
The Point Where Reality Meets Fantasy: Mixed Adversarial Generators for Image Splice Detection

Vladimir V. Kniaz^{1,2}, Vladimir A. Knyaz^{1,2}

¹State Res. Institute of Aviation Systems (GosNIIAS)
125319, 7, Victorenko str., Moscow, Russia
{vl.kniaz, knyaz}@gosniias.ru

²Moscow Institute of Physics and Technology (MIPT)
141701, 9 Institutskiy per., Dolgoprudny, Russia

Fabio Remondino

Fondazione Bruno Kessler (FBK)
Via Sommarive 18, Trento, Italy
remondino@fbk.eu

Abstract

Modern photo editing tools allow creating realistic manipulated images easily. While fake images can be quickly generated, learning models for their detection is challenging due to the high variety of tampering artifacts and the lack of large labeled datasets of manipulated images. In this paper, we propose a new framework for training of discriminative segmentation model via an adversarial process. We simultaneously train four models: a generative retouching model G_R that translates manipulated image to the real image domain, a generative annotation model G_A that estimates the pixel-wise probability of image patch being either real or fake, and two discriminators D_R and D_A that qualify the output of G_R and G_A . The aim of model G_R is to maximize the probability of model G_A making a mistake. Our method extends the generative adversarial networks framework with two main contributions: (1) training of a generative model G_R against a deep semantic segmentation network G_A that learns rich scene semantics for manipulated region detection, (2) proposing per class semantic loss that facilitates semantically consistent image retouching by the G_R . We collected large-scale manipulated image dataset to train our model. The dataset includes 16k real and fake images with pixel-level annotations of manipulated areas. The dataset also provides ground truth pixel-level object annotations. We validate our approach on several modern manipulated image datasets, where quantitative results and ablations demonstrate that our method achieves and surpasses the state-of-the-art in manipulated image detection. We made our code and dataset publicly available ¹.

1 Introduction

While every image captured by the human eye is real, digital photos can be easily manipulated to present scenes that never existed in reality. Such manipulated image can be easily generated by copying the part of one image into another. This image manipulation is called an image splice and can be used maliciously to create fake news or change historical photos [1]. Recent research [1, 2] suggests that training a model for splice localization is more challenging than other types of object detection problems as the domain of manipulated images is extensive and diverse. Therefore, the collection of the representative training dataset is difficult. Moreover, the forger can adapt to the detection algorithm by changing the manipulation technique. This principle is used in Generative Adversarial Networks (GANs) to train a generator network to synthesize images from noise [3], text descriptions [4], scene graphs [5] or by image-to-image translation [6, 7, 8, 9]. Fake images produced by the generator are evaluated against real images by an adversarial discriminator network that learns

¹<http://zefirus.org/MAG>

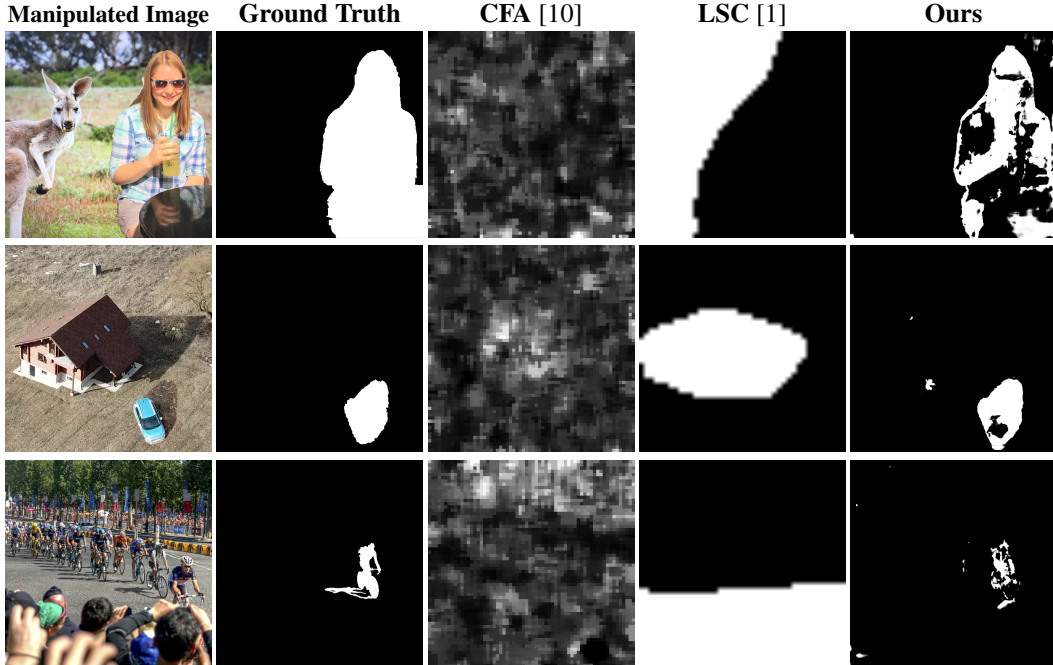


Figure 1: Comparison against two state-of-the-art methods on our *FantasticReality* dataset (Section 3.3). Our results are shown in the last column. Zoom in for details.

to classify them as ‘real’ or ‘fake.’ Even though no ‘fake’ images exist in the training dataset, the discriminator successfully learns to detect them during the training process. We hypothesize that adversarial training of an image-to-image translation generator against a splice localization generator can improve the splice localization accuracy.

In this paper, we propose a Mixed Adversarial Generators (MAG) framework, in which we simultaneously train four models: a generative retoucher G_R , an adversarial generative annotator G_A , and two discriminators D_R and D_A that qualify the output of G_R and G_A . The aim of our retoucher G_R is suppressing image tampering artifacts in the input image splices from the training dataset. We train our adversarial annotator G_A to predict splice localization masks in the ‘retouched’ images generated by the retoucher G_R . The adversarial loss provided by the annotator G_A forces the retoucher G_R to mask those particular tampering artifacts that allow G_A to detect the image splice. Unlike other splice detection models, our annotator G_A learns to adapt to changing tampering techniques of the retoucher G_R . Therefore, our annotator G_A receives a new sample from the manipulated image domain every iteration. Moreover, with the increasing epoch samples are becoming more complex. To further increase the splice localization rate, we train our annotator G_A to predict object classes for the input image. Resulting semantic labeling is used to provide a semantic consistency loss for the retoucher G_R . The semantic loss forces the output of the retoucher G_R to present objects of the same semantic classes as the input image.

Our adversarial generators extend the GAN framework with two key contributions: (1) training of a generative model G_R against a deep semantic segmentation network G_A that learns rich scene semantics for manipulated region detection, (2) proposing per class semantic loss that facilitates semantically consistent image retouching by the G_R . Unlike the recently proposed Sem-GAN model [11], we do not use the pertained segmentation model but train it adversarially. We perform a comprehensive evaluation of our MAG framework, where quantitative results and ablations demonstrate that our annotator G_A achieves and surpasses the state-of-the-art in splice localization on several challenging image splice datasets (see Figure 1 and 3).

We evaluate our retoucher G_R on image-to-image translation tasks to demonstrate that our MAG framework is not limited to the splice localization task. Semantic loss function allows us to train challenging image-to-image translation tasks that are unfeasible for baselines. We also introduce a new *FantasticReality* dataset that includes 16k image splices with pixel-level ground truth annotations

of manipulated areas, and instance and class labels for ten object categories. We made our code and the dataset publicly available.

2 Related Work

Splice detection. Modern splice detection methods fall into three categories: tampering artifacts-based approaches leverage local discrepancies in image noise [12, 13, 14, 15, 16, 17, 18], compression artifacts [19, 20, 21, 22], or camera’s color filter array inconsistencies [10, 23, 24, 25, 26, 27, 28, 29, 30, 31] to detect tampered image regions; consistency-based methods [32, 1] compare pairs of local image patches to localize image areas, where predicted camera model [33, 32] or image metadata [1] are inconsistent with the rest of the image; deep learning-based methods [34, 35, 1, 2, 36] detect image splice regions either by comparison of local patches in a siamese network [1] or using fully convolutional networks [2] to predict labeling of the tampered regions. While many digital image forensic datasets were introduced recently [37, 38, 39, 40, 41], they usually include only several hundreds of photos and do not provide enough of training data for modern methods. Related to our multi-task annotation prediction, Salloum et al. [2] have proposed multi-task training to localize tampered regions and their edges.

Image-to-image translation. Modern methods for image generation conditioned by an input image are trained in either supervised [6, 42, 43, 11, 44], unsupervised [7, 8, 9, 45, 46, 47] or mixed [48] setting. Unsupervised approaches are trained on an unpaired dataset leveraging the latent space assumption [8], the cycle consistency loss [7] or other criteria to learn a mapping from source to target domain. Recent research demonstrates exciting progress in multimodal image-to-image translation [49, 42]. Related to our semantic consistency loss function are the loss functions proposed in Sem-GAN [11] and InstaGAN [50] models. Unlike our MAG framework Sem-GAN model leverages a pretrained segmentation model to provide semantic loss. Unlike InstaGAN [50] model, our retoucher generator G_R does not require instance masks as an input. Closely related to our retoucher generator G_R , Mejjati et al. [9] propose to use attention guided training to perform translation only for the target object.

Most of the modern approaches in the image-to-image translation are based on Generative Adversarial Networks [3], which can capture the sample distribution in the target domain using an adversarial game of two players. Recent research demonstrates that GANs can solve more challenging tasks than image-to-image translation. They can learn complex transforms between physically different domains such as image-to-thermal translation [51, 52, 53, 54, 55], image-to-voxel model transformation [56, 57], and image synthesis from audio data [58]. In our MAG framework, we replace the discriminator network with an adversarial annotator generator G_A . While the discriminator predicts a scalar probability of an input image being either real or fake, our annotator generator G_A predicts a pixel-level probability map of an image patch being either authentic image or splice.

3 Mixed Adversarial Generators

Our goal is training two generator networks adversarially: a splice retoucher G_R and a splice localization annotator G_A . We consider three domains: the input domain $\mathcal{A} \in \mathbb{R}^{W \times H \times 3}$ of potentially manipulated images, the authentic domain $\mathcal{B} \in \mathbb{R}^{W \times H \times 3}$ of untampered images, and the output domain $\mathcal{C} \in [0, 1]^{W \times H \times (2+K)}$ of splice localization and class segmentation masks, where K is the number of predicted object classes. While an image $\mathbf{A} \in \mathcal{A}$ may be either authentic or tampered, all images $\mathbf{B} \in \mathcal{B}$ are authentic, $\mathcal{B} \subset \mathcal{A}$. We use assumptions made by Salloum et al. [2] as the starting point for our generator G_A . Specifically, we train our generator G_A for multi-task prediction of splice segmentation mask $\mathbf{C}_m \in [0, 1]^{W \times H}$, splice edge mask $\mathbf{C}_e \in [0, 1]^{W \times H}$, and object class segmentation $\mathbf{C}_s \in [0, 1]^{W \times H \times K}$. Therefore, we learn a mapping $G_A : (\mathbf{A}) \rightarrow \mathbf{C}$, where $\mathbf{A} \in \mathcal{A}$ is an input potentially manipulated image, $\mathbf{C} \in \mathcal{C}$ is an output tensor obtained by concatenation of $\mathbf{C}_m, \mathbf{C}_e, \mathbf{C}_s$. The goal of our retoucher generator G_R is learning a mapping from manipulated image domain \mathcal{A} to the authentic domain \mathcal{B} . To this end, the aim of adversarial training of the G_R is maximizing the probability of an annotator G_A making a mistake in splice detection of the retouched image $\hat{\mathbf{B}}$. We believe that the retoucher G_R in the training loop facilitates our annotator G_A to learn complicated splice retouching approaches. We use attention-guided learning assumption made by Mejjati et al. [9] as the starting point for our retoucher G_R . We observe the similarity between the attention map proposed by Mejjati et al. [9] and the alpha channel used for the splice

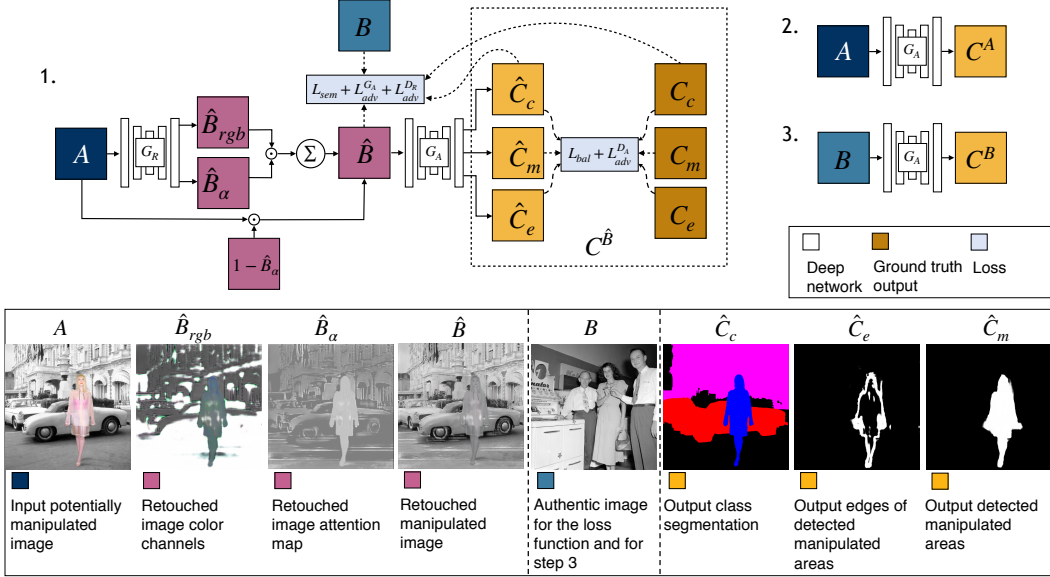


Figure 2: **Our proposed pipeline:** We want our annotator G_A to predict annotations correctly for three kinds of images: retouched spliced images $\hat{B} = G_R(A)$ (1), original spliced images $A \in \mathcal{A}$ from the training dataset (2), and authentic images $B \in \mathcal{B}$ (3). Our retoucher G_R learns to hide a wide range of tampering artifacts such as modern-to-retro photo translation, blurring of tampering edges, and compensating light source inconsistencies. During the training, we feed manipulated images A , retouched images \hat{B} , and authentic images B to our splice localization annotator G_A .

generation. We hypothesize that attention-guided learning of our retoucher G_R allows us to model splice generation with layers in photo-editing applications, e.g., GIMP or Photoshop. We learn a mapping $G_R : (A) \rightarrow (\hat{B}_{rgb}, \hat{B}_\alpha)$, where $\hat{B}_{rgb} \in \mathbb{R}^{W \times H \times 3}$ is an image with the retouched splice area, and $\hat{B}_\alpha \in [0, 1]^{W \times H}$ is the attention map. We obtain the target retouched splice image \hat{B} similarly to [9] by

$$\hat{B} = \hat{B}_\alpha \odot \hat{B}_{rgb} + (1 - \hat{B}_\alpha) \odot A, \quad (1)$$

where \odot is an element-wise product. Our proposed pipeline is presented in Figure 2. We train two discriminator networks D_R, D_A to provide adversarial losses for the output of our generators G_R and G_A . The architecture and the loss function of the retoucher G_R are presented in Section 3.1, whereas the structured loss function of the annotator G_A is described in Section 3.2.

3.1 Retoucher Generator G_R

Architecture. We use the U-Net generator architecture [59] as the starting point for our retoucher G_R . While skip connections of the U-Net generator facilitate robust learning of tampering techniques by our retoucher G_R , deconvolutional layers often introduce checkerboard artifacts in output images. Our annotator G_A quickly learns checkerboard features to detect images produced by our retoucher G_R . To avoid such a scenario, we replaced deconvolutional layers with an upsample layer followed by a convolutional layer, inspired by the architecture proposed in [60]. We term the resulting architecture that is free from the checkerboard artifacts as U-Net-UC (see supplementary material Table 1).

Loss function. Three loss functions govern the training process for our retoucher G_R : $\mathcal{L}_{sem}^{G_A}$, $\mathcal{L}_{adv}^{G_A}$, and $\mathcal{L}_{adv}^{D_R}$, where a superscript indicates the network providing the loss. The aim of our semantic

consistency loss function $\mathcal{L}_{sem}^{G_A}$ is to make the classes of objects in the output image $\hat{\mathbf{B}}$ recognizable by our annotator G_A

$$\mathcal{L}_{sem}^{G_A}(\mathbf{C}_s, \hat{\mathbf{C}}_s) = \mathbb{E}_{\mathbf{B} \sim p(\mathbf{B})} \left[\left\| \mathbf{C}_s - \hat{\mathbf{C}}_s \right\|_1 \right], \quad (2)$$

where $\hat{\mathbf{C}}_s = G_A(\hat{\mathbf{B}})_s$ is the class segmentation produced by our annotator G_A , \mathbf{C}_s is the ground truth class segmentation. Our adversarial annotator loss $\mathcal{L}_{adv}^{G_A}$ stimulates our retoucher G_R to mask tampering artifacts in the input sliced images. In other words, we want to maximize the probability of our annotator G_A making a mistake in splice localization $\hat{\mathbf{C}}_m = G_A(\hat{\mathbf{B}})_m$

$$\mathcal{L}_{adv}^{G_A}(\mathbf{C}_m, \hat{\mathbf{C}}_m) = \mathbb{E}_{\mathbf{B} \sim p(\mathbf{B})} \left[\left\| \mathbf{0}_{W,H} - \hat{\mathbf{C}}_m \right\|_1 \right], \quad (3)$$

where $\mathbf{0}_{W,H}$ is the a splice localization filled with zeros. Finally, we use a discriminator's D_R adversarial loss function to make our image realistic globally

$$\mathcal{L}_{adv}^{D_R}(\hat{\mathbf{B}}) = \mathbb{E}_{\mathbf{B} \sim p(\mathbf{B})} \left[\log(1 - D_R(\hat{\mathbf{B}})) \right]. \quad (4)$$

We obtain the final energy to be optimized by combining all losses

$$\mathcal{L}_R(\mathbf{C}_s, \hat{\mathbf{C}}_s, \mathbf{C}_m, \hat{\mathbf{C}}_m, \hat{\mathbf{B}}) = \lambda_{sem}^{G_A} \cdot \mathcal{L}_{sem}^{G_A} + \lambda_{adv}^{G_A} \cdot \mathcal{L}_{adv}^{G_A} + \lambda_{adv}^{D_R} \cdot \mathcal{L}_{adv}^{D_R}, \quad (5)$$

where we use the loss hyper-parameters $\lambda_{sem}^{G_A} = 10$, $\lambda_{adv}^{G_A} = 10$, $\lambda_{adv}^{D_R} = 0.25$ in our experiments.

3.2 Annotator Generator G_A

Loss function. We train our annotator G_A utilizing a combination of our balanced L^1 loss function \mathcal{L}_{bal} and an adversarial loss $\mathcal{L}_{adv}^{D_A}$. We observe that training our annotator G_A using the L^1 distance $\|\mathbf{C} - \hat{\mathbf{C}}\|$ between the ground-truth and predicted annotations results in a large number of false negatives in splice localizations. We hypothesize that making the penalty for false negatives and false positives equal for each image can improve the overall splice localization score. We implement this hypothesis in our balanced loss function based on the Dice loss [61]

$$\mathcal{L}_{bal}(\mathbf{C}, \hat{\mathbf{C}}) = \underbrace{\sum_{i=1}^{2+K} \frac{|\mathbf{C}_i \cap (1 - \hat{\mathbf{C}}_i)|}{|\mathbf{C}_i|}}_{\text{False negatives}} + \underbrace{\sum_{i=1}^{2+K} \frac{|(1 - \mathbf{C}_i) \cap \hat{\mathbf{C}}_i|}{|1 - \mathbf{C}_i|}}_{\text{False positives}}, \quad (6)$$

where i is the index of an annotation channel. Channel \mathbf{C}_1 provides a splice mask annotation \mathbf{C}_m , channel \mathbf{C}_2 provides a splice edges annotation \mathbf{C}_e . The predicted class labels are given by \mathbf{C}_i for $i \in \{3, 4, \dots, 2 + K\}$, where K is the number of classes ($K = 10$ in our experiments). The area of predicted annotations (white area) in the channel \mathbf{C}_i is given by $|\mathbf{C}_i|$, the background area (black area) in the channel \mathbf{C}_i is given by $|1 - \mathbf{C}_i|$.

We want our annotator G_A to predict annotations correctly for three kinds of images: original spliced images $\mathbf{A} \in \mathcal{A}$ from the training dataset, retouched spliced images $\hat{\mathbf{B}} = G_R(\mathbf{A})$, and authentic images $\mathbf{B} \in \mathcal{B}$. Therefore, for each iteration, we evaluate the loss \mathcal{L}_{bal} on three pairs of ground truth and predicted annotations: $(\mathbf{C}^{\mathbf{A}'}, \hat{\mathbf{C}}^{\mathbf{A}'})$, $(\mathbf{C}^{\mathbf{A}'}, \hat{\mathbf{C}}^{\hat{\mathbf{B}}})$, $(\mathbf{C}^{\mathbf{B}}, \hat{\mathbf{C}}^{\mathbf{B}})$. We use the superscript to denote the corresponding color image for the annotation. Please, note that both original spliced image \mathbf{A} and the retouched spliced image $\hat{\mathbf{B}}$ have the same annotation $\mathbf{C}^{\mathbf{A}'}$ with an adversarial class segmentation mask $\mathbf{C}_s^{\mathbf{A}'} = \mathbf{C}_s^{\mathbf{A}} \odot (1 - \mathbf{C}_m^{\mathbf{A}})$. We want to train our annotator G_A to predict class segmentation adversarially: it must generate the correct class annotations only for authentic image areas and predict empty class annotations for manipulated regions. Specifically, we multiply our ground truth semantic segmentation $\mathbf{C}_s^{\mathbf{A}}$ by an inverted splice localization mask $(1 - \mathbf{C}_m^{\mathbf{A}})$. The

multiplication by the inverted mask leaves authentic areas untouched and removes annotations for manipulated regions.

The aim of our adversarial loss $\mathcal{L}_{adv}^{D_A}$ is to avoid blurry output splice localization masks [6]. It is provided by a conditional discriminator D_A with PatchGAN architecture [6]

$$\mathcal{L}_{adv}^{D_A}(\hat{\mathbf{B}}, \hat{\mathbf{C}}^{\hat{\mathbf{B}}}) = \mathbb{E}_{\mathbf{B} \sim p(\mathbf{B})} \left[\log(1 - D_A(\hat{\mathbf{B}}, \hat{\mathbf{C}}^{\hat{\mathbf{B}}})) \right]. \quad (7)$$

We obtain the resulting energy to optimize by combining four loss functions

$$\mathcal{L}_A = \lambda_{bal} \left(\mathcal{L}_{bal}(\mathbf{C}^{A'}, \hat{\mathbf{C}}^A) + \mathcal{L}_{bal}(\mathbf{C}^{A'}, \hat{\mathbf{C}}^{\hat{\mathbf{B}}}) + \mathcal{L}_{bal}(\mathbf{C}^B, \hat{\mathbf{C}}^B) \right) + \lambda_{adv}^{D_A} \cdot \mathcal{L}_{adv}^{D_A}(\hat{\mathbf{B}}, \hat{\mathbf{C}}^{\hat{\mathbf{B}}}), \quad (8)$$

where we use the loss hyper-parameters $\lambda_{bal} = 1, \lambda_{adv}^{D_A} = 1$ in our experiments.

3.3 FantasticReality Dataset

We collected large-scale image tampering dataset with 16k authentic and 16k tampered images to perform extensive training and evaluation of our MAG model. Compared to previous datasets [37, 38, 39, 40], our *FantasticReality* dataset is more extensive in terms of scene variety and image count. To the best of our knowledge, it is the first tampering dataset that provides both tampering masks and instance and class labels for each image. For each authentic and tampered image, we manually generated instance and class segmentation for ten object classes: person, car, truck, van, bus, building, cat, dog, tram, boat. Examples from the dataset are presented in Figure 1 in the supplementary material.

4 Experiments

We perform extensive experiments to evaluate our MAG model on splice localization. We compare our model to three modern state-of-the-art deep learning splice detection frameworks: ManTra [62], LSC [1], MFCN [2]. We provide a comparison to non-deep learning methods to be consistent with LSC: NOI [18], CFA [10], DCT [19]. ManTra-Net (ManTra) [62] is a self-supervised model that learns to classify 385 image manipulation types. Learned Self-Consistency (LSC) [1] is a self-supervised model. Multi-Task Fully Convolutional Network (MFCN) [2] leverages a deep two-stream architecture to predict splice mask and splice edge mask. Noise Variance (NOI) [18] leverages wavelet analysis to detect inconsistency in noise patterns. Color Filter Array (CFA) [10] searches for inconsistencies in artifacts of demosaicking algorithm to detect tampered regions. JPEG DCT [19] leverages inconsistencies of JPEG blocking artifacts to detect tampered image regions. For the LSC algorithm, we use a pertained model provided by authors. We implemented the MFCN model and train it on the training split of our *FantasticReality* dataset. We train our MAG model on the ‘Rough’ split of our *FantasticReality* dataset. We use a batch size of one and an Adam solver with initial learning rate of $2 \cdot 10^{-4}$. We trained our MAG model for 400 epochs.

We perform evaluation on five manipulated image datasets *CASIA v2.0* [37], *Carvalho* [38], *Columbia* [39], *Realistic Tampering* [40] and our *FantasticReality* dataset. For the fair evaluation, we downscale all images to match the input size 512×512 of our annotator generator G_R . We use the downscaled images to evaluate all baselines and our framework. If two images are used for splice generation, the choice of ‘authentic’ and ‘tampered’ regions is ambiguous. To avoid ambiguity, we follow the method proposed in [1]. Namely, we compare the areas of the ‘background’ image and the ‘pasted’ images. We define the smaller region as the tampered. If the regions are equal, we calculate the mAP score for the original tampering mask and an inverted mask. We use the higher score and term it permuted mAP (p-mAP) similar to [1]. For additional details on the evaluation protocol, please, refer to the supplementary material. Furthermore, we perform ablation studies to demonstrate the influence of each component of our framework on the resulting performance.

4.1 Annotator Generator G_A Evaluation

Splice Localization. We evaluate our model and baselines on the task of splice localization using ground-truth masks of spliced regions. Specifically, we want our model to predict a per-pixel

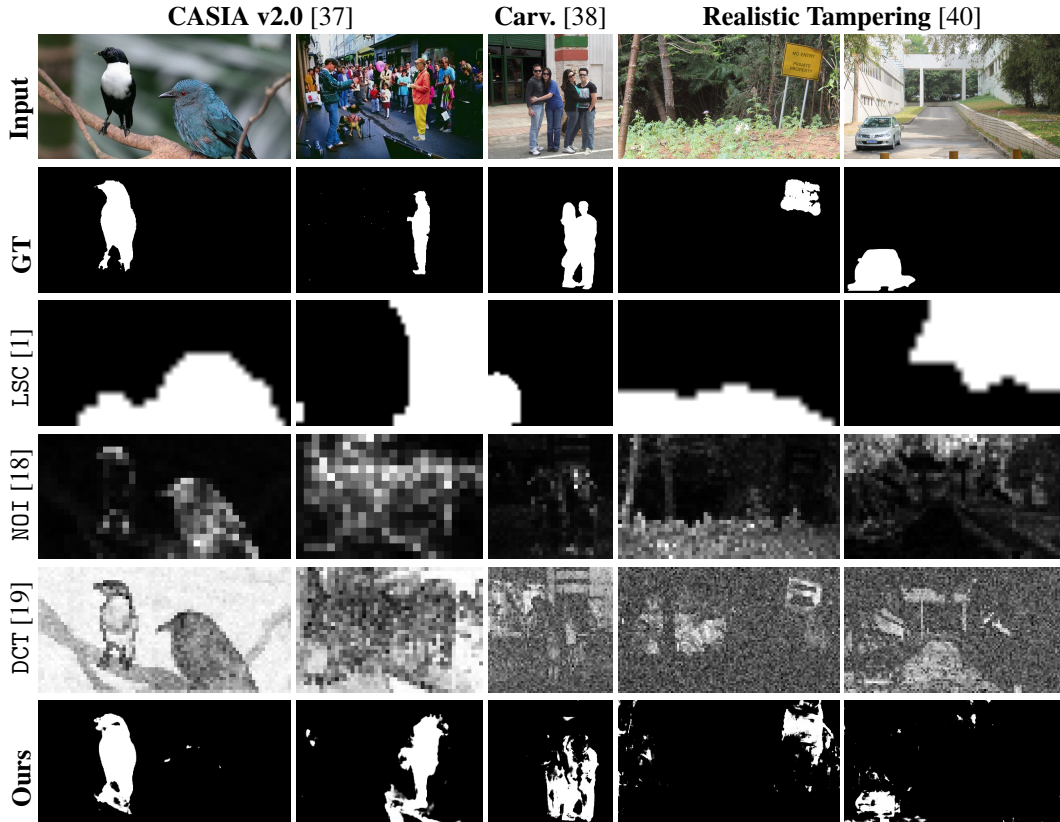


Figure 3: Comparison against the state-of-the-art methods on image splices from CASIAv2, Carvalho and Realistic Tampering datasets. Our results are shown in the last row. Zoom in for details.

Dataset	CASIA v2.0 [37]			Columbia [39]			RT [40]			Carvalho [38]			FantasticReality		
Metric	mAP	p-mAP	cIOU	mAP	p-mAP	cIOU	mAP	p-mAP	cIOU	mAP	p-mAP	cIOU	mAP	p-mAP	cIOU
LSC	0.32	0.41	0.47	0.25	0.44	0.41	0.33	0.47	0.52	0.15	0.33	0.24	0.17	0.45	0.36
CFA	0.37	0.40	0.42	0.39	0.44	0.44	0.45	0.48	0.49	0.32	0.32	0.33	0.45	0.50	0.48
NOI	0.29	0.45	0.46	0.26	0.43	0.40	0.48	0.45	0.50	0.21	0.31	0.21	0.18	0.49	0.29
LSC	0.35	0.39	0.35	0.22	0.42	0.43	0.37	0.49	0.48	0.16	0.37	0.25	0.26	0.51	0.41
MFCN	0.36	0.41	0.48	0.27	0.45	0.42	0.41	0.51	0.36	0.42	0.36	0.37	0.40	0.51	0.46
ManTra	0.40	0.40	0.45	0.48	0.48	0.58	0.50	0.50	0.54	0.33	0.33	0.38	0.57	0.57	0.73
No G_R	0.41	0.35	0.12	0.34	0.32	0.29	0.28	0.35	0.18	0.20	0.37	0.29	0.27	0.45	0.36
Single-task	0.12	0.15	0.24	0.11	0.19	0.21	0.17	0.16	0.14	0.19	0.15	0.18	0.12	0.17	0.21
Ours	0.74	0.74	0.76	0.69	0.69	0.77	0.50	0.51	0.55	0.48	0.48	0.56	0.61	0.61	0.76

Table 1: **Splice Localization:** We evaluate our model on 5 datasets using mean average precision (mAP, permuted-mAP) over pixels and per class IOU (cIOU).

probability of an image patch being tampered. We present results in terms of mAP, permuted mAP [1], and per class Intersection over Union (cIOU) in Table 1 and in Figure 3. Our MAG model achieves state-of-the-art in splice localization on all datasets. The LSC model fails to detect splices when authentic and spliced regions originate from the same camera model and share similar camera metadata.

Ablation Study. We evaluate the necessity of all components of our MAG framework by comparing the splice localization accuracy of several ablated versions of our model presented in Table 1 and Figure 4. Firstly, we evaluate the performance of annotator G_A trained without retoucher G_R (No G_R). Both qualitative and quantitative results demonstrate that the competition of two generators is the critical

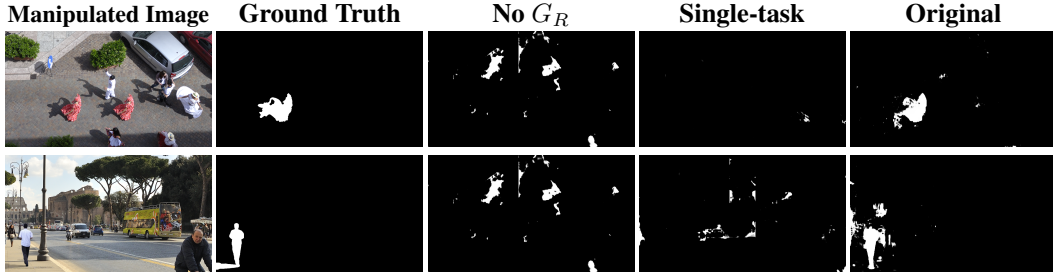


Figure 4: Qualitative results for ablated versions of our MAG framework evaluated on Realistic Tampering dataset.

component of our MAG framework. Secondly, we evaluate our framework trained for the single task of predicting splice area annotations. The results prove that multi-task training outperforms the single-task version of our model (Single-task).

4.2 Semantic-guided Retoucher Generator G_R Evaluation

Examples in Figures 5 and 6 demonstrate how our retoucher G_R gradually removes the tampering artifacts in the input splice A with an increasing epoch. While other deep learning splice detection methods receive both realistic and rough splices from the first training epoch, our annotator G_A sees only rough splices at the first epoch. With an increasing epoch, retoucher G_R produces more complicated splices, which allows G_A to focus attention on the sophisticated tampering techniques that could appear in real splices. We believe that this is the main reason why our MAG framework achieves state-of-the-art results and outperforms other deep learning methods.

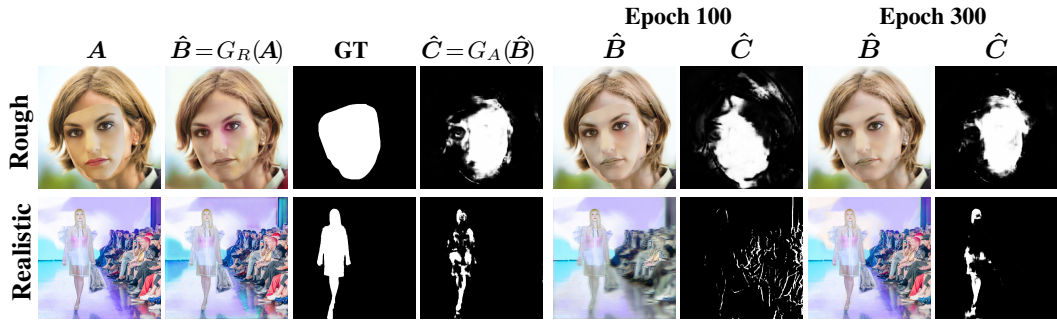


Figure 5: Performance for retoucher G_R on rough and realistic splices. Figure 6: Adaptation of annotator G_A over time.

5 Conclusion

We showed how adversarial training based on a learning retoucher generator in the loop could help a splice localization model to learn a wide range of image manipulations. Our mixed adversarial generators extend the generative adversarial networks framework by replacing a scalar value fake prediction discriminator with a pixel-level fake region annotator. The proposed retoucher generator is trained simultaneously with an annotator generator trying to maximize the probability of the annotator to make a mistake. Such adversarial training improves the annotator splice localization rate as it observes changing image manipulation techniques through the training process. Furthermore, the competition of two generators allows the retoucher generator to achieve the state-of-the-art performance in image-to-image translation tasks. Our main observation is that semantic-guided training allows our splice localization annotator to reason explicitly about splices and their semantic consistency, and achieve and surpass the state-of-the-art methods in splice localization on several challenging datasets.

Acknowledgments

The reported study was funded by the Russian Science Foundation (RSF) according to the research project N° 19-11-11008 and the Russian Foundation for Basic Research (RFBR) according to the research project N° 17-29-04509 . We want to thank Belgian Surrealist artist René Magritte for teaching us through his art how to find the point where fantasy meets reality.

References

- [1] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting Fake News: Image Splice Detection via Learned Self-Consistency. In *The European Conference on Computer Vision (ECCV)*, September 2018.
- [2] Ronald Salloum, Yuzhuo Ren, and C C Jay Kuo. Image Splicing Localization using a Multi-task Fully Convolutional Network (MFCN). *Journal of Visual Communication and Image Representation*, 51:201–209, February 2018.
- [3] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative Adversarial Networks. *CoRR*, 2014.
- [4] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiao lei Huang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [5] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image Generation From Scene Graphs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, 2017.
- [7] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.
- [8] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 700–708. Curran Associates, Inc., 2017.
- [9] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3693–3703. Curran Associates, Inc., 2018.
- [10] P. Ferrara, T. Bianchi, A. De Rosa, and A. Piva. Image forgery localization via fine-grained analysis of cfa artifacts. *IEEE Transactions on Information Forensics and Security*, 7(5):1566–1577, Oct 2012.
- [11] Anoop Cherian and Alan Sullivan. Sem-gan: Semantically-consistent image-to-image translation. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pages 1797–1806, 2019.
- [12] Bo Liu and Chi-Man Pun. Splicing Forgery Exposure in Digital Image by Detecting Noise Discrepancies. *International Journal of Computer and Communication Engineering*, 4(1):33–38, 2015.
- [13] Bo Liu, Chi-Man Pun, and Xiao-Chen Yuan. Digital Image Forgery Detection Using JPEG Features and Local Noise Discrepancies. *The Scientific World Journal*, 2014(6):1–12, 2014.
- [14] Chi-Man Pun, Bo Liu, and Xiao-Chen Yuan. Multi-scale Noise Estimation for Image Splicing Forgery Detection. *Journal of Visual Communication and Image Representation*, 38(C):195–206, July 2016.
- [15] Thibaut Julliard, Vincent Nozick, and Hugues Talbot. Image Noise and Digital Image Forensics. In Yun-Qing Shi, Hyoung Joong Kim, Fernando Pérez-González, and Isao Echizen, editors, *Digital-Forensics and Watermarking*, pages 3–17, Cham, 2016. Springer International Publishing.
- [16] Wu-Chih Hu, Wei-Hao Chen, Deng-Yuan Huang, and Ching-Yu Yang. Novel Detection of Image Forgery for Exchanged Foreground and Background Using Image Watermarking Based on Alpha Matte. In *2012 Sixth International Conference on Genetic and Evolutionary Computing, ICGEC 2012, Kitakyushu, Japan, August 25-28, 2012*, pages 245–248, 2012.
- [17] Miroslav Goljan, Jessica J. Fridrich, and Rémi Cogramne. Rich model for steganalysis of color images. In *2014 IEEE International Workshop on Information Forensics and Security, WIFS 2014, Atlanta, GA, USA, December 3-5, 2014*, pages 185–190, 2014.

- [18] Babak Mahdian and Stanislav Saic. Using noise inconsistencies for blind image forensics. *Image and Vision Computing*, 27(10):1497 – 1503, 2009. Special Section: Computer Vision Methods for Ambient Intelligence.
- [19] S. Ye, Q. Sun, and E. Chang. Detecting digital image forgeries by measuring inconsistencies of blocking artifact. In *2007 IEEE International Conference on Multimedia and Expo*, pages 12–15, July 2007.
- [20] Y Su, J Zhang, and J Liu. Exposing Digital Video Forgery by Detecting Motion-Compensated Edge Artifact. In *2009 International Conference on Computational Intelligence and Software Engineering*, pages 1–4, December 2009.
- [21] Wu-Chih Hu and Wei-Hao Chen. Effective forgery detection using DCT+SVD-based watermarking for region of interest in key frames of vision-based surveillance. *IJCSE*, 8(4):297–305, 2013.
- [22] Ashima Gupta, Nisheeth Saxena, and Sunil Kumar. Detecting Copy move Forgery using DCT. *International Journal of Scientific and Research Publications*, 3:2250–3153, May 2013.
- [23] Harpreet Kaur, Jyoti Saxena, and Sukhjinder Singh. Simulative comparison of copy-move forgery detection methods for digital images. *International Journal of Electronics, Electrical and Computational System*, 4:62–66, 2015.
- [24] Bo Liu and Chi-Man Pun. HSV Based Image Forgery Detection for Copy-Move Attack. In *Mechatronics Engineering, Computing and Information Technology*, pages 2825–2828. Trans Tech Publications Ltd, July 2014.
- [25] Feng Zeng, Wei Wang, Min Tang, and Zhanghua Cao. Exposing Blurred Image Forgeries through Blind Image Restoration. In *10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 3PGCIC 2015, Krakow, Poland, November 4-6, 2015*, pages 466–469, 2015.
- [26] Wu-Chih Hu, Wei-Hao Chen, Deng-Yuan Huang, and Ching-Yu Yang. Effective image forgery detection of tampered foreground or background image based on image watermarking and alpha mattes. *Multimedia Tools and Applications*, 75(6):3495–3516, March 2016.
- [27] Xin Wang, Bo Xuan, and Si-long Peng. Digital Image Forgery Detection Based on the Consistency of Defocus Blur. In *2008 Fourth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, pages 192–195. IEEE, July 2008.
- [28] Aniket Roy, Rahul Dixit, Ruchira Naskar, and Rajat Subhra Chakraborty. Copy-Move Forgery Detection with Similar But Genuine Objects. In *Digital Image Forensics: Theory and Implementation*, pages 65–77. Springer Singapore, Singapore, 2020.
- [29] Irene Amerini, Lamberto Ballan, Roberto Caldelli, Alberto Del Bimbo, Luca Del Tongo, and Giuseppe Serra. Copy-move forgery detection and localization by means of robust clustering with J-Linkage. *Signal Processing : Image Communication*, 28(6):659–669, July 2013.
- [30] Longyin Wen, Honggang Qi, and Siwei Lyu. Contrast enhancement estimation for digital image forensics. *ACM Transactions on Multimedia Computing, Communications and Applications*, 14(2):1–21, May 2018.
- [31] I-Cheng Chang, J Cloud Yu, and Chih-Chuan Chang. A forgery detection algorithm for exemplar-based inpainting images using multi-region relation. *Image and Vision Computing*, 31(1):57–71, January 2013.
- [32] L. Bondi, S. Lameri, D. Güera, P. Bestagini, E. J. Delp, and S. Tubaro. Tampering detection and localization through clustering of camera-based cnn features. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1855–1864, July 2017.
- [33] L. Bondi, L. Baroffio, D. Güera, P. Bestagini, E. J. Delp, and S. Tubaro. First steps toward camera model identification with convolutional neural networks. *IEEE Signal Processing Letters*, 24(3):259–263, March 2017.
- [34] J. H. Bappy, A. K. Roy-Chowdhury, J. Bunk, L. Nataraj, and B. S. Manjunath. Exploiting spatial structure for localizing manipulated image regions. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4980–4989, Oct 2017.
- [35] Belhassen Bayar and Matthew C. Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, IH&MMSec 2016, Vigo, Galicia, Spain, June 20-22, 2016*, pages 5–10, 2016.
- [36] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning Rich Features for Image Manipulation Detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [37] Jing Dong, Wei Wang, and Tieniu Tan. CASIA image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing, ChinaSIP 2013, Beijing, China, July 6-10, 2013*, pages 422–426, 2013.
- [38] T. J. d. Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, and A. d. R. Rocha. Exposing digital image forgeries by illumination color classification. *IEEE Transactions on Information Forensics and Security*, 8(7):1182–1194, July 2013.

- [39] Tian-Tsong Ng and Shih-Fu Chang. A data set of authentic and spliced image blocks. Technical report, Columbia University, June 2004.
- [40] P. Korus and J. Huang. Multi-scale analysis strategies in prnu-based tampering localization. *IEEE Trans. on Information Forensics & Security*, 2017.
- [41] B Wen, Y Zhu, R Subramanian, T Ng, X Shen, and S Winkler. COVERAGE — A novel database for copy-move forgery detection. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 161–165. IEEE, September 2016.
- [42] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 465–476. Curran Associates, Inc., 2017.
- [43] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [44] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [45] Amelie Royer, Konstantinos Bousmalis, Stephan Gouws, Fred Bertsch, Inbar Mosseri, Forrester Cole, and Kevin Murphy. XGAN: Unsupervised image-to-image translation for many-to-many mappings, 2018.
- [46] Matthew Amodio and Smita Krishnaswamy. Travelgan: Image-to-image translation by transformation vector learning. *CoRR*, abs/1902.09631, 2019.
- [47] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Transgaga: Geometry-aware unsupervised image-to-image translation. *CoRR*, abs/1904.09571, 2019.
- [48] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Learning image-to-image translation using paired and unpaired training samples. *arXiv preprint arXiv:1805.03189*, 2018.
- [49] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018.
- [50] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Instagan: Instance-aware image-to-image translation. In *International Conference on Learning Representations*, 2019.
- [51] He Zhang, Vishal M Patel, Benjamin S Riggan, and Shuowen Hu. Generative adversarial network-based synthesis of visible faces from polarimetric thermal faces. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 100–107. IEEE, 2017.
- [52] Teng Zhang, Arnold Wiliem, Siqi Yang, and Brian C Lovell. TV-GAN: Generative Adversarial Network Based Thermal to Visible Face Recognition. December 2017.
- [53] Vladimir V. Kniaz, Vladimir A. Knyaz, Jiří Hladůvka, Walter G. Kropatsch, and Vladimir Mizginov. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 606–624, Cham, 2019. Springer International Publishing.
- [54] V. V. Kniaz and A. N. Bordodimov. Long wave infrared image colorization for person re-identification. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W12:111–116, 2019.
- [55] Vladimir V. Kniaz and Vladimir A. Knyaz. Chapter 6 - multispectral person re-identification using gan for color-to-thermal image translation. In Michael Ying Yang, Bodo Rosenhahn, and Vittorio Murino, editors, *Multimodal Scene Understanding*, pages 135 – 158. Academic Press, 2019.
- [56] Vladimir A. Knyaz, Vladimir V. Kniaz, and Fabio Remondino. Image-to-voxel model translation with conditional adversarial networks. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 601–618, Cham, 2019. Springer International Publishing.
- [57] V. V. Kniaz, F. Remondino, and V. A. Knyaz. Generative adversarial networks for single photo 3d reconstruction. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W9:403–408, 2019.
- [58] Chia-Hung Wan, Shun-Po Chuang, and Hung-yi Lee. Towards audio to scene image synthesis using generative adversarial network. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 496–500, 2019.
- [59] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. Springer International Publishing, Cham, 2015.

- [60] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.
- [61] W. R. Crum, O. Camara, and D. L. G. Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, 25(11):1451–1461, Nov 2006.
- [62] Wael AbdAlmageed Yue Wu and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgerieswith anomalous features. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.