Composition-Guided Neural Network for Image Cropping Aesthetic Assessment

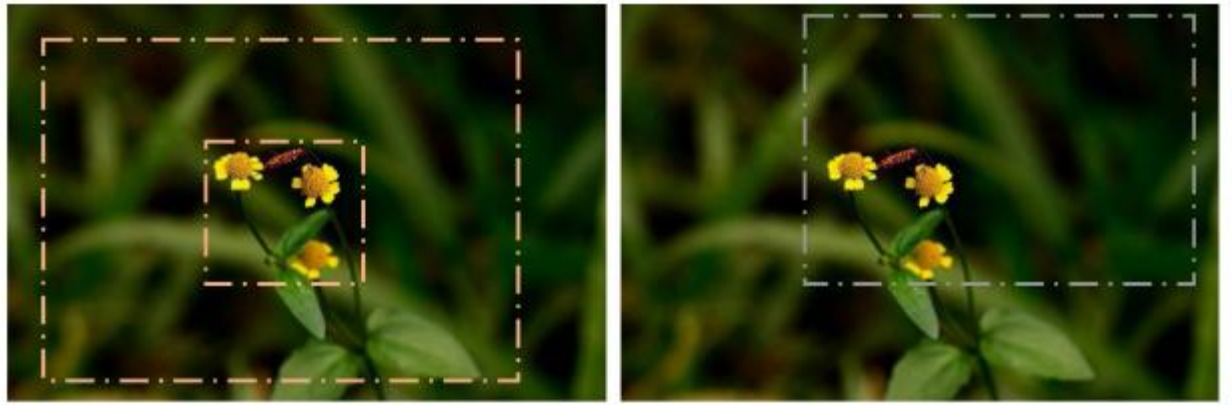
Shijia Ni , Feng Shao , Member, IEEE, Xiongli Chai , Hangwei Chen , and Yo-Sung Ho , Fellow, IEEE

IEEE Transactions on Multimedia, 2022

Introduction

Background:

- How to explore the interaction between image aesthetic rules and crops is the key to finding views with good composition.
- It is subjective to evaluate candidate crops, which mainly depends on aesthetic knowledge, but it is not an easy task for people without extensive photography experience.
- Existing methods mostly find good views by extracting general aesthetic features of crops without fully exploring the aesthetic rules.



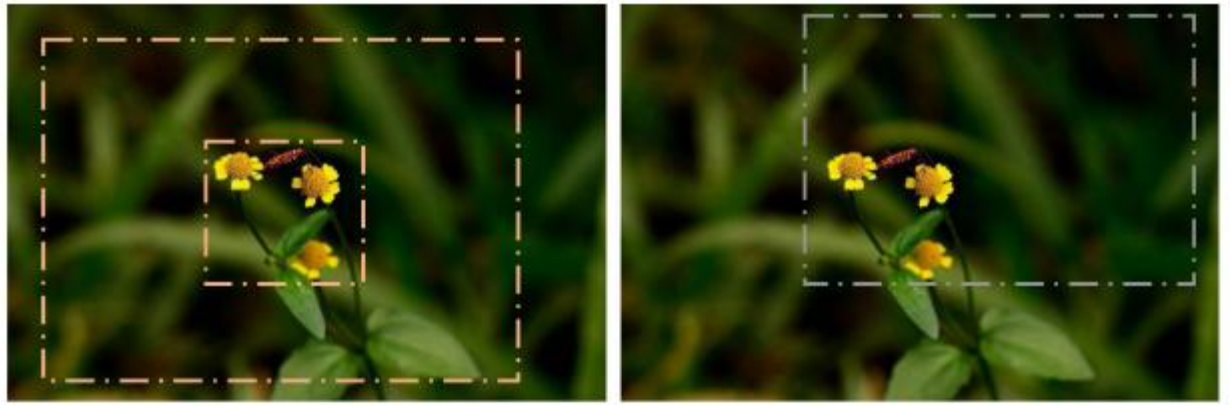
(a) Good composition(rule of central) **(b) Bad composition**

Fig. 1. Example of images crop with composition rules. The cropped region in the left image. (a) Denotes a good crop based on human-defined composition rules, while the cropped region in the right image. (b) Denotes a bad crop without considering composition rules.

Introduction

Contribution:

- We propose a Composition-Guided Image Cropping Aesthetic Assessment Network (CGICAANet), which automatically finds the best crop with good composition and enhances the consistency between the predicted crops and labels.
- We construct a composition pattern module (CPM) to adaptively explore suitable composition rules for the images in a direct and interpretable way.
- To enhance the consistency between candidates and annotations, we design an effective multi-task loss to train the CGICAANet so that the predicted scores, sorting order and characteristics of the candidates can forcefully approximate the annotations simultaneously.



(a) Good composition(rule of central) **(b) Bad composition**

Fig. 1. Example of images crop with composition rules. The cropped region in the left image. (a) Denotes a good crop based on human-defined composition rules, while the cropped region in the right image. (b) Denotes a bad crop without considering composition rules.

Overview of the CGICAANet

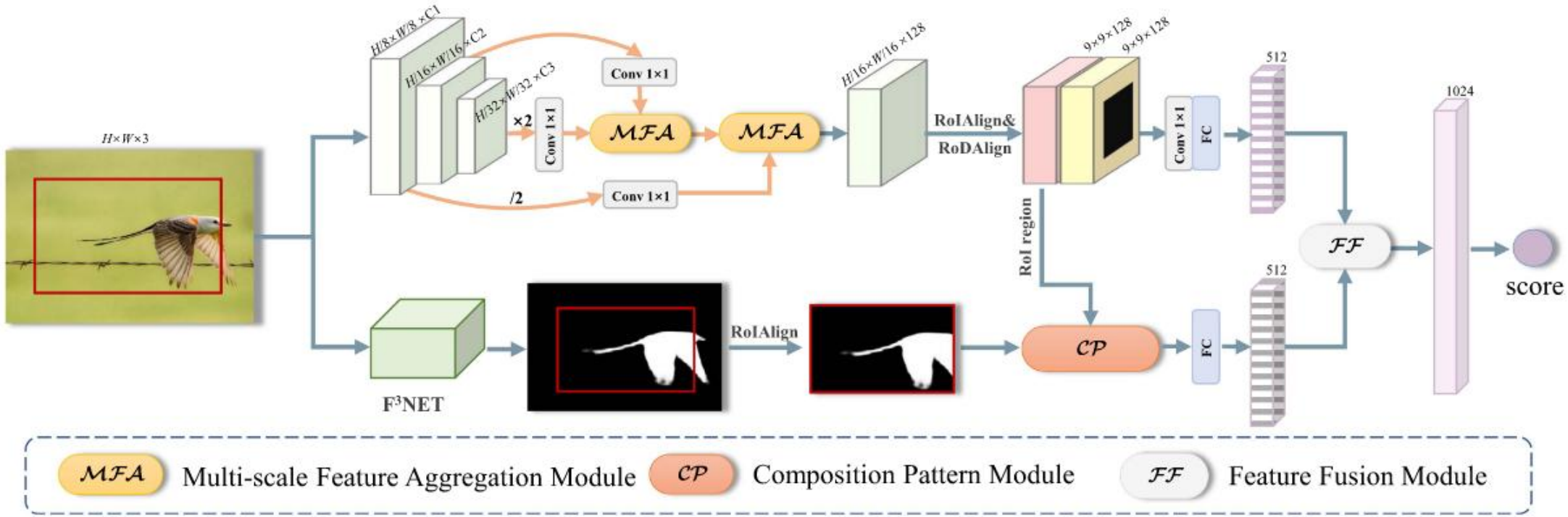
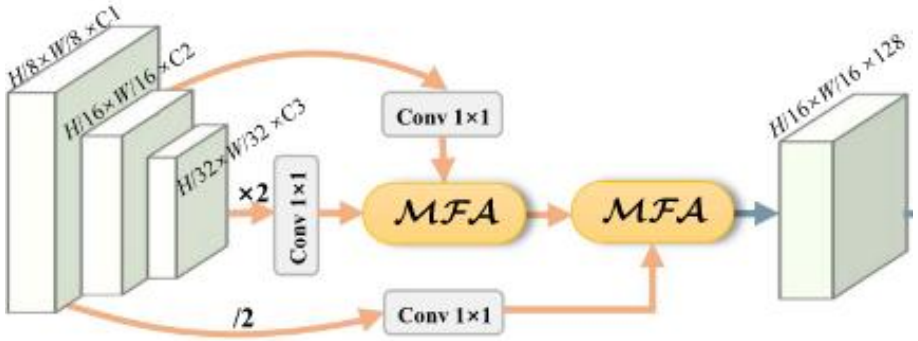


Fig. 2. Overview of the proposed network (CGICAANet). The detailed structures of our feature extraction with a multi-scale feature aggregation module, composition pattern module and feature fusion module are illustrated in Figs. 3–5, respectively. Notations "×2" and "/2" denote bilinear upsampling and downsampling, respectively.

Feature Extraction

- Backbone network



- Low-level features usually contain much detailed information, while the high-level features contain rich semantic information.
- By cascading MFAM, high-level and low-level features at different scales can be effectively aggregated in a progressive manner.
- Compared with single-scale features, the multi-scale features are beneficial to suppress noise distractions and boost the discrimination of the features.

- Multi-scale feature aggregation module (MFAM)

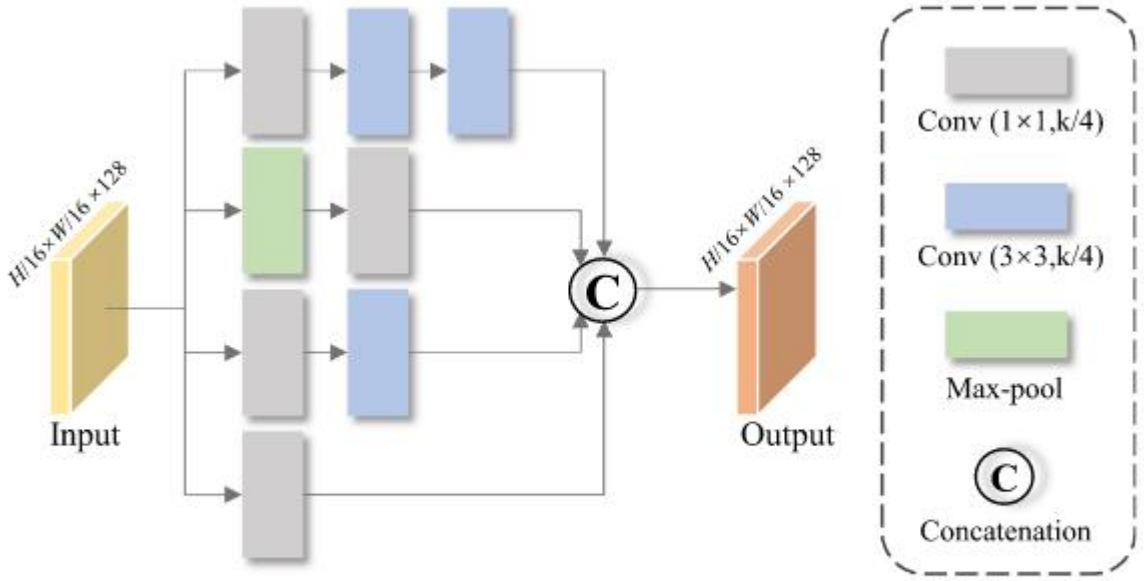


Fig. 3. Detailed illustration of Multi-scale Feature Aggregation Module (MFAM), where *k* denotes the dimension of the feature map.

Composition Pattern Module (CPM)

- Each composition pattern contains several non-overlapping partitions.
- For the p-th pattern, we divide the input global feature map F and saliency map S into Kp non-overlapping partitions

$$\theta\left(\chi_k^p\right)=\frac{1}{|\chi_k^p|}\sum\nolimits_{(i,j)\in\chi_k^p}\mathrm{x}_{i,j}\in\mathbb{R}^C$$

$$\mathrm{f}_{cp}=\sum_{p=1}^{P}\omega_p\mathrm{f}_{cp}^p$$

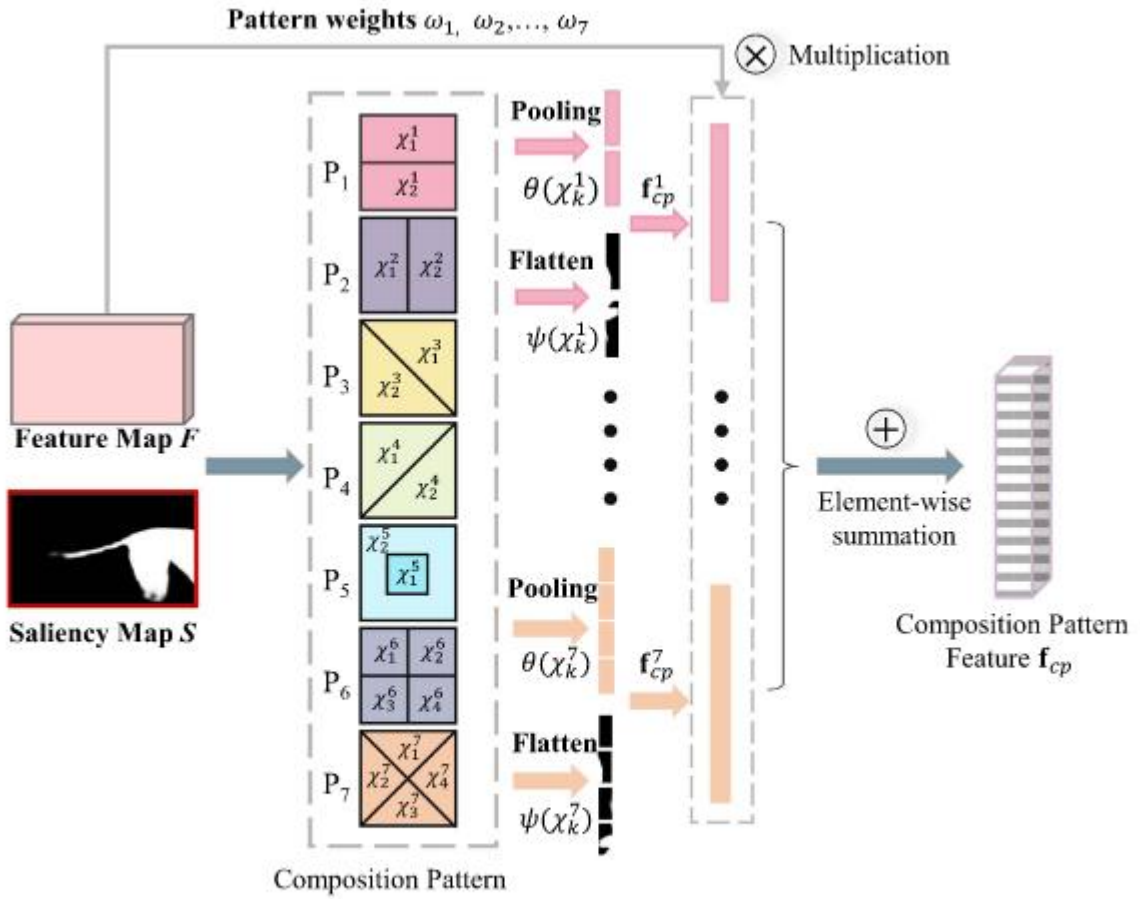


Fig. 4. Composition Pattern Module (CPM) contains seven Saliency-augmented composition patterns.

Feature Fusion Module (FFM)

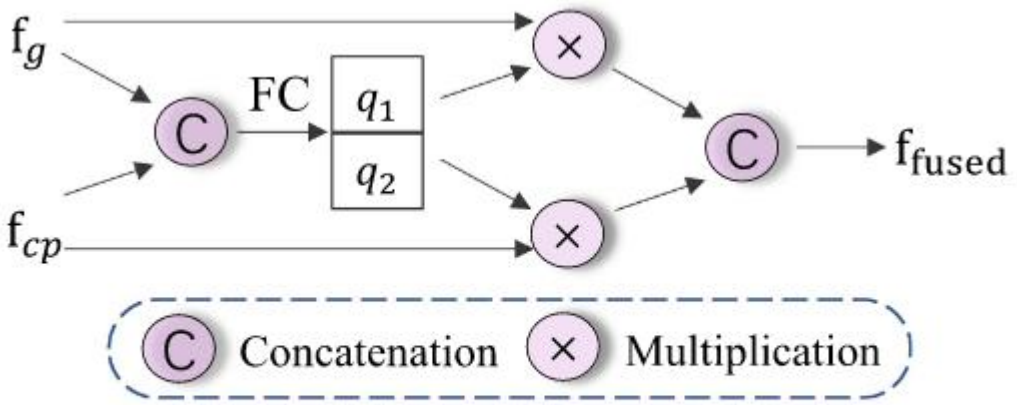


Fig. 5. Feature Fusion Module (FFM) contains a FC layer with sigmoid activation.

- We dynamically adjust the weights of the two features through learning.

$$\mathbf{f}_{\text{fused}} = [q_1\mathbf{f}_g, q_2\mathbf{f}_{cp}] \in \mathbb{R}^{C'}$$

Loss Function

- The Smooth L1 loss is broadly used in regression problems due to its robustness to outliers.

$$L_1(x) = \begin{cases} 0.5x_{ij}^2 & \text{if } x < 1, \\ |x_{ij}| - 0.5 & \text{otherwise.} \end{cases}$$

- Since the regression loss implicitly models the sorting order of different candidate boxes, we also utilize a ranking loss to explicitly reflect the score gap between different regions.

$$L_{rank} = \frac{\sum_{i,j} \max\left(0, -\varphi\left(g_i - g_j\right)\left(\left(p_i - p_j\right) - \left(g_i - g_j\right)\right)\right)}{N\left(N-1\right)/2}$$

- The CIoU loss is used to make the predicted candidate crop close to the labeled box, which considers the overlap area, center point and aspect ratio of crops.

$$L_{CIoU} = 1 - \frac{1}{\mathrm{M}} \sum_{i=1}^{M} \left(1 - IoU_i + R_{CIoU_i}\right)$$

$$R_{CIoU} = \frac{\rho^2\left(\mathbf{b}, \mathbf{b}^{gt}\right)}{c^2} + \alpha v \qquad \alpha = \frac{v}{(1 - IoU) + v}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{\omega^{gt}}{h^{gt}} - \arctan\frac{\omega}{h}\right)^2$$

$$Loss = L_1 + \beta L_{rank} + \gamma L_{CIoU}$$

Experiment

TABLE II
QUANTITATIVE COMPARISON BETWEEN THE STATE-OF-THE-ART METHODS ON THE TESTING SET OF GAICD [30] DATASET

Method	Baseline_L	A2-RL	VPN	VFN	VEN	GAIC	GAIC	GAIC	CGS	TransView	Ours	Ours
backbone	-	Alexnet	VGG16	Alexnet	VGG16	VGG16	Mobile-V2	Shuffle-V2	VGG16	Mobile-V2	Mobile-V2	Shuffle-V2
$Acc_{1/5}$ ↑	26.5	23.2	36.0	26.6	37.5	67.2	68.2	68.0	67.4	69.0	69.2	68.4
$Acc_{2/5}$ ↑	-	-	-	26.5	35.0	62.7	64.3	64.1	63.2	66.9	64.9	65.4
$Acc_{3/5}$ ↑	-	-	-	26.7	35.3	58.9	61.3	60.7	60.1	61.9	62.1	62.1
$Acc_{4/5}$ ↑	-	-	-	25.7	34.2	55.9	58.5	56.6	57.0	57.8	58.0	58.3
$Acc_{1/10}$ ↑	44.0	39.5	48.5	40.6	50.5	84.0	84.4	85.8	83.0	85.4	85.2	84.4
$Acc_{2/10}$ ↑	-	-	-	40.2	49.2	81.4	82.7	82.5	80.3	84.1	83.0	83.3
$Acc_{3/10}$ ↑	-	-	-	40.3	48.4	79.2	80.7	80.5	78.3	81.3	81.7	81.2
$Acc_{4/10}$ ↑	-	-	-	39.3	46.4	76.8	78.7	77.8	76.7	78.6	79.5	78.8
$Acc^{w}_{1/5}$ ↑	16.4	15.1	19.1	18.0	20.2	48.2	48.8	49.2	49.4	-	51.0	51.3
$Acc^{w}_{2/5}$ ↑	-	-	-	13.3	15.2	46.4	46.8	47.6	47.0	-	48.9	49.6
$Acc^{w}_{3/5}$ ↑	-	-	-	12.3	14.1	44.5	45.4	46.0	45.1	-	47.2	47.8
$Acc^{w}_{4/5}$ ↑	-	-	-	11.3	13.4	43.2	44.1	43.2	43.5	-	44.5	45.4
$Acc^{w}_{1/10}$ ↑	27.7	25.6	29.4	27.9	30.1	63.7	64.2	65.1	64.3	-	66.1	66.8
$Acc^{w}_{2/10}$ ↑	-	-	-	22.9	25.4	62.9	63.6	64.5	62.9	-	65.0	65.9
$Acc^{w}_{3/10}$ ↑	-	-	-	21.8	24.1	61.9	62.8	63.4	61.7	-	64.3	64.1
$Acc^{w}_{4/10}$ ↑	-	-	-	20.6	23.3	60.8	61.8	61.4	60.6	-	62.8	62.5
SRCC ↑	-	-	-	0.485	0.616	0.842	0.849	0.850	0.854	0.857	0.858	0.855
PCC ↑	-	-	-	0.503	0.662	0.866	0.874	0.872	0.879	0.880	0.880	0.876
Runtime	-	274ms	12ms	1092ms	623ms	16ms	7ms	7ms	8ms	-	20ms	17ms

↑ Denote larger is better. The top two results are highlighted in red and blue respectively.

Experiment

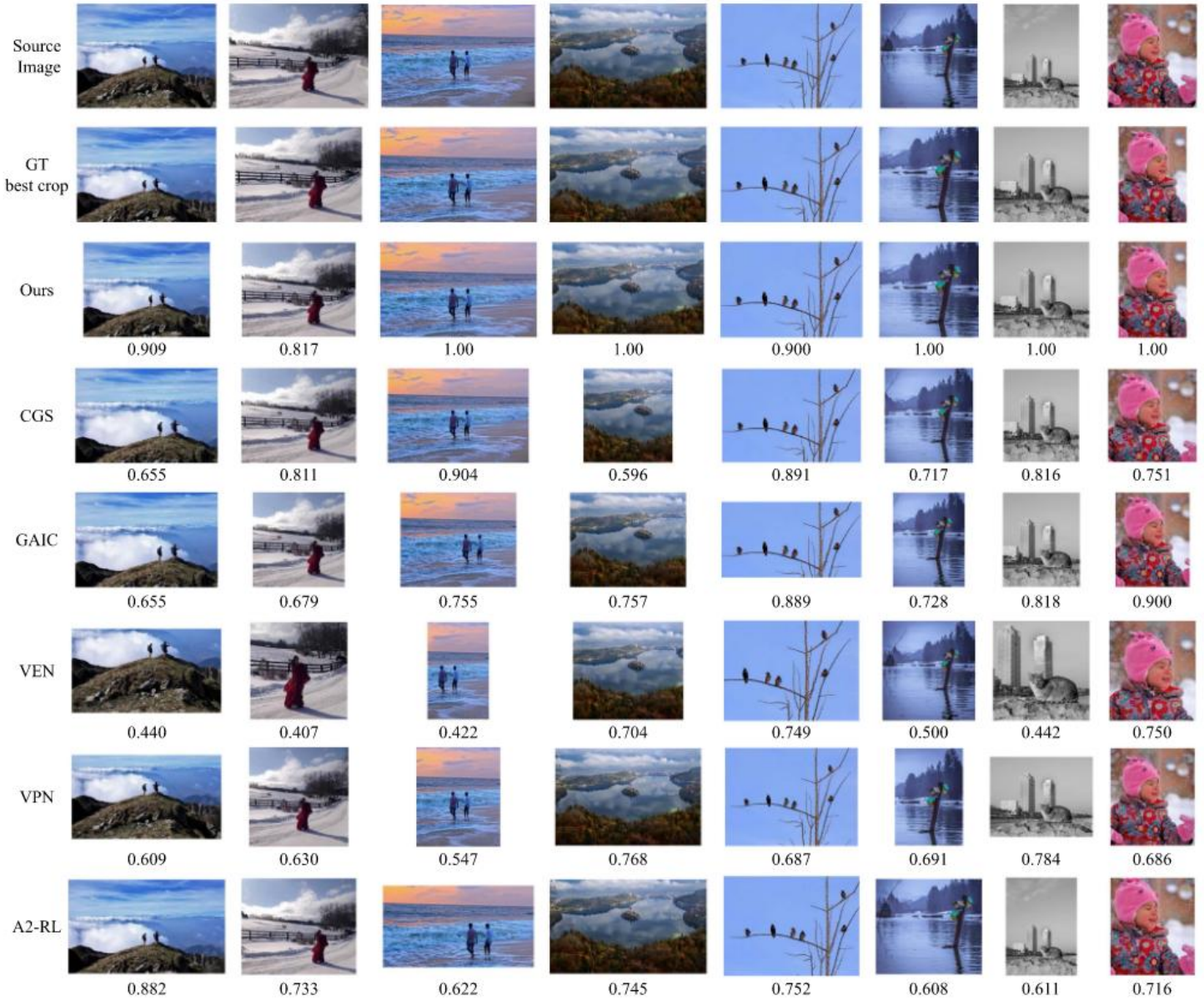


Fig. 6. Illustration of qualitative comparison of returned top-1 views of all the methods on GAICD dataset. The IoU value represents the degree of overlap between the returned top-1 view of the methods and the ground truth.

Experiment

Fig. 7. Illustration of qualitative comparison of returned top-1 views of all the methods on HCDB dataset. The IoU value represents the degree of overlap between the returned top-1 view of the methods and the ground truth.

Experiment

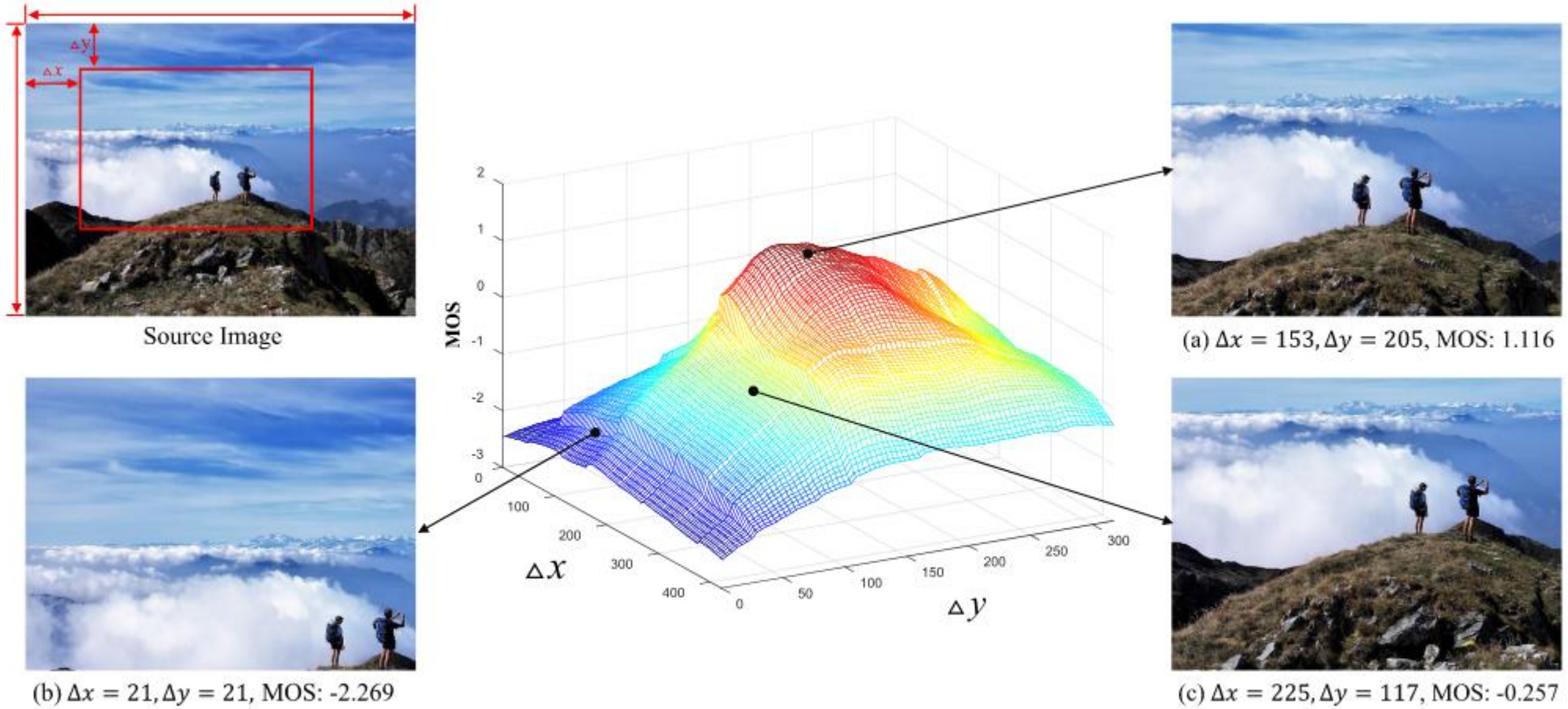


Fig. 9. "Mountaintop" image sequence cropped by sliding over the source image using the window with fixed width and height over an entire range of Δx and Δy values. The higher surface in the middle plot manifests the better quality of the crops predicted via the CGICAANet model. The upper-left image corresponds to the source image, while the image (a)–(c) denotes the crops with different quality predicted by our model.

Thanks