Good View Hunting: Learning Photo Composition from Dense View Pairs

Zijun Wei[1], Jianming Zhang[2], Xiaohui Shen[2], Zhe Lin[2], Radomír Měch[2],
Minh Hoai[1], Dimitris Samaras[1]
[1]Stony Brook University, [2]Adobe Research

Abstract

Finding views with good photo composition is a challenging task for machine learning methods. A key difficulty is the lack of well annotated large scale datasets. Most existing datasets only provide a limited number of annotations for good views, while ignoring the comparative nature of view selection. In this work, we present the first large scale Comparative Photo Composition dataset, which contains over one million comparative view pairs annotated using a cost-effective crowdsourcing workflow. We show that these comparative view annotations are essential for training a robust neural network model for composition. In addition, we propose a novel knowledge transfer framework to train a fast view proposal network, which runs at 75+ FPS and achieves state-of-the-art performance in image cropping and thumbnail generation tasks on three benchmark datasets. The superiority of our method is also demonstrated in a user study on a challenging experiment, where our method significantly outperforms the baseline methods in producing diversified well-composed views.

1. Introduction

Photo composition is one of the essential ingredients in the craft of photography. In this work, we aim to train models to find views with good composition (Fig. 1). Solving this problem can benefit many applications such as image cropping [44, 10], image thumbnailing [18], view recommendation [7, 38], and autonomous photo taking [4].

Finding good views in a scene is challenging for both machines [15, 10] and humans [20]. Recently, with the emergence of deep learning, many works attempt to train photo composition models directly from data [42, 18, 10, 31, 16]. However, the scarcity of large scale datasets limits further progress in this direction.

In this paper, we present the first large scale Comparative Photo Composition (CPC) dataset, from which we could harvest over 1 million comparative view pairs. Each view pair contains two views generated from the same image and their relative ranking crowdsourced under a cost-effective workflow. This type of comparative annotation is almost completely missing in existing datasets (see Table 1). Composition models trained on these view pairs perform significantly better than those trained on existing datasets.

Figure 1: Given an input image shown on the left, our View Proposal Network (VPN) can generate a set of diversified well-composed views (the top three view shown on the right) in less than 15 ms. The generated views can be used in a wide range of applications such as image cropping, image thumbnailing, image re-targeting and real-time view recommendation.

Another major contribution of this work is a novel knowledge transfer framework to train a real-time anchor-box-based view proposal model [17, 35]. Unlike object proposal networks, the label assignment for our view proposal model is quite challenging. First, assigning an anchor box to a ground truth view, based on an overlap metric (*e.g.* IOU), is fragile for training composition models; a slight adjustment of the view can often make a big difference in composition quality. Moreover, the annotation is not exhaustive: most of the anchor boxes will not be annotated. In contrast to the object detection scenario, we cannot assume that they are negative samples (see Fig. 2).

Our knowledge transfer framework obviates the above issues. In our teacher-student [21] framework, we first train a view evaluation model on the view pairs using a Siamese architecture [12]. Then we deploy this model as a teacher to score the candidate anchor boxes on a large variety of images. These teacher scores train the view proposal net

Dataset	Images	Views/Image	View Pairs
AVA [32]	250K	N/A	N/A
FastAT [18]	28064	≤ 3	N/A
FLMS [19]	500	≤ 10	N/A
ICDB [44]	1000	3	N/A
FCDB [9]	3359	10	33590
CPC (Ours)	10800	24	$>$ 1,000,000

Table 1: Comparison with the current image cropping/aesthetics datasets. Our CPC dataset contains much more annotated views for each image and provides over one million comparative view pairs.

as a student model to output the same anchor box score rankings. To train the student, we propose a Mean Pairwise Squared Error (MPSE) loss. We show that the proposed framework trains significantly better view proposal networks than the standard object proposal training protocol. An example result of our view proposal network is shown in Fig. 1.

Our view proposal model runs at over 75 FPS and achieves state-of-the-art performance on three benchmark datasets for image cropping and thumbnailing. To further test the robustness of our method, we designed a user study for a challenging experiment where each compared method must produce a diverse set of views for each test image. Experimental results show that our method outperforms other baseline methods.

In summary, our main contributions are:

- We present the Comparative Photo Composition dataset (CPC), the first large scale dataset with over 1M comparative view pairs.
- We propose a novel knowledge transfer framework for training a fast View Proposal Net, which obviates the label assignment issue by using a teacher network for dense view scoring.
- We evaluate extensively our View Proposal Net and demonstrate that it outperforms the state-of-the-art models in both accuracy and speed[1].

2. Related Work

Photo Composition Datasets. Comparative composition annotations are essential for training composition models [10]. However, most existing datasets for image cropping [18, 44] do not provide such comparative view annotations. A recent dataset [8] attempted to address this issue; annotators were asked to compare randomly sampled views from the same image, thus constructing a moderate-sized dataset. However, the data collection protocol is not every efficient, hence it's only possible to collect about 30K view pairs (see Tab. 1).

There are also a number of datasets for image aesthetics [32, 23, 14, 30], some of which provide image-level scores or attributes for photo composition [23, 32, 6]. However, models trained on such datasets may not perform well in distinguishing the composition quality of views from the same scene [10, 9].

Photo Composition Models. Many methods have been developed to find well composed views for various applications, such as image cropping [48, 19, 8, 44, 10], image thumbnail generation [18], view recommendation [7, 11, 38] and automatic triage that groups similar photos and decides which ones to keep [6]. Many of these methods rely on predefined rules or manually crafted features to evaluate the composition of an image crop [33, 19, 11, 44]. With the advent of deep learning, end-to-end models [22, 23, 27, 28] can be trained without explicitly modeling composition. Although they achieve state-of-the-art performance, these deep learning models must be executed in an inefficient sliding window fashion. Two more recent works [18, 42] address the speed issue by adopting fast object detection frameworks [13, 35] for image thumbnailing and image cropping. Compared with [18, 42], our work is more focused on the general composition ranking rather than a specific application. Thus, our method can be useful in a wider range of applications such as view recommendation and autonomous photo taking, where the content preservation constraint used in image cropping and thumbnailing does not apply.

Knowledge Transfer. Our framework for training a real-time view proposal network is inspired by the success of knowledge distillation that transfers the knowledge learned through an ensemble of models into a smaller single model [21]. Different from recent work [36, 2] that is geared towards model compression or domain transfer [41], we use a knowledge distillation framework to tackle the difficulty of label assignment (Fig. 2) when training an anchor-box-based view proposal network.

3. The CPC Dataset

We present the Comparative Photo Composition (CPC) dataset. Unlike the common data collection protocols for image cropping and thumbnailing, where only a few positive (good) views are drawn in each image, we aim to collect both positive and hard negative view annotations by relative ranking. We believe that such annotations are important in training a discriminative composition model.

To this end, we propose a cost-effective crowdsourcing workflow for data collection. With a budget of $3,000, we have collected 10,800 images. For each image we generated 24 views, where are ranked by 6 Amazon Mechanical Turk (AMT) workers. These rankings correspond to **more than**

[1]Our models and datasets collected in this work and supplementary material can be found at http://www.cs.stonybrook.edu/~cvl/projects/wei2018goods/VPN_CVPR2018s.html

1 Million effective comparisons[2] between different views from the same image.

3.1. Image Sources

To cover a wide range of scene categories, we randomly selected candidate images from multiple datasets: AVA [32], MS-COCO [25], AADB [23] and the Places dataset [49]. The collected images include not only professional photos but also everyday photos with varied image quality. To further ensure image diversity, we run the salient object subitizing model [46] on our collected images and sampled approximately equal number of images with 0, 1, 2 or 3+ dominant objects. Therefore, a substantial portion of images in our dataset contain two or more salient objects, in contrast to the previous datasets where most images contain only one dominant object [44, 29]. We then manually removed image collages and drawings. Duplicates and overlaps with test datasets that we used [9, 18, 19] were also removed using image hashing [45].

3.2. View Sampling

Directly asking AMT workers to draw crop windows can be a very inefficient way of collecting view samples. On the other hand, randomly sampled views are easy to generate but the view samples may be dominated by obviously poor compositions. Therefore, we used existing composition algorithms to generate a pool of candidate views that is more likely to include good ones.

Specifically, we pooled candidate views from multiple image re-composition and cropping algorithms [10, 23, 19] over a set of aspect ratios (1:1, 3:4, 4:3 and 16:9) and scales (0.5, 0.6, $\cdots$, 1.0). For each of our predefined aspect ratios, we randomly selected 4 views uniformly from the candidate pool. We also randomly sampled two additional views different from the selected ones to address the potential risk of view bias caused by the existing view-generating methods. In the end, we have $(4+2) \times 4 = 24$ candidate views for each image.

3.3. Two-Stage Annotation

Pairwise labeling the candidate views will lead to a quadratic number of annotations with respect to the candidate views. Thus, we designed a two-stage annotation protocol that greatly reduces the cost and simplifies the task for the AMT workers.

Stage One: Aspect-ratio-wise View Selection. The six views in each aspect ratio group are presented to an annotator at the same time with the request to select 2-5 good views. At the end of this stage, the annotators will select a total of 8-20 views from the four aspect ratio groups.

Stage Two: Overall Top View Selection. We show all the views selected in stage one and ask the annotator to select the best three views from them. We also randomly mix some unselected views as a test for quality control. Annotations that contradict the stage one selections are rejected.

Compared to the alternative annotation protocol that asks an annotator to directly rank 24 views in a single stage, our proposed two-stage work flow reduces the cognitive load by splitting a large task into simpler and smaller ones. Moreover, the quality control test in stage two can help reduce the noise of the dataset. We assign each image to 6 annotators.

The two-stage annotation protocol is general: it can also be applied to rank crops generated by professional photographers.

3.4. Ranking Pair Generation

After the annotation collection, the comparative view pairs are generated in two ways. We first generate ranking pairs of the same aspect ratio based on the averaged votes collected in Stage One, because the views of the same aspect ratios have been directly compared. Second, the views that are selected as best ones by more than 3 annotators in Stage Two are considered as the overall best views and will be paired with the remaining views that have been directly or indirectly compared with them. On average, we generate over 100 view pairs for each image, leading to more than 1 million comparative view pairs. Note that the above procedure may not be the optimal way to generate view pairs. We encourage users to explore various pair generation methods. The effectiveness of our CPC dataset is demonstrated in Sec 5.1.

4. Training View Proposal Networks

Inspired by the success of fast detection frameworks [26, 35, 34], we want to train a real-time anchor-box-based View Proposal Network, which takes an image as input and outputs scores corresponding to the list of predefined anchor boxes. Then we need to figure out if we can directly train such a VPN as in the object detection task. In object detection, an anchor box with a high IOU (Intersection Over Union) score with any annotation box will be labeled as positive; otherwise it is assumed to be a negative sample. However, such label assignment scheme is problematic in our task.

As shown in Fig. 2, there is a significant chance that an anchor box might be a good one even if it does not significantly overlap with any of the views annotated as good. On the other hand, anchor boxes that are close to a good view can still be a bad view as a subtle shift of the crop may change the composition drastically.

[2]For an image with $n = 24$ annotated views, there will be $n \times (n-1)/2 = 276$ view comparisons, leading to about $276 \times 10800 \approx 3$M view pairs. After pruning ambiguous view pairs, we have more than 1M view comparisons.

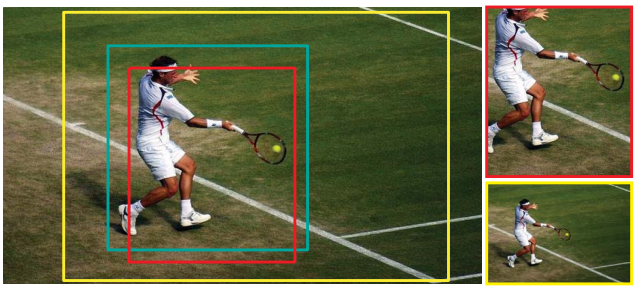

Figure 2: The label assignment problem. Given a good view annotation (cyan), there can be many bad views with high IOU (Intersection Over Union) with the annotation (*e.g.* the view in red), and many good views with low IOU (*e.g.* the view in yellow) with the annotation. Traditional IOU-based label assignment schemes for object detection will fail for our problem.

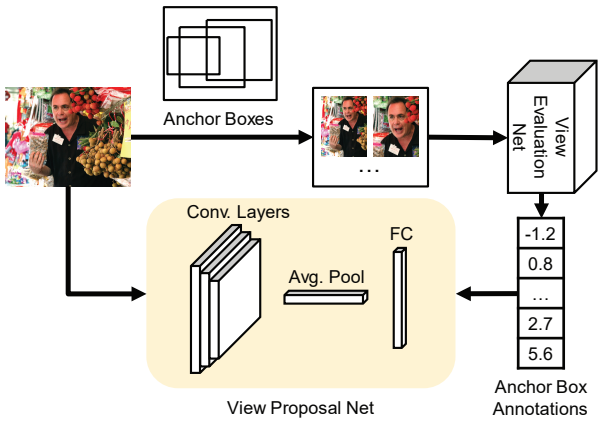

Figure 3: Overview of the knowledge transfer framework for training the View Proposal Network. See text for details.

To tackle this challenge, we propose a knowledge transfer framework inspired by [21] to train our View Proposal Net (VPN) as a student model under the supervision of a teacher model. The teacher model is a View Evaluation Net (VEN) that takes a view as input and predicts a composition score, and thus it is straightforward to train on our CPC dataset. To transfer the knowledge, we run the VEN on the anchor boxes for a given image, and then use the predicted scores to train the VPN using a novel Mean Pairwise Squared Error (MPSE) loss. The training framework is illustrated in Fig. 3 and will be detailed in the next section.

4.1. Knowledge Transfer by MPSE

Formally, a VPN is an anchor-box-based proposal network that takes an image as input and outputs a score list $\mathbf{q} = [q_1, \ldots, q_n]$ for a set of predefined anchor boxes $\mathbf{B} = [B_1, \ldots, B_n]$. A VEN is an evaluation network that predicts the composition score for a given image crop. The VEN is able to rank all the anchor boxes for an image and generate a score list $\mathbf{y} = [y_1, \ldots, y_n]$ by evaluating each anchor box. The problem of knowledge transfer is how to train the VPN to produce the same anchor box rankings as the VEN.

We first try the simple point-wise approach by regressing the Mean Squared Error (MSE) to minimize the mean absolute difference between $\mathbf{q}$ and $\mathbf{y}$. However, we empirically find that the training under the MSE loss is unstable. Inspired by the pair-wise and list-wise approaches for learning-to-rank models [3, 5], we propose a simple yet effective loss, called the Mean Pairwise Square Error (MPSE) loss, to jointly consider all the pair-wise ranking orders in a score list. The MPSE loss is defined as

$$l(\mathbf{y}, \mathbf{q}) = \frac{\sum_{i,j=1\ldots n, i\neq j}((y_i - y_j) - (q_i - q_j))^2}{n(n-1)/2}. \quad (1)$$

The MPSE loss can be manipulated algebraically to vectorized operations to be efficiently implemented [1]. We find that the MPSE loss makes the VPN training converge faster and provides better performance than MSE (see Sec. 5.2).

4.2. Implementation

View Proposal Net (VPN): Our VPN is adapted from two recent successful real-time object detection frameworks: SSD [26] and MultiBox [17]. The back-bone network is based on SSD (truncated after Conv9). On top of the backbone network we add a convolutional layer, an average pooling layer and a fully connected layer that outputs N scores corresponding to N predefined anchor boxes. Similar to [17], we predefine the set of anchor boxes by densely sliding over different scales and aspect ratios of an normalized image, resulting in a set of $N = 895$ predefined anchor boxes (similar to [17]). For training and testing, we resize the image to 320×320 regardless of its original size.

View Evaluation Net (VEN): Following recent works [23, 10], given the pair-wise annotations in the CPC dataset, we train the View Evaluation Net using the Siamese architecture [12]. The Siamese architecture is composed of two weight-shared VENs, each of which outputs a score for one of the two input image pairs, say I_i and I_j. Assuming that I_i is preferred more than I_j, the following loss is used to train the Siamese network:

$$l(I_i, I_j) = max\{0, 1 + f(I_j) - f(I_i)\}, \quad (2)$$

where $f(I_i)$ and $f(I_j)$ denote the outputs of the Siamese network.

The VEN in our work is based on VGG16 [37] (truncated after the last max pooling layer) with two new Fully Connected (FC) layers and a new output layer. We reduce the channels of the FC layers to 1024 and 512 respectively since the our model only outputs one ranking score instead of the probability distribution over 1000 classes.

Training Details. We initialized the VEN by the weights of the VGG16 model trained on ImageNet and trained for 60 epochs on image pairs from the CPC dataset with a starting learning rate of 0.001 that decays by 0.1 every 20 epochs using Stochastic Gradient Descent (SGD) with momentum of 0.9. Early stopping was applied based on results on a validation set from XPView.

To train the VPN, theoretically we can use an unlimited number of images. In our implementation, we use the images in our CPC dataset plus 40,000 images from the AADB [23] dataset. Adding more images does not seem to further improve the result. The VPN is initialized by the SSD model [26] for object detection and is trained using the same hyperparameters as the VEN.

5. Dataset and Method Analyses

In this section we analyze the effectiveness of our CPC dataset (Sec. 5.1). We do so by training VEN baselines on the existing datasets and comparing their performance on a test set annotated by experts. We further evaluate the knowledge transfer framework by showing how well the VPN approximates the ranking produced by VEN (Sec. 5.2).

5.1. Validating the CPC Dataset

The Expert View Dataset (XPView). To evaluate the ranking performance of models trained on the various datasets, we collected a new dataset consisting of 992 images with dense compositions, annotated by three experts (graduate students in visual-arts) instead of AMT workers. We generated the candidate views as described in Sec. 3.2 but with more diverse aspect ratios. For each image we pooled 24 candidate crops from 8 aspect ratios. Each view was annotated with one of the three labels: good, fair, or poor. We additionally asked the annotators to draw good compositions that should have been included in the set of candidate views. As analyzed in supplementary material, our inter- and intra- annotator consistency analysis shows significant agreement. Keep only unanimously labeled ones, we get 1830 good, 4915 fair and 1875 bad views. A total of 18229 comparisons of views from the same images are generated. We randomly select a subset of 200 images from the XPView dataset as the test set.

Baselines and setup. We consider the following VEN baselines trained on 1) the remaining 792 images of XPView, 2) the AVA dataset [32], with the classification training strategy used in [42, 27, 28, 32], 3) the FCDB dataset with the ranking strategy used in [9], 4) unlabeled images for unsupervised learning following [9]. For our CPC dataset, we train two VENs: one based on the ranking loss described in Eq. 2 and the other based on the softmax classification loss. For the classification loss, views selected by 3 or more annotators out of 6 in Stage Two (Sec. 3.3) are considered positive samples and those selected by less than 3 annotators in Stage One are considered negative.

To control for the model capacity and overfitting issues on different datasets, for each baseline VEN model we train three VEN variants with 41M, 27M and 14M parameters and report the best performance The architectures of the three variants are described in the supplementary material. The major differences are the sizes of the FC layers.

For evaluation, we report the swap error (SW) [10], which is the ratio of erroneously ranked (swapped) pairs over all valid pairs on the XPView test set.

Results. The results are presented in Tab. 2. The VEN model with the ranking loss (Eq. 2) trained on CPC achieves the best performance (0.22 swap error). Training the VEN under the classification framework yields slightly worse performance. The performance gap between models trained on our CPC dataset and the one trained on the FCDB dataset reveals that the density of annotations and the size of the dataset are more critical to the performance than the choice of loss. Note that training VENs on the AVA dataset with classification loss yields the worst result. We have tried a ranking loss as well but did not observe improvement (0.45 swap error). This is consistent with the findings in [10] that models trained on aesthetic relations derived from distinct images do not necessarily perform well in ranking views from the same image. Even though models trained on the 792 images from XPView achieve better performance than previous existing datasets, the cost of drawing views and the demand on expert annotators makes the dataset difficult to scale. We have also tried to combine the CPC, FCDB [9], and XPView datasets to train the VENs but did not observe better performance. Further analysis is presented in supplementary material.

Training	Loss	#images	Swap error ↓
AVA[32]	Classification	250K	0.37
FCDB[9]	Ranking	3359	0.32
Unlabeled[10]	Ranking	-	0.32
XPView*	Ranking	792	0.28
CPC*	Classification	10.5K	0.24
CPC*	Ranking	10.5K	**0.22**

Table 2: Performance of the View Evaluation Nets trained on different datasets: models trained on the CPC dataset perform best; the VEN trained with a ranking loss moderately outperforms the one trained with a classification loss. All of the models are evaluated on the same test set from XPView. * indicates datasets collected in our paper.

In the rest of the paper, if not otherwise specified, the VEN refers to the best model that is trained on the whole CPC dataset with ranking loss (the last row of Tab. 2).

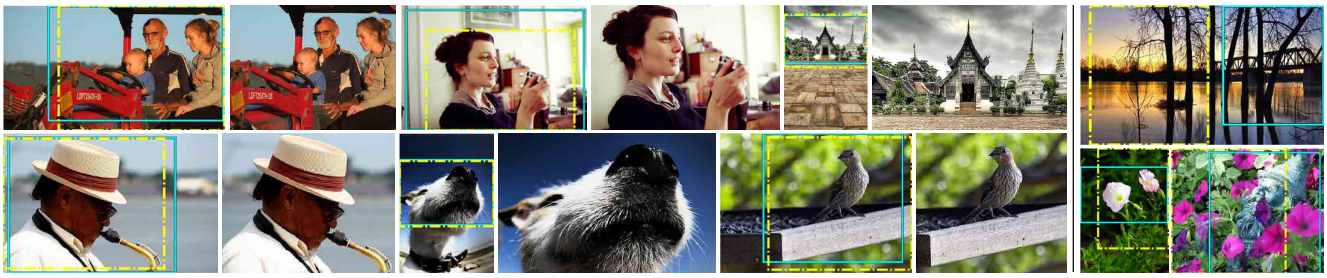

Figure 4: Qualitative Results of the View Proposal Net. (a) shows some results of the VPN in image cropping [9] (top row) and image thumbnailing [18] (bottom row). The ground truth annotations are in cyan and the predicted views are in yellow. (b) provides a few examples receiving poor IoU metric scores on the ground truth annotations. As we can see, the top and bottom left predicted views are still well composed even though they do not match well with the ground truth views.

5.2. Analysis on Knowledge Transfer

We investigate the influence of the training loss and the size of the transfer sets (the set of images that are labeled by the teacher net for training the student net [21]) on the VPN trained via the knowledge transfer framework. We compare the swap error between the VPN and the teacher model VEN over the predefined 895 anchor boxes.

We evaluate VPNs trained with the following configurations: 1) using only the CPC dataset images with the Mean Squared Error (MSE) loss, 2) using the same set of images with the Mean Pairwise Squared Error (MPSE) loss, and 3) using an additional set of 40K images from AADB [23] with the MPSE loss (CPC+AADB).

As shown in Tab. 3, compared to the MSE loss, the MPSE loss slightly improves performance. The MPSE loss also leads to faster convergence as we observed during training. Increasing the size of the transfer set also considerably reduces the swap error.

Transfer set	Loss	VPN-VEN swap error ↓
CPC	MSE	0.1601
CPC	MPSE	0.1554
CPC+AADB	MPSE	0.1312

Table 3: VPN-VEN Comparisons: Overall the VPN trained with the knowledge transfer framework performs very similar to the VEN. The VPN trained with the MPSE loss with additional unlabeled data has the closest performance to the VEN (see CPC+AADB).

6. Experiments

We evaluate the View Proposal Net and the View Evaluation Net by quantitatively comparing them to state-of-the-art models on benchmark datasets. We also compare the speed of the VPN on a GPU to existing real-time models designed for particular tasks in Tab. 5.

6.1. Quantitative Evaluation

We quantitatively evaluate our proposed VPN and VEN on three benchmark datasets for the image cropping and image thumbnailing tasks and compare them with state-of-the-art models on these tasks. In addition to the VPN trained under our knowledge transfer framework, we investigate an alternative for training the VPN: following object detection works [34, 26], we assign the positive and negative labels defined in Sec. 5.1 to each anchor box and train the VPN with a binary classification loss directly without knowledge transfer. We denote the VPN trained this way as **VPN-direct**. More alternatives of training the VPN are explored in the supplementary material.

6.1.1 Image Cropping

We evaluate the performance of our VPN and VEN for image cropping tasks.

Dataset. We evaluate on two datasets: (1) FLMS [19] containing 500 images that have 10 sets of cropping ground truth by 10 different expert annotators from Amazon Mechanical Turk and (2) the FCDB [9] containing 348 testing images[3].

Setup and metrics. To show the generalization ability of our models, we test the VPN and VEN trained with CPC on the two datasets with no additional training. For image cropping tasks, there is a content preservation prior that the cropped view should cover most of the major content. Therefore, we perform a simple post-processing by discarding small views. We cross-validated the small size threshold on the training set of FCDB. We report the average overlap ratio (IoU) and average boundary displacement error (Disp.) used in previous works [10, 44, 42] as performance evaluation metrics:

[3] At the time of our downloading, only 343 images were accessible

FCDB [9]							
	VEN	VPN	VPN-direct	VFN [10]	MNA-CNN [31]	AesRankNet [23]	RankSVM [9]
IoU ↑	**0.7349**	0.7109	0.6429	0.6842	0.5042	0.4843	0.6020
Disp.↓	**0.072**	0.073	0.092	0.084	0.136	0.140	0.106

FLMS [19]							
	VEN	VPN	VPN-direct	Wang *et al* [42]	Suh *et al* [39]	Fang *et al* [19]	Chen *et al* [8]
IoU ↑	**0.8365**	0.8352	0.7875	0.8100	0.7200	0.74	0.64
Disp. ↓	**0.041**	0.044	0.051	0.057	0.063	-	0.075

Table 4: Image cropping results on the FCDB [9] and FLMS [19] dataset. Both our VPN and VEN models outperform state-of-the-art models in terms of IoU and Displacement Error on the two datasets substantially. The diminished performance of VPN-direct (Sec. 6.1) reveals the effectiveness of our knowledge transfer framework (VPN).

$$IoU = \frac{Area^{gt} \cap Area^{pred}}{Area^{gt} \cup Area^{pred}}, \quad (3)$$

$$Disp.Error = \sum_{k \in \{boundaries\}} ||B_k^{gt} - B_k^{pred}||/4, \quad (4)$$

where $Area^{gt}$ is the area of the ground truth crop view and $Area^{pred}$ is the area of the predicted view; B_k is the *normalized* boundary coordinate.

Results. The results are summarized in Tab. 4. The baseline model scores are from [42] for the FLMS dataset and [10] for the FCDB dataset, respectively. Both our VPN and VEN models outperform state-of-the-art models on the two datasets substantially. Notably, the lower performance of VPN-direct reveals the effectiveness of our knowledge transfer framework.

6.1.2 Image Thumbnailing

We evaluate the performance of VPN and VEN on the image thumbnailing task.

Dataset. We evaluate on the test set of the FAST-AT dataset [18] that contains 3910 images with 2 to 3 ground truth thumbnails of different aspect ratios.

Setup and metrics. To match the aspect ratio of the target thumbnail, we pick the top view from the subset of 895 candidates that have an aspect ratio mismatch less than 0.01 as in [18] and then shrink it to fit the aspect ratio of the target thumbnail to avoid spilling out of the image border. Following [18], we use the IoU metric from Sec. 6.1.1, and the offset error, which is the distance between the centers of the ground truth bounding box and the predicted bounding box.

Results. As shown in Tab. 6. Both the VPN and the VEN outperform the state of the art models in terms of IoU and offset metrics. The baselines are from [18].

6.2. Qualitative Visualization

Some qualitative results are shown in Fig. 4. Moreover, since the VPN is general, it also works on images with extreme aspect ratios such as panoramas, as in Fig. 5, even though it was not trained for them.

Figure 5: View suggestion for panoramic images. For each panoramic image, we define an aspect ratio and our VPN returns the top 3 ranked views after a non-maximum-suppression threshold of 0.2. Our VPN produces satisfactory results even though it is not trained on any panoramic images.

7. User Study

The subjective nature of image composition requires validating composition models through human surveys. We demonstrate the accuracy and diversity of our VPN and VEN models through a challenging user study experiment: we randomly sample 200 images from XPView. For each image, we pick the top 5 views generated by each model. We enforce the diversity of outputs of each model by applying a non-maximum suppression with a 0.6 IoU threshold.

Figure 6: User study results. nDCG (normalized Discounted Cumulative Gain, ↑), Average Good View ratio (↑), Average Bad View ratio (↓) of different models. Our VEN and VPN outperform all baselines by a large margin.

Model	Type	Speed
View Evaluation Net	Sliding Window	0.2 FPS
Wang *el al* [42]	Two Stage	5 FPS
Fast-AT [18]	Detection [13]	10 FPS
View Proposal Net	Proposal	75+ FPS

Table 5: Speed comparison (on GPU). The VPN is as accurate and 150× faster than the VEN. The VPN is significantly more accurate and 7-15× faster than state-of-the-art models [42, 18] that are designed for particular view suggestion tasks.

Fast-AT Thumbnail Dataset [18]				
Method	VEN	VPN	Fast-AT [18]	SOAT [40]
IoU↑	0.7021	**0.7079**	0.6800	0.5200
offset ↓	49.9	**48.2**	55.0	80.5

Table 6: Thumbnailing results on the Fast-AT dataset [18]. Both VPN and VEN outperform the baseline models.

The top anchors are then shrunk to the closest normal aspect ratio(e.g., $1:1, 3:4$ ...) for VPN. We mix and pool the outputs from different models together and ask 6 graduate students majoring in art (different from the annotators of the XPView dataset) to exhaustively select the good and bad views. In addition to the average good view rate and average bad view rate, we also report the nDCG (normalized Discounted Cumulative Gain, [43]) that measures the consistency between the model's rankings and user preferences:

$$nDCG = \frac{DCG}{Ideal(DCG)}, \tag{5}$$

where $DCG = \sum_{i=1}^{k} \frac{rel_i}{log_2(i+1)}$ in which rel_i is the relevance score of the i-th view (1 for good, -1 for bad and 0 for fair). The $Ideal(DCG)$ is the DCG score when all k ($k = 5$ in our case) scores are 1.

We carefully implemented the following baselines and validated them on the benchmark datasets that the original implementations reported on.

Sal+Face. We reimplemented [19] using a state-of-the-art saliency method [47] and added a face detector. Our implemented Sal+Face baseline achieves an IoU score of 0.82 on FLMS [19], which is even higher than the current state-of-the-art model [42].

VPN-direct. Described in Sec. 6.1, the VPN-direct model is essentially an alternative to the Fast-AT model [18] and slightly outperforms it as shown in Tab. 6.

VFN*. we implemented the View Finding Network [10], a state-of-the-art model based on aesthetics. For fair comparison to our model, we replaced the AlexNet [24] backbone in [10] with VGG [37]. Our implementation of VFN, denoted as **VFN***, achieves a comparable IoU score (0.68) on FCDB-test [10].

AVA-Cls. We trained a VEN with classification loss using the AVA dataset following the Aesthetic Assessment network in [42].

Results are shown in Fig. 6. The proposed VEN and VPN outperform the baselines by a large margin. Notably, VPN, while trained under the supervision of VEN, performs even better than the VEN. We posit that the reason is that the VPN takes the whole image as input and considers global information during learning.

8. Conclusions

We have presented the CPC dataset, a large scale photo composition dataset with more than 1 million view pairs and proposed a knowledge transfer framework to train the real-time View Proposal Net. We have shown that the View Proposal Net trained on the CPC dataset using knowledge transfer achieves state-of-the-art performance and runs at 75+ FPS. In the future work we plan to explore the effect of professional versus AMT annotations in the quality of the predictions.

Acknowledgment. This project was partially supported by a gift from Adobe, NSF CNS-1718014, the Partner University Fund, and the SUNY2020 Infrastructure Transportation Security Center.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. 4

[2] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1613–1622. JMLR.org, 2015. 2

[3] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96. ACM, 2005. 4

[4] Z. Byers, M. Dixon, K. Goodier, C. M. Grimm, and W. D. Smart. An autonomous robot photographer. In *Proceedings IEEE nternational Conference on Intelligent Robots and Systems*, 2003. 1

[5] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136. ACM, 2007. 4

[6] H. Chang, F. Yu, J. Wang, D. Ashley, and A. Finkelstein. Automatic triage for a photo series. *ACM Transactions on Graphics (TOG)*, 35(4):148, 2016. 2

[7] Y.-Y. Chang and H.-T. Chen. Finding good composition in panoramic scenes. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2225–2231. IEEE, 2009. 1, 2

[8] J. Chen, G. Bai, S. Liang, and Z. Li. Automatic image cropping: A computational complexity study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 507–515, 2016. 2, 7

[9] 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 *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 226–234. IEEE, 2017. 2, 3, 5, 6, 7

[10] Y.-L. Chen, J. Klopp, M. Sun, S.-Y. Chien, and K.-L. Ma. Learning to compose with professional photographs on the web. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 37–45. ACM, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[11] B. Cheng, B. Ni, S. Yan, and Q. Tian. Learning to photograph. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 291–300. ACM, 2010. 2

[12] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 539–546. IEEE, 2005. 1, 4

[13] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016. 2, 8

[14] R. Datta, D. Joshi, J. Li, and J. Wang. Studying aesthetics in photographic images using a computational approach. *Computer Vision–ECCV 2006*, pages 288–301, 2006. 2

[15] Y. Deng, C. C. Loy, and X. Tang. Aesthetic-driven image enhancement by adversarial learning. *arXiv preprint arXiv:1707.05251*, 2017. 1

[16] Y. Deng, C. C. Loy, and X. Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4):80–106, 2017. 1

[17] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2147–2154, 2014. 1, 4

[18] S. A. Esmaeili, B. Singh, and L. S. Davis. Fast-at: Fast automatic thumbnail generation using deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4178–4186. IEEE, 2017. 1, 2, 3, 6, 7, 8

[19] C. Fang, Z. Lin, R. Mech, and X. Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *ACM Multimedia*, 2014. 2, 3, 6, 7, 8

[20] K. A. Hamblen. Approaches to aesthetics in art education: A critical theory perspective. *Studies in Art Education*, 29(2):81–90, 1988. 1

[21] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. 1, 2, 4, 6

[22] L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, 2014. 2

[23] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*, pages 662–679. Springer, 2016. 2, 3, 4, 5, 6, 7

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 8

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 3

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016. 3, 4, 5, 6

[27] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 457–466. ACM, 2014. 2, 5

[28] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 990–998, 2015. 2, 5

[29] W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2206–2213. IEEE, 2011. 3

[30] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. *Computer Vision–ECCV 2008*, pages 386–399, 2008. 2

[31] L. Mai, H. Jin, and F. Liu. Composition-preserving deep photo aesthetics assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 497–506, 2016. 1, 7

[32] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2408–2415. IEEE, 2012. 2, 3, 5

[33] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato. Sensation-based photo cropping. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 669–672. ACM, 2009. 2

[34] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 3, 6

[35] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 2, 3

[36] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4, 8

[38] H.-H. Su, T.-W. Chen, C.-C. Kao, W. H. Hsu, and S.-Y. Chien. Preference-aware view recommendation system for scenic photos based on bag-of-aesthetics-preserving features. *IEEE Transactions on Multimedia*, 14(3):833–843, 2012. 1, 2

[39] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs. Automatic thumbnail cropping and its effectiveness. In *Proceedings of the 16th annual ACM symposium on User interface software and technology*, pages 95–104. ACM, 2003. 7

[40] J. Sun and H. Ling. Scale and object aware image thumbnailing. *International journal of computer vision*, 104(2):135–153, 2013. 8

[41] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015. 2

[42] W. Wang and J. Shen. Deep cropping via attention box prediction and aesthetics assessment. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2, 5, 6, 7, 8

[43] Y. Wang, L. Wang, Y. Li, D. He, W. Chen, and T.-Y. Liu. A theoretical analysis of ndcg ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013)*, 2013. 8

[44] J. Yan, S. Lin, S. Bing Kang, and X. Tang. Learning the change for automatic image cropping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 971–978, 2013. 1, 2, 3, 6

[45] C. Zauner. Implementation and benchmarking of perceptual image hash functions. 2010. 3

[46] J. Zhang, S. Ma, M. Sameki, S. Sclaroff, M. Betke, Z. Lin, X. Shen, B. Price, and R. Mech. Salient object subitizing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4045–4054, 2015. 3

[47] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech. Minimum barrier salient object detection at 80 fps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1404–1412, 2015. 8

[48] L. Zhang, M. Song, Q. Zhao, X. Liu, J. Bu, and C. Chen. Probabilistic graphlet transfer for photo cropping. *IEEE Transactions on Image Processing*, 22(2):802–815, 2013. 2

[49] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014. 3