IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 14, 2021, accepted July 26, 2021, date of publication July 28, 2021, date of current version August 6, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3100816*

# An Experience-Based Direct Generation Approach to Automatic Image Cropping

**CASPER L. CHRISTENSEN AND ANEESH VARTAKAVI**
Gracenote, Emeryville, CA 94608, USA
Corresponding author: Aneesh Vartakavi (aneesh.vartakavi@nielsen.com)

**ABSTRACT** Automatic Image Cropping is a challenging task with many practical downstream applications. The task is often divided into sub-problems - generating cropping candidates, finding the visually important regions, and determining aesthetics to select the most appealing candidate. Prior approaches model one or more of these sub-problems separately, and often combine them sequentially. We propose a novel convolutional neural network (CNN) based method to crop images directly, without explicitly modeling image aesthetics, evaluating multiple crop candidates, or detecting visually salient regions. Our model is trained on a large dataset of images cropped by experienced editors and can simultaneously predict bounding boxes for multiple fixed aspect ratios. We consider the aspect ratio of the cropped image to be a critical factor that influences aesthetics. Prior approaches for automatic image cropping, did not enforce the aspect ratio of the outputs, likely due to a lack of datasets for this task. We, therefore, benchmark our method on public datasets for two related tasks - first, aesthetic image cropping without regard to aspect ratio, and second, thumbnail generation that requires fixed aspect ratio outputs, but where aesthetics are not crucial. We show that our strategy is competitive with or performs better than existing methods in both these tasks. Furthermore, our one-stage model is easier to train and significantly faster than existing two-stage or end-to-end methods for inference. We present a qualitative evaluation study, and find that our model is able to generalize to diverse images from unseen datasets and often retains compositional properties of the original images after cropping. We also find that the model can generate crops with better aesthetics than the ground truth in the MIRThumb dataset for image thumbnail generation with no fine tuning. Our results demonstrate that explicitly modeling image aesthetics or visual attention regions is not necessarily required to build a competitive image cropping algorithm.

**INDEX TERMS** Automatic image cropping, convolutional neural networks, image enhancement, image processing.

## I. INTRODUCTION

With the proliferation of devices like smartphones, smart televisions, and tablets, imagery in different aspect ratios is necessary for a user interface to comply with responsive web design standards. These images are often manually cropped, which can be very laborious to perform for a large number of images. Automatic Image Cropping, therefore, has great practical significance for large catalogs of images. An effective aesthetic cropping algorithm could be helpful to industries and applications that store and display large amounts of media, such as social networks or image sharing platforms, image galleries, surveillance systems, photography and graphic design software.

Image cropping is often performed to highlight visual attention regions discarding unwanted regions in the process. Alternatively or in conjunction, cropping can be performed to improve or maintain the aesthetics of an image. Experienced users, including trained photographers, may use composition concepts such as the rule of thirds or the golden ratio to maximize aesthetics while deciding how to crop images. The aspect ratio of the final cropped image is also essential when performing this task, as it affects the aesthetics and the framing of the image. For example, selecting a portrait crop from a landscape image with multiple subjects can only include a subset of them, and the final crop should not include any partially cropped faces for aesthetic reasons. Advances in image cropping methods could therefore inform and guide research in visual perception and aesthetics.

The detection of visual attention regions in images has been an active area of research for some time [33]. Attention-based automatic cropping approaches build on it by drawing bounding boxes around the image's salient regions, assuming that the best crop should include the salient region.

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Olague.

Assessment of image aesthetics is also an active research area, starting with low-level rules and features, which are difficult to formulate and do not generalize, to recent deep learning approaches [9]. The aspect ratio of an image is also essential to the perceived aesthetics, recognized by some recent aesthetics assessment approaches [5], [35]. However, it is rarely mentioned as a requirement or concern in prior approaches to automatic image cropping. Although some techniques can output bounding boxes in different aspect ratios [32], [43], they do so by evaluating multiple candidates and are therefore inefficient. Image cropping is a typical first stage for thumbnail generation approaches, which try to create smaller representations of images. These strategies often create thumbnails in fixed aspect ratios but usually do not consider image aesthetics [3], [9].

Early approaches to automatic image cropping tended to focus on either the aesthetics or the visual attention regions. More recent solutions try to incorporate both by modeling the cropping process in two stages. First, they determine a visual attention region or a Region Of Interest (ROI), and then, they draw a bounding box to maximize aesthetics. This two-stage approach has some disadvantages: when the image has no salient regions [24], [32], or when it has multiple salient subjects, some of which may need to be excluded for aesthetic reasons [22], [32].

We believe that a single-stage approach that implicitly models the image's aesthetics and attention regions can overcome some of the drawbacks of existing image cropping techniques. Our proposed model is less susceptible to failure cases that occur when attention or aesthetics are modeled explicitly, such as, when no salient region is found or when the ground truth for aesthetic assessment is ambiguous due to neutral image aesthetics [9]. We evaluate several common CNN architectures in a transfer learning framework and find that a WideResNet50-2 [42] backend achieves the best overall performance on our dataset with an IoU of 0.867. This model is more lightweight and efficient than two-stage approaches and is simpler to train. Without any model optimization or pruning, our model can process over 600 images/sec, or over 3000 crops/sec as each image is cropped in 5 aspect ratios on a single Nvidia Tesla V100 GPU during inference. This is significantly faster than existing approaches [3], [22]–[25].

To the best of our knowledge, this work is the first attempt at addressing the problem of image cropping directly, without explicitly modeling visual attention or aesthetics.Due to a lack of public datasets to support our approach, we train our CNN-based model using a large internal dataset of images cropped by experienced editors in fixed aspect ratios, who simultaneously maintain image aesthetics and important image content. We propose an efficient architecture that predicts bounding boxes for multiple aspect ratios simultaneously, without evaluating multiple crop candidates. Prior approaches for image cropping did not enforce the aspect ratio of their outputs. We, therefore, benchmark our task on datasets for two related tasks - FCDB [6] for aesthetic image cropping without regard to aspect ratio, and MIR-Thumb [3] for thumbnail generation in fixed aspect ratios where aesthetics are not crucial. Our model with a WideResNet50-2 backend, modified to generate outputs in any aspect ratio, is competitive with and more efficient than existing approaches on FCDB, achieving an IoU of 0.692. We also achieve state-of-the-art performance on the MIR-Thumb dataset at an IoU of 0.741 with no fine-tuning. This demonstrates that explicitly modeling aesthetics or attention regions is not strictly required for accurate and efficient image cropping. Finally, we include a qualitative evaluation, where we investigate the generalization ability of the model on the FCDB and MIRThumb datasets without fine tuning. Where also observe that the model can generate more aesthetic crops on MIR-Thumb than the original ground truth. This finding highlights some challenges in the objective evaluation of image cropping systems, such as the reliance on crowd-sourced workers to gather ground truth and using a single reference for the IoU metric when several equally good crops may exist.

In summary:

- We are the first work to attempt aesthetic image cropping directly and show that explicitly modeling visual attention or image aesthetics is not necessary to build a competitive image cropping algorithm.
- We propose a simple architecture, with no bells and whistles that is easier to train compared to recent state-of-the-art approaches, such as separated network branches for bounding box prediction [3], ROI-aware pooling operations [3], [22], [25], human-defined composition patterns [32], and custom loss functions [22], [32].
- Our proposed single-stage model is efficient and able to output bounding boxes of multiple fixed aspect ratios, without evaluating multiple candidates, which is novel for aesthetic aware image cropping approaches.

## II. RELATED WORK

Prior approaches to solve the automatic image cropping problem can be distinguished by how the cropping candidates are initially determined and how they are evaluated to get the final crop. The task of selecting cropping candidates is generally solved by a few different approaches:

- *Sliding-Judging* - These techniques generate a large number of candidates by moving windows of varying sizes and aspect ratios over the original image, each of which is then evaluated against some criterion such as image aesthetics or attention regions to find the best candidate [11], [28], [31], [43]. These strategies are generally computationally inefficient as the search space spans the entire image [22], [37]. Some authors have developed strategies to mitigate this by exploiting properties such as local redundancy [43] or by eliminating candidates that do not encompass the entire region of interest [40]. Other authors suggest more efficient solutions that evaluate fewer candidates, but without regard to aesthetics [4].

- *Determining-Adjusting* - These methods try to first determine an ROI in the image. They then generate many candidates around that region by adjusting the position, height, or aspect ratio of the bounding boxes, and evaluate each of them to find the best cropping candidate [37], [38]. They are more efficient than sliding-judging approaches because they generate fewer candidates, but they struggle when no ROI is found [22], [24].
- *Finding-Generating* - These methods aim to predict a single crop region by calculating a bounding box that includes the visual attention region in the image. This is then fed into a regression network that predicts the optimal bounding box [22], [24]. These strategies are efficient because they generate a single candidate, instead of generating and evaluating multiple candidates as in determining-adjusting approaches. However, these methods also struggle when no ROI is found [22], [24].

Once the candidates are generated, prior approaches evaluate them in a few different ways:

- *Saliency* or attention-based methods assume that the best crop will generally contain the most salient regions. The techniques for finding the salient regions range from signal processing [16] to deep learning methods [20], [34], [36]. *Determining-adjusting* approaches often use these methods to find an ROI [37], [38]. Other saliency-based cropping methods include Ardizzone *et al.* [1], Ciocca *et al.* [8], and Sun and Ling [31].
- *Aesthetic evaluation* methods try to quantify and score images or crop candidates based on their aesthetic qualities. A comprehensive review of these methods is presented by Deng *et al.* [9]. Aesthetic image cropping algorithms sometimes use features inspired by composition rules such as the rule of thirds and visual balance [17], [32], [40]. Datasets such as AVA [27] enable learning aesthetics using deep learning methods [26]. Other aesthetics-based image cropping approaches include Nishiyama *et al.* [28], Zhang *et al.* [45], and Chen *et al.* [7].
- *Fusion* methods try to combine attention and aesthetic methods in two stages and harness the advantages of both. Some approaches use a determining-adjusting strategy by first predicting the attention region, then generating a small number of candidates around it, and finally selecting the one with the best aesthetic evaluation score [37], [38]. Finding-generation strategies try to regress the bounding box after detecting salient regions in an image [22], [24]. Other fusion approaches include Tu *et al.* [32], Guo *et al.* [13], and Li *et al.* [19].
- *Experience-based* methods try to predict a bounding box using a dataset of images cropped by humans. Prior methods that follow this strategy design handcrafted features such as sharpness and color distance that are then used for regressing the bounding boxes [40], [41].

Some image cropping methods do not easily fit into this framework: a reinforcement learning framework [19], rank-based evaluation metrics on a densely annotated dataset [43], [44], weakly supervised learning [21], and rank-based learning approaches [6], [25].

We propose an *experience-based direct generation* strategy, which has not been attempted for aesthetic image cropping to the best of our knowledge. We propose the term *direct generation* to represent methods that predict the bounding box directly from an input image, without the overhead of detecting visual attention regions or evaluating multiple cropping candidates. These methods do not suffer from the same drawbacks as *finding-generating* approaches, such as when an ROI is absent, and are more efficient than *sliding-judging* and *determining-adjusting* methods. Our model is trained to directly predict the bounding boxes for different aspect ratios simultaneously, using a shared feature extractor for efficiency. We build a large internal dataset to train our experience-based approach with no handcrafted features, overcoming the limitations that restricted other methods [13], [22], [24], [39].

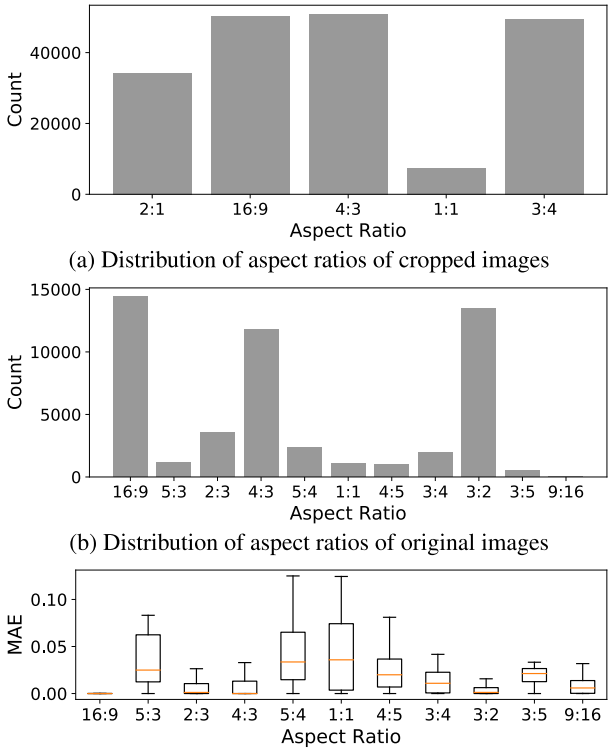
There are some use cases where efficiency is not as important and evaluating multiple candidates may be desired, for example when presenting multiple candidates to a user and allowing them to pick the best candidate based on their preferences. However, in this work, we continue a trend in prior work that focuses on applications that benefit from reducing the number of evaluated candidates for efficiency reasons.

Image cropping is related to thumbnail generation, which aims to create smaller representative versions of the original images by preserving the most useful content from the original image and discarding the background. In contrast, image cropping approaches try to create new images, balancing aesthetic quality while including visually salient regions. Our approach is similar to some recent thumbnail generation approaches such as FastAT [10], and CropNet [3] in that they predict an output without generating multiple candidates. CropNet uses a similar strategy of a shared feature extractor and dedicated branches to predict multiple bounding boxes of fixed aspect ratios but follows a different strategy of predicting bounding boxes. CropNet is trained on MIR-Thumb, a smaller crowd-sourced dataset annotated by non-experienced workers, in contrast with our larger dataset annotated by trained experts who also pay attention to image aesthetics. In Section IV, we benchmark our approach on the MIR-Thumb test set and achieve state-of-the-art performance with no fine-tuning. We also demonstrate that our algorithm can produce more aesthetically pleasing images and display some examples.

## III. OUR APPROACH

### A. DATASET

Prior approaches [13], [22], [24], [39] often cite the lack of large datasets for effective image cropping and design workarounds to overcome this limitation. We were unable to




(a) Distribution of aspect ratios of cropped images

(b) Distribution of aspect ratios of original images

(c) Mean Absolute Error (MAE) between the aspect ratio of the original image and the closest aspect ratio

**FIGURE 1. Aspect ratio distributions.**

find existing datasets that support our experience-based direct generation approach to image cropping with strict aspect ratio requirements. We therefore collect an internal dataset of about 51, 000 images for this study, serving as iconic imagery for TV programs and movies. The images usually include the lead characters along with a background that conveys context relevant to the program. Each image was manually cropped in up to 5 aspect ratios (16:9, 4:3, 2:1, 3:4, and 1:1) by a large group of experienced editors who were asked to retain important image content, preserve the aesthetics, and adhere to strict aspect ratio requirements. Unlike some datasets such as FCDB, we did not rate or discard images based on their aesthetics in an effort to mitigate subjective bias. Some prior datasets suggested bounding boxes for their workers to rank or annotate, citing efficiency reasons [6], [39]. In contrast, we allowed our editors to crop the images directly to avoid bias. Only a single editor was allowed to crop a given image in one aspect ratio, and no ranking or rating information was collected. We present some examples of these images in Section IV-D. Of all the images in the resulting dataset, not every image was cropped in every aspect ratio, as can be seen in Figure 1a. The dataset is also diverse in the aspect ratios of the original images, as illustrated in Figure 1b. A successful model would have to generalize to input images of many sizes. We consider other common aspect ratios in addition to those mentioned above for this visualization. We also compute the mean absolute error of each image to the closest aspect ratio in Figure 1c, similar to the analysis performed by Celona *et al.* [2].

### B. PRE-PROCESSING AND AUGMENTATION

We resize our images to (224, 224) and pad with zeros where necessary, to retain the aspect ratio of the original image. We store the bounds locations and the bounding boxes annotated by the editors as normalized coordinates of the top left and the bottom right corners. We then augment the image by randomly applying horizontal flips or color transformations such as changing the brightness or saturation or converting to grayscale. We do not apply any spatial transformation such as rotation or vertical flipping because these affect the composition of the image [43].

### C. MODEL

Our proposed model can be conceptually divided into two modules - a shared CNN-based feature extractor as the backbone, and multiple parallel regression heads, one for each aspect ratio. We illustrate this in Figure 2a. This design allows us to add predictor heads for new aspect ratios without having to retrain the rest of the network from scratch, or significantly increasing the time for inference. As illustrated in IV-C, we are also able to generate crops for unseen aspect ratios without pre-training by leveraging the predictions from similar aspect ratios, which is helpful when training data is scarce.

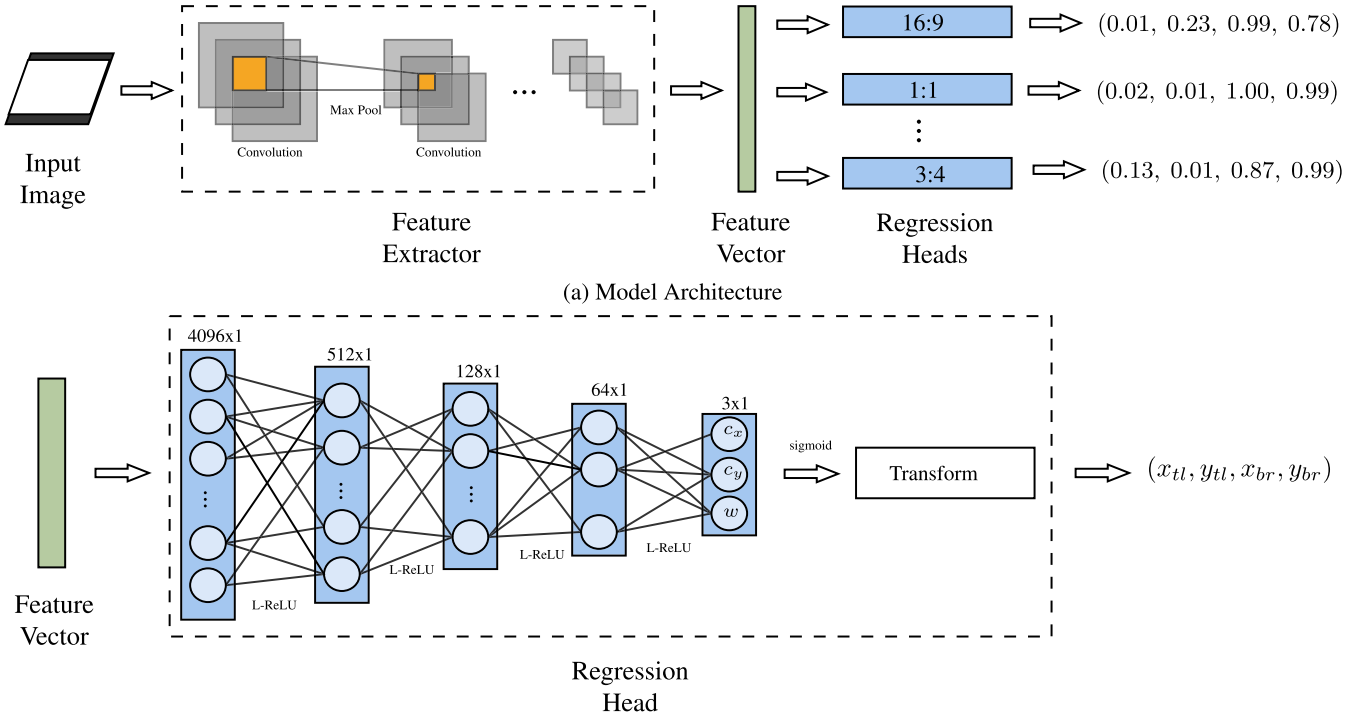
#### 1) FEATURE EXTRACTOR

The feature extractor is designed to output a fixed-length feature vector for each input image, which is subsequently fed to the regression heads. We use a shared feature extractor because the regression head for each aspect ratio needs similar information to make a prediction, such as the important regions and their locations in the image. We try a few common CNN architectures for the backbone, including VGG [30], ResNet [14], DenseNet [15], WideResNet [42] and MobileNet-v2 [29]. These architectures have been used for many computer vision tasks, including some previous solutions to automatic image cropping [25], [32]. Another advantage of using common architectures is the wide availability of pre-trained network weights on Image Classification and related tasks. We study the effect of transfer learning using these pre-trained networks in Section IV-A.

#### 2) REGRESSION HEAD

As shown in Figure 2b, each regression head is a densely connected neural network, with Leaky ReLU as the activation function for the intermediate layers, and sigmoid activation at the output. Each regression head is dedicated to predicting a bounding box of a single aspect ratio, often represented as coordinates of the top-left $(x_{tl}, y_{tl})$ and the bottom-right corners $(x_{br}, y_{br})$. However, predicting the bounding box using this representation does not guarantee that the output would correspond to the desired fixed aspect ratio.

We use an alternate regression head to predict images with a fixed aspect ratio, which we call an aspect ratio




(a) Model Architecture

(b) Aspect Ratio Enforced Regression Head, where the transform generates the bounding box in the correct aspect ratio.

**FIGURE 2.** Proposed model framework.

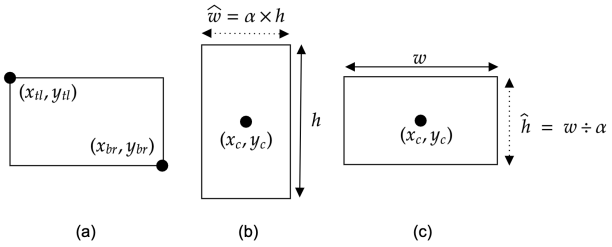



**FIGURE 3.** (a) Non-enforced bounding box prediction. (b) Enforced bounding box prediction where width $\hat{w}$ is inferred from the aspect ratio $\alpha$ (c) Enforced bounding box prediction where height $\hat{h}$ is inferred from the aspect ratio $\alpha$.

enforced regression head. For a landscape or square aspect ratio, we predict the coordinates of the center $(x_c, y_c)$ and the width $w$, using the aspect ratio $\alpha$ to predict the height. For a portrait aspect ratio, we predict the center coordinates $(x_c, y_c)$ and the height $h$, using $\alpha$ to predict the width. We illustrate this in Figure 3. Since the aspect ratio, $\alpha$, is fixed for a given regression head, we can draw a bounding box by calculating the remaining dimension, represented by the transform operation in Figure 2b. At run time, we clip the prediction bounding box to the largest possible bounding box for the given image and the predicted center coordinates $(x_c, y_c)$ to avoid invalid output. We use the Smooth L1 loss between the annotated and the predicted bounding box coordinates, similar to Fast R-CNN [12]. Our experiments below were performed with models with a single enforced regression head per aspect ratio. The architecture could be extended to include multiple regression heads per aspect ratio if desired. This could be useful in cases where, for example, a close-up version and a zoomed out version for each aspect ratio are needed.

## IV. EXPERIMENTS AND ABLATION STUDY

We perform a 60/20/20 split on our dataset to create training, validation, and test sets. We use the ADAM [18] optimizer with a default learning rate of 0.0001 and a batch size of 128, and use early stopping with the validation set.

We use the Boundary Displacement Error (BDE) and the Intersection over Union (IoU) to evaluate cropping, in line with previous approaches [2], [13], [22], [24], [32]. As we did not have editors rank or rate different crops, we cannot compute ranking metrics or leverage ranked learning approaches. All metrics in the tables in this section are averaged across all the aspect ratios in our test set.

### A. EVALUATION ON OUR DATASET

We compare our models with a baseline method that predicts bounding boxes of varying sizes in the correct aspect ratio around the center of the image [2]. We denote this family of methods, *Baseline-s*, where *s* represents the scaling factor of the bounding box as a fraction of the largest possible bounding box for that aspect ratio. We also compare our method with GAIC [43], a recent method capable of cropping images in fixed aspect ratios. This is enabled by selecting the aspect ratio of the generated candidates, different from

**TABLE 1. Model evaluation on our dataset.**

| Model | Pre-Training | Train Set | Enforced | IoU ↑ | BDE ↓ |
|---|---|---|---|---|---|
| Baseline-0.8 | - | - | - | 0.645 | 0.076 |
| Baseline-0.9 | - | - | - | 0.697 | 0.061 |
| Baseline-1.0 | - | - | - | **0.728** | **0.053** |
| GAIC [43] | ImageNet | GAIC [43] | True* | 0.723 | 0.058 |
| Ours (WideResNet-50-2) | ImageNet | Ours | True | **0.867** | **0.023** |
| Ours (WideResNet-50-2) | ImageNet | Ours | False | 0.855 | 0.025 |
| Ours (WideResNet-50-2) | None | Ours | True | 0.832 | 0.030 |

**TABLE 2. CNN Backbone Architecture Comparison, pre-trained on ImageNet and fine-tuned on our dataset.**

| Model | Size | IoU ↑ | BDE ↓ |
|---|---|---|---|
| VGG16 | 138.3M | 0.854 | 0.025 |
| WideResNet-50-2 | 68.8M | **0.867** | **0.023** |
| ResNet-50 | 25.5M | 0.861 | 0.025 |
| ResNeXt-50 | 25.0M | 0.846 | 0.027 |
| Densenet-121 | 7.9M | 0.860 | 0.025 |
| MobileNet-v2 | 3.5M | 0.854 | 0.025 |

our enforced predictor head method which is more efficient. We use the trained models and code released by the authors, but are unable to fine-tune GAIC on our training set as GAIC relies on densely annotated images which are not available in our dataset. Our consolidated results, evaluated on our dataset can be seen in Table 1.

We also study the influence of pre-training on the ImageNet dataset for the feature extractor component of the model. We find that pre-training offers significant performance improvements, and present our results in Table 1, likely because both tasks require the model to learn the position and the type of objects in an image. We use transfer learning for all subsequent experiments.

#### 1) ENFORCED ASPECT RATIO PREDICTION

We test our method of enforcing the aspect ratio of the bounding box, and report the results with non-enforced and enforced predictions in Table 1. The aspect ratio enforced prediction method improves model performance while also satisfying the exact aspect ratio requirement.

#### 2) CNN BACKBONE ARCHITECTURE

We experiment with various common CNN architectures pre-trained on ImageNet for the feature extractor. We use enforced aspect ratio regression heads, keep all other hyper-parameters such as learning rate constant and present the metrics in Table 2. We find that the WideResNet-50-2 architecture performs the best on our test set overall. We also find that MobileNet-v2 performs very well, considering its smaller size in terms of the number of trainable parameters.

### B. EVALUATION ON FCDB

The datasets most commonly used to evaluate automatic image cropping methods like FCDB [6] do not impose any requirements on aspect ratios. Since our model is designed to predict fixed aspect ratios, this makes an accurate benchmark difficult. Nevertheless, we modify our model for this experiment to produce bounding boxes in any aspect ratio. Specifically, we remove the aspect ratio enforced regression heads and attach a single non-enforced regression head to the trained feature extractor. We further split the FCDB training set 80/20 into a training and validation split, and then fine-tune our modified model on the resulting training split using a batch size of 128, and an ADAM optimizer with a learning rate of $1 \times 10^{-5}$ for 300 epochs. We use early stopping on the validation split, similar to the previous experiments.

We present metrics on the FCDB test set in Table 3. We report the metrics of VFN [7] and VPN [39] as in Lu *et al.* [25], without including the ground truth window as a candidate view for VFN, and without the post processing step in VPN.

The results demonstrate that our approach is competitive with other models that explicitly model image aesthetics or visual attention regions without evaluating multiple crop candidates. Our model achieves a higher IoU score than the end-to-end model by Lu *et al.* [22], with a more straightforward training approach that does not require the identification of visual attention regions. LVRN [25] achieves a slightly higher IoU score, but evaluates an average of 1,745 candidates per image, which is inefficient. The ASM-Net [32] achieves a higher IoU score, but uses an inefficient two-stage searching step and derives composition patterns from human-defined composition rules that may not generalize. Out of these, VFN [7], LVRN [25], Wang *et al.* [38] and Lu *et al.* [22] perform fine-tuning on the FCDB training set, while the authors of the other approaches only use the FCDB test set for evaluation purposes. We are not able to find a significant influence of fine-tuning on the metrics, likely because of the relatively small size of FCDB (1395 train and 348 test images) and the wide differences between individual approaches. Our model's performance is comparable to or better than the subset of models that fine-tune on FCDB, and is significantly more efficient.

We also study the impact of transfer learning on our dataset, by initializing the feature extractor using the weights from ImageNet and training the model as described above. The resulting model labeled "Ours-ImageNet Only", achieves an IoU of 0.679, slightly worse than the model initialized with the weights learned on our dataset, but better than Lu *et al.* [22], VFN [7] and VPN [39]. This implies that the proposed architecture is a more significant contributor to

**TABLE 3.** Evaluation on FCDB, where * highlight models that explicitly model aesthetics and/or attention regions. FPS refers to the number of input frames per second.

| Model | Fine Tuning | Avg. Candidates | IoU ↑ | BDE ↓ | FPS↑ | GPU Hardware |
|---|---|---|---|---|---|---|
| VFN [7] | Yes | 137 | 0.632 | 0.098 | 0.78 [21] | N/A |
| A2-RL [19] | No | 13.56 | 0.663 | 0.089 | 4.08 | Nvidia Titan X |
| VPN [39] | No | 895 | 0.664 | 0.085 | 75 | N/A |
| Wang et al.* [38] | Yes | 1296 | 0.65 | 0.08 | - | - |
| LVRN [25] | Yes | 1745 | 0.7100 | 0.0735 | 125 | Nvidia 1080 |
| ASM-Net * [32] | No | N.A. | 0.748 | 0.068 | - | - |
| Lu et al.* [22] | Yes | 1 | 0.673 | 0.058 | 50 | Nvidia 2080 Ti |
| Lu et al.* [23] | No | 1 | 0.673 | 0.058 | 50 | Nvidia 2080 Ti |
| Lu et al.* [21] | No | 1 | 0.681 | 0.084 | 285 | Nvidia 1080 Ti |
| **Ours-ImageNet Only** | Yes | 1 | 0.679 | 0.067 | 606 | Nvidia Tesla V100 |
| **Ours** | Yes | 1 | 0.692 | 0.064 | 606 | Nvidia Tesla V100 |

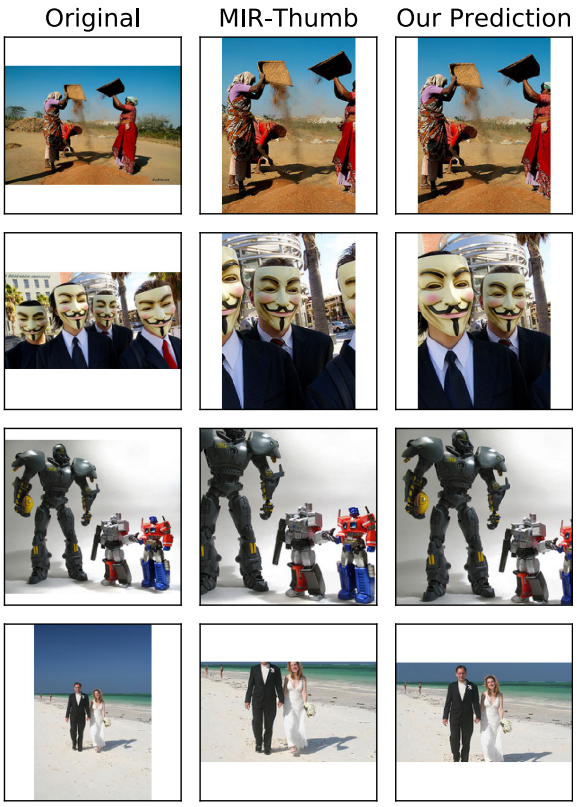



**FIGURE 4.** Our model can generate crops with better aesthetics than the original annotations in MIR-Thumb without fine-tuning.

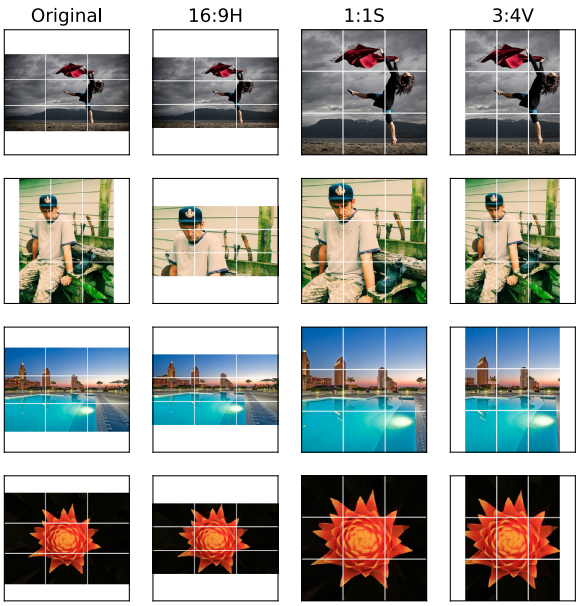



**FIGURE 5.** Illustrating the model's ability to retain aesthetic and composition properties (eg. rule of thirds) of the original image, evaluated on images sourced from the MIR-Thumb dataset without fine-tuning.

the performance on FCDB, compared to pre-training on our dataset.

Prior approaches measured their efficiency during using the time to crop a single image as a metric, which is dependent on hardware, input image size and implementation details, making a fair comparison difficult. Nevertheless, we present the number of crops per second and hardware self-reported by the authors in Table 3 as a measure of efficiency. Amongst the approaches compared, sliding-judging methods such as VFN [7] have the lowest efficiency. More recent approaches from Lu *et al.* [22], [23] report an overall processing speed of 50fps for their image cropping solution an Nvidia 2080Ti GPU. The weakly supervised approach by Lu *et al.* [21] is able to crop images at 285 fps. In contrast, our model with a WideResNet-50-2 backbone can crop 606 input frames per second on a single Nvidia Tesla V100 GPU in 5 different aspect ratios simultaneously, resulting in over 3000 output crops per second. This time is calculated for inference only without any optimizations, and does not include time to load and pre-process the images, or save the final cropped images in order to be consistent and enable comparisons with prior work [3].

### C. EVALUATION ON MIR-THUMB

Even though the goals of image cropping methods differ from those of thumbnail generation, datasets like MIR-Thumb used by CropNet [3] are similar to ours, in that they annotate the same image with bounding boxes of different aspect ratios. We, therefore, evaluate our model on the MIR-Thumb test set to test its generalization ability. We also include our baseline methods from Section IV-A2 for comparison.

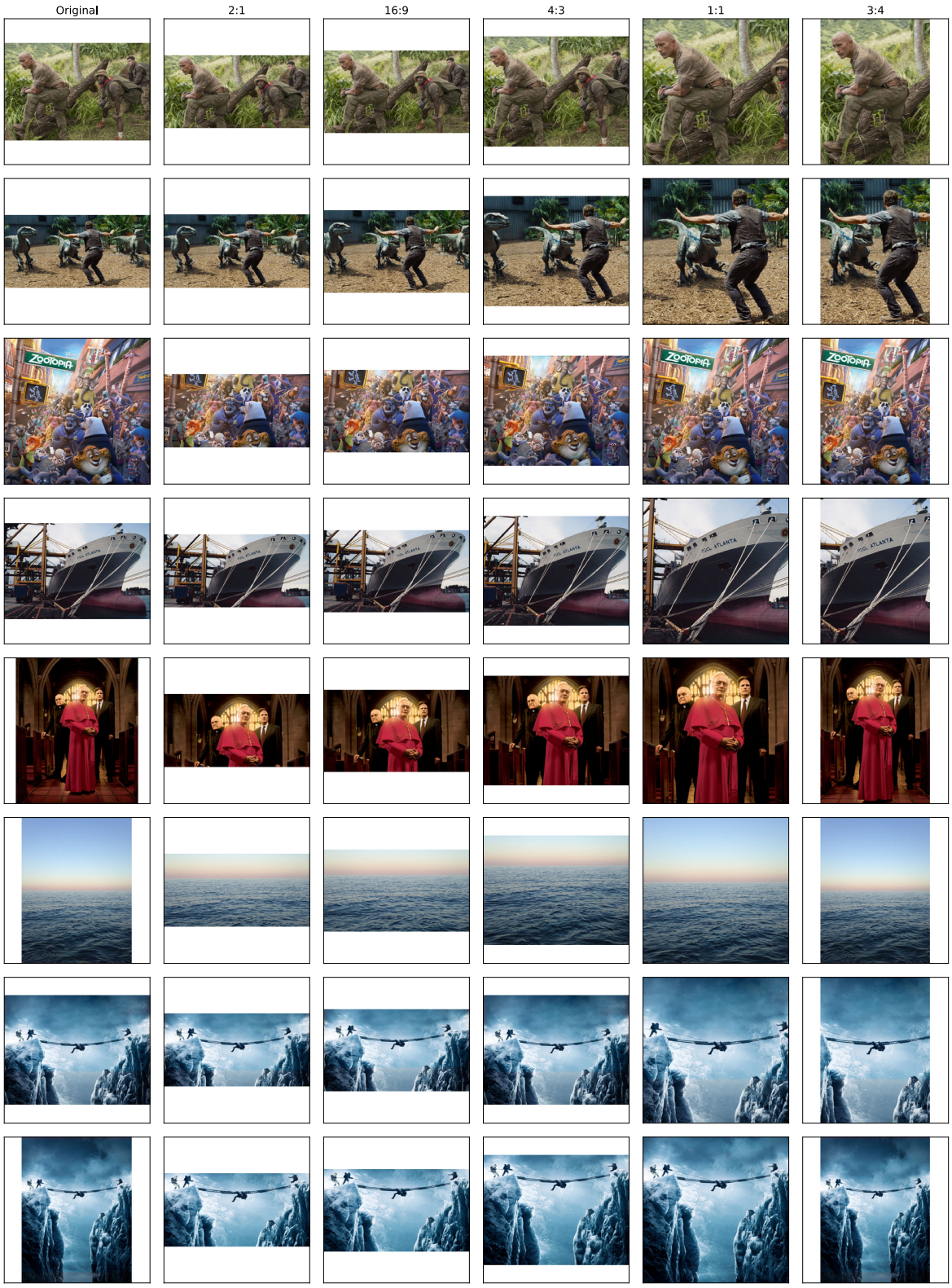



**FIGURE 6.** Results on images from our test dataset.

MIR-Thumb includes some aspect ratios that were not present in our dataset, namely 21:9, 9:16, and 9:21. We synthesize the model predictions for these aspect ratios by adjusting the bounding boxes of the closest aspect ratio in our model. To generate the target aspect ratio of 21:9, we reduce the height uniformly around the center of the 2:1 prediction in our model and keep the width constant. The closest aspect ratio to 9:16 and 9:21 was 3:4,

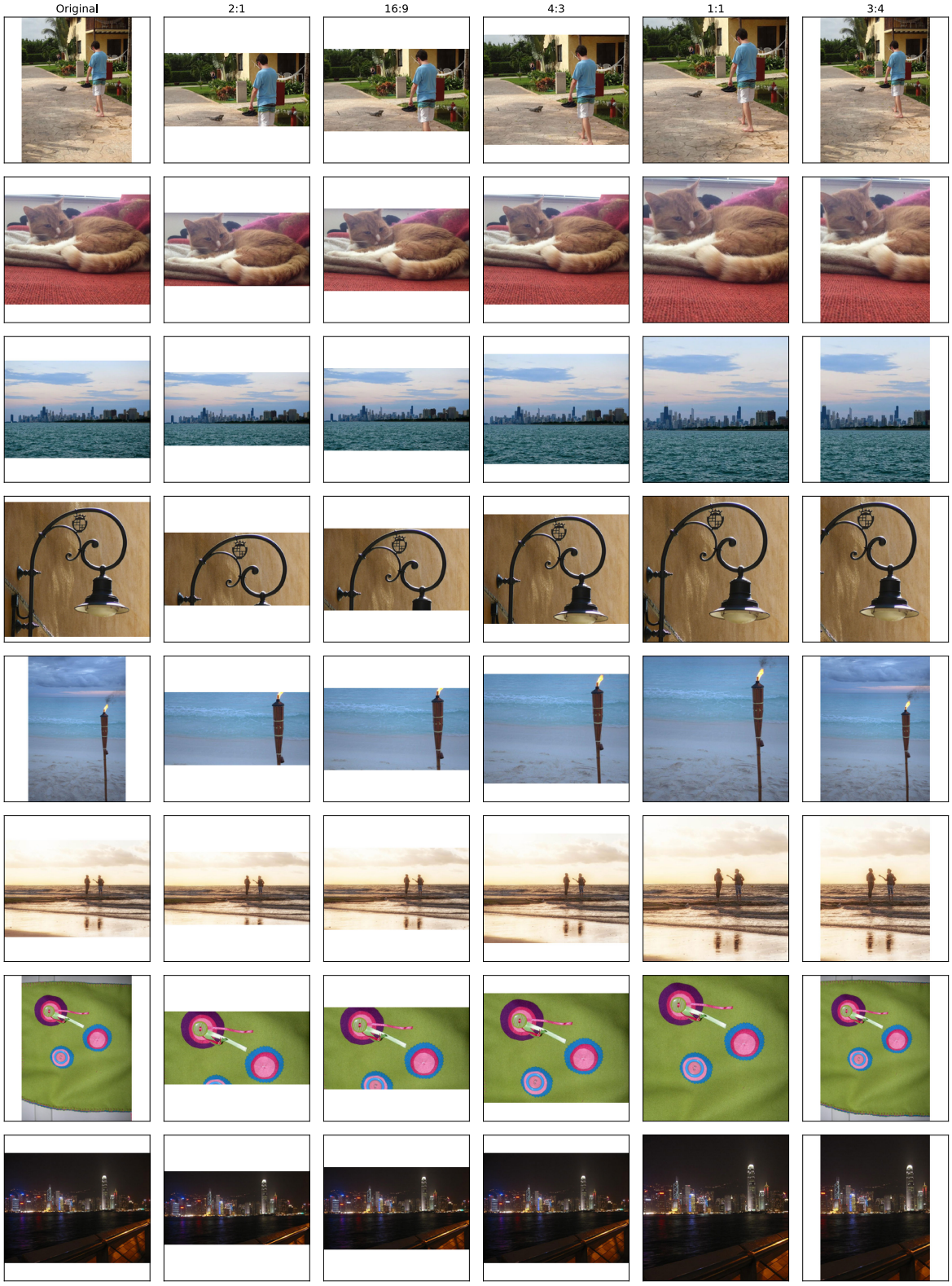



**FIGURE 7.** Results from our model trained on our dataset and evaluated on images in FCDB with no fine tuning.

where we keep the height constant and shrink the width. The results are shown in Table 4, where our model achieves state-of-the-art performance with no fine-tuning. We achieve a significantly higher IoU of 0.770 when we only consider aspect ratios that were in our training set namely 16:9, 1:1, 4:3, and 3:4. These results indicate that our learned model generalizes well to other similar datasets and tasks.

**TABLE 4.** Evaluation on MIR-Thumb.

| Model | Train Set | Test Set | IoU ↑ |
|---|---|---|---|
| Baseline-0.8 | None | MIR-Thumb | 0.488 |
| Baseline-0.9 | None | MIR-Thumb | 0.505 |
| Baseline-1.0 | None | MIR-Thumb | 0.506 |
| CropNet | FAT-Clean [3] | MIR-Thumb | 0.672 |
| CropNet | MIR-Thumb | MIR-Thumb | 0.711 |
| **Ours** | Ours | MIR-Thumb | **0.741** |

### D. QUALITATIVE ASSESSMENT

The perception of image aesthetics is inherently subjective. We, therefore, provide visual examples of some results of our algorithm on images from MIR-Thumb, FCDB, and our dataset. We first illustrate a few cases where our model with no fine-tuning produced crops with better aesthetics than even the annotated images in the MIR-Thumb test set, as seen in Figure 4. Most of these cases involve partially cropped human subjects which are rare in our training dataset but appear more frequently in the MIR-Thumb dataset, resulting in our model predictions getting a low IoU score, even though the predicted crops have arguably better aesthetics. This finding reveals some open challenges in the objective evaluation of image cropping systems, such as using the IoU as a metric, the reliance on a single reference annotation and using inexperienced crowd-sourced workers. Future research in these areas is critical in order to build reliable and robust image cropping systems.

Additionally, we find that our model predictions appear to retain some composition aspects from the original image without explicitly modeling aesthetics during training. To illustrate this, we draw a rule-of-thirds grid over some images from MIR-Thumb and the resulting predictions in Figure 5. This behavior is consistent even when the aspect ratios of the source images and the target crop are quite different. We believe this behavior is a likely result of our dataset that includes well-composed source images cropped by editorial experts, unlike other datasets for image cropping such as FLMS that exclude well-composed images, assuming that they do not require further cropping [11].

Furthermore, we include some of the model predictions from our test set in Figure 6. The model can identify the main subject in the image, even if the subject is relatively small, facing away from the camera or is inanimate. The last two rows in Figure 6 are intended to display predictions when we input two images with similar content but different aspect ratios. In both cases, the model can preserve the regions of interest while producing aesthetically similar crops for many of the output aspect ratios.

We finally include some examples of the model predictions on FCDB without any fine tuning in Figure 7, to illustrate the generalization ability of our model on a different dataset. The model is able to perform well on challenging and diverse images such as close up images of pets, day and night time landscapes, abstract patterns and inanimate objects. We observe that the model is able to retain important image content even in difficult cases when a large portion of the image has to be excluded, such as choosing a 2:1 crop of a portrait image (as seen in rows 1, 5, and 7 of Figure 7).

## V. CONCLUSION

We proposed a novel *experience-based direct generation* strategy for image cropping. The model was designed to directly predict bounding boxes for a fixed aspect ratio, without explicitly modeling image aesthetics or visual attention regions. The model was trained on a large dataset of images annotated by experts, who tried to maintain image aesthetics and visual attention regions in the cropped images. We designed an efficient, straightforward architecture with a shared feature extractor and multiple dedicated regression heads to simultaneously predict the bounding box for different aspect ratios. Our model is easier to train than existing multi-stage approaches, and more efficient for inference as it does not evaluate multiple candidates.

Due to a lack of public datasets for our task, we benchmarked our model on two related datasets - FCDB for aesthetic image cropping without regard to aspect ratio, and MIR-Thumb for image thumbnail generation in fixed aspect ratios where aesthetics are not crucial. Our model, modified to generate outputs without defined aspect ratios, achieved results comparable to existing approaches, while being more efficient and easier to train. We achieved state-of-the-art results on the MIR-Thumb dataset without fine-tuning. Finally, we displayed some examples where our model generates more aesthetic crops than the ground truth annotations in MIRThumb. We also performed a qualitative evaluation and showed that our model is able to generalize across multiple datasets without fine-tuning, and also frequently retain aesthetic properties of the source image in the final crops.

## REFERENCES

[1] E. Ardizzone, A. Bruno, and G. Mazzola, "Saliency based image cropping," in *Image Analysis and Processing—ICIAP*, A. Petrosino, Ed. Berlin, Germany: Springer, 2013, pp. 773–782.

[2] L. Celona, G. Ciocca, P. Napoletano, and R. Schettini, "Autocropping: A closer look at benchmark datasets," in *Image Analysis and Processing—ICIAP*, E. Ricci, S. R. Bulò, C. Snoek, O. Lanz, S. Messelodi, and N. Sebe, Eds. Cham, Switzerland: Springer, 2019, pp. 315–325.

[3] H. Chen, B. Wang, T. Pan, L. Zhou, and H. Zeng, "CropNet: Real-time thumbnailing," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 81–89.

[4] J. Chen, G. Bai, S. Liang, and Z. Li, "Automatic image cropping: A computational complexity study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 507–515.

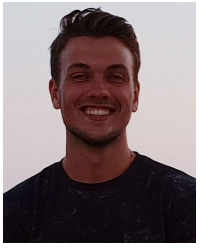
[5] Q. Chen, W. Zhang, N. Zhou, P. Lei, Y. Xu, Y. Zheng, and J. Fan, "Adaptive fractional dilated convolution network for image aesthetics assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14114–14123.

[6] Y.-L. Chen, T.-W. Huang, K.-H. Chang, Y.-C. Tsai, H.-T. Chen, and B.-Y. Chen, "Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 226–234.

[7] Y.-L. Chen, J. Klopp, M. Sun, S.-Y. Chien, and K.-L. Ma, "Learning to compose with professional photographs on the web," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 37–45.

[8] G. Ciocca, C. Cusano, F. Gasparini, and R. Schettini, "Self-adaptive image cropping for small displays," *IEEE Trans. Consum. Electron.*, vol. 53, no. 4, pp. 1622–1627, Nov. 2007.

[9] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 80–106, Jul. 2017.
[10] S. A. Esmaeili, B. Singh, and L. S. Davis, "Fast-at: Fast automatic thumbnail generation using deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4622–4630.
[11] C. Fang, Z. Lin, R. Mech, and X. Shen, "Automatic image cropping using visual composition, boundary simplicity and content preservation models," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 1105–1108.
[12] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[13] G. Guo, H. Wang, C. Shen, Y. Yan, and H.-Y.-M. Liao, "Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2073–2085, Aug. 2018.
[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[15] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[16] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[17] M. B. I. C. T. Khuan and M. E. R. M. K. Islam, "Aaics: Aesthetics-driven automatic image cropping and scaling," in *Proc. Int. Conf. Data Mining, Multimedia, Image Process. Their Appl. (ICDMMIPA)*, 2016, p. 8.
[18] P. D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 2015, pp. 1–15.
[19] D. Li, H. Wu, J. Zhang, and K. Huang, "A2-RL: Aesthetics aware reinforcement learning for image cropping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8193–8201.
[20] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 362–370.
[21] P. Lu, J. Liu, X. Peng, and X. Wang, "Weakly supervised real-time image cropping based on aesthetic distributions," in *Proc. 28th ACM Int. Conf. Multimedia (MM)*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 120–128.
[22] P. Lu, H. Zhang, X. Peng, and X. Jin, "An end-to-end neural network for image cropping by learning composition from aesthetic photos," *ArXiv*, vol. abs/1907.01432, 2019.
[23] P. Lu, H. Zhang, X. Peng, and X. Jin, "Learning the relation between interested objects and aesthetic region for image cropping," *IEEE Trans. Multimedia*, early access, Oct. 9, 2020, doi: 10.1109/TMM.2020.3029882.
[24] P. Lu, H. Zhang, X. Peng, and X. Peng, "Aesthetic guided deep regression network for image cropping," *Signal Process., Image Commun.*, vol. 77, pp. 1–10, Sep. 2019.
[25] W. Lu, X. Xing, B. Cai, and X. Xu, "Listwise view ranking for image cropping," *IEEE Access*, vol. 7, pp. 91904–91911, 2019.
[26] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 497–506.
[27] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2408–2415.
[28] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato, "Sensation-based photo cropping," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 669–672.
[29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 2015, pp. 1–14.
[31] J. Sun and H. Ling, "Scale and object aware image thumbnailing," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 135–153, Sep. 2013.
[32] Y. Tu, L. Niu, W. Zhao, D. Cheng, and L. Zhang, "Image cropping with composition and saliency aware aesthetic score map," in *Proc. 34th AAAI Conf. Artif. Intell., 32nd Innov. Appl. Artif. Intell. Conf., 10th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, New York, NY, USA, Feb. 2020, pp. 12104–12111.
[33] A. Shashua and S. Ullman, "Structural saliency: The detection of globally salient structures using a locally connected network," in *Proc. 2nd Int. Conf. Comput. Vis.*, 1988, pp. 321–327.
[34] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2798–2805.
[35] L. Wang, X. Wang, T. Yamasaki, and K. Aizawa, "Aspect-ratio-preserving multi-patch image aesthetics score prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.
[36] W. Wang, Q. Lai, H. Fu, J. Shen, and H. Ling, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.
[37] W. Wang and J. Shen, "Deep cropping via attention box prediction and aesthetics assessment," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2186–2194.
[38] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1531–1544, Jul. 2019.
[39] Z. Wei, J. Zhang, X. Shen, Z. Lin, R. Mech, M. Hoai, and D. Samaras, "Good view hunting: Learning photo composition from dense view pairs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5437–5446.
[40] J. Yan, S. Lin, S. B. Kang, and X. Tang, "Learning the change for automatic image cropping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 971–978.
[41] J. Yan, S. Lin, S. B. Kang, and X. Tang, "Change-based image cropping with exclusion and compositional features," *Int. J. Comput. Vis.*, vol. 114, no. 1, pp. 74–87, 2015.
[42] S. Zagoruyko and N. Komodakis, "Wide residual networks," *ArXiv*, vol. abs/1605.07146, 2016.
[43] H. Zeng, L. Li, Z. Cao, and L. Zhang, "Grid anchor based image cropping: A new benchmark and an efficient model," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 15, 2019, doi: 10.1109/TPAMI.2020.3024207.
[44] H. Zeng, L. Li, Z. Cao, and L. Zhang, "Reliable and efficient image cropping: A grid anchor based approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5942–5950.
[45] L. Zhang, M. Song, Q. Zhao, X. Liu, J. Bu, and C. Chen, "Probabilistic graphlet transfer for photo cropping," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 802–815, Feb. 2013.



**CASPER L. CHRISTENSEN** was born in Copenhagen, Denmark, in 1994. He received the B.S. and M.S. degrees in computer science from the University of Copenhagen, in 2017 and 2020, respectively, spend a semester abroad at the National University of Singapore, in 2019.

From 2019 to 2020, he spent eight months at San Francisco Bay Area as a Research Engineer at Gracenote, where he focused on applying novel machine learning methods within image processing. He is currently working at Trustpilot, Copenhagen, Denmark, as a Machine Learning Engineer. His research interests include natural language processing, probabilistic models, and MLOps.

**ANEESH VARTAKAVI** received the B.E. degree in electronics and communication engineering from Manipal Institute of Technology, India, in 2012, and the master's degree in music technology from Georgia Institute of Technology, in 2014.

After graduating, he worked at Gracenote, where he focused on applied machine learning for computer vision, natural language processing, and music information retrieval applications. Here, he pioneered interactive machine learning tools that improve and extend the quality, creativity, and productivity of human editors. He has contributed to eight patents. He is interested in applying multi-modal machine learning to help people discover and connect with new entertainment content they love.