneberger, Fischer, and Brox 2015; Chen et al. 2017; Sun et al. 2019) while MINet (Pang et al. 2020) beats all of the semantic segmentation methods in AP metric. In S2AM (Cun and Pun 2020), they predict an inharmonious region mask as side product to indicate the region to be harmonized. Unfortunately, the quality of inharmonious mask is far from satisfactory since image harmonization is their main focus. Another interesting observation is that typical segmentation methods achieve the most competitive performance among the methods that are not specifically designed for inharmonious region localization. It might be attributed to that semantic segmentation methods are originally designed in a general framework and generalizable to inharmonious region localization task.

Methods	Evaluation Metrics		
	AP(%) ↑	F_1 ↑	IoU(%) ↑
UNet	74.90	0.6717	64.74
DeepLabv3	75.69	0.6902	66.01
HRNet-OCR	75.33	0.6765	65.49
MFCN	45.63	0.3794	28.54
MantraNet	64.22	0.5691	50.31
MAGritte	71.16	0.6907	60.14
H-LSTM	60.21	0.5239	47.07
SPAN	65.94	0.5850	54.27
F3Net	61.46	0.5506	47.48
GATENet	62.43	0.5296	46.33
MINet	77.51	0.6822	63.04
S2AM	43.77	0.3029	22.36
DIRL	80.02	0.7317	67.85
MadisNet(UNet)	81.15	0.7372	67.28
MadisNet(DIRL)	**85.86**	**0.8022**	**74.44**

Table 1: Quantitative comparison with baseline methods on iHarmony4 dataset. The best results are denoted in boldface.

Since our framework can accommodate any region localization network, we explore using UNet and DIRL under our framework, which are referred to as MadisNet(UNet) and MadisNet(DIRL) respectively. It can be seen that MadisNet(DIRL) (*resp.*, MadisNet(UNet)) outperforms DIRL (*resp.*, UNet). MadisNet(DIRL) beats the existing inharmonious region localization method and all of the state-of-the-art methods from related fields by a large margin, which verifies the effectiveness of our framework. *In the remainder of experiment section, we use DIRL as our default region localization network (i.e., "MadisNet" is short for "MadisNet(DIRL)"), unless otherwise specified.*

Qualitative Comparison We show the visualization results as well as baselines in Figure 3, which shows that our method can localize the inharmonious region correctly and

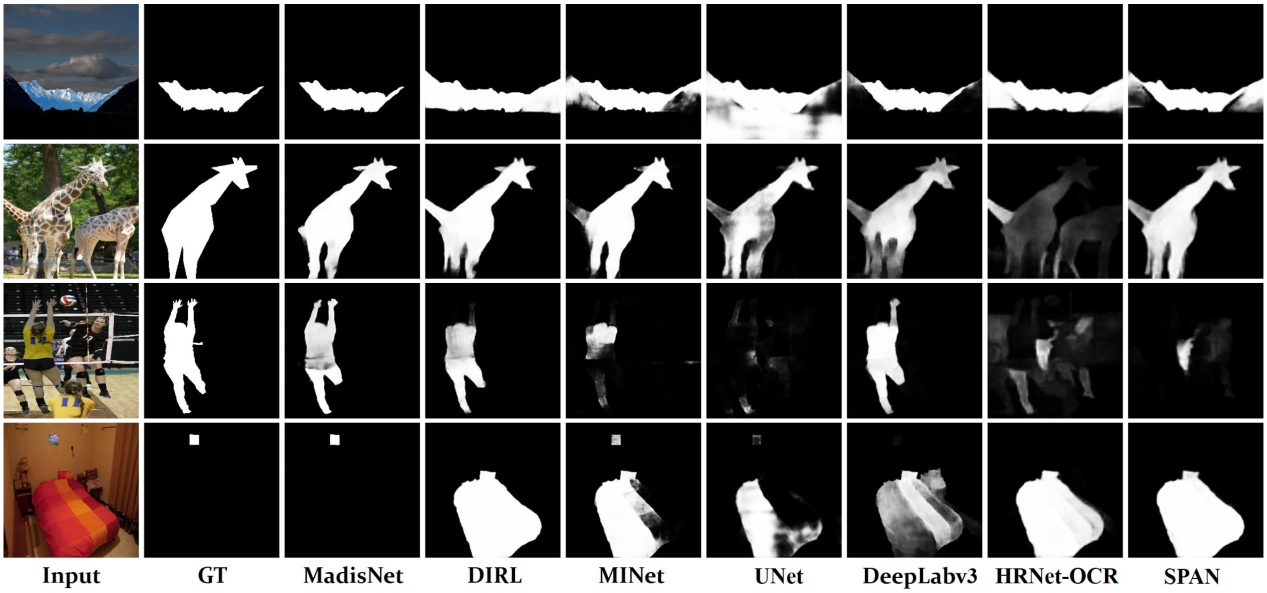

Figure 3: Qualitative comparison with baseline methods. GT is the ground-truth inharmonious region mask.

Components		Evaluation Metrics		
Encoder	Self Attention	AP ↑	F_1 ↑	IoU ↑
VC		81.05	0.7508	69.43
VC	✓	83.54	0.7749	72.08
CDC		82.80	0.7697	71.64
CDC	✓	**85.86**	**0.8022**	**74.44**

Table 2: Ablation study on the components of improved HDRNet. “VC” denotes the vanilla convolution layer and “CDC” means the central difference convolution layer.

Loss Terms	Evaluation Metrics		
	AP(%) ↑	F_1 ↑	IoU(%) ↑
$\mathcal{L}_{loc}$	80.95	0.7401	68.81
$\mathcal{L}_{loc} + \mathcal{L}_{ddm}$	81.86	0.7533	69.84
$\mathcal{L}_{loc} + \mathcal{L}_{di}$	83.18	0.7701	71.67
$\mathcal{L}_{loc} + \mathcal{L}_{ddm} + \mathcal{L}_{di}$	**85.86**	**0.8022**	**74.44**

Table 3: The comparison among different loss terms.

preserve the boundaries accurately. In comparison, the baseline methods may locate the wrong object (row 4) or only detect an incomplete region (row 3). More visualization results can be found in Supplementary.

Ablation Studies

Loss Terms First, we analyze the necessity of each loss term in Table 3. One can learn that our proposed $\mathcal{L}_{ddm}$ and $\mathcal{L}_{di}$ are complementary to each other. Without our proposed $\mathcal{L}_{ddm}$ and $\mathcal{L}_{di}$, the performance is significantly degraded, which proves that $\mathcal{L}_{ddm}$ and $\mathcal{L}_{di}$ play important roles in inharmonious region localization.

iHDRNet Then, we conduct ablation study to validate the effectiveness of CDC layer and self-attention layer in our iHDRNet. The results are summarized in Table 2. By comparing row 1 (*resp.*, 3) and row 2 (*resp.*, 4), we can see that it is useful to employ self-attention layer to capture the long-range dependencies with promising improvement. The comparison between row 2 and row 4 demonstrates that CDC layer performs more favorably than vanilla convolution layer, since CDC layer can capture both intensity-level information and gradient-level information.

Study on Color Manipulation Approaches

To find the best color manipulation approach for inharmonious region localization, we compare our color mapping module iHDRNet with non-learnable color transformation and learnable color transformation.

For non-learnable color transformation, we transform the input RGB image to other color spaces (HSV, YCrCb). Besides, one might concern that whether the learnable color mapping is equivalent to applying random color jittering to the input image, thereby we also take the color jittering augmentation into account. Because the above color transformation approaches do not involve learnable model parameters, we simply apply them to the input images and feed transformed images into the region localization network, during which DDM loss and DI loss are not used.

For learnable color transformation, we compare with LUTs (Zeng et al. 2020), DCENet (Guo et al. 2020), and HDRNet (Gharbi et al. 2017). We directly replace iHDRNet with these color transformation approaches and the other components of our proposed framework remain the same, in which DDM loss and DI loss are used.

The results are summarized in Table 4. We also include the RGB baseline, which means that no color mapping is applied, and the result is identical with DIRL in Table 1. One

Color Mapping	Evaluation Metrics		
	AP(%) ↑	F_1 ↑	IoU(%) ↑
RGB(Baseline)	80.02	0.7317	67.85
HSV	79.86	0.7282	67.40
YCrCb	81.07	0.7484	69.35
ColorJitter	77.50	0.7068	65.40
LUTs	78.39	0.7181	66.16
DCENet	81.90	0.7623	70.92
HDRNet	81.05	0.7508	69.43
iHDRNet	**85.86**	**0.8022**	**74.44**

Table 4: The comparison among different color mapping methods. RGB(baseline) means that no color mapping is applied.

	$d_{f,b} + m < d'_{f,b}$	$d_{f,b} < d'_{f,b}$
Training set	76.22%	99.74%
Test set	77.38%	99.68%

Table 5: The percentage of images whose domain discrepancy is enlarged after color mapping. $d_{f,b}$ is short for $d(\boldsymbol{z}_f, \boldsymbol{z}_b)$ and $d'_{f,b}$ is short for $d(\boldsymbol{z}'_f, \boldsymbol{z}'_b)$. Here $m = 0.01$ as described in section .

can observe that the non-learnable color mapping methods achieve comparable or even worse results compared with RGB baseline. We infer that they are unable to reveal the relationship between inharmonious region and background through simple traditional color transformation. In learnable color mapping methods, LUT achieves even worse scores than RGB baseline. This might be that LUT only learns a global transformation for the whole image without considering local variation. HDRNet and DCENet slightly improve the performance. One possible explanation is that both HDRNet and DCENet are region-specific color manipulation methods, so they could learn color transformation for different regions adaptively to make downstream localization module easily discover the inharmonious region. Our iHDRNet achieves the best results, because the central difference convolution (Yu et al. 2020) can help identify the color inconsistency in synthetic images and the self-attention layer can capture long-range dependencies between distant pixels.

Analyses of Domain Discrepancy

We report the percentage of images whose domain discrepancy is magnified after color transformation in Table 5. For both training set and test set, we report two results: the percentage of $d(\boldsymbol{z}_f, \boldsymbol{z}_b) + m < d(\boldsymbol{z}'_f, \boldsymbol{z}'_b)$ and the percentage of $d(\boldsymbol{z}_f, \boldsymbol{z}_b) < d(\boldsymbol{z}'_f, \boldsymbol{z}'_b)$, in which the latter one is a special case of the former one by setting $m = 0$. From Table 5, we can see that the color mapping module learnt on the training set can generalize to the test set very well. In test set, the domain discrepancy of 77.38% images is enlarged by at least a margin m after color transformation. When we relax the requirement, *i.e.*, $m = 0$, the percentage is as high as 99.68% on the test set.

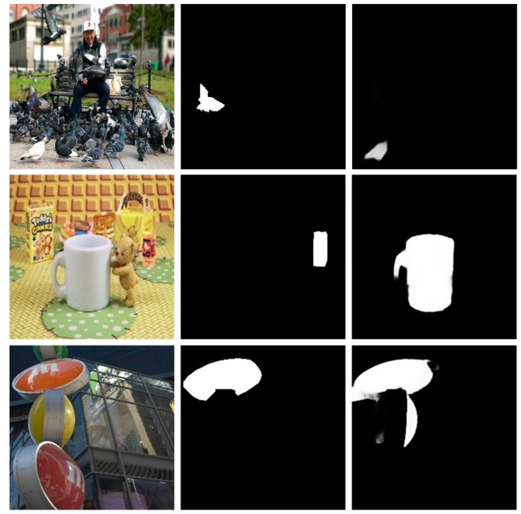

Figure 4: Failure cases of our method. "GT" is the ground-truth inharmonious region mask.

Discussion on Limitation

Figure 4 shows three failure cases of our model. In row 1, our model treats the white pigeon at the bottom left of image as the inharmonious region. We conjecture that the inharmonious region has similar dark tone with surrounding pigeons so that our model is misled by the white pigeon. In row 2, the white cup is recognized as the inharmonious region, probably because the ground-truth inharmonious region and background share warm color tone. In the last row, our model views the yellow light sign as inharmonious region too, because the inharmonious region is brighter than the background. In summary, our model may be weak when the target inharmonious region is surrounded by objects with similar color or intensity.

Results on Four Sub-datasets and Multiple Inharmonious Regions

Because iHarmony4 (Cong et al. 2020) contains four sub-datasets, we show the results on four sub-datasets in Supplementary. Furthermore, this paper mainly focuses on one inharmonious region, but there could be multiple disjoint inharmonious regions in a synthetic image. Therefore, we also demonstrate the ability of our method to identify multiple disjoint inharmonious regions in Supplementary.

Conclusion

In this paper, we have proposed a novel framework to resolve the inharmonious region localization problem with color mapping module and our designed domain discrepancy magnification loss. With the process of color mapping module, the inharmonious region could be more easily discovered from the synthetic images. Extensive experiments on iHarmony4 dataset have demonstrated the effectiveness of our approach.

Acknowledgements

This work is partially sponsored by National Natural Science Foundation of China (Grant No. 61902247), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Municipal Science and Technology Key Project (Grant No. 20511100300).

References

Amerini, I.; Becarelli, R.; Caldelli, R.; and Del Mastio, A. 2014. Splicing forgeries localization through the use of first digit features. In *IEEE WIFS*.

Bappy, J. H.; Simons, C.; Nataraj, L.; Manjunath, B.; and Roy-Chowdhury, A. K. 2019. Hybrid LSTM and encoder–decoder architecture for detection of image forgeries. *IEEE Transactions on Image Processing*, 28(7): 3286–3300.

Bayar, B.; and Stamm, M. C. 2016. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *IH&MMSec*.

Beniak, M.; Pavlovicova, J.; and Oravec, M. 2008. Automatic face detection based on chrominance components analysis. In *IWSSIP*.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Cho, H.; Sung, M.; and Jun, B. 2016. Canny text detector: Fast and robust scene text localization algorithm. In *CVPR*.

Cohen-Or, D.; Sorkine, O.; Gal, R.; Leyvand, T.; and Xu, Y.-Q. 2006. Color harmonization. In *ACM SIGGRAPH*.

Cong, W.; Niu, L.; Zhang, J.; Liang, J.; and Zhang, L. 2021. Bargainnet: Background-Guided Domain Translation for Image Harmonization. In *ICME*.

Cong, W.; Zhang, J.; Niu, L.; Liu, L.; Ling, Z.; Li, W.; and Zhang, L. 2020. Dovenet: Deep image harmonization via domain verification. In *CVPR*.

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.

Ferrara, P.; Bianchi, T.; De Rosa, A.; and Piva, A. 2012. Image forgery localization via fine-grained analysis of CFA artifacts. *IEEE Transactions on Information Forensics and Security*, 7(5): 1566–1577.

Gal, R.; Patashnik, O.; Maron, H.; Chechik, G.; and Cohen-Or, D. 2021. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. arXiv:2108.00946.

Gharbi, M.; Chen, J.; Barron, J. T.; Hasinoff, S. W.; and Durand, F. 2017. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4): 1–12.

Guo, C.; Li, C.; Guo, J.; Loy, C. C.; Hou, J.; Kwong, S.; and Cong, R. 2020. Zero-reference deep curve estimation for low-light image enhancement. In *CVPR*.

Guo, Z.; Zheng, H.; Jiang, Y.; Gu, Z.; and Zheng, B. 2021. Intrinsic Image Harmonization. In *CVPR*.

Hao, G.; Iizuka, S.; and Fukui, K. 2020. Image Harmonization with Attention-based Deep Feature Modulation. In *BMVC*.

Hu, X.; Zhang, Z.; Jiang, Z.; Chaudhuri, S.; Yang, Z.; and Nevatia, R. 2020. SPAN: Spatial pyramid attention network for image manipulation localization. In *ECCV*.

Huh, M.; Liu, A.; Owens, A.; and Efros, A. A. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *ECCV*.

Jia, J.; Sun, J.; Tang, C.-K.; and Shum, H.-Y. 2006. Drag-and-drop pasting. *ACM Transactions on graphics (TOG)*, 25(3): 631–637.

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

Kniaz, V. V.; Knyaz, V.; and Remondino, F. 2019. The point where reality meets fantasy: Mixed adversarial generators for image splice detection. In *NeurIPs*.

Liang, J.; Niu, L.; and Zhang, L. 2021. Inharmonious Region Localization. In *ICME*.

Ling, J.; Xue, H.; Song, L.; Xie, R.; and Gu, X. 2021. Region-aware Adaptive Instance Normalization for Image Harmonization. In *CVPR*.

Liu, G.; Reda, F. A.; Shih, K. J.; Wang, T.; Tao, A.; and Catanzaro, B. 2018. Image Inpainting for Irregular Holes Using Partial Convolutions. In *ECCV*.

Pang, Y.; Zhao, X.; Zhang, L.; and Lu, H. 2020. Multi-scale interactive network for salient object detection. In *CVPR*.

Panzade, P. P.; Prakash, C. S.; and Maheshkar, S. 2016. Copy-move forgery detection by using HSV preprocessing and keypoint extraction. In *PDGC*.

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. *arXiv preprint arXiv:1912.01703*.

Pérez, P.; Gangnet, M.; and Blake, A. 2003. Poisson image editing. In *ACM SIGGRAPH*.

Pun, C.-M.; Liu, B.; and Yuan, X.-C. 2016. Multi-scale noise estimation for image splicing forgery detection. *Journal of visual communication and image representation*, 38: 195–206.

Rao, Y.; and Ni, J. 2016. A deep learning approach to detection of splicing and copy-move forgeries in images. In *IEEE WIFS*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

Roy, S.; and Bandyophadyay, S. K. 2013. Face detection using a hybrid approach that combines HSV and RGB. *International Journal of Computer Science and Mobile Computing*, 2(3): 127–136.

Salloum, R.; Ren, Y.; and Kuo, C.-C. J. 2018. Image splicing localization using a multi-task fully convolutional network (MFCN). *Journal of Visual Communication and Image Representation*, 51: 201–209.

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

Sofiiuk, K.; Popenova, P.; and Konushin, A. 2021. Foreground-aware Semantic Representations for Image Harmonization. In *WACV*.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *CVPR*.

Sunkavalli, K.; Johnson, M. K.; Matusik, W.; and Pfister, H. 2010. Multi-scale image harmonization. *ACM Transactions on Graphics (TOG)*, 29(4): 1–10.

Tao, M. W.; Johnson, M. K.; and Paris, S. 2010. Error-tolerant image compositing. In *ECCV*.

Tsai, Y.-H.; Shen, X.; Lin, Z.; Sunkavalli, K.; Lu, X.; and Yang, M.-H. 2017. Deep image harmonization. In *CVPR*.

Wang, R.; Zhang, Q.; Fu, C.-W.; Shen, X.; Zheng, W.-S.; and Jia, J. 2019. Underexposed photo enhancement using deep illumination estimation. In *CVPR*.

Wei, J.; Wang, S.; and Huang, Q. 2020. F^3Net: Fusion, Feedback and Focus for Salient Object Detection. In *AAAI*.

Wu, Y.; AbdAlmageed, W.; and Natarajan, P. 2019. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *CVPR*.

Xia, X.; Zhang, M.; Xue, T.; Sun, Z.; Fang, H.; Kulis, B.; and Chen, J. 2020. Joint bilateral learning for real-time universal photorealistic style transfer. In *ECCV*.

Yang, C.; Li, H.; Lin, F.; Jiang, B.; and Zhao, H. 2020. Constrained R-CNN: A general image manipulation detection model. In *ICME*.

Yu, Z.; Zhao, C.; Wang, Z.; Qin, Y.; Su, Z.; Li, X.; Zhou, F.; and Zhao, G. 2020. Searching Central Difference Convolutional Networks for Face Anti-Spoofing. In *CVPR*.

Zeng, H.; Cai, J.; Li, L.; Cao, Z.; and Zhang, L. 2020. Learning image-adaptive 3D lookup tables for high performance photo enhancement in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, H.; Goodfellow, I.; Metaxas, D.; and Odena, A. 2019. Self-Attention generative adversarial networks. In *ICML*.

Zhao, X.; Pang, Y.; Zhang, L.; Lu, H.; and Zhang, L. 2020. Suppress and balance: A simple gated network for salient object detection. In *ECCV*.

Zhou, P.; Chen, B.-C.; Han, X.; Najibi, M.; Shrivastava, A.; Lim, S.-N.; and Davis, L. 2020. Generate, segment, and refine: Towards generic manipulation segmentation. In *AAAI*.

Zhou, Y.; Barnes, C.; Shechtman, E.; and Amirghodsi, S. 2021. TransFill: Reference-guided Image Inpainting by Merging Multiple Color and Spatial Transformations. In *CVPR*.