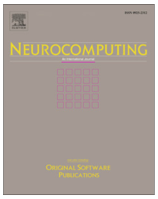
Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Semantic inpainting on segmentation map via multi-expansion loss

Jianfeng He [a], Xuchao Zhang [b], Shuo Lei [a,*], Shuhui Wang [c], Chang-Tien Lu [a], Bei Xiao [d]

[a] *Sanghani Center for Artificial Intelligence and Data Analytics, Virginia Tech, Falls Church, VA, USA*
[b] *NEC Laboratories America, Princeton, NJ, USA*
[c] *Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, China*
[d] *Department of Computer Science, American University, Washington, DC, USA*

ARTICLE INFO

Article history:
Received 17 January 2022
Revised 31 March 2022
Accepted 6 June 2022
Available online 08 June 2022
Communicated by Zidong Wang

Keywords:
Segmentation manipulation
Global and local GAN

ABSTRACT

Semantic Inpainting on Segmentation Map (SISM) aims to manipulate segmentation maps by semantics. Recent works show SISM provides semantic-aware auxiliary information for better style or structure manipulations. Providing structural assistance, segmentation maps have been broadly used as an intermediate interface to achieve better image manipulation. Mainstream solutions of image manipulation use Generative Adversarial Net (GAN) globally, locally or jointly. It is also applicable to SISM. However, the discriminator of global GAN is easier fooled, because the majority of its input is the same as the ground-truth, which is hard to fully mitigate the inconsistency between inpainted areas and the context. The inconsistency is more difficult for local GAN to address, due to the lack of context in its input. To mitigate the inconsistency, we propose a novel Multi-Expansion (MEx) loss. It is implemented by the adversarial loss on MEx areas. Each MEx area has the inpainted area as dominance and keeps knowledge of the scene context, so the consistency of the SISM results can be boosted. We propose an approximation of MEx loss, i.e., A-MEx loss, to further enhance the stability and usability. Besides performing well on SISM tasks, MEx loss also performs impressively on natural image inpainting. Extensive experiments on the two tasks demonstrate the advantages of our model over existing methods on four challenging datasets, such as a 2.59% increase in Hamm on SISM in Cityscape and a decrease of 5.00% FID on natural image inpainting in CMP Facade. The code of our work is available at: https://github.com/he159ok/AMEx-MEx-Loss.

1. Introduction

Image manipulation, transforming or editing images, is a popular topic with many applications [3–6]. Though its final outputs are manipulated natural images, many works have shown segmentation maps provide auxiliary information for better style [7,8,5] or structure [6,3,9] manipulation. Since manual editing on segmentation maps is time-consuming, and the pre-existing segmentation maps are not always adapted to the context, we focus on Semantic Inpainting on Segmentation Map (SISM) [3], which automatically manipulates segmentation maps given a semantic (e.g. name of an object to be inpainted) and a mask area. Fig. 1 shows SISM and its applications. Via SISM, users manipulate segmentation maps by desired semantics and locations. Many recent works show segmentation maps provide auxiliary information for better style or structure manipulation. Segmentation maps provide spatial instructions for style manipulation in image translation [7,10,11]. They also benefit the semantic image manipulation since the users can edit objects by semantics directly on the segmentation maps [3,9,12,13]. Therefore, SISM has many applications for various downstream tasks, such as image translation(e.g. generating more diverse natural images in terms of image textures and additional layouts, shown in the top row in Fig. 1) and semantic inpainting (e.g. inpainting part of an image for certain semantics, shown in the second row in Fig. 1). This work focuses on improving SISM, rather than improving its downstream tasks. The inpainted area in a SISM output should not only be consistent with its context, but also reflect the given semantics (shown in the left-most of Fig. 1). The dual goals make SISM unique in comparison to the traditional image inpainting [14,15]. Because the dual goals achieve both context consistency and the given semantics on segmentation maps, we call our task SISM.

Little has been done on improving the performance of SISM. Previous works on image inpainting [16–18] applied global and local GAN, which has broad applications [19–25], to synthesize a more detailed appearance. Specifically, the global discriminator

[*] Corresponding author.
E-mail addresses: jianfenghe@vt.edu (J. He), slei@vt.edu (S. Lei), wangshuhui@ict.ac.cn (S. Wang), ctlu@vt.edu (C.-T. Lu), bxiao@american.edu (B. Xiao).

https://doi.org/10.1016/j.neucom.2022.06.025
0925-2312/© 2022 Elsevier B.V. All rights reserved.

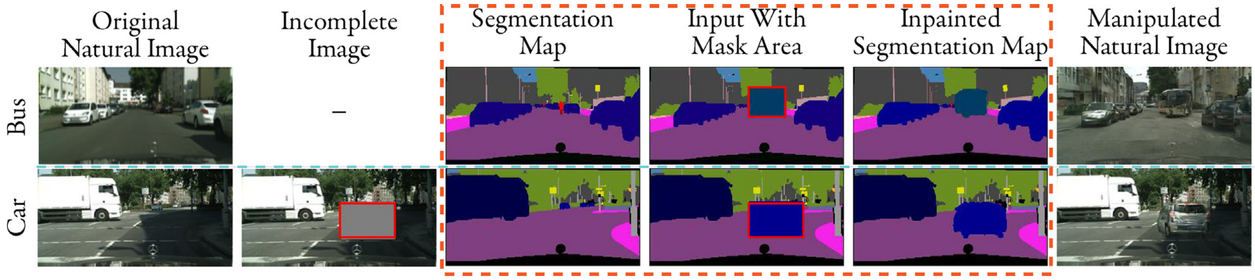


Fig. 1. Examples of SISM model and its applications on Cityscape [1]. The three columns in orange rectangle show SISM task: given a mask area (red rectangle) defined by a bounding box with a target label (leftmost vertical text), a SISM model outputs an inpainted segmentation map. The inpainted segmentation map is then inputted to a downstream model to generate manipulated natural images. For example, the top row uses image translation [2] (translating from the whole inpainted segmentation map) as a downstream task, where "-" means unnecessary in the input. For the downstream application in the top row, its goal is not to preserve the appearance of the non-inpainted area, but to translate the manipulated SISM into a new natural image. The bottom row uses semantic inpainting [3] (first concatenating the whole inpainted segmentation map and incomplete image, then the concatenated result is translated into the manipulated natural image) as a downstream task. Different from the downstream application in the first row, the downstream application in the second row preserves the non-inpainted area (context in the original natural image). This paper focuses on improving SISM, instead of improving its do.wnstream tasks.

takes a whole image as input to penalize inconsistency of the scene, while the local discriminator focuses on an inpainted area to judge its detailed appearance. Thus, we further improve the inpainted results by analyzing the drawbacks of global GAN and local GAN separately, and proposing Multi-Expansion (MEx) loss to boost the performance of SISM, which is also effective on natural image inpainting. We hypothesize that the problem with global GAN is inadequate learning of image consistency. The reason is that most of its input (which is the whole image) is the same as the ground-truth, easier fooling the discriminator compared with local GAN. Fig. 2 shows the training losses for image inpainting on the CMP Facade dataset [26] [1]. The left panel in Fig. 2 shows that the discriminator loss of global GAN converges earlier than that of local GAN, especially after 400 epochs. The earlier convergence shows the discriminator of global GAN can be easier fooled than that of local GAN. The right panel shows that the generator loss of global GAN cannot decrease as much as that of the local GAN. Since the losses for global and local GANs both come from the same mask areas computed with regard to the same ground truths, we expect the training losses converge to the same level. Thus, the global GAN cannot adequately learn the image consistency in the generation. The image consistency is also difficult for local GAN, due to lack of context [2] in its input. Thus, even we apply the global and local GAN, *the texture and layout consistency between an inpainted area and its context* (abbreviated as image consistency) cannot be fully achieved, due to the inadequate learning of global GAN and no context information to the local GAN.

To address this issue, we propose a novel Multi-Expansion (MEx) loss to boost the learning of image consistency while keeping the advantages of global and local GAN. Specifically, we construct MEx areas by expanding the inpainted areas multiple times, similar to a cascade of crops. Due to the expansion operations, each MEx area has the inpainted area as a principal component and the partial original context as an assisting component. We show that MEx loss is not only effective on SISM, but can also improve natural image inpainting. Our main contributions are summarized below.

Analyzing the drawbacks of global and local GAN. We perform experimental analysis to illustrate the drawback of global and local GAN, which does not adequately learn the image consistency between an inpainted area and its context.

Proposing MEx loss. We propose MEx loss to boost image consistency. MEx loss is the adversarial loss on MEx areas. Each MEx area has the inpainted area as the majority and its context as the minority.

Proposing A-MEx loss. A-MEx Loss is proposed to boost the convenience and stability of MEx loss. A-MEx loss is effective on SISM and natural image inpainting.

Improving performance of SISM and natural image inpainting. Our proposed MEx loss and A-MEx loss achieve significant improvements on both SISM and natural image inpainting on four datasets, such as a 2.59% increase in Hamm in Cityscape[1] by MEx loss on SISM, and a further decrease of 5.00% FID on natural image inpainting by A-MEx loss.

2. Related work

Image manipulation has tremendous progress through the rapid development of GAN [27]. Previous works mainly manipulate image attributes (e.g. styles [28,2,7,8,10,29]), structures (e.g. shapes [4]), or attributes and structures together (e.g. semantics [3,30,12], image inpainting [31], object removal [32]). Besides direct manipulation on natural images [31,33,34], it is a new trend to manipulate images with auxiliary information from other domains, such as sketches [35,29,36,37], segmentation maps [38,39,7,40,2,11,41–43], scene graphs [44–46] and other natural images [47–50,10].

Since the datasets for segmentation maps are abundant and the segmentation accuracy has been continuously increasing, it is popular to use segmentation maps as an intermediate step in image manipulation [51–54]. In addition to providing spatial instructions for style manipulation in image translation [7,8,5], segmentation maps are also beneficial on semantic manipulation, where objects can be added or revised based on semantics. For example, [55,6,3] generate the segmentation maps of target objects, which are later translated into natural images for semantic image manipulation. Similar to the frameworks used in the above studies, [9,13] focus on image inpainting. Considering better-manipulated segmentation maps lead to better-manipulated natural images, we focus on improving the manipulation on segmentation maps. Like [3], we set mask areas and semantics as guides for the manipulation as it is easy for a user to choose a mask area and input a semantic ID.

The difference between MEx Loss and other common techniques is obvious. Generally, an attention mechanism learns weighted scores at either pixel-level [56] or feature-level [57] for the inputs. Partial convolution [58] filters the information from mask areas in convolution operations to reduce noisy impact from mask areas. Multi-scale resizes a complete image into various

[1] Please refer to experiment section for more experimental settings.
[2] The context refers to the originally unmasked part in the image inpainting input.

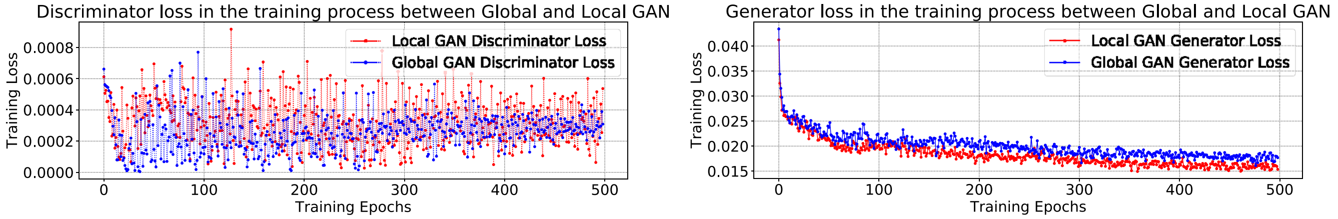


Fig. 2. Diagram of training loss in training the global GAN and the local GAN separately.

scales for boosting feature representation [59,60] or high-resolution [61]. In contrast, MEx loss focuses on boosting the image consistency by adversarial loss on MEx areas, which is a different goal from that of multi-scale. To achieve the different goals, MEx loss applies only parts of an image, while multi-scale uses a complete image. PatchGAN [39] designs a patch-based discriminator to penalize realism at the patch-level, and Pix2PixHD [2] further improves the discriminator of PatchGAN by adding multi-scale into the patch-based discriminator. But the two discriminators take the whole images as the input, where the image consistency is still not adequately learned, because most of the input is the same as ground-truth in the image inpainting. Plus, unlike the static methods of the conventional attention mechanism and partial convolution, MEx loss is dynamic, changing the radius for each expansion. This provides MEx loss with multiple mid-level views.

3. Model

3.1. A basic SISM framework

Given a complete segmentation map $\mathbf{S}^c \in \mathbb{R}^{H\times W\times 1}$ and its colorful version $\mathbf{Y}^c \in \mathbb{R}^{H\times W\times 3}$ (where $\mathbf{S}, \mathbf{Y}, H$ and W represent the single-channel segmentation map, RGB-channel segmentation map, height and width, respectively), our goal is to synthesize the inpainted segmentation map $\widehat{\mathbf{Y}} \in \mathbb{R}^{H\times W\times 3}$ reflecting a semantic defined by a target label l^t.

We choose an object bounding box B, which guides us to generate $\mathbf{A}^u$ by the area of B on $\mathbf{A}^c$, where $\mathbf{A}$ can be $\mathbf{S}$ or $\mathbf{Y}$. Specifically, we define an object bounding box $B = \left\{\mathbf{b}, l^t\right\}$, as a combination of box corner $\mathbf{b} \in \mathbb{R}^4$ and a target label l^t. A user can semantically manipulate a segmentation map by setting an arbitrary bounding box with a target label. Fig. 3 shows the overall pipeline of SISM, where the output is an inpainted segmentation map $\widehat{\mathbf{Y}}$. First, we construct an incomplete segmentation map $\mathbf{S}^u \in \mathbb{R}^{H\times W\times 1}$ by copying $\mathbf{S}^c$ and masking all pixels in the bounding box B as l^t, which informs the model of the location and semantics to generate. We apply the global observation of $\mathbf{S}^c$, while [3] only applies a local squared observation of $\mathbf{S}^c$, omitting much of the original context that belongs to the same image. Its omitting partial context reduces the quality of the manipulated results (verified in Table 2).

Then, a learnable basic structure generator G^Y generates an initial segmentation map $\widetilde{\mathbf{Y}}$ by $\widetilde{\mathbf{Y}} = G^Y(\mathbf{S}^u, B)$. After $\widetilde{\mathbf{Y}}$, we construct the inpainted segmentation map $\widehat{\mathbf{Y}}$ by fusing $\widetilde{\mathbf{Y}}$ and $\mathbf{Y}^u$. This is given by,

$$\widehat{\mathbf{Y}} = \widetilde{\mathbf{Y}} \cdot \mathbf{M} + \mathbf{Y}^u \cdot (\mathbf{1} - \mathbf{M}) \tag{1}$$

where $\cdot$ is element-wise multiplication, $\mathbf{M} \in \mathbb{R}^{H\times W}$ is a binary matrix specified by B, with all elements inside the bounding box B as 1; and $\mathbf{1} \in \mathbb{R}^{H\times W}$ is a matrix with all elements as 1. Based on Eq. 1, $\widehat{\mathbf{Y}}$ not only has the inpainted area from $\widetilde{\mathbf{Y}}$, but also keeps the rest the same as $\mathbf{Y}^u$. As the final output of SISM, $\widehat{\mathbf{Y}}$ should reflect the class-specific structure of the object defined by l^t and should have high image consistency between the inpainted area and its surrounding context. Considering the two requirements, the loss function $\mathscr{L}_{\text{B}}$ to train a basic G^Y and its respective discriminators is given by,

$$\mathscr{L}_{\text{B}} = \mathscr{L}_{\text{adv}_\text{G}} + \mathscr{L}_{\text{adv}_\text{L}} + \lambda_1 \mathscr{L}_{\text{other}} \tag{2}$$

where $\mathscr{L}_{\text{adv}_\text{G}}$ and $\mathscr{L}_{\text{adv}_\text{L}}$ are the conditional global and local adversarial losses defined on $\mathbf{S}^u$ and $\mathbf{Y}^c$ for ensuring the perceptual quality of predicted segmentation maps, which are,

$$\begin{aligned}\mathscr{L}_{\text{adv}_\text{G}} = {} & \mathbb{E}_{(\mathbf{Y}^c,\mathbf{S}^u)\sim p_{r_1}}\left[log\left(D^G(\mathbf{Y}^c,\mathbf{S}^u)\right)\right]+ \\ & \mathbb{E}_{\left(\widehat{\mathbf{Y}},\mathbf{S}^u\right)\sim p_{g_1}}\left[1-log\left(D^G\left(\widehat{\mathbf{Y}},\mathbf{S}^u\right)\right)\right]\end{aligned} \tag{3}$$

and

$$\begin{aligned}\mathscr{L}_{\text{adv}_\text{L}} = {} & \mathbb{E}_{(\mathbf{Y}^c\cdot\mathbf{M},\mathbf{S}^u)\sim p_{r_2}}\left[log\left(D^L(\mathbf{Y}^c\cdot\mathbf{M},\mathbf{S}^u)\right)\right]+ \\ & \mathbb{E}_{\left(\widehat{\mathbf{Y}}\cdot\mathbf{M},\mathbf{S}^u\right)\sim p_{g_2}}\left[1-log\left(D^L\left(\widehat{\mathbf{Y}}\cdot\mathbf{M},\mathbf{S}^u\right)\right)\right]\end{aligned} \tag{4}$$

where the complete segmentation map $\mathbf{Y}^c$ is the ground-truth, and D^G and D^L are learnable discriminators for the global GAN and the local GAN respectively. The p_r and p_g are distributions of real and generated data respectively. The $\mathscr{L}_{\text{other}}$ in Eq. 2 refers to the weighted sum of other losses [3], such as perception loss [62] applied in [2,7]. After getting $\widehat{\mathbf{Y}}$, we classify it into $\widehat{\mathbf{S}} \in \mathbb{R}^{H\times W\times 1}$, which is the input of the downstream model for natural images. The classification uses the smallest distance between a pixel value in $\widehat{\mathbf{Y}}$ and RGB values of all semantic IDs.

3.2. Multi-expansion (MEx) loss

As discussed in Section 1, even though the global and local GAN learns more details, the image consistency is still not fully achieved due to the insufficient learning of global GAN and no context to local GAN. Based on Eq. 1, we know that the inconsistency exists between the inpainted area in $\widetilde{\mathbf{Y}}$ and its context provided by $\mathbf{Y}^u$, which is the same as the ground truth.

To improve the image consistency, we propose MEx loss. It calculates the adversarial losses by $(q+1)$ MEx discriminators, $D^E_0, D^E_1, \ldots, D^E_q$ for respective MEx areas. The key idea of the MEx loss is building MEx areas $\mathbf{Z}$, as shown in Fig. 4. Recall that $\mathbf{M}$ is constructed by box corners $\mathbf{b} = [b_1, b_2, b_3, b_4]$ from B, where (b_1, b_2) and (b_3, b_4) are the coordinates of the top-left corner and the bottom-right corner for B. Then, at the j-th level expansion, we have the j-level masked area $\mathbf{M}^E_j \in \mathbb{R}^{H\times W}$ constructed by $\mathbf{b}^E_j = [b_1 - j\times\alpha, b_2 - j\times\beta, b_3 + j\times\alpha, b_4 + j\times\beta]$, where α and β are step lengths in the vertical and horizontal directions respectively. If any coordinate in $\mathbf{b}^E_j$ is smaller than 0 or greater than $H-1$ or $W-1$, it is set to 0 or $H-1$ or $W-1$ respectively. Based on $\mathbf{M}^E_j$,

[3] More details about $\mathscr{L}_{\text{other}}$ for each baseline are shown in Section 4.1.3.

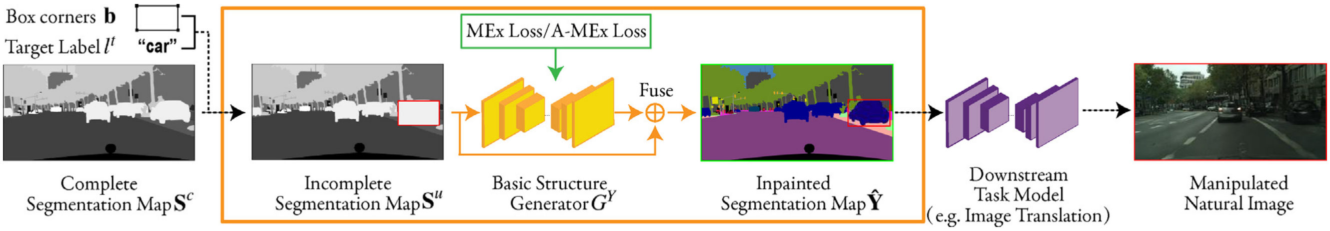

Box corners **b**
Target Label l^t
"car"
MEx Loss/A-MEx Loss
Fuse
Complete Segmentation Map $\mathbf{S}^c$
Incomplete Segmentation Map $\mathbf{S}^u$
Basic Structure Generator G^Y
Inpainted Segmentation Map $\hat{\mathbf{Y}}$
Downstream Task Model (e.g. Image Translation)
Manipulated Natural Image

Fig. 3. Diagram of the our framework for SISM. With complete segmentation map $\mathbf{S}^c$, we mask all pixels in an object bounding box $B = \{\mathbf{b}, l^t\}$ as l^t to build an incomplete segmentation map $\mathbf{S}^u$. Conditioned on $(\mathbf{S}^u, B)$ and supervised by the complete segmentation map $\mathbf{Y}^c$ in color, the basic structure generator G^Y outputs the initial segmentation map $\widetilde{\mathbf{Y}}$, which is fused with $\mathbf{Y}^u$ to obtain the inpainted segmentation map $\hat{\mathbf{Y}}$. Finally, the $\hat{\mathbf{Y}}$ is transferred to a one-channel segmentation map $\hat{\mathbf{S}}$, followed by a downstream task (e.g. image translation) model. To boost image consistency, we apply MEx loss (shown as Fig. 4)) or A-MEx loss to G^Y. The $\mathbf{Y}^c, \widetilde{\mathbf{Y}}, \mathbf{Y}^u$ and $\hat{\mathbf{S}}$ are not draw.n in the figure.

Table 2
The tIOU and Hamm on the Cityscape for SISM results.

Methods	Cityscape	
	tIOU↑	Hamm↑
Pix2PixHD [2]	0.7350	0.7483
Pix2PixHD + MEx	0.7493	0.7677
Pix2PixHD + A-MEx	**0.7652**	**0.7734**
SPADE [7]	0.7670	0.7721
SPADE + MEx	0.7699	0.7769
SPADE + A-MEx	**0.7744**	**0.7820**
TwoSM (local observation) [3]	0.7064	0.7443
TwoSM	0.7793	0.7855
TwoSM + MEx	0.7885	0.7977
TwoSM + A-MEx	**0.8072**	**0.8245**

we have the j-level MEx area $\mathbf{Z}_j^c$ for the ground-truth and $\hat{\mathbf{Z}}_j$ for the inpainted segmentation map as,

$$\mathbf{Z}_j^c = Crop\left(\mathbf{Y}^c \cdot \mathbf{M}_j^E\right), \hat{\mathbf{Z}}_j = Crop\left(\hat{\mathbf{Y}} \cdot \mathbf{M}_j^E\right) \tag{5}$$

where *Crop* is an operator to crop the input and keep only the area with non-zero elements, as we want to avoid the impact from all-zero areas.

From Fig. 4 and Eq. 5, we see that the j-level MEx area $\hat{\mathbf{Z}}_j$ consists of two parts: the inpainted area from the inpainted segmentation map $\hat{\mathbf{Y}}$ and its context from the ground-truth. Unlike the input from global GAN, the MEx loss takes MEx areas as input, such that the discriminators D^E are not so easily fooled. Since the inpainted area in each MEx area is dominated by setting α and β as smaller values than H and W, it differs from global GAN, where the original context is dominant. Moreover, though the inpainted area is dominated, D^E has knowledge from the context of $\mathbf{Y}^u$ by setting q greater than 0 (when $q = 0$, MEx loss is equivalent to the local adversarial loss), which solves the problems of the local GAN. Thus, with the MEx areas sent to D^E in MEx loss, the D^E is less likely fooled by the context and more likely to penalize inconsistency of the inpainted area of $\hat{\mathbf{Y}}$. In this way, the image consistency is improved. The MEx loss function is formulated as,

$$\begin{aligned}\mathscr{L}_{\mathrm{MEx}} = \sum_{j=0}^{q} \mathbb{E}_{\mathbf{Z}^c}\left[log\left(D_j^E\left(\mathbf{Z}_j^c, Crop\left(\mathbf{S}^u \cdot \mathbf{M}_j^E\right)\right)\right)\right] + \\ \textstyle\sum_{j=0}^{q} \mathbb{E}_{\hat{\mathbf{Z}}}\left[1 - log\left(D_j^E\left(\hat{\mathbf{Z}}_j, Crop\left(\mathbf{S}^u \cdot \mathbf{M}_j^E\right)\right)\right)\right]\end{aligned} \tag{6}$$

where the condition $\left(\mathbf{S}^u \cdot \mathbf{M}_j^E\right)$ is also cropped.

We design MEx areas rather than only one expansion area for two reasons. First, we do not know which size of the MEx area is optimal beforehand, thus, we explore MEx areas with different sizes. Second, the MEx areas provide multiple extra mid-level views (1-th level to the q-th level), which are similar to the ensemble learning [63], MEx loss can improve the inpainted results via multiple mid-level views [64], besides global view and local view. We have our final objective function with MEx loss as,

$$\mathscr{L}_{\mathrm{Final}} = \mathscr{L}_{\mathrm{adv_G}} + \lambda_1 \mathscr{L}_{\mathrm{other}} + \lambda_2 \mathscr{L}_{\mathrm{MEx}} \tag{7}$$

where $\mathscr{L}_{\mathrm{adv_L}}$ is included in $\mathscr{L}_{\mathrm{MEx}}$, when $q \geqslant 0$ in Eq. 6.

3.3. Approximated MEx (A-MEx) loss

We further boost the convenience and stability of MEx loss by proposing an approximation, A-MEx loss, which uses same-sized inputs and only one D^E. Initially, we find that the *Crop* operation leads to inputs in various sizes, which might require a special design of the D^E. Thus, in A-MEx loss, we assume the impact of all-zero areas is negligible, and remove all *Crop* in Eq. 5 and Eq. 6, which keeps all-zero areas like [16]. Then, all MEx Areas sent to D^E are the same size as $H \times W \times \{C\}$, where $\{C\}$ is their original channel numbers. The stability of the MEx loss improves as a result of using same-sized inputs. However, removing the *Crop* operation enlarges input sizes. Thus, for less memory consumption, in A-MEx loss, we do not apply $(q+1)$ MEx discriminators D^E, by applying only one D^E to save memory. Specifically, the A-MEx loss function can be formulated as,

$$\begin{aligned}\mathscr{L}_{\mathrm{A-MEx}} = \sum_{j=0}^{q} \mathbb{E}_{\mathbf{Z}^c}\left[log\left(D^E\left(\mathbf{Z}_j^c, \left(\mathbf{S}^u \cdot \mathbf{M}_j^E\right)\right)\right)\right] + \\ \textstyle\sum_{j=0}^{q} \mathbb{E}_{\hat{\mathbf{Z}}}\left[1 - log\left(D^E\left(\hat{\mathbf{Z}}_j, \left(\mathbf{S}^u \cdot \mathbf{M}_j^E\right)\right)\right)\right]\end{aligned} \tag{8}$$

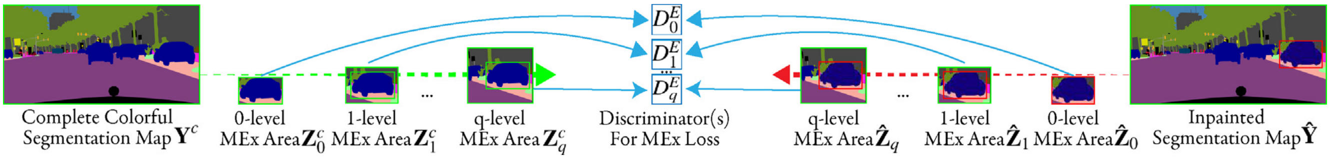

D_0^E
D_1^E
...
D_q^E
Complete Colorful Segmentation Map $\mathbf{Y}^c$
0-level MEx Area $\mathbf{Z}_0^c$
1-level MEx Area $\mathbf{Z}_1^c$
...
q-level MEx Area $\mathbf{Z}_q^c$
Discriminator(s) For MEx Loss
q-level MEx Area $\hat{\mathbf{Z}}_q$
...
1-level MEx Area $\hat{\mathbf{Z}}_1$
0-level MEx Area $\hat{\mathbf{Z}}_0$
Inpainted Segmentation Map $\hat{\mathbf{Y}}$

Fig. 4. Diagram of MEx areas for MEx loss. Given a complete colorful segmentation map $\mathbf{Y}^c$, the green dotted arrow shows MEx areas $\mathbf{Z}^c$ for the ground-truth from the 0-th to the q-th level. Given an inpainted segmentation map $\hat{\mathbf{Y}}$, the red dotted arrow represents the process to obtain MEx areas $\hat{\mathbf{Z}}$ for the inpainted segmentation map. The green border and red border show the ground-truth and the inpainted area respectively. The pairs of MEx areas are inputted into respective discriminators for the MEx loss.

where the $\widehat{\mathbf{Z}}_j$ and the $\mathbf{Z}_j^c$ are calculated by Eq. 5 without *Crop*. Because the current inputs to D^E still have the inpainted areas as the majority and their context as the minority, A-MEx loss also boosts the image consistency. We can therefore replace $\mathscr{L}_{\text{MEx}}$ by $\mathscr{L}_{\text{A-MEx}}$ in Eq. 7.

3.4. Choices of MEx loss or A-MEx loss

The difference between the two losses is that MEx loss has less memory consumption in sample size due to cropped operation in Eq. 5 and Eq. 6, but more memory consumption in model parameters due to multiple discriminators. And A-MEx loss has more memory consumption in sample size but less memory consumption in model parameters. Based on the difference, we can choose either MEx loss or A-MEx loss based on the two factors: data size and GPU memory.

In terms of data size, if we have a large data set, such as COCO [65], we use MEx loss because it has more parameters for sufficient learning; otherwise, we should use A-MEx loss for a small dataset, such as Cityscape, to avoid over-fitting, because A-MEx loss has fewer parameters. In terms of GPU memory, if we are limited by GPU resources, we should choose the method that has less GPU consumption. Concretely, if the GPU memory consumption of the total image size of a batch is larger than the GPU memory consumption of model parameters, we should use MEx loss, which has less GPU memory consumption in sample size; otherwise, we should use A-MEx loss, which has less GPU memory consumption in model parameters.

4. Experiments

Our experiments are majorly conducted on the SISM task and minorly conducted on the image inpainting. The reasons for the settings are that, our focused task is SISM with segmentation maps, and image inpainting with natural images is a popular task as our additionally minor task. Therefore, we first show the experiment settings for the two tasks at first. Then the quantity analysis (including parameter sensitivity analysis), followed by the quality analysis (including human evaluation) and the SISM applications are given.

4.1. Experiment setup

4.1.1. Datasets

Datasets for SISM. We conduct SISM experiments on three datasets: street scenes using Cityscape [1], indoor scenes using NYU V2 (NYU) [66], and face images using rectified Helen Face [26,67]. Specifically, we apply 2975 training images and 500 testing images for Cityscape; 1200 training images and 249 testing images for NYU; and 1800 training images and 299 testing images for Helen Face. We choose 8 categories of objects (‘person’, ‘rider’, ‘car’, ‘truck’, ‘bus’, ‘train’, ‘motorcycle’, and ‘bicycle’) to inpaint on Cityscape from 35 available object categories. For NYU, which has 895 semantic labels, we choose 7 categories (‘cabinet’, ‘chair’, ‘floor’, ‘table’, ‘wall’, ‘window’, ‘picture’). Helen Face has 11 categories, from which, we choose 8 categories(‘left brow’, ‘right brow’, ‘left eye’, ‘right eye’, ‘nose’, ‘upper lip’, ‘inner mouth’, ‘lower lip’).

Datasets for natural image inpainting. To demonstrate the generality of MEx loss in natural images, we also conduct image inpainting on architecture images from CMP Facade [68], from which 598 images are used for training and 8 images are applied for testing. During testing, each natural image is randomly masked 20 times.

4.1.2. Evaluation metrics

Metrics for SISM. We use three metrics for SISM results, the inpainted segmentation maps: (1) Target Intersection-Over-Union (tIOU) [69], which is the ratio of overlapped area for target objects to their union area; (2) Hamming Distance (Hamm) [70], which is the ratio of the number of pixels that have the same values as the ground-truth, to the number of total pixels; (3) Human Evaluation (HuEv), which is the result of a two-alternative forced-choice experiment, where users compare two inpainted segmentation maps (ours and a baseline’s) and choose the better one with reference to the ground-truth. We report the mean percentage of the times that a model was chosen, averaged over all observers. We also apply (4) Frechet Inception Distance (FID) [71] for the reality of the translated natural images from the inpainted segmentation maps by pretrained models [2].

Metrics for natural image inpainting. For natural image inpainting, we apply two commonly-used metrics in addition to FID and HuEv: (5) Structural Similarity Index (SSIM) [72], (6) L1 reconstruction loss (L1).

4.1.3. Baselines and ablation setting

Baselines for SISM. We compare our model with three baselines on SISM: (1) Pix2PixHD [2], outputting RGB-channel segmentation maps, with $\mathscr{L}_{\text{other}}$ as the feature match loss and perception loss [62]; (2) SPADE [7], which outputs RGB-channel segmentation maps, has spatially-adaptive normalization layers with the same $\mathscr{L}_{\text{other}}$ to Pix2PixHD; (3) We refer the SISM model in [3] with global observation as Two-Stream Model (TwoSM), which outputs one-hot segmentation maps from its context stream, its $\mathscr{L}_{\text{other}}$ is the reconstruction loss.

Baselines for natural image inpainting. The four additional baselines on natural image inpainting are: (4) Global and Local GAN (GL) [16], which applies global and local GAN; (5) Partial-Convolution (PC) [58], which designs a novel convolution to decrease the impact of all-zero areas; (6) Gated-Convolution (GC) [73], which has a learnable dynamic feature selection mechanism; (7) External-Internal Learning (EIL) [74], which externally reconstructs the grey images in the first stage, and propagates colors within the single image via internal learning in the second stage. We use GC as its first stage (EIL-S1), and represent its second stage as EIL-S2.

Ablation setting for two tasks. Different from “TwoSM”, which applies global observation, we apply “TwoSM (local observation)” to represent the originally local observation in [3], which omits much of the context. To show the effect of MEx loss, our experiments are designed by ablation settings on two tasks: we apply {a baseline name + MEx (or A-MEx)} to represent a method combined by a baseline and MEx (or A-MEx) loss.

4.1.4. Implementation details

Parameter setting for two tasks. We set $\lambda_1 = \lambda_2 = 1$ for all experiments. We apply MEx (A-MEx) loss on baselines using $q = 4$ on both Pix2PixHD and SPADE, and $q = 2$ on TwoSM. We set $\alpha = \beta = 5$ by default on all other SISM models, except $\alpha = \beta = 20$ on SPADE. Follow [3], the bounding box is randomly selected for each image at each epoch in the training, while in the testing, we compare the two methods by choosing the same bounding box for each image. We set $q = \alpha = \beta = 4$ for all natural image inpainting. We apply Adam [75] as our optimizer with an initial learning rate of 0.0002, decaying the learning rate from the 100-th epoch to the last epoch (200-th epoch) ending with a learning rate of 0.

Image sizes for two tasks. The training and generated image resolutions are 256×128, 192×144 and 256×256 respectively for Cityscape, NYU and Helen Face. As for the natural image inpainting task, we set 248 × 242 as the training and generated image resolutions.

4.2. Quantitative analysis

4.2.1. Quantitative results

Table 1, Table 2 show results for SISM. And Table 3 shows results of natural image inpainting. These allow us to conclude the following.

(1) MEx loss is effective on the SISM model outputting RGB-channel segmentation maps. From Table 1, MEx loss improves results in each metric. For the inpainted segmentation map, the Hamm has improved 24.45% in NYU with MEx loss, and the tIOU has improved 32.61% in Helen Face by MEx loss. For the translated natural image, we also see a slight improvement in FID, which shows that our inpainted segmentation maps are closer to the ground-truth. Table 2 shows MEx loss and A-Mex loss are also effective on Pix2PixHD and SPADE, such as 1.28% improvement in Hamm on SPADE.
(2) MEx loss is also effective on the SISM model outputting one-hot segmentation maps. For example, TwoSM + A-MEx improves 4.96% in Hamm in Table 2.
(3) A-MEx loss performs better than MEx loss in our cases, such as 3.36% improvement in Hamm by TwoSM + A-MEx and 4.11% improvement in tIOU by Pix2PixHD + A-MEx, compared with that of using MEx loss in Table 2.
(4) A-MEx loss also improves natural image inpainting. Table 3 shows improvements in SSIM, L1 and FID on CMP Facade. In particular, the FID of generated images decreases by 13.13%, which is a great improvement in image quality. Moreover, the MEx loss is complementary to SOTA image inpainting methods, by further decreasing 6.55% and 5.00% FID applied to GL + PC and GL + GC respectively. Plus, A-MEx further decreases 5.00% and 5.09% in FID for stage 1 and stage 2 of EIL respectively.
(5) Our improvement, replacing local observation with global observation, is effective. Table 2 shows that TwoSM + MEx has an obvious improvement over TwoSM (local observation). For example, the tIOU improves 11.62% and Hamm improves 10.32%. These demonstrate that the removed context, which is kept in the global observation, is important in improving the performance of SISM.

4.2.2. Parameter sensitive analysis

We conduct experiments for the MEx times q (applied in Eq. 5, Eq. 6 and described in Fig. 4). We compare the tIOU and Hamm for different q by Pix2PixHD + MEx on Cityscape, where the q values are set at 0, 1, 2, 4, 8. Plus, the α and β are both fixed at 5 in the analysis. The result of the sensitivity analysis is shown in Fig. 5. From the figure, we conclude as below:

(1) MEx loss is effectively complementary to global and local GAN, as tIOU and Hamm in $q = 0$ (equivalent to the GL) are significantly lower than those in $q > 0$. Since $q = 0$ is equivalent to the GL, this shows that our multiple mid-level views are effectively complementary to global and local views.
(2) The tIOU and Hamm both fluctuated proportionally to q. When $q = 2$, we have the highest tIOU and Hamm, while the tIOU and Hamm in $q = 4$ decrease around 1% in comparison to $q = 2$. However, when $q = 8$, the tIOU and Hamm increase noticeably compared to $q = 4$. The fluctuation shows that tIOU and Hamm are sensitive to q.

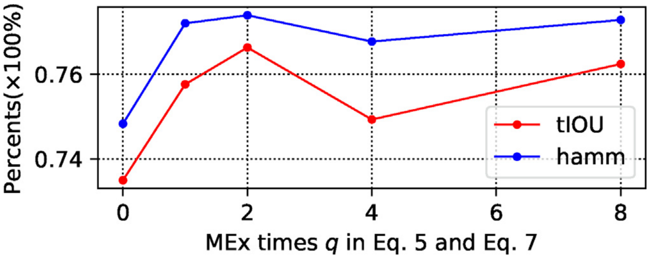


Fig. 5. Parameter sensitivity analysis on MEx times q on Cityscape.

Table 1
The tIOU, Hamm, and FID on three datasets, as well as HuEv on Cityscape.

Methods	tIOU↑	Hamm↑	FID↓	HuEv↑
Pix2PixHD [2] (Cityscape)	0.7350	0.7483	139.47	0.4704
Pix2PixHD + MEx(Cityscape)	**0.7493**	**0.7677**	**139.14**	**0.5296**
Pix2PixHD (NYU)	0.4222	0.3763	138.16	0.2511
Pix2PixHD + MEx(NYU)	**0.4851**	**0.4683**	**137.44**	**0.7489**
Pix2PixHD (Helen Face)	0.2806	0.4753	64.91	0.3980
Pix2PixHD + MEx(Helen Face)	**0.3721**	**0.5278**	**63.90**	**0.6020**

Table 3
The statistics on the CMP Facade for natural image inpainting.

Methods	CMP Facade			
	SSIM↑	L1↓	FID↓	HuEv↑
GL [16]	0.3827	21.59	63.14	0.4727
GL + A-MEx	**0.3912**	**20.22**	**54.84**	**0.5273**
GL + PC [58]	0.4194	17.06	51.46	-
GL + PC + A-MEx	**0.4248**	**16.84**	**48.09**	-
GL + GC [73]	0.4098	**16.63**	52.55	-
GL + GC + A-MEx	**0.4195**	16.88	**49.92**	-
EIL [74]-S1	0.4223	16.75	44.05	-
EIL [74]-S1 + A-MEx	**0.4376**	**15.51**	**41.50**	-
EIL [74]-S2	0.5751	9.76	41.49	-
EIL [74]-S2 + A-MEx	**0.5864**	**8.94**	**39.38**	-

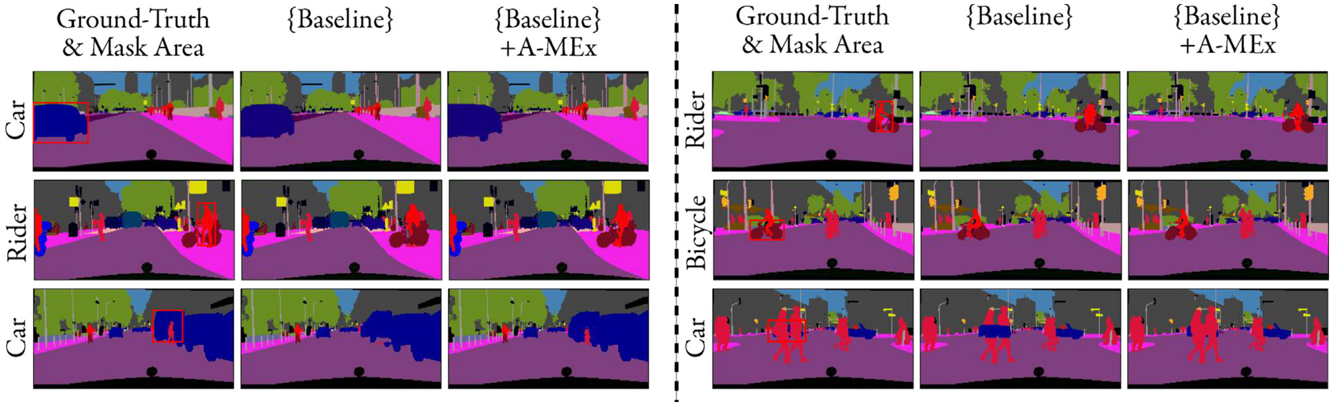


Fig. 6. Examples of results of different baselines with A-MEx loss on Cityscape. The two panels show 6 different examples and are arranged in the same way. The ground-truth and mask areas (red rectangles) are shown in the first column. The right two columns show the inpainted segmentation maps of different baselines without/with A-MEx loss, where the {Baseline} is Pix2PixHD, SPADE and TwoSM for the first, second, and third rows respectively. The target labels are shown on the leftmost columns next to the images. Please zoom in for better visualization.

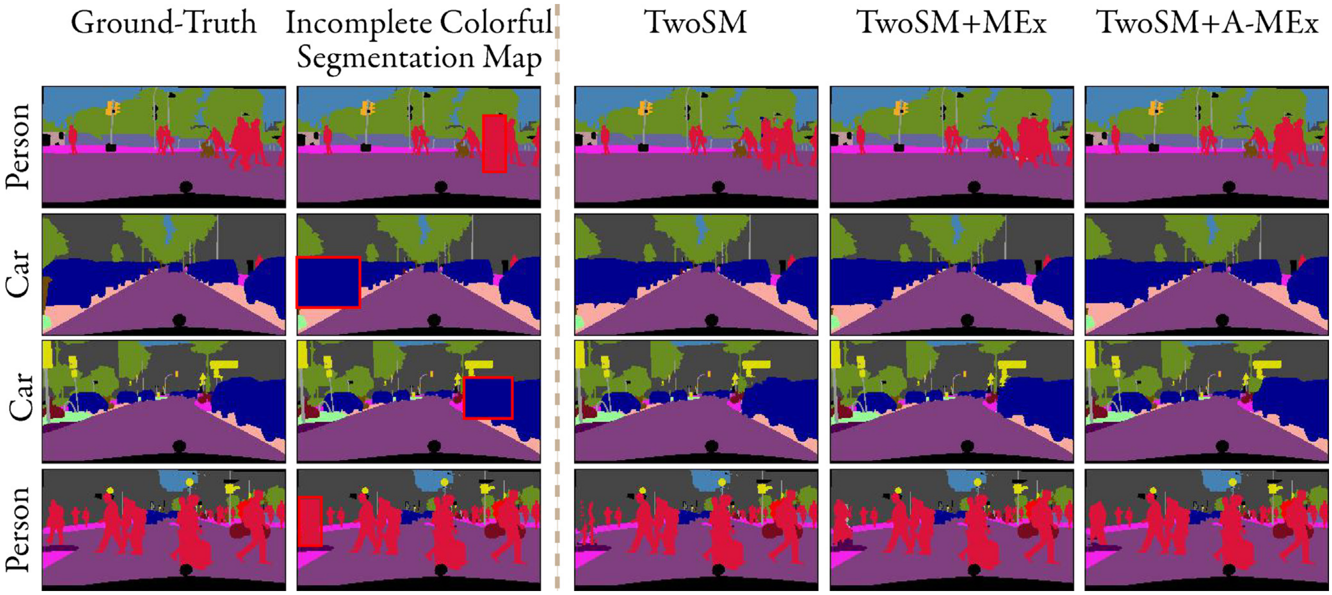


Fig. 7. Examples of inpainted segmentation maps of TwoSM + Full and TwoSM + Full + A-MEx on the Cityscape. The ground-truth segmentation maps, and the incomplete color segmentation maps are shown on the left. The right two columns show the inpainted segmentation maps of TwoSM + Full and TwoSM + Full + A-MEx respectively. The leftmost vertical texts are their target labels. Please zoom in for better visualization.

4.3. Qualitative analysis

4.3.1. Qualitative results

From Fig. 1, Fig. 6, Fig. 7, Fig. 8, Fig. 9, Fig. 10 and Fig. 12, we can conclude as below.

(1) Fig. 6 shows the inpainted segmentation maps generated by A-MEx loss are better consistent with both context and plausibility of object shapes. For example, in the right panel of the third row, a car is inpainted behind the pedestrians with higher consistency, instead of cutting them through. The better consistency also results in improved shapes. For example, the added car in the first row of the left panel has a more plausible shape when A-MEx loss is used in comparison to the baseline.
(2) Fig. 7 shows additional cases where TwoSM + A-MEx has better qualitative performance than TwoSM in different semantics. For example, its first row shows a better generation in person from TwoSM + A-MEx, because of its more pronounced contour of the person. Its second row shows the better car, because the car generated by TwoSM + A-MEx has more clear tires compared with the one in TwoSM.
(3) Fig. 8 shows additional qualitative results, which demonstrate the effectiveness of MEx loss on Pix2PixHD. For example, its first row shows MEx loss can help generate a more competitive rider. The car in the second row is generated more reasonably by MEx loss.
(4) From Fig. 9, we can see that the generated left eye (first row) and rear tire (second row) by Pix2PixHD + MEx in a natural image appear more natural than the ones generated by Pix2PixHD. These results show the effectiveness of MEx loss in Pix2PixHD.
(5) Fig. 10 shows that A-MEx loss significantly boosts image quality, especially the image consistency between the inpainted areas and the context, in natural image inpainting on each baseline.

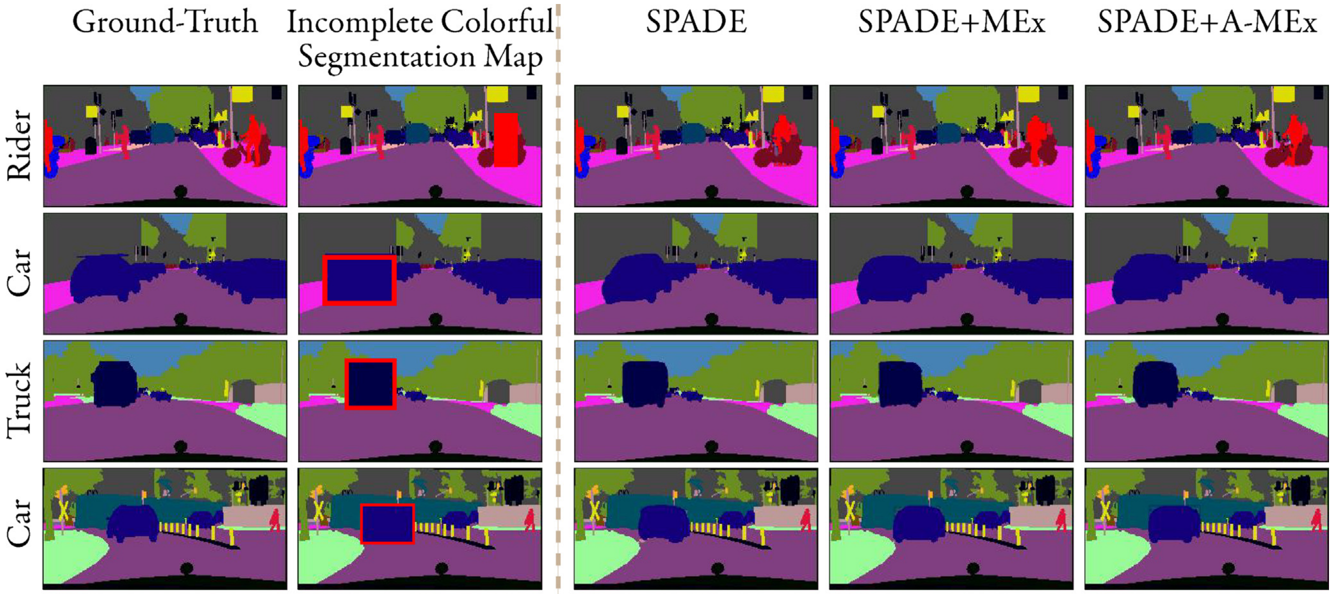


Fig. 8. Examples of inpainted segmentation maps of SPADE, SPADE + MEx loss and SPADE + A-MEx loss on the Cityscape. The ground-truth segmentation maps, and the incomplete color segmentation maps are shown on the left. The right three columns show the inpainted segmentation maps of SPADE, SPADE + MEx and SPADE + A-MEx respectively. The leftmost vertical texts are their target labels. Please zoom in for better visualization.

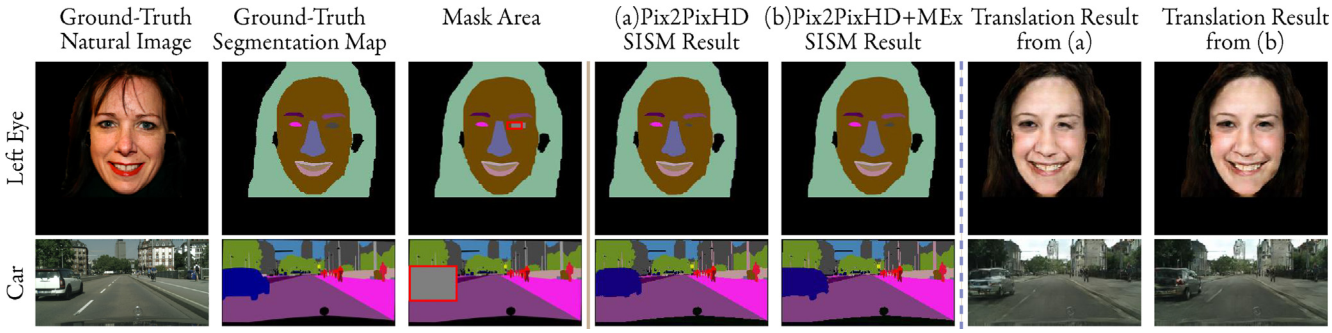

Ground-Truth Natural Image
Ground-Truth Segmentation Map
Mask Area
(a)Pix2PixHD SISM Result
(b)Pix2PixHD+MEx SISM Result
Translation Result from (a)
Translation Result from (b)
Left Eye
Car

Fig. 9. Examples of results of Pix2PixHD and Pix2PixHD + MEx on the Helen Face and Cityscape. The ground truths and mask areas are shown on the left. The right two panels show the inpainted segmentation maps (SISM results) and their respective translated images. Please zoom in for better visualization..

4.3.2. Human evaluations

We conduct user evaluation by recruiting naive observers from other organizations to complete the two-alternative forced choices experiments. The observers are provided with a web-based interface for each experiment. A screenshot of the experiment website is shown as Fig. 11, where the observers are shown the ground-truth, and incomplete segmentation map, which shows the object bounding box in certain color respecting to the target label. On the right side of the interface, the randomly arranged image results (choices A and B) from the models are displayed for comparison. The users are asked to choose which image (choice A or B) that best inpaints the mask area in consideration of the ground-truth and the incomplete map shown on the left. At each trial, the choices A and B randomly represent different model results.

As discussed in Section 4.1.2., HuEv calculates the mean percentage of the times that a model output was preferred, averaged over all observers. We use HuEv to represent user preference. In terms of SISM (Table 1), the HuEv results (from 11 users on each dataset) show that Pix2PixHD + MEx generates better inpainted segmentation maps for all three datasets. In Cityscape, the HuEv results show a 52.96% user preference for Pix2PixHD + MEx and a 47.04% for Pix2PixHD, respectively. In NYU, HuEv results show Pix2PixHD + MEx generates better inpainted segmentation maps by having significantly higher user preference for the MEx result, which is 74.89% user preference for Pix2PixHD + MEx and only 25.11% for Pix2PixHD, respectively. Also in Helen Face, HuEv results show 60.20% user preference for Pix2PixHD + MEx and 39.80% for the baseline, which shows Pix2PixHD + MEx improves the SISM results. In terms of natural image inpainting (Table 3), HuEV results (from 12 users) also show A-MEx loss improves natural image inpainting with a 52.73% user preference for GL + A +MEx.

4.4. Applications of SISM

Each row in Fig. 1 and Fig. 12 shows a different SISM application. In their first rows, given a semantic, SISM provides diverse

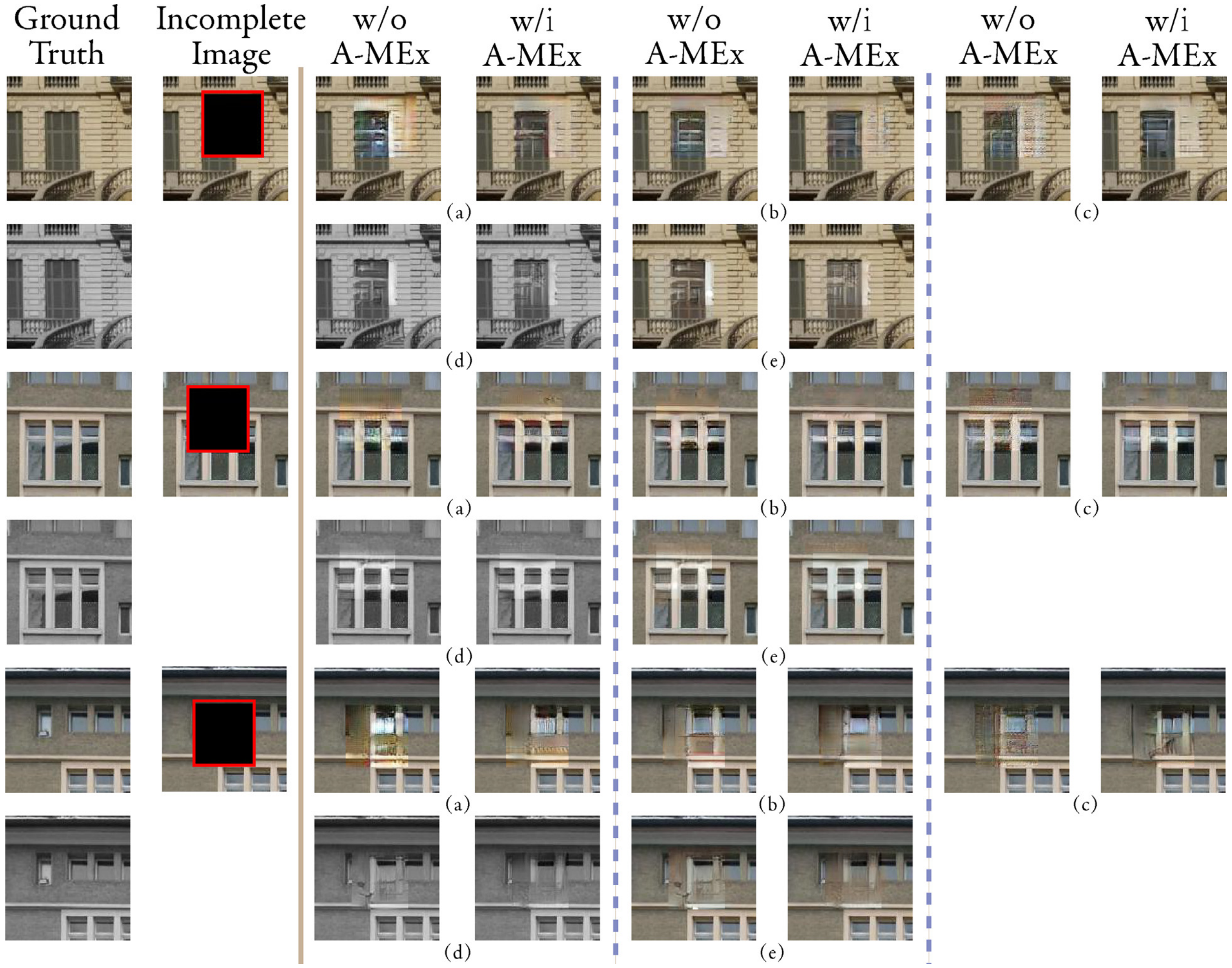
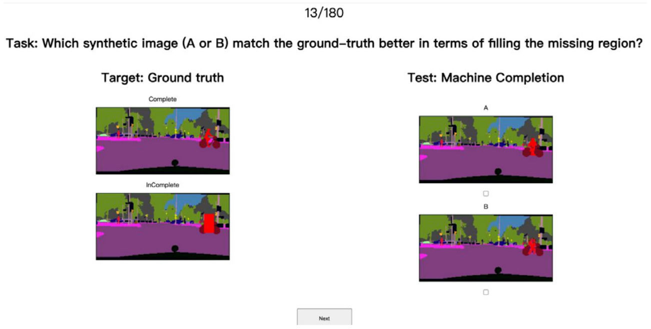


Fig. 10. Examples of natural image inpainting results on CMP Facade. The leftmost lists the ground truths for both RGB images and grey images (only compared to EIL [74]-S1). The right three panels show the generated results without A-MEx loss (w/o A-MEx) or with A-MEx loss (w/i A-MEx). The five comparison groups are: (a):GL [16], (b): GL + PC [58], (c):GL + GC [73], (d):EIL [74]-S1, and (e):EIL [74]-S2. Please zoom in for better .visualization..


13/180

Task: Which synthetic image (A or B) match the ground-truth better in terms of filling the missing region?

Target: Ground truth

Complete

InComplete

Test: Machine Completion

A

□

B

□

Next

Fig. 11. A screenshot for the website used for user evaluation. In the top, we show the task question as "Task: Which synthetic image (A or B) matches the ground-truth better in terms of filling the missing region?" In the left part, the complete segmentation map, and incomplete segmentation map with an object bounding box in certain color respecting to the target label are shown. For each turn, we provide randomly arranged results by choices A and B in the right part.

structures for image translation. In their second rows, we inpaint natural images by semantics, where SISM manipulates structures before style translation. We apply Pix2PixHD + A-MEx for the first application, and TwoSM + A-MEx for the second. The downstream model for the image translation is from [2] and the downstream model for the semantic image inpainting is from [3].

5. Conclusion

In this paper, we first analyze the drawback of global and local GAN and find that global GAN is hard to fully learn the image consistency between the inpainted areas and the context. We then propose a novel loss function, MEx loss, to further improve the image consistency in image manipulation. MEx loss is implemented by adversarial losses on MEx areas, which have the inpainted areas as the majority and their context as the minority. To enhance stability and usability, we further propose A-MEx loss. Experimental results on four datasets demonstrate that MEx loss and A-MEx loss have achieved superior performances on the SISM models with various output representations and natural image inpainting.

CRediT authorship contribution statement

Jianfeng He: Conceptualization, Investigation, Methodology, Formal analysis, Software, Validation, Writing – original draft, Visualization. **Xuchao Zhang:** Conceptualization, Resources, Supervision, Investigation, Writing – review & editing. **Shuo Lei:** Conceptualization, Supervision, Investigation, Formal analysis, Software, Writing – review & editing. **Shuhui Wang:** Conceptualization, Resources, Writing – review & editing. **Chang-Tien Lu:** Funding acquisition, Resources, Writing – review & editing. **Bei**

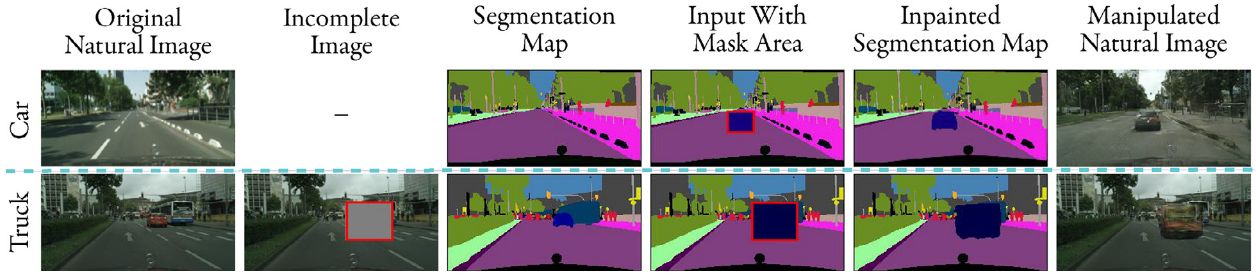


Fig. 12. Two applications of SISM with image manipulation as a downstream task. Top panel: given a semantic, SISM provides diverse structures for image translation. Bottom panel: we inpaint a natural image by a semantic, where we use SISM to get the inpainted segmentation map; then, we concatenate the inpainted segmentation map and the original context from the natural image (shown as ''Incomplete Image''); finally, the concatenation is used for a style translation. The vertical texts on the left are their target labels. The symbol ''-'' means it is not applicable to the task. Please note that since we focus on SISM here, the image translation can be improved by other SOTA methods. Please zoom in for better visualization.

Xiao: Conceptualization, Supervision, Software, Resources, Writing – review & editing, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Shuhui Wang was supported in part by National Natural Science Foundation of China: 62022083.

References

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3213–3223.

[2] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional gans, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8798–8807.

[3] S. Hong, X. Yan, T.S. Huang, H. Lee, Learning hierarchical semantic image manipulation through structured representations, Advances in Neural Information Processing Systems (2018) 2708–2718.

[4] X. Han, Z. Wu, W. Huang, M. R. Scott, L. S. Davis, Finet: Compatible and diverse fashion image inpainting, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4481–4491..

[5] P. Zhang, B. Zhang, D. Chen, L. Yuan, F. Wen, Cross-domain correspondence learning for exemplar-based image translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5143–5153.

[6] E. Ntavelis, A. Romero, I. Kastanis, L. Van Gool, R. Timofte, Sesame: semantic editing of scenes by adding, manipulating or erasing objects, European Conference on Computer Vision, Springer (2020) 394–411.

[7] T. Park, M.-Y. Liu, T.-C. Wang, J.-Y. Zhu, Semantic image synthesis with spatially-adaptive normalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2337–2346.

[8] P. Zhu, R. Abdal, Y. Qin, P. Wonka, Sean: Image synthesis with semantic region-adaptive normalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5104–5113..

[9] L. Liao, J. Xiao, Z. Wang, C.-W. Lin, S. Satoh, Guidance and evaluation: Semantic-aware image inpainting for mixed scenes, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16, Springer, 2020, pp. 683–700..

[10] L. Zhao, Q. Mo, S. Lin, Z. Wang, Z. Zuo, H. Chen, W. Xing, D. Lu, Uctgan: Diverse image inpainting based on unsupervised cross-space translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5741–5750.

[11] Z. Zhu, Z. Xu, A. You, X. Bai, Semantically multi-modal image synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5467–5476.

[12] S. Mo, M. Cho, J. Shin, Instagan: Instance-aware image-to-image translation, arXiv preprint arXiv:1812.10889..

[13] Y. Song, C. Yang, Y. Shen, P. Wang, Q. Huang, C.-C. J. Kuo, Spg-net: Segmentation prediction and guidance network for image inpainting, arXiv preprint arXiv:1805.03356..

[14] D. Bau, H. Strobelt, W. Peebles, J. Wulff, B. Zhou, J.-Y. Zhu, A. Torralba, Semantic photo manipulation with a generative image prior, ACM Transactions on Graphics (TOG) 38 (4) (2019) 1–11.

[15] Z. Yi, Q. Tang, S. Azizi, D. Jang, Z. Xu, Contextual residual aggregation for ultra high-resolution image inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7508–7517.

[16] S. Iizuka, E. Simo-Serra, H. Ishikawa, Globally and locally consistent image completion, ACM Transactions on Graphics (ToG) 36 (4) (2017) 1–14.

[17] Y. Wang, X. Tao, X. Qi, X. Shen, J. Jia, Image inpainting via generative multi-column convolutional neural networks, in: Advances in neural information processing systems, 2018, pp. 331–340..

[18] Z. Hui, J. Li, X. Wang, X. Gao, Image fine-grained inpainting, arXiv preprint arXiv:2002.02609..

[19] J. Chen, H. Fang, Z. Liu, The application of a deep convolutional generative adversarial network on completing global tec maps, Journal of Geophysical Research: Space, Physics 126 (3) (2021), e2020JA028418.

[20] J. Tang, H. Hu, Q. Zhou, H. Shan, C. Tian, T.Q. Quek, Pose guided global and local gan for appearance preserving human video prediction, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 614–618.

[21] Y. Lu, T. Sun, X. Jiang, K. Xu, B. Zhu, Frontal view synthesis based on a novel gan with global and local discriminators, in: 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), IEEE, 2019, pp. 1–5..

[22] Y. Liu, W. Cai, X. Yuan, J. Xiang, Gl-gan: Adaptive global and local bilevel optimization model of image generation, arXiv preprint arXiv:2008.02436..

[23] P. Li, Y. Hu, R. He, Z. Sun, Global and local consistent wavelet-domain age synthesis, IEEE Transactions on Information Forensics and Security 14 (11) (2019) 2943–2957.

[24] B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp, P. Bourgeat, Sample-adaptive gans: Linking global and local mappings for cross-modality mr image synthesis, IEEE transactions on medical imaging 39 (7) (2020) 2339–2350.

[25] W. Wang, S. Wang, W. Fan, Z. Liu, J. Tang, Global-and-local aware data generation for the class imbalance problem, in: Proceedings of the 2020 SIAM International Conference on Data Mining, SIAM, 2020, pp. 307–315.

[26] V. Le, J. Brandt, Z. Lin, L. Bourdev, T. S. Huang, Interactive facial feature localization, in: European conference on computer vision, Springer, 2012, pp. 679–692..

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680..

[28] X. Wang, A. Gupta, Generative image modeling using style and structure adversarial networks, European Conference on Computer Vision, Springer (2016) 318–335.

[29] H. Qiu, C. Xiao, L. Yang, X. Yan, H. Lee, B. Li, Semanticadv: Generating adversarial examples via attribute-conditional image editing, arXiv preprint arXiv:1906.07927..

[30] S. Yao, T. M. Hsu, J.-Y. Zhu, J. Wu, A. Torralba, B. Freeman, J. Tenenbaum, 3d-aware scene manipulation via inverse graphics, in: Advances in neural information processing systems, 2018, pp. 1887–1898..

[31] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, Generative image inpainting with contextual attention, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5505–5514.

[32] R.R. Shetty, M. Fritz, B. Schiele, Adversarial scene editing: Automatic object removal from weak supervision, Advances in Neural Information Processing Systems (2018) 7706–7716.

[33] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, A.A. Efros, Generative visual manipulation on the natural image manifold, European Conference on Computer Vision, Springer (2016) 597–613.

[34] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.
[35] W. Chen, J. Hays, Sketchygan: Towards diverse and realistic sketch to image synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9416–9425..
[36] P. Sangkloy, J. Lu, C. Fang, F. Yu, J. Hays, Scribbler: Controlling deep image synthesis with sketch and color, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5400–5409..
[37] Y. Jo, J. Park, Sc-fegan: face editing generative adversarial network with user's sketch and color, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1745–1753.
[38] Q. Chen, V. Koltun, Photographic image synthesis with cascaded refinement networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 1511–1520.
[39] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.
[40] H. Tang, D. Xu, Y. Yan, P.H. Torr, N. Sebe, Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7870–7879.
[41] V. Sushko, E. Schönfeld, D. Zhang, J. Gall, B. Schiele, A. Khoreva, You only need adversarial supervision for semantic image synthesis, arXiv preprint arXiv:2012.04781..
[42] X. Zhan, X. Pan, B. Dai, Z. Liu, D. Lin, C.C. Loy, Self-supervised scene de-occlusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3784–3792.
[43] H. Dhamo, A. Farshad, I. Laina, N. Navab, G.D. Hager, F. Tombari, C. Rupprecht, Semantic image manipulation using scene graphs, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5213–5222.
[44] J. Johnson, A. Gupta, L. Fei-Fei, Image generation from scene graphs, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1219–1228.
[45] G. Mittal, S. Agrawal, A. Agarwal, S. Mehta, T. Marwah, Interactive image generation using scene graphs, arXiv preprint arXiv:1905.03743..
[46] P. Ardino, Y. Liu, E. Ricci, B. Lepri, M. De Nadai, Semantic-guided inpainting network for complex urban scenes manipulation, arXiv preprint arXiv:2010.09334..
[47] C.-H. Lin, E. Yumer, O. Wang, E. Shechtman, S. Lucey, St-gan: Spatial transformer generative adversarial networks for image compositing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9455–9464..
[48] T. Xiao, J. Hong, J. Ma, Elegant: Exchanging latent encodings with gan for transferring multiple face attributes, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 168–184..
[49] Y. Zhao, B. Price, S. Cohen, D. Gurari, Guided image inpainting: Replacing an image region by pulling content from another image, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019, pp. 1514–1523.
[50] E. Collins, R. Bala, B. Price, S. Susstrunk, Editing in style: Uncovering the local semantics of gans, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5771–5780.
[51] A. Tao, K. Sapra, B. Catanzaro, Hierarchical multi-scale attention for semantic segmentation, arXiv preprint arXiv:2005.10821..
[52] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, Q. V. Le, Rethinking pre-training and self-training, arXiv preprint arXiv:2006.06882..
[53] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha, et al., Resnest: Split-attention networks, arXiv preprint arXiv:2004.08955..
[54] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, H.-M. Gross, Efficient rgb-d semantic segmentation for indoor scene analysis, arXiv preprint arXiv:2011.06961..
[55] Y. Ding, G. Teng, Y. Yao, P. An, K. Li, X. Li, Context-aware natural integration of advertisement object, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 4689–4693.
[56] T. Zhou, C. Ding, S. Lin, X. Wang, D. Tao, Learning oracle attention for high-fidelity face completion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7680–7689.
[57] R. Zhou, Y.-D. Shen, End-to-end adversarial-attention network for multi-modal clustering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14619–14628.
[58] G. Liu, F.A. Reda, K.J. Shih, T.-C. Wang, A. Tao, B. Catanzaro, Image inpainting for irregular holes using partial convolutions, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 85–100.
[59] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.
[60] A. Karnewar, O. Wang, Msg-gan: Multi-scale gradients for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7799–7808..
[61] K. Jiang, Z. Wang, P. Yi, C. Chen, B. Huang, Y. Luo, J. Ma, J. Jiang, Multi-scale progressive fusion network for single image deraining, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8346–8355.
[62] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4681–4690.
[63] O. Sagi, L. Rokach, Ensemble learning: A survey, Wiley Interdisciplinary Reviews, Data Mining and Knowledge Discovery 8 (4) (2018) e1249.
[64] D. Tao, Y. Guo, B. Yu, J. Pang, Z. Yu, Deep multi-view feature learning for person re-identification, IEEE Transactions on Circuits and Systems for Video Technology 28 (10) (2017) 2657–2666.
[65] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755..
[66] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, in: European conference on computer vision, Springer, 2012, pp. 746–760.
[67] J. Lin, H. Yang, D. Chen, M. Zeng, F. Wen, L. Yuan, Face parsing with roi tanh-warping, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5654–5663.
[68] R. Tyleček, R. Šára, Spatial pattern templates for recognition of objects with regular structure, German Conference on Pattern Recognition, Springer (2013) 364–374.
[69] X. Liu, G. Yin, J. Shao, X. Wang, et al., Learning to predict layout-to-image conditional convolutions for semantic image synthesis, Advances in Neural Information Processing Systems (2019) 568–578.
[70] J. Mun, W.-D. Jang, D.J. Sung, C.-S. Kim, Comparison of objective functions in cnn-based prostate magnetic resonance image segmentation, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 3859–3863.
[71] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: Advances in neural information processing systems, 2017, pp. 6626–6637..
[72] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing 13 (4) (2004) 600–612.
[73] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, Free-form image inpainting with gated convolution, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4471–4480.
[74] T. Wang, H. Ouyang, Q. Chen, Image inpainting with external-internal learning and monochromic bottleneck, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5120–5129.
[75] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980..

Jianfeng He is a fourth-year Ph.D. candidate in Computer Science Department at Virginia Tech, VA, USA. He received his M.S. degree in computer technology at the University of the Chinese Academy of Sciences, Beijing, China, in 2017, and received his B.E. degree in digital media from Central China Normal University, Wuhan, China, in 2014. His research interests include image manipulation, image and text understanding, and uncertainty analysis.

Xuchao Zhang is a researcher at NEC Laboratories America. He obtained his PhD degree in 2019 from Computer Science Department at Virginia Tech and received his B.E. degree in Software Engineering from Shanghai Jiao Tong University. His research interests include natural language processing, artificial intelligence, and data mining, with special interests in trustworthy prediction in sequential data, multi-lingual representation learning and robust learning in noisy data.

Shuo Lei is a fourth-year Ph.D. candidate majoring in computer science, and her supervisor is Dr. Chang-Tien Lu. She obtained her Bachelor's degree and Master's degree from Beihang University in 2015 and 2018, respectively. Her research interests include learning with limited labeled data in computer vision and natural language processing.

Shuhui Wang received the BS degree in electronics engineering from Tsinghua University, Beijing, China, in 2006, and the PhD degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2012. He is currently an associate professor with the Institute of Computing Technology, Chinese Academy of Sciences. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include semantic image analysis, image and video retrieval and large-scale Web multimedia data mining. He is a member of the IEEE.

Chang-Tien Lu is a Professor of Computer Science and Associate Director of the Sanghani Center for AI and Data Analytics at Virginia Tech. He received his Ph.D. from the University of Minnesota at Twin Cities in 2001. He served as General Chair of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems in 2009, 2020, and 2021, and the International Symposium on Spatial and Temporal Databases in 2017. He also served as Secretary (2008-2011) and Vice Chair (2011-2014) of the ACM Special Interest Group on Spatial Information (ACM SIGSPATIAL). His research interests include spatial databases, data mining, urban computing, and intelligent transportation systems. He has published over 170 articles in top rated journals and conference proceedings. He is an ACM Distinguished Member and IEEE Senior Member.

Bei Xiao is an associate professor in Computer Science at American University, Washington, DC, USA. She received her Ph.D. degree in neuroscience from University of Pennsylvania in 2009. She received her postdoctoral training at Massachusetts Institute of Technology between 2010-2013. She joined American University as an assistant professor in 2014 and was promoted to Associate Professor in 2021. Her research interests are in the fields of human vision, computer vision, machine learning, computational modeling of cognition, Virtual Reality, and computer graphics. She has received research funding from National Science Foundation. She has authored more than twenty articles in top journals and proceedings in human vision, neuroscience, computer vision and graphics. She is directing the computational material perception lab.