Understanding Compositional Structures in Art Historical Images using Pose and Gaze Priors

Towards Scene Understanding in Digital Art History

Prathmesh Madhu[1[0000−0003−2707−415X]⋆], Tilman Marquart[1⋆], Ronak Kosti[1[0000−0003−2453−7876]], Peter Bell[2[0000−0003−4415−7408]], Andreas Maier[1[0000−0002−9550−5284]], and Vincent Christlein[1[0000−0003−0455−3799]]

[1] Pattern Recognition Lab, https://lme.tf.fau.de/
[2] Institute for Art History, https://www.kunstgeschichte.phil.fau.de/
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
Email: prathmesh.madhu@fau.de

Abstract. Image compositions as a tool for analysis of artworks is of extreme significance for art historians. These compositions are useful in analyzing the interactions in an image to study artists and their artworks. Max Imdahl in his work called *Ikonik*, along with other prominent art historians of the 20th century, underlined the aesthetic and semantic importance of the structural composition of an image. Understanding underlying compositional structures within images is challenging and a time consuming task. Generating these structures automatically using computer vision techniques (1) can help art historians towards their sophisticated analysis by saving lot of time; providing an overview and access to huge image repositories and (2) also provide an important step towards an understanding of man made imagery by machines. In this work, we attempt to automate this process using the existing state of the art machine learning techniques, without involving any form of training. Our approach, inspired by Max Imdahl's pioneering work, focuses on two central themes of image composition: (a) detection of action regions and action lines of the artwork; and (b) pose-based segmentation of foreground and background. Currently, our approach works for artworks comprising of protagonists (persons) in an image. In order to validate our approach qualitatively and quantitatively, we conduct a user study involving experts and non-experts. The outcome of the study highly correlates with our approach and also demonstrates its domain-agnostic capability. We have open-sourced the code: https://github.com/image-compostion-canvas-group/image-compostion-canvas

Keywords: compositional structures, art history, computer vision

1 Introduction

Understanding narratives present in an artwork has always been a challenge and is strongly researched since the late 19th century [29]. A high-level interpretation

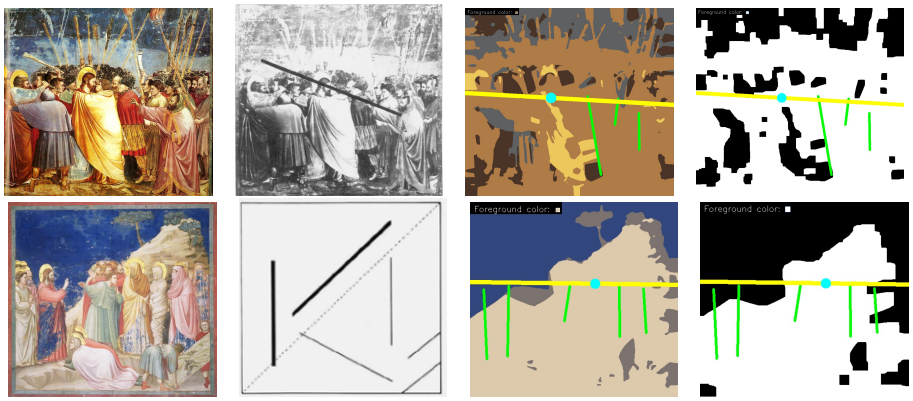
⋆ equal contribution

for the given scene is more ambiguous than a low-level interpretation [3] meaning it is easy to recognize and interpret small objects and characters in the image rather than presenting a high level abstract description of any scene. In the recent times of deep learning, numerous highly successful supervised techniques have been presented for various applications like object detection [11], person identification [22], segmentation [26], retrieval [16] etc. However, these methods have some major drawbacks. First, even though they perform extremely well on the benchmark datasets, they fail to generalize across other real-world datasets. The reason being that it is extremely difficult to capture the real world distribution and it's complexity within a single large dataset. Second, these deep models are very sensitive and error-prone compared to humans who naturally adapt to the visual context as stated in [30].

One solution suggested in the DeepNets [30] paper is the idea of using the compositionality of the images to generate models that can be trained on finite datasets, and then can be generalized across unseen datasets. The assumption is that the structures within an image are composed of various substructures following a grammatical set of rules, which can be learned from a finitely annotated datasets. These compositional models can be used to reason and diagnose the system, extrapolate beyond data and answer varied questions based on learned knowledge structure. On a parallel front, from a theoretical perspective, Bienenstock *et al.* [3] suggested that the hierarchical compositional structure of natural visual scenes can be reduced down to a collection of drastic combinatorial restrictions; meaning one can break down any scene into 2D projections of objects present in the scene. From this perspective, it can be understood that *composition* is one of the fundamental aspects of human cognition.

Motivated by this understanding of compositional structures within images, in this paper, we investigate the problem of determining and analyzing the composition of various scenes in art history. Specifically motivating is the work of Max Imdahl called *Ikonik* [12], where he formulates a methodology using an artwork's structure to determine its significance. His work is considered as a model of image analysis, which can be considered as a complement to Panofsky's iconology [14]. He argues that visual cues or the "visual seeing" overpowers biblical or textural references by giving references to artworks by Giotto, cf. Fig. 1. Giving the examples of *Ascent to Cavalry* and *The kiss of Judas* (Fig. 1a), he explains how the structural relation between foreground and background, the distribution of colors and dynamism between the characters can help in understanding artists and their artworks.

Motivated by his work, we propose an non-supervised computational approach to find compositional structures in an image. Our algorithm is driven by the human perception that posture and gaze of the main protagonists help to identify the region of interest (action regions) of any scene (Figures 1c and 1d). We enable pre-trained OpenPose [4] framework to detect pose-keypoints of protagonists and further exploit them to get a pseudo gaze-estimate without using any existing state of the art gaze detection methods which would have required largely annotated data. We combine all these compositional elements and plot



(a) **Original Art** (b) **Imdahl's Take** (c) **ICC** (d) **FG/BG Sep.**

Fig. 1: (a) Giotto's Original Paintings, (top) "The kiss of Judas", (bottom) "Raising of Lazarus"; (b) Imdahl's compositional analysis using structural elements for Giotto's works; (c) Their corresponding Image Composition Canvas (ICC); (d) Binarized ICC to highlight the foreground/background separation.

it on an empty canvas which gives an estimate of the representation of the underlying structure and composition within a given scene, which we call *Image Composition Canvas* (ICC). These compositional elements of ICC can later be used as visual features to cross-retrieve images from various datasets. Our method generates *ICCs* that contains two important aspects of image composition: (1) constructing the global action lines and action regions (2) pose-based semantic separation of foreground and background. The detected pose-keypoints [4] not only assist in foreground/background extraction, but also help in generating the global action lines and action regions. The result is a "visual seeing" which concentrates strictly on the compositional structure of an image and is, thereby, complementary to human perception usually driven by the semantics of the narration.

2 Related Work

Semblances in Art History Max Imdahl showed that discovering similar painting structures or compositional elements is an important aspect in the analysis of artworks [13]. However, detecting the relevant features for the same is a relatively novel task for computer vision. Previous works mainly focus on features for image retrieval, for example, in *SemArt* [10] the paintings are aligned with its attributes such as title, author, type, time and captions by a neural network which learns a common embedding space between them. In another example, these attributes (also called contextual information) are trained jointly in a multi-task manner to achieve a context-aware embedding [9], thereby improving the retrieval performance through a knowledge graph that is created using

attributes as prior information. In *Artpedia* [27], the authors are able to align visual and textual content of the artwork without paired supervision. Jenicek *et al.* come closest to our approach in that they link paintings through a two step retrieval process, by finding similarity-of-pose across different motives [16,2]. They also explore the underlying visual correlation between artworks.

Human Pose and Gaze Estimation Human posture is a strong marker for compositional element of the image. It helps to understand the visual relationships depicted in the image. For the semantic understanding of a scene, it becomes important to analyze the poses of all persons in the given scene. Multi-Person Pose Estimation (MPPE) is a challenging task due to occlusions, varying interactions between different people and objects. MPPE could be divided into *top-down* (detect person → predict pose point for each) [24] and *bottom-up* (detect body joints for all persons → join the points to detect poses) [4] approaches. The current state-of-art method uses *PoseRefiner* [8] as a post-processing technique to refine the pose estimation to get the best results on the MPII dataset [1]. Empirical tests showed OpenPose [4] performed well on our art-historical dataset, so we use it for all our experiments.

Importance of human gaze is evident in the related work of detecting human gazes for automatic driver analysis and understanding attention spans. Gaze360 [17] works on a diverse set of environments, and it is well suited for video or multi-frame inputs depicting temporal relationship which is absent in our data. Another method, *Where are they looking?* [25], uses head location and the image to predict the gaze direction in the image. This method requires gaze annotations (eye locations and gaze directions), however, since our data is a specialized collection of images chosen to study the compositions in a non-supervised manner, this approach could not be applied for our work. Hence, we derive gaze-bisection-vectors from pose keypoints for gaze priors.

Foreground/Background Separation. Detecting foreground and background has been one of the important pre-processing task for computer vision before the deep learning techniques came into practice. Separating foreground from background helps to focus on where the object/region of interest lies at. For example, in image-reconstruction-guided landmark detection, the algorithms are quite often assisted by fine-grained separation of background and foreground techniques [7,21,15]. Specifically, [7] factorizes the reconstruction task to achieve better landmark-detection for foreground objects by simultaneously improving the background rendering.

Exploiting this prior information about the separation of foreground and background is very useful even for existing state of the art methods. For example, in single stage object detectors the high number of candidate locations ($\approx$100 k) creates huge bias towards background classes [20]. Dynamically weighted focal loss [18] mitigates this imbalance by heavily penalizing the learning model for incorrect foreground objects, thereby forcing the model to attend more to the foreground objects. Our approach inherently uses pose estimates of the protago-

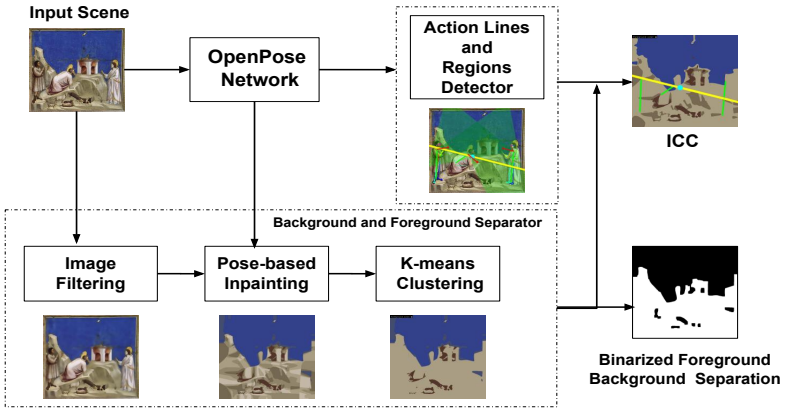


Fig. 2: Generating Image Composition Canvas (*ICC*): Proposed *ICC* pipeline along with foreground/background separation from images. First row details in Fig. 3

nists. We use an enhanced k-means clustering with the pose estimates to achieve foreground background separation.

Our algorithm is motivated by how a human understands the composition of scenes in paintings. In brief, our main contributions include: (1) Generation of Image Composition Canvas (ICC) which showcases the global and local action lines and action regions, (2) Semantic separation of foreground/background, (3) Generizability of the approach to images leading a step towards domain-agnostic modeling.

3 Methodology

We propose a non-supervised approach to generate *ICC* to assist our understanding of underlying structures in art historical images. Our approach uses a pre-trained OpenPose network [4], image processing techniques and a modified k-means clustering method. The pipeline of the proposed algorithm is shown in Fig. 2 (Fig. 3 shows the visual counterpart). Our method consists of two main branches: (1) a detector for action lines and action regions (Sec. 3.2) and (2) the foreground/background separator (Sec. 3.3). We use the estimated poses to detect pose triangles and propose a simple technique which we call gaze cones to obtain gaze directions estimates without involving any training or fine-tuning. We also use the detected keypoints for foreground/background separation (Sec. 3.3). We combine this information and draw a final ICC that estimates the image composition of the given scene under study (Sec. 3.4). We evaluate the proposed method by performing a user-study where we ask domain experts and non-experts to annotate action lines (global and local) and action regions, details of which are mentioned in Sec. 4.1

3.1 Data Description

The dataset contains 20 images of fresco paintings from the $13th$ century. Figures 1a, 4b and 5a show some examples of Giotto's famous fresco paintings. The paintings come from the Scrovegni chapel also known as arena chapel in Padua, Italy and were made by the painter Giotto di Bondone (considered the decisive pioneer of the Italian Renaissance). Most of these images have been analyzed by Max Imdahl in his book *Ikonik Arenafresken* [13]. We also use 10 random images containing people from COCO test dataset, 5 "Annunciation of Our Lady" and 5 "Baptism of Christ" images for evaluation and to show that our approach is domain-agnostic.

3.2 Action Lines and Action Regions

Global *Action line* (AL) is the line that passes through the main *activity* in the scene, which normally is also aligned with the central protagonists. Local *Action Line* or *Pose Line* (PL) represent the poses of the protagonists, abstracted in the form of a line. *Action region*(s) (AR) is(are) the main *region(s) of interest*, more often than not it is the region where gazes of all the characters theoretically meet, or the focus of their attention.

In order to detect ALs and PLs, we first pass the image through pre-trained OpenPose network. The output is a 25-dimensional keypoint vector shown in Fig. 3a. It is easy for OpenPose to detect all the keypoints for a fully visible human body. However, sometimes the full body is not visible, as can be seen for the character in the middle in Fig. 3c. We correct such poses using a pose corrector (Sec. 3.2a). We then detect the gazes of these characters using three major keypoints detected by OpenPose (Sec. 3.2b). The output of the gaze estimator gives us the gaze directions in the form of cones as visualized in Fig. 3d. The intersecting region of all the gaze cones is considered as the region of interest, where the main ARs would be present. Based on the intersection pattern of the cones, there can be multiple ARs. Also, a line representing the direction and intersection of all the views is considered to be the global AL. The slope of this line is calculated by combining all the gaze slopes (Sec. 3.2c).

(a) Pose Correction For pose estimation, we tried a few SOTA methods, including associative embedding for joint detection and grouping [23], however OpenPose [4] gave the best results. Bottom-up approaches have the overhead of detecting persons in art historical images [22], which is an entirely different problem altogether. Hence, we use OpenPose [4] pre-trained on the Human Foot keypoint dataset [4] in combination with the 2017 version of the COCO [19] dataset. The network takes an entire image as input and returns fully connected body poses for all humans detected in an image as output.

Often OpenPose would predict incorrect keypoints, which were difficult to interpret, motivating us to come up with the pose-triangles approach which is robust for person interpretation irrespective of the occluded keypoints. With the detected keypoints, we create a pose triangle. For the triangle corners, we split

(a) (b) (c) (d) (e)

Fig. 3: (a) 25 pose-keypoints used by OpenPose [4]; (b) Pose triangle using groups of keypoints; (c) OpenPose generated keypoints, thicker lines have higher confidence. Generated *Pose Triangles* (Sec. 3.2a) in blue and the final abstracted result in bold green line; (d) Gaze-bisection-vectors (Sec. 3.2c) in bold red, cones in light transparent green, the intersection of all cones in purple; (e) Overlapping intersections in purple

the 25 keypoints into three body regions as shown in Fig. 3b. The region from shoulders towards the head forms the top triangle corner (brown trapezoid). The left (green trapezoid) and the right leg (yellow trapezoid) regions form the left and the right corners of the pose triangle, respectively. The specific keypoints associated with these regions are $\text{top}_C \leftarrow$ [0,1,2,5,15,16,17,18], $\text{left}_C \leftarrow$ [9,10,11,22,23,24], $\text{right}_C \leftarrow$ [12,13,14,19,20,21]. The keypoint from each region with the highest confidence score is chosen to be the representative point from which the pose triangle will be formed. For each of the pose triangles, we create a bisector line called *pose line* (PL) from the top corner of the triangle to the line segment formed by the other two corners (bold green lines as seen in Fig. 3c). The keypoints' splits (3 regions for 3 corners of the pose triangle) are chosen heuristically.

(b) Gaze Estimation + Correction Our data did not have annotations in order to fine-tune an existing gaze estimation model. Also, our goal was to get a rough estimate of the gazes for all the persons in order to detect the central focus within the image. Looking at the pose keypoints, we can argue that the gaze direction can roughly be estimated from the face and neck keypoints. Here, we exploit this hypothesis and use the pose keypoints $0, 1$ and 8 to generate a bisection vector as the first step (red vector, Fig. 3d). The bisection vector bisects the line segment joining keypoints 0 and 8. In Fig. 3d, we observe that these estimates are few degrees off of the original gaze and hence we also apply a correction to the gaze vector, denoted as correction angle. We skip those cases when one of the three keypoints is missing. Rather than just viewing the exact gaze, we represent gazes using gaze cones with an opening angle of 50 (25 degrees +/- direction) degrees in order to compensate the error estimate (green transparent cones, Fig. 3d). Experiments with various angles however did not affect the ARs and hence, we heuristically chose to keep the value 50[3].

The intersection of these cones are the probable regions where everyone in the scene is looking at. Hence, the centroids of these intersections become the

[3] similar results were achieved by using various angles: 10, 20, 60, 80.

(a) (b) (c) (d)

Fig. 4: (a) Gaze-Bisection-Vectors (red vectors) that point towards 2nd and 3rd quadrants are mapped to 4th and 1st quadrants (green vectors), respectively. (b), (c) show an example where there are 2 global action lines and 2 action regions detected. (d) is the eye-fixation map [5] of the image.

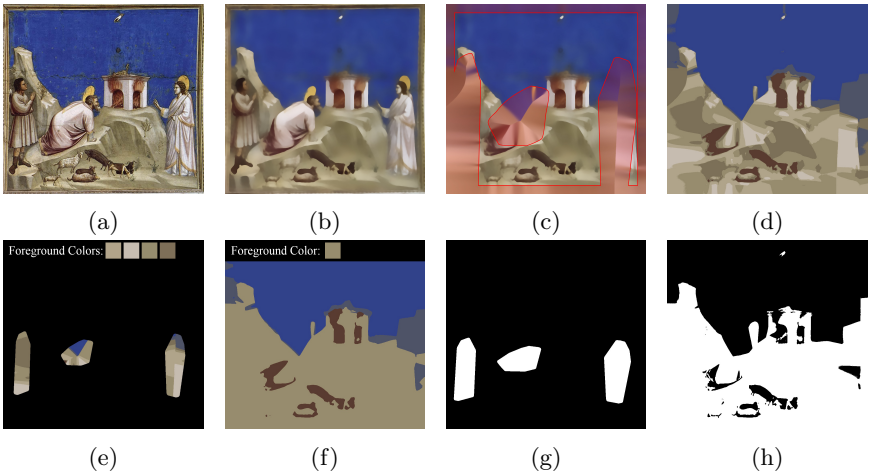
starting points of our global ALs. The cases with no intersection are skipped. In cases where there is more than one intersecting area, we proceed with two or more ARs (cf. Figures 4b and 4c). In order to plot the AL on our canvas, we require its slope. Therefore, we aggregate all the individual gaze directions in order to find the slope of the global AL.

(c) Combination of all gaze-bisection-vector slopes For calculating the slope of our global ALs, we aggregate all individual gaze directions (average of all the individual gaze slope) and find a common slope, where only the slopes of the gaze-bisection-vectors are considered. The slope value is with respect to the x-axis (Cartesian Coordinates) in the positive horizontal direction. Since the slope remains the same when we rotate the gaze-bisection-vector by 180 degrees, we map all the gaze vectors pointing towards the 2nd and 3rd quadrant to 4th and 1st quadrants respectively (Fig. 4a). We then draw a line in our output canvas through every centroid of the previously calculated intersection areas and use this aggregated slope as the slope of this new line. Figures 3d and 3e show how the aggregated gaze vectors form the global AL.

3.3 Semantic Foreground/Background separation

While analyzing human behavior within the semantics of scene understanding, humans present in the scene usually form the foreground. We exploit this property of scene semantics in our approach and consider the humans as our foreground (FG). Additionally, we define the objects near the human poses and the immediate surroundings as part of their foreground and the rest of the image as background (BG). The pipeline for FG and BG separation is shown in Fig. 2. It consists of the following three steps, which are explained in more details below: image filtering, keypoint-based inpainting, and pose- and color-informed k-means clustering.

(a) Image filtering. We observe that historical paintings show chipped paint, cracks or weathering, so we first apply a median filter that helps in reducing



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 5: Visual steps of *ICC*: (a) Original image, (b) Cracks filtered, (c) Bodies and picture frame inpainted, (d) Colors clustered using k-means, (e) Dominant colors under body positions selected as foreground colors, blue colors are below the threshold, (f) Detected foreground colors replaced and output filters applied, (g) Binarization applied to (e), (h) Binarized foreground/background separation.

these artifacts. We apply a bilateral filter to preserve the sharp edges as they are most crucial to separate major objects and protagonists as seen in Fig. 5b.

(b) Keypoint-based Inpainting. Next, we inpaint all pixels covered by detected human poses. Therefore, we generate the convex hull of all pose keypoints and scale the hull up by 70 % in x-direction and 40 % in y-direction, to ensure the mask is covering a little more than the human body in the painting. Using this mask, we inpaint the previously filtered image using the fast marching method [28]. Fig. 5c shows human bodies being replaced with a color mix marked by the red mask. These colors are considered as the potential foreground colors.

(c) Pose- and Color-informed k-means clustering. The inpainted image is then clustered into a smaller amount of colors using k-means clustering. This creates smaller color clusters (cf. Fig. 5d). The colors near the body postures of the characters are associated with the foreground by virtue of being related to the bodies of the characters. Next, we generate a second mask around the human poses, but scale each hull down by 30 % in y-direction to ensure that our mask is only placed over the core of our previously generated color mix (cf. Fig. 5c) using the inpainting (cf. Fig. 5e). All colors under this mask covering more than 6 % of the mask are then considered as foreground color. For experiments, we tried k=3, 5, 7, 9 and found k=7 to generate better qualitative results. The scaling factors for the convex hull were chosen heuristically.

3.4 Generating Output Canvas – Bringing It All Together

We generate two types of ICC, one colored ICC and a binary one denoting foreground/background pixels. In the colored version, all detected foreground colors are replaced with the most dominant foreground color. We apply median filter to remove small leftover color blobs/fragments as a post-processing step. For the binary version, each pixel is set to one if it is one of the foreground colors and zero otherwise. We apply morphological dilation and erosion filtering on the binary version to perform the closing of small fragments as post-processing. Finally, a morphological filter opening is applied to remove small blobs in the background that are typically due to the k-means results on the clustered cracks. We then gather the ALs, ARs (Sec. 3.2) and overlay it on the canvas with FG/BG separation (Sec. 3.3) to obtain our ICC. We can observe colored canvas in the Fig. 1c; and the binarized canvas in Fig. 1d.

4 Evaluation and Experiments

In this section, we first discuss the user study done for quantitative evaluation and the metrics being used for the same. Then we do a qualitative interpretation of our method, followed by general validity of our method by testing on cross-domain data. In the last part, we give performance evaluation of our method based on the user study.

4.1 User Study – Design and Evaluation Metrics

Design We evaluate the performance of ICC quantitatively and qualitatively through a user study. A set of 11 images were collected, 6 are Giotto's paintings, an annunciation scene, a baptism scene and 3 are chosen from COCO [19] with people present in them. We asked the annotators to label global action lines (AL), action regions (AR) and pose lines (PL) on these images. For quality control, a demo of the labeling process with an example and associated instructions were continuously available to the annotators while doing the task. This video (https://streamable.com/wk8ol8) describes how the user-study was conducted, including the labeling interface, the definition of each label and instructions to generate them with a sample example. We have 2 set of annotators, domain experts (E) and non-experts (NE). An *E* has a background in art history or its methodology; *NE*s are all university graduates. A total of 72 annotators (10 Es and 62 NEs) were presented with this set of images and asked to label the AL, AR and PL in them.

Evaluation Metrics. We evaluate each of AL, AR and PL individually and for all the annotators. We compare the performances between: E *vs.* ICC, NE *vs.* ICC and E *vs.* NE. The standard deviation (SD_{AR}) in the labeling of ARs is a measure of agreement between the annotators, lower SD_{AR} means they agree more where as higher SD_{AR} means they disagree more. For comparison

(a) **Original** (b) **AR** (c) **AL** (d) **PL** (e) **Eye-Fix**^{n}

Fig. 6: User Study analysis: 1st row shows an example of *The kiss of Judas*, 2nd of *annunciation* and 3rd of *COCO*[19]. (a) column are original images; (b) shows the Action Region (AR) by ICC (cyan), all Evaluators (red) and the centroid (white) of all annotators; (c) and (d) show AL and PL by ICC (yellow, green) and all annotators (red, red) respectively; and (e) highlights the eye-fixation regions [5]

of labeled ARs and predicted ARs, the L_2 distance is considered between E/NE and ICC. For AL and PL, the lines are considered as sets of points in the image. We find the distance between these two sets using Hausdorff Distance (HD) [6]. We also calculate the angular deviation (AD) between the AL of the E/NE and the ICC. Higher angular deviation means higher value of AD.

4.2 Results and Discussion

Action Lines. Each AL represents an important action or geometry in the original painting, however, both differ in their presentation (cf. Max Imdahl's *et al.* manually created analysis in Fig. 1b). It is not easy to quantify such expressions since they are more visually interpretive than quantitatively. In 1st row of Fig. 1a, Giotto's "Kiss of Judas", Judas is hugging Jesus. Many people are directly looking at them. This structure is apparent in other works of Giotto. The average slope of all these gazes creates the AL (yellow line in Fig. 1c). This AL gives us information about the direction from which these important points have been viewed. Sometimes, however, OpenPose fails to detect human poses, especially the poses of Jesus and Judas were not recognized (due to occlusions), which could have had a strong impact on the AL. Giotto's paintings sometimes also have a line of heads, with one head next to each other and most of them on the same height in the image (Fig. 1a). In this case, we see that our method does a very good job of detecting the AL (Fig. 1c).

Action Regions. The gaze structure in Giotto's paintings also motivates us to use their gaze directions in finding the approximate region of action. Since the gaze directions are conditioned on the body keypoints (Sec. 3.2.b) they also have pose context. We use the intersection of all gazes to identify important regions in the picture. In Fig. 1c, we see the locations of the ARs are very interesting. These regions are the center of high activity, which follows our definition of ARs (Sec. 3.2). We also observe that generally the AL passes through ARs. Figures 4b to 4d show an interesting case with multiple ARs and ALs. Our ICC is capable of predicting 2 ALs and their corresponding ARs. We see that some people focus on the 2 central characters accounting for the 1st AL, while the 2 central characters are looking at the sitting character creating the 2nd AL. The corresponding eye-fixation map in Fig. 4d highlights the regions where the 2 ALs are passing through and where the 2 ARs are located.

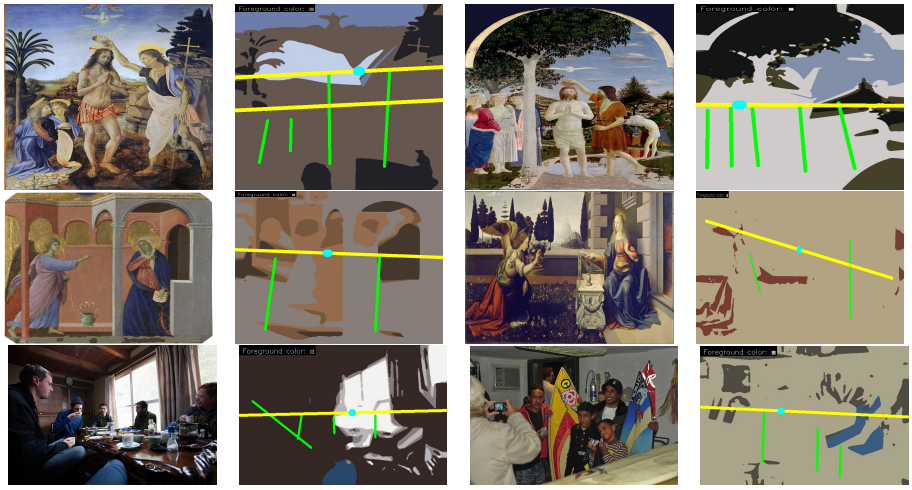
Semantic Foreground/Background separation. The main characters usually form the foreground of the image. Additionally, using their poses, we detect which of the areas in the image constitute foreground/ background. We in-paint the image at their body positions. In order to achieve this, K-means clustering is applied with a small k-value (good clustering and a filled, crack-less foreground surface). The small k-value could lead, in some cases, to a completely incorrect separation if the same colors were present in the foreground/background. Giotto's paintings also show the geometry of foreground/ background separation (cf. Fig. 1b) to be an important underlying structuring element. This separation is apparent in Figures 1c and 3e as the big light brown region covering half of the image and the remaining blue backdrop as the background.

4.3 Cross-Domain Adaptability

The first 2 source (Fig. 7, 1st & 2nd row), images are taken from 2 different iconographies: *Baptism* and *Annunciation*. For the third, (Fig. 7, 3rd row), we use images from COCO – a dataset of images from everyday life. We choose images that contains people doing different activities.

Figures 7a and 7b depict examples where our method is able to locate the AR, AL and PL very effectively. In addition, the foreground/background separation is also clearly visible in the ICC. For example in 2nd row of Fig. 7b, we see that the PLs are correct, the AL passes through the main activity and the AR is localized between the two characters. In 1st row of Fig. 7b, our method is able to detect the presence of 2 ALs: one aligns with the gaze of the female and the waist of the male, and the other aligns through the heads of the males. The AR is also very well localized, with good foreground/background separation.

In Figures 7c and 7d, our method fails either for ALs, or the foreground/ background separation. However, sometimes the ARs are very well localized even in these examples. In 1st row of Fig. 7d, we can observe that the main action region and background-foreground is incorrect, while the AL seems to be acceptable. Similarly, in 2nd and 3rd rows of Fig. 7d, the foreground/background



(a) **Original** (b) **ICC** (c) **Original** (d) **ICC**

Fig. 7: Cross domain analysis: 1st row depicts *Baptism*, 2nd shows *annunciation*, and the 3rd are images from *COCO* [19]. (a) and (c) columns are original images, (b) and (d) are their corresponding *ICC*.

Table 1: Quantitative Evaluation. 1st and 2nd columns show the Standard Deviation (SD_{AR}) for AR amongst Experts (E) and Non-Experts (NE). 3rd, 4th and 5th columns show the L_2 distance between centroids of E/ICC, NE/ICC and E/NE. 6th column shows the HD between all annotators (ALL) and ICC. 7th column shows cosine angular deviation (degrees) between ALL/ICC.

$\mathbf{SD_{AR}}$		$\mathbf{L_2}$			**HD**	**AD**
E	**NE**	**E/ICC**	**NE/ICC**	**E/NE**	**ALL/ICC**	**ALL/ICC**
0.054	0.089	0.187	0.180	0.035	664.35	36.57°

separation is not good and neither is the AL, but the AR is localized very well between the protagonists.

4.4 Quantitative Evaluation of User Study

For comparing action regions (AR) we reduce the image sizes to 1×1. From Tab. 1, we can see that the Es have more agreement among-st themselves for ARs as compared to NEs. However, the Euclidean distance (L_2) between the centroid of the Es and the ICC (E/ICC) has a similar value to that of the NEs and the ICC (NE/ICC), indicating that the region of ARs has high agreement for all groups, i. e., E, NE and ICC (people *vs. ICC*). The low value of L_2 between E/NE shows that they concur about the ARs (an error of 3.5% for the localization of AR). This fact is also recognizable in Fig. 6b, where the centroids

of all annotators lie very close to that proposed by our ICC. The eye (or gaze)-fixation maps of these images are also shown in Fig. 6e. These maps show those regions of the image where the gaze of an observer would focus. These maps are complementary to the compositional elements like ALs and ARs. The structure of these maps follows the slope of ALs and are distributed along the AL and the AR.

HD between all the annotators and ICC (ALL/ICC) is quite low (38 %) than the worst HD distance (ALL/Infinity: 1070.2 for the case of end points of the diagonal) showing that the ALs of our ICC has very good correlation to all annotators. We also see that the angular distance (AD) of all ALs with that of our ICC has an average value of 36.57° (ALL/ICC).

We observed that there is a positive correlation (0.2) between the images with low SD_{AR} and their corresponding $L_{2(AR)}$, meaning that when the annotators agreement for the position of AR was higher, our method predicted AR closer to the labeled ones. Similarly, we observed opposite trend for high SD_{AR} images, with correlation of -0.790, meaning that when there is higher disagreement between the annotators about the location of the AR, our method predicted AR farther away. Similar trend is observed when we correlate SD_{AR} with HD. When annotators agree (low SD_{AR}), the predicted AL is closer to the labeled ones, where as when annotators disagree (high SD_{AR}), the predicted AL is farther to the labeled one.

5 Conclusion and Future Work

With the help of existing machine learning tools, state-of-the-art pose estimation framework and image analysis, we have presented a novel no-training approach to understand underlying structures in art historical scenes. We show that recognizing key elements within a scene, such as ALs, ARs and FG/BG separation, helps in understanding the composition of any scene. Apart from using a pre-trained OpenPose framework, our method does not need any training. This makes our method very light, robust and applicable to any scene for analysis, especially when there is no ground truth available, which is often the case in art historic images. It can be used to pre-compare images for image retrieval and analysis. We also showed that our method works extremely well for images from related domains as well as images from everyday life.

Our method is a complementary approach to the perception of semantics in artworks [10] and iconography in general. It is one of the first attempts at understanding scenes or paintings in art history grounded in compositional elements. Future development could include training OpenPose on art historical data, and including the scene information to explore the role of scene narratives on the underlying compositions. At the same time our algorithm can be applied to various target domains where composition lines and narratives are used.

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2014)
2. Bell, P., Impett, L.: Ikonographie und Interaktion. Computergesttzte Analyse von Posen in Bildern der Heilsgeschichte. Das Mittelalter **24**(1), 31–53 (Jul 2019). https://doi.org/10.1515/mial-2019-0004, http://www.degruyter.com/view/j/mial.2019.24.issue-1/mial-2019-0004/mial-2019-0004.xml
3. Bienenstock, E., Geman, S., Potter, D.: Compositionality, mdl priors, and object recognition. In: Advances in neural information processing systems. pp. 838–844 (1997)
4. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. arXiv:1812.08008 [cs] (May 2019)
5. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. IEEE Transactions on Image Processing **27**(10), 5142–5154 (2018)
6. Dubuisson, M.., Jain, A.K.: A modified hausdorff distance for object matching. In: Proceedings of 12th International Conference on Pattern Recognition. vol. 1, pp. 566–568 vol.1 (Oct 1994). https://doi.org/10.1109/ICPR.1994.576361
7. Dundar, A., Shih, K.J., Garg, A., Pottorf, R., Tao, A., Catanzaro, B.: Unsupervised disentanglement of pose, appearance and background from images and videos. arXiv preprint arXiv:2001.09518 (2020)
8. Fieraru, M., Khoreva, A., Pishchulin, L., Schiele, B.: Learning to refine human pose estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2018)
9. Garcia, N., Renoust, B., Nakashima, Y.: Context-aware embeddings for automatic art analysis. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval. pp. 25–33 (2019)
10. Garcia, N., Vogiatzis, G.: How to read paintings: semantic art understanding with multi-modal retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)
11. Gonthier, N., Gousseau, Y., Ladjal, S., Bonfait, O.: Weakly supervised object detection in artworks. In: Computer Vision – ECCV 2018 Workshops. pp. 692–709. Springer International Publishing (2018)
12. Imdahl, M.: Giotto, Arenafresken: Ikonographie-Ikonologie-Ikonik. Wilhelm Fink (1975)
13. Imdahl, M.: Giotto, Arenafresken: Ikonographie, Ikonologie, Ikonik. W. Fink, München (1980), oCLC: 7627867
14. Ionescu, V.: What do you see? the phenomenological model of image analysis: Fiedler, husserl, imdahl. Image and Narrative **15**(3), 93–110 (2014)
15. Jakab, T., Gupta, A., Bilen, H., Vedaldi, A.: Unsupervised learning of object landmarks through conditional image generation. In: Advances in Neural Information Processing Systems. pp. 4016–4027 (2018)
16. Jenicek, T., Chum, O.: Linking art through human poses. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1338–1345 (Sep 2019). https://doi.org/10.1109/ICDAR.2019.00216
17. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6912–6921 (2019)

18. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
19. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft COCO: Common Objects in Context. arXiv:1405.0312 [cs] (Feb 2015)
20. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
21. Lorenz, D., Bereska, L., Milbich, T., Ommer, B.: Unsupervised part-based disentangling of object shape and appearance. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10955–10964 (2019)
22. Madhu, P., Kosti, R., Mührenberg, L., Bell, P., Maier, A., Christlein, V.: Recognizing characters in art history using deep learning. In: Proceedings of the 1st Workshop on Structuring and Understanding of Multimedia HeritAge Contents. p. 1522. SUMAC 19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3347317.3357242, https://doi.org/10.1145/3347317.3357242
23. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: Advances in neural information processing systems. pp. 2277–2287 (2017)
24. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4903–4911 (2017)
25. Recasens, A., Khosla, A., Vondrick, C., Torralba, A.: Where are they looking? In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28, pp. 199–207. Curran Associates, Inc. (2015)
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
27. Stefanini, M., Cornia, M., Baraldi, L., Corsini, M., Cucchiara, R.: Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In: Ricci, E., Rota Bulò, S., Snoek, C., Lanz, O., Messelodi, S., Sebe, N. (eds.) Image Analysis and Processing – ICIAP 2019. pp. 729–740. Springer International Publishing, Cham (2019)
28. Telea, A.: An image inpainting technique based on the fast marching method. Journal of graphics tools **9**(1), 23–34 (2004)
29. Volkenandt, C.: Bildfeld und Feldlinien. Formen des vergleichenden Sehens bei Max Imdahl, Theodor Hetzer und Dagobert Frey, pp. 407 – 430. Wilhelm Fink, Leiden, The Netherlands (2010), https://www.fink.de/view/book/edcoll/9783846750155/B9783846750155-s021.xml
30. Yuille, A.L., Liu, C.: Deep nets: What have they ever done for vision? arXiv preprint arXiv:1805.04025 (2018)