FACULTY OF ENGINEERING
AND ARCHITECTURE

Improving Pose Estimation on Art Collections with Style Transfer

Tristan Verheecke
Student number: 20043518

Supervisors: Prof. dr. ir. Dieter De Witte, Prof. dr. Steven Verstockt
Counsellors: Kenzo Milleville, Ravi Khatri

Master's dissertation submitted in order to obtain the academic degree of
Master of Science in Information Engineering Technology

Academic year 2023-2024

GHENT
UNIVERSITY

Preface

I've been interested in Art my entire life. In fact, I've a degree in the Fine Arts from LUCA School of Arts. There, I was known for my technological ability and one of my professors at the time asked me why I didn't do anything with that in my artworks. That remark has since stuck with me and was part of my motivation to apply for readmission for my Master of Science. With all the advancements in AI, I started thinking more and more about doing work with that. Like Matisse and Turner, I'm not satisfied with the tools available, but want to create my own.

It was therefor to my delight that I was able to work on this thesis which has provided me the opportunity to acquire more insight in the subject. I would like to thank my supervisors Dieter De Witte and Steven Verstockt for this wonderful opportunity, and my counsellor Kenzo Milleville for his great guidance. As well as all the other people at IDLab for their feedback. I also want to thank Karine Lacaracina, Lies Van De Cappelle and the other people at RMFAB for providing help with the artistic sensibilities of the thesis.

Enjoy the read,

Tristan Verheecke
Ghent, June 2023

Conference Paper Title*

*Note: Sub-titles are not captured in Xplore and should not be used

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

This document is a model and instructions for LaTeX. Please observe the conference page limits.

II. Ease of Use

A. Maintaining the Integrity of the Specifications

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

III. Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections III-A–III-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: “Wb/m^2” or “webers per square meter”, not “webers/m^2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.
- Use a zero before decimal points: “0.25”, not “.25”. Use “cm^3”, not “cc”.)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not

Identify applicable funding agency here. If none, delete this.

"Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is . . ."

D. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Fig. 1. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks …”. Instead, try “R. B. G. thanks…”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first …”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
[4] K. Elissa, “Title of paper if known,” unpublished.
[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove the template text from your paper may result in your paper not being published.

Contents

List of Figures

List of Tables

List of Acronyms

A

AdaIN	Adaptive Instance Normalization , 12
AIC-HKD	AI Challenger Human Keypoint Detection , 4
ASMs	Active Shape Models , 3

B

BN	Batch Normalization viii, 10, 11

C

CBIR	Content Based Image Retrieveal viii, 14, 15
cGAN	conditional Generative Adversarial Network , 7, 12
CIN	Conditional Instance Normalization , 11, 12
CIR	Category Image Retrieval , 14
CNN	Convolutional Neural Network , 2, 5, 8, 9
COCO	Common Object in Context , 3, 10
CPMs	Convolutional Pose Machines viii, 6, 9
CPN	Cascaded Pyramid Network , 8

F

FLIC	Frames Labeled In Cinema , 4

G

GAN	Generative Adversarial Network , 7, 12, 14

H

HPE	Human Pose Estimation , 2–5, 7, 10, 19

I

IIR	Instance Image Retrieval , 14
ILP	Integer Linear Programming , 8
IN	Instance Normalization viii, 10, 11

L

LSP	Leeds Sports Pose , 3

M

MPII	Max Planck Institute for Informatics viii, 3, 4
MSE	Mean Square Error , 12, 14

N

NMS	Non-Maximum-Suppression , 8
NST	Neural Style Transfer , 2, 12

P

PAF	Part Affinity Field , 9
PAF	Part Association Fields , 9
PIF	Part Intensity Fields , 9

R

ResNet	Residual Network , 7, 8
RMFAB	Royal Museums of Fine Arts of Belgium , 1, 20
RPME	Regional Multi-person Pose Estimation , 8

S

SAHR	Scale-adaptive Heatmap Regression , 9
SIFT	Scale-Invariant Feature Transform , 15
SMPL	Skinned Multi-Person Linear , 3

V

VAE	Variational Autoencoder , 14

W

WAHR	Weight-adaptive Heatmap Regression , 9

List of Code Fragments

1

Introduction

1.1 Problem definition

To make art collections more accessible, museums put a huge effort in digitalizing their catalogue. However, they don't contain much metadata about the content and it is time-consuming to enhance them manually. To make ths process easier, they want to utilize computer vision. Art collections (paintings, statues, drawings, etc.) turn out to be less interpretable by the algorithms that were developed for photography over the last few decades. These scan the images in search of recognizable objects and add their labels to the metadata. Even the latest state-of-the-art technology, struggles to recognize objects when pointed at a painting in a museum. A solution may be to start over and have paintings annotated by humans.

This has been done in 2 recent projects: Saint-George-On-A-Bike [17] and INSIGHT [18]. However, paintings are very complex and manual annotation doesn't scale and is very expensive. For example, 10,000 paintings were annotated by Royal Museums of Fine Arts of Belgium (RMFAB) with no clear return on investment. They spent a year on this and this is not something they want to repeat. How can we automate this process and ensure that state-of-the-art computer vision models give good results on paintings and artworks?

Specifically for this thesis, pose estimation will be investigated.

1.2 Proposed solution

(dirty version) We will first examine the effectiveness of existing models on a collection of paintings from 2 different movements. For this we will need to have a pose estimator, a style transformer and a collection of test data.

A first method: We will first convert the test data with the style transformer to a painting and then we will apply pose estimation. The test data will have coordinates of the joints, which we will compare with the results of the pose estimation. However, the joints are of the original image. How do we convert those coordinates to map to the styled image? Problem: This method does not use any real paintings and will be susceptible to the accuracy of the style transformer.

A second method: We can apply pose estimation to real paintings and then convert them to a realistic image with style transfer. We can then use pose estimation to the realistic images and compare them with the style transformed results. This will also require a way to map the results of the real painting to that of the style transformed. Problem: While we're using real paintings now, the results will still depend on the accuracy of style transformer.

A third method: We can annotate the paintings ourselves and use pose estimation to assess the pose estimation algorithms. Problem: We must annotate the paintings ourselves.

There are several things that can be improved: The dataset, the algorithm, the input

2

Literature study

In order to correctly implement a solution, we need to understand the fundamentals. These consist of 2 research fields: Human Pose Estimation (HPE) and Neural Style Transfer (NST). The former will be used to detect poses in the art collections, but not before the latter has tried to make an improvement. Following will be an overview of the available research in these domains. Discussing what the goals of them are, how they achieve it, what their challenges are and their limitations.

2.1 Human Pose estimation

HPE aims to detect human features from input data such as images and videos. It's an elementary part of computer vision with many applications among which are human action recognition (sign language), human tracking (surveillance), and human-computer interaction (video games). This is an extensively researched area with a diverse range of different techniques. This chapter will try to give an overview of all the many challenges and proposed solutions. The focus will be on deep learning models, which have surpassed classical solutions significantly. Specifically, around 2D monocular HPE eg., [19][3][20][21].

The human body has a high degree-of-freedom due to all the limbs, self-similar parts and body types, which may cause self-occlusion or rare/complex poses. The variations in configuration are made even larger due to clothing, lighting, foreground occlusion, as well as viewing angles and truncation, among others, as shown in fig. **??**. This makes HPE one of the most difficult tasks in computer vision [22][2].

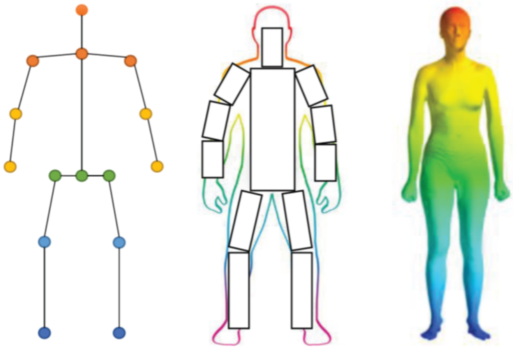
2.1.1 Representation

An important factor in HPE is how the pose will be represented. Depending on the needs of the problem you can have a skeleton-base, contour-base, or volume-base solution [2] as seen in Fig. **??**.

Skeleton-based model

The skeleton is build of a tree-structured set of keypoints that represent the joints of the human body. These can be explicitly described by their coordinates in 2D or 3D space [4]. More suitable for a Convolutional Neural Network (CNN) however is a heatmap which constructs a 2D Gaussian kernel around a keypoint [20][23]. They are easily implemented and became the dominant representation. While the skeleton-based model is a compact and flexible representation it suffers in this aspect by not being able to hold texture or shape information [3].

Figure 2.1: The various challenges HPE solutions face. Images from MPII dataset. [1][2]



(a) Skeleton (b) Contour (c) Volume

Figure 2.2: Models for pose representation [3]

Contour representation

To capture the shape of the body parts, contour representation uses rectangles to estimate the body contours. These methods include cardboard models [24] and Active Shape Models (ASMs) [25] and were mainly in use in earlier HPE methods [2].

Volume representation

Volumetric geometric shapes can also be used as a method of representation. Earlier methods used simple shapes like cylinders, conics, and other shapes [26]. Volume representation is a 3D mesh that represents the human body. The most used model is Skinned Multi-Person Linear (SMPL), which includes natural pose-dependent deformations imitating soft-tissue dynamics [27].

For the purpose of our research, a simple model is the only thing we need. We only need to be aware of the most essential joints to label a pose. This makes the skeleton-based model the ideal representation to work with and will be the focus of further study.

2.1.2 Datasets

There are several publicly available datasets. There are some that are outdated and we will leave those out, focusing only on datasets used for deep learning.

Leeds Sports Pose (LSP) Dataset [28] contains 2,000 images found on Flickr using 8 different tags looking for sport activities (athletics, badminton, baseball, gymnastics, parkour, soccer, tennis, and volleyball). Each person has 14 keypoints. An extended version was later introduced [29], now consisting of 10,000 images. For this set they only focused on the more challenging tags (parkour, gymnastics, and athletics).

MPII Human Pose Dataset [1] contains 24,290 images with 40,522 labeled people. They were extracted from YouTube videos found by querying for physical activities. Each person has 16 keypoints and also includes occlusion labels.

Common Object in Context (COCO) Dataset [30] is a large-scale dataset for a wide range of computer vision algorithms.

For HPE, the set contains more than 200,000 images in which 250,000 persons are annotated. Each person has 17 keypoints, a bounding-box and visibility labels. This dataset has become the most popular for benchmarking.

Frames Labeled In Cinema (FLIC) Dataset [31] contains 5,003 images extracted from Hollywood movies. They ran a person detector which collected 20,000 images from 30 movies. Occluded and difficult poses were then removed leaving only 5,000 images to be annotated. Only the upper body received 10 keypoints.

AI Challenger Human Keypoint Detection (AIC-HKD) Dataset [31] contains 300,000 images found using Internet search engines. In these, over 700,000 humans are annotated. Each person has 14 keypoints, a bounding-box, as well as visibility and left/right labels.

CrowdPose Dataset [32] puts an emphasis on crowded images. 30,000 images from MPII, glsCOCO and glsAIC-HKD were measured with a Crowd Index, which evaluates the crowdedness. Finally, 20,000 images are selected and 80,000 persons annotated. Each person has 14 keypoints and a full-body bounding box.

Human-Art Dataset [33] bridges the gap between natural and artificial images. The set contains 50,000 high-quality images with 123,000 annotated humans. Each person has 17 keypoints, bounding boxes, self-contact points, and text information.

2.1.3 Discriminative Methods and Generative Methods

Before deep learning became prominent in HPE there were already a number of different methods in use. Some of these methods are compatible with the deep learning methods and were thus adopted. An early distinction is between generative and discriminative methods.

Generative Model

A generative method will work with prior beliefs about the pose. More information about this can be found in the section about representation 2.1.1. It will project the pose on the image and verify it with the image data. If they don't comply, the pose is adjusted using the descent direction found by minimizing an error function [34].

Discriminative Model

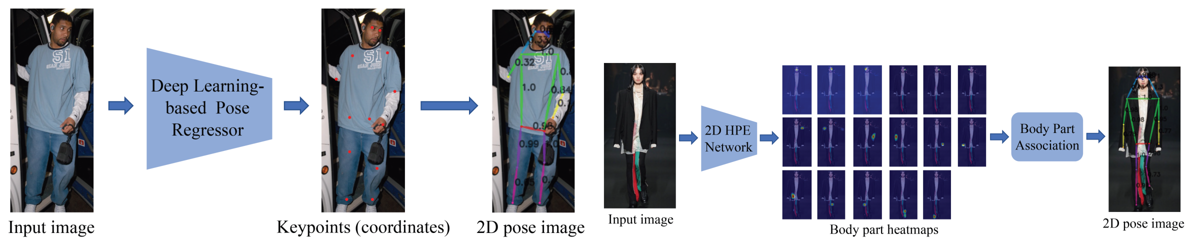
Discriminative methods on the other hand, try to map the pose on the image data with learned models. There are several methods in this category, among which are the deep learning-based methods. The deep-learning methods are further categorized by the following sections.

2.1.4 Single-Person Methods

Single-person pose estimation will try to evaluate only one pose from an image. There are 2 major methods that are in use: regression methods and detection-based methods.

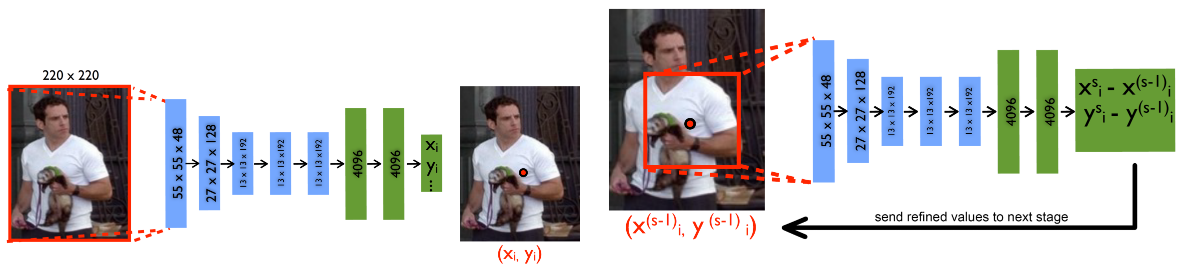
Regression-based Methods

The regression-based methods learn a network that maps all the body keypoints to the image-data directly as show in 2.3a.



(a) Regression Methods (b) Heatmap-based Methods

Figure 2.3: The different methods of single-person human pose estimation.[3]



(a) Initial stage (b) Stage s

Figure 2.4: Convolution layers in blue and fully connected layers in green. The initial stage is applied to the whole images, while in stage s it will work on a sub-image based on the result of the previous stage.[4]

The first successful deep learning model came from Toshev and Svegedy [4] and is considered the switch in paradigm from classic approaches to deep learning HPE. Toshev et al. uses a 7-layered model with 5 convolution layers and 2 fully-connected layers for the pose regressor, based on AlexNet for its simple but effective architecture [35]. They then cascade the resulting found keypoints of this model to itself where it refines it using the area around the keypoints. While the network is the same, the different stages will have different learned parameters. With every stage the found keypoints become more accurate. A illustration of this can be found in Fig. 2.4.

Carreira et al. [36] introduce an Iterative Error Feedback which is a self-correcting model using top-down feedback. Using the image-data and a starting pose modeled as a heatmap, the model, based on GoogLeNet [37], will predict an error for each keypoint. The pose is then corrected based on the error and fed back into the model as a heatmap with the image. With each iteration it converges towards the solution instead of making the prediction in one go. Regression-based methods map the keypoints directly on the image, making it a non-linear problem. This will cause less robust generalization however [20].

Heatmap/Detection-based Methods

The detection-based methods will first estimate the individual body parts using heatmaps, which leads to an easier optimization and a more robust generalization [21]. Most of the latest HPE methods use heatmaps because of this. After the joints are found they are then assembled to fit a human skeleton. This process is shown in 2.3b.

Tompson et al. [38] proposed a hybrid architecture where the detection of body parts is handled by a CNN and a Spatial-

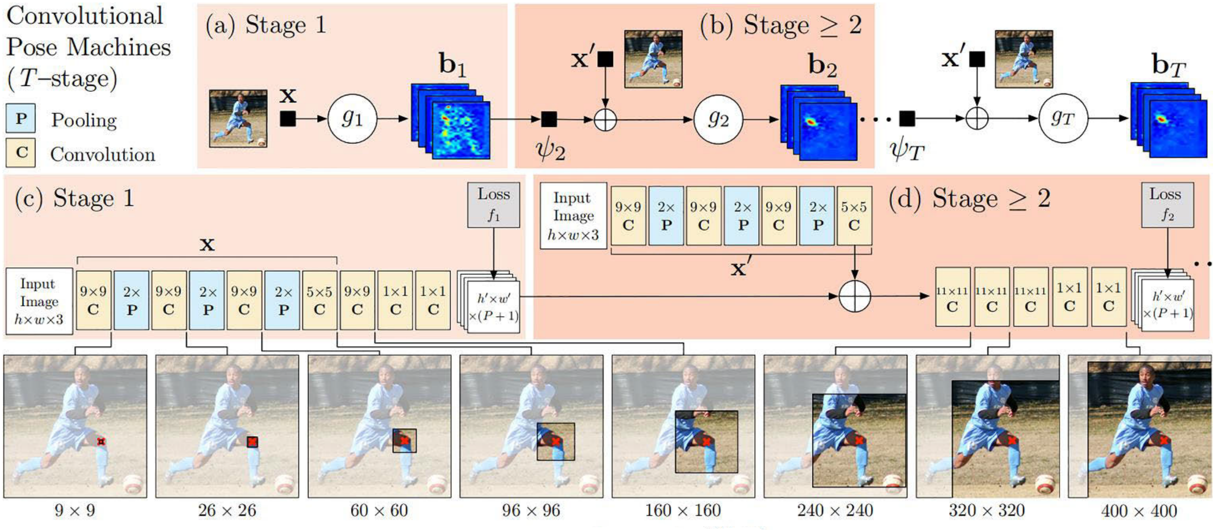
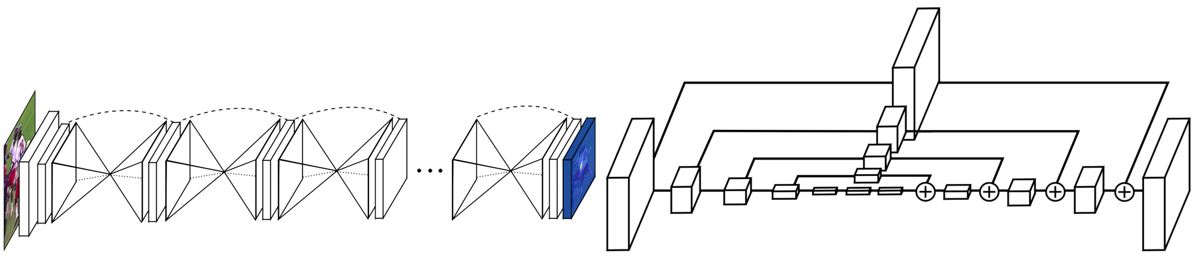


Figure 2.5: Architecture and receptive fields of CPMs. (a) and (b) represent the pose machine architecture.[5] (c) and (d) show the corresponding convolutional networks used by CPMs.[6]



(a) Stacked Hourglass

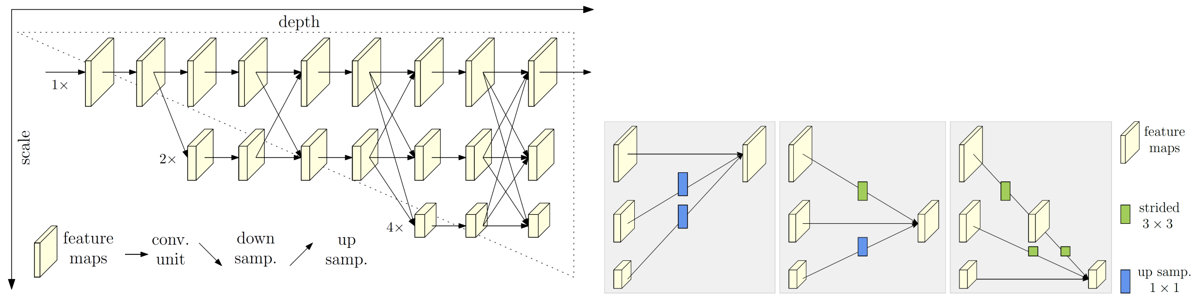
(b) Hourglass Module

Figure 2.6: The structure of a "stacked hourglass" network and a single "hourglass" module.[7]

Model to bring those together. The first step produces many false-positives and these are removed in the second step by restricting joint inter-connectivity to enforce correct anatomy. They build on this in [39], where they used a cascade to refine predictions.

A fundamental work written by Wei et al. [6] combines convolution networks with Pose Machines [5]. Pose Machines is an iterative architecture which consists of 2 models: the first is used for stage 1 where it extracts potential heatmaps for the joints. The second model is used for subsequent stages where the result of the previous stage is fed in together with the results of its own convolution network on the input image. This gradually refines the predictions for the joints and their positioning. 2.5 shows this process.

Another influential work was being written at the same time by Newell et al. [7]. Similar to CPMs, this is also an iterative architecture. They suggest what they call a "stacked hourglass" network, where "hourglass" modules are repeated 2.6a. In an "hourglass" module, first, the features are downsampled and afterwards upsampled again 2.6b. This network captures different spatial relationships between joints at different resolutions. Several other works [40][**?**][41] have since improved on the network design.



(a) HRNet

(b) Multi-Scale Fusion

Figure 2.7: The architecture of the High-Resolution network and how it applies multi-scale fusion.[8]

Both these use intermediate supervision to tackle the problem of vanishing gradients. This still doesn't build a deep sub-network for feature extraction which limits the estimations. This has become less of a problem with the emergence of Residual Network (ResNet)[42] which allows better back-propagation at deeper levels through shortcuts.

A more recent work by Sun et al. [8] maintains the high-resolution representations instead of working the high-resolution from the low-to-high sub-network. After a first high-resolution sub-network, it gradually adds high-to-low sub-networks in parallel to predict multi-resolution features. Before each branch, they apply multi-scale fusion, which joins the predicted features from each scale on each scale. Both are shown in 2.7. This network has proven very effective and inspired several variations [43][44][45].

With the emergence of neural networks also came Generative Adversarial Networks (GANs) [46], which proved useful for HPE. They are employed to improve constraints of joint inter-connectivity and infer occluded body parts.

Chen et al. [9] propose a structure-aware convolution network using a stacked hourglass as generator which generates heatmaps for each joint. They use 2 discriminators, one to discriminate between low- and high-confidence predictions, another for real and fake poses. The network is designed as a conditional Generative Adversarial Network (cGAN) [47], which allows it to generate pose heatmaps as well as occlusion heatmaps.

A more classic GAN is used by Chou et al. [48], where they use a stacked hourglass network for both the generator as the discriminator. The generator predicts the heatmaps for each joint and the discriminator distinguished between the real and fake ones.

2.1.5 Multi-Person Methods

With multi-person methods comes an extra layer of difficulty: they need to be able to detect each person separately. To solve this problem multi-person methods propose several solutions. The 2 most popular are top-down and bottom-up methods.

Top-Down Methods

This method will first try to detect all persons in the image with a human detector. Each person is cropped by the bounding box and a single-person estimator predicts a pose for each person.

Occlusion and truncation are a regular occurrence in multi-person scenes and inevitable problem. One of the early multi-person models, by Iqbal et al. [49], works towards creating a robust model against occlusion. It uses Faster RCCN [50]

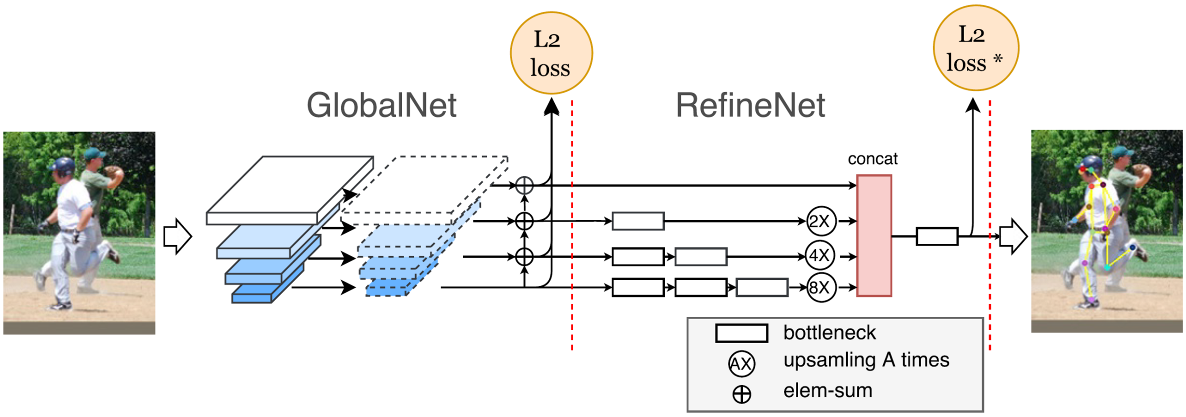


Figure 2.8: Cascaded Pyramid Network. "L2 loss*" means L2 loss with online hard keypoints mining.[9]

to detect the human boundaries. After which, it applies integer linear programming for each person's fully connected graph. This technique is similar to [51], but instead of working on all globally found joints it only considers local joints. It can also handle any kind of occlusion or truncation.

The use of a human detector comes with its own sort of problems. Fang et al. [52], with Regional Multi-person Pose Estimation (RPME), try to remedy these with 2 components: They try to tackle inaccurate bounding boxes with Symmetric Spatial Transformer Network, redundant detections with Parametric Pose Non-Maximum-Suppresion. They also propose a 3rd component, Pose-Guided Proposals Generator, which can augment training samples.

Papandreou et al. [53] use a 2 stage pipeline. In the first stage, they employ the Faster RCNN detector [50]. In the second stage, they estimate the pose in each found bounding box using their own network. It predicts heatmaps using a fully convolutional ResNet and use their own novel aggregation procedure. Afterwards, they do post-processing using keypoint-based Non-Maximum-Suppression (NMS) a method of their own making.

A continuous effort is taken by Chen et al. [9] to deal with occlusion and truncation. They suggest a 2 stage architecture, a Cascaded Pyramid Network (CPN) as seen in 2.8, where first the "simple" keypoints are captured with GlobalNet, a feature pyramid network based on [54], and the "hard" keypoints are handled by their RefineNet, based on the upsampling and concatenating of HyperNet [55] and using an adapted stacked hourglass. They achieved great results and several others improved on their work [56][57].

In more recent research, a new method was become more powerful than CNNs. The Transformer [**?**], based on attention mechanisms which are used to optimize recurrent networks [58], eliminates the use of recurrent layers, keeping only the attention mechanisms. Yang et al. [59] use this architecture because allows for better understanding of the spatial dependencies and learns at a higher rate.

Bottom-Up Methods

A different approach is taken with bottom-up methods. They first locate all joints in the image and then assemble them in potential humans.

DeepCut by Pishchulin et al. [51], one of the first multi-person models using CNNs. Using Fast R-CNN [50], it detects the body parts and labels each. With the joints found, it then uses Integer Linear Programming (ILP) to assemble them. This

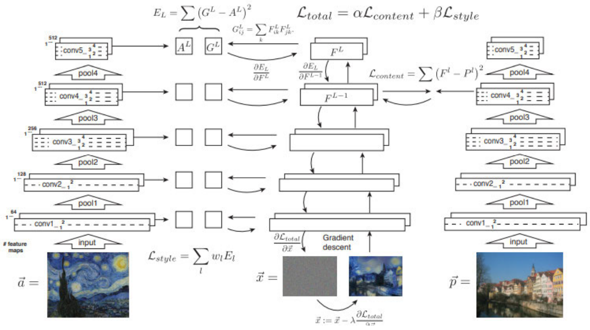


Figure 2.9: Style transfer algorithm. (Gatys et al. [10]).

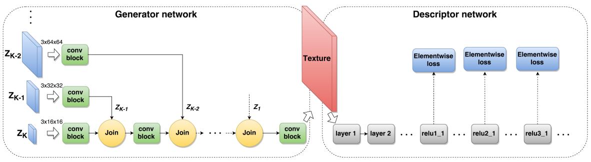


Figure 2.10: A texture network by Ulyanov et al. [11]. The generator network (left) is the only one that changes. The descriptor network (right) is used to calculate the loss.

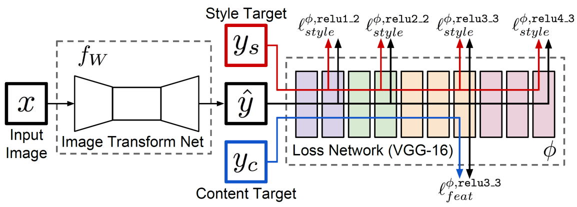


Figure 2.11: An image transformation network by Johnson et al. [12]. The image transform network (left) is the only one that changes. A loss network (right) is used to define perceptual loss functions.

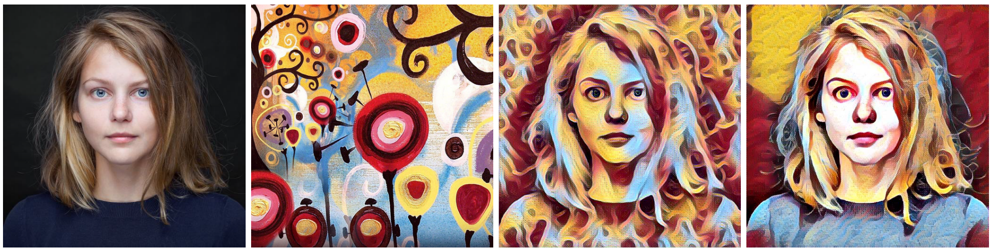
method is very computationally expensive; NP-hard. Insafutdinov et al. [60] therefor introduce a stronger part detector and better optimization strategy with DeeperCut.

CPMs make a return with OpenPose by Cao et al. [61], they're used to predict the joints with heatmaps and Part Affinity Fields (PAFs). A part affinity field also encodes the position and orientation of the limb which makes the assembly of joints into different poses possible. They can achieve real-time results with this method, and several others have improved on their design [62][63][57]. The high performance is only applicable to high-resolution images. Low-resolution images or images with occlusions perform poorly.

Kreiss et al. [64] continue on the idea of fields and introduce the Part Intensity Fields (PIF) and Part Association Fields (PAF). First, they predict the location of the different joints with PIF. Afterwards, they use PAF to find the inter-joint relationships. They are able to outperform any previous OpenPose-based proposals on low-resolution and occlusions.

Newell et al. [65] introduce a new method called associative embedding for supervising CNNs both detection and grouping. This is a single-stage architecture as opposed to the two-staged architectures previously discussed. They make use of the stacked hourglass network from [7] with some small modifications.

Continuing on the idea of associative embedding, Cheng et al. [43] use HRNet [8] as backbone for their HigherHRNet. Their method focuses on the scale-variance problem; a problem which hasn't been studied much, so it can localize keypoints for small persons better. Lou et al. [66] introduce Scale-adaptive Heatmap Regression (SAHR) and Weight-adaptive Heatmap Regression (WAHR) to the scale-variance problem. SAHR adaptively adjusts the standard deviation of each heatmap corresponding with the scale of the person. WAHR rebalances the foreground and background samples, so SAHR can work to its fullest extent.



(a) Content Image (b) Style Image (c) StyleNet with BN (d) StyleNet with IN

Figure 2.12: A comparison between (c) BN and (d) IN.[13]

Summary

An important challenge for HPE is making predictions in scenes with hight occlusions. Top-down models achieve state-of-the art performance in almost all benchmark datasets [2]. Top-down models has difficulty with overlapping bodies and human detectors might fail finding humans there. To the same extent, bottom-up models will have greater inaccuracy with grouping in occluded scenes. Computationally, the top-down model's speed is limited by the number of people found. The higher efficiency of bottom-up models, make them more suitable for real-time applications.

2.2 Image Style Transfer

Image Style Transfer is the technique of applying the style of one image to the content of another. Classically this was a problem reserved for only artists, but more recently this has also interested computer scientists. There are several different ideas on how this can be achieved, ranging from how to separate the style from the content, to how well an algorithm can generalize. An overview of all the different challenges and solutions will be given in this chapter.

2.2.1 Datasets

Due to a lack of benchmark datasets, multiple papers will mix and match from different datasets, like COCO or ImageNet [67].

Cityscape Dataset [68] consists of 2975 images of cityscapes with semantic annotations.

Facades Dataset [69] consists of 400 images of building facades with architectural annotations.

Maps Dataset [70] consists of 1096 images of maps and areal photos gathered from Google Maps around New York City.

Edges2shoes Dataset [71] consists of 50,000 paired images between edges and photos of shoes.

Edges2handbags Dataset [72] consists of 137,000 paired images between edges and photos of handbags.

Season transfer Dataset [62] consists of 2127 images of Yosemite during summer and winter downloaded from Flickr.

Night2Day Dataset [73] consists of 20,000 images taken from time-lapse datasets and annotated through crowd-sourcing.

WikiArt Dataset [74] consists of 80,000 fine-art paintings. All are annotated for 27 styles, 60,000 are annotated for 20 genres and 20,000 for 23 artists.

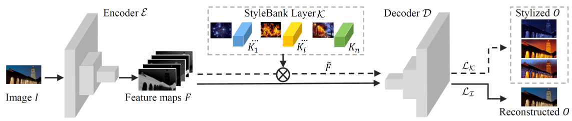


Figure 2.13: The stylebank network by Chen et al. [9].

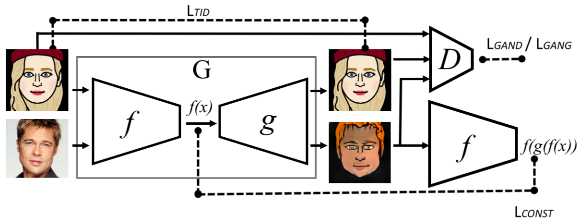


Figure 2.14: The domain transfer network by Taigman et al. [14].

2.2.2 Optimization-based Networks

Gatys et al. [10] introduce deep neural networks to image style transfer. Using a modified VGG-network [75], they extract the features of an image by reconstructing the content from the feature maps in the higher layers on a white noise image. The same is done for the style of the other image. It extracts the style representation of the image by using the Gram matrix to represent style features of the image and then reconstructs it on the same white noise image. The Gram matrix is the vector product of two sets of vectorized feature maps. This method is shown in 2.9. They remark that the resolution of the images affects the performance of the algorithm and is thus restricted to low resolutions. At the same time, the synthesized images contain some low-level noise, but this can possibly be removed with a denoiser.

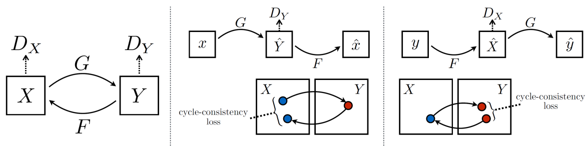
2.2.3 Feed-forward Generation Networks

To improve the performance, Ulyanov et al. [11] suggest the use of a feed-forward generation network instead of back-propagation Backpropagation requires an iterative process to change the pixel values to match the desired statistics. A feed-forward network can do this in a single evaluation. To train such a network they use a pre-trained network for image classification, and calculate a texture and content loss like [10], as shown in 2.9. Johnson et al. [12] propose a very similar method as can be seen in 2.9

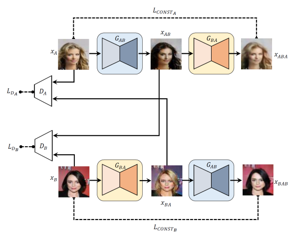
Since their contribution did increase the speed, but at the expense of quality, Ulyanov et al. [13] suggest further improvements to their network. First, they replace BN [76] with IN which alone has a significant impact on quality as can be seen in 2.12. Second, they learn the generator to sample from the Julesz ensemble [77] which improves variation in the outputs.

Dumoulin et al. [78] note that previous feed-forward networks are limited to one style. In order to facilitate many different styles, there would need to be a network trained separately for each which limits the applications for mobile devices. In order to make the network more memory efficient, they propose a conditional style transfer network; given a content image and a style name, it transforms the image to the corresponding style. They argue that after normalization each style can be distinguished by specializing scaling and shifting parameters. They call this Conditional Instance Normalization (CIN). Since it only changes the scale and shift parameters for different styles, the network requires fewer parameters. Of the 1.6M parameters, only 3K are needed for the different styles.

Another network that puts a focus on multiple styles comes from Chen et al. [9]. They propose a StyleBank, as seen in Fig. **??**, which can store multiple convolution filter banks each representing a different style. They use an auto-encoder network with in between a StyleBank layer. During training, for each $T+1$ iterations the entire network is first trained

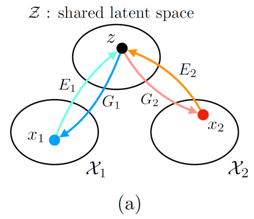


(a) As illustrated by Zhu et al. [62].

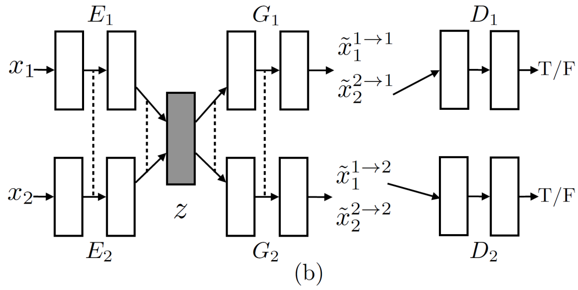


(b) As illustrated by Kim et al. [58].

Figure 2.15: The cycle-consistent network.



(a) The shared latent space assumption.



(b) The unsupervised image-to-image translation network.

Figure 2.16: Liu et al. [15].

with a perception loss for the first T iterations. Then only the auto-encoder network is trained with a Mean Square Error (MSE) loss. This way the auto-encoder only retains the content and the StyleBank layer only the different styles. This also allows to lock the encoder and decoder to learn a new style afterwards.

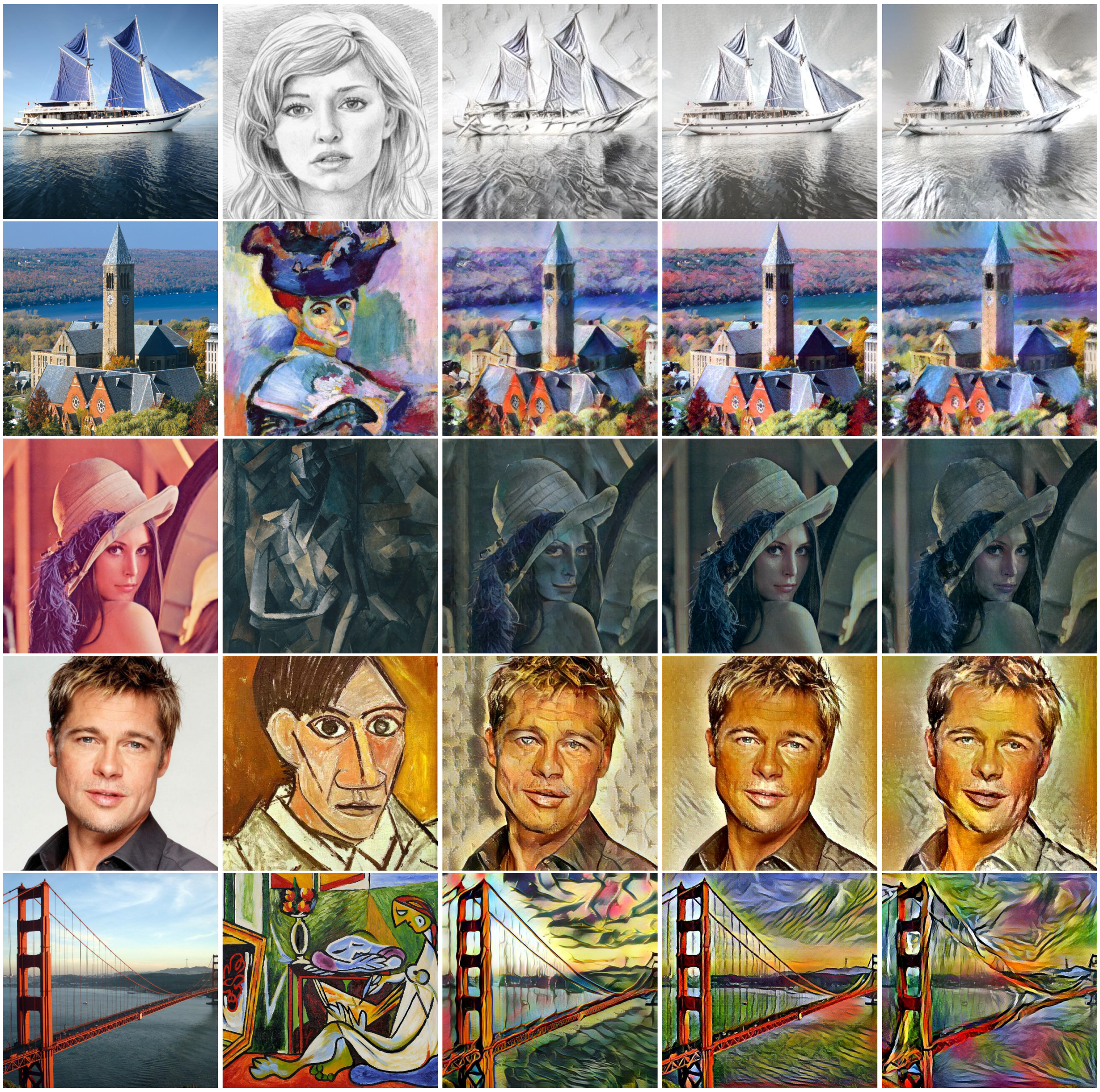
While CIN allows for multiple styles, it's still limited to the ones that were seen during training. Huang et al. [79] try to remedy this by introducing an Adaptive Instance Normalization (AdaIN) layer. Unlike the other normalization techniques, AdaIN does not have affine parameters, and will adaptively compute these from the style image. 2.17 shows how well the different networks can handle unseen styles.

2.2.4 Generative Adversarial Networks

With the introduction of GANs, the quality of generative models have greatly increased. It is not surprising then that this got picked up in research for NST.

Among the first was Isola et al. [70] who use a cGAN. With cGAN, the generator network has an extra input which here is the image to be translated. They use the network from [80] which uses modules of the form convolution-BatchNorm-ReLu[76]. Additionally, in order to pass shared features in the generator they add skip connections like with "U-Net" [81]. For the discriminator, which they call PatchGAN, they validate $N \times N$ patches and take the average as output. They take this loss together with the $L1$ loss because $L2$ loss produces blurry results.

This still requires paired training samples, while Taigman et al. [14] are doing research in unsupervised domain transfer. Domain transfer can be used for NST, but this is not possible the other way around. Their network uses a encoder-decoder as the generator and they assume that $f(x)$ is constant between 2 domains. The discriminator has a ternary output and distinguishes between real, fake and reconstruction. They add several new loss functions which check the consistency between the 2 domains (consistency loss) and if G performs perfect reconstruction (reconstruction loss). This can be seen in



(a) Content Image (b) Style Image (c) Huang et al. (d) Ulyanov et al. (e) Gatys et al.

Figure 2.17: A comparison between different style transfers where the style was not seen during training.

??. For f, they use a pre-trained network that is trained on paired samples.

In order to make the network completely unsupervised, Yi et al.[82] propose DualGAN, Kim et al. [58] DiscoGAN and Zhu et al. [62] CycleGAN, which are all 3 essentially the same proposal. The entire model consists of 2 cycle-consistent networks where each translates from one domain to the other. A cycle-consistent network will first translate the input to target domain and then back to the original domain. Each domain has a discriminator which compares the real input from one network with the fake from the other; the adversarial loss. As seen in 2.15b. In addition to this there's a cycle-consistency loss, which is the MSE between the input and the reconstructed image as you can see in 2.15a. The goal is to minimize the adversarial and cycle-consistency losses, while maximizing the discriminators' accuracy. Zhu et al. [62] also introduce an identity loss.

Liu et al. [15] introduce the latent space assumption which assumes that paired images from different domains can be mapped to a shared latent space with the same latent representation. The network consists of 2 domain image encoders E_1 and E_2, 2 domain image generators G_1 and G_2, and 2 domain discriminators D_1 and D_2. As can be seen in **??**. The encoders and generators are paired and form a Variational Autoencoder (VAE) [83]. The encoder maps the input to latent space, and the generator reconstructs the image. This is the reconstruction loss. They use weight-sharing, which shares the weight of the last 2 layers of the encoders and of the first 2 layers of the generators. The generators and discriminators are paired to form a GAN. The generator can also construct an image from the latent code from the other encoder's input. This image is used to train the GAN. They also show that the shared-latent space assumption implies cycle-consistency, which is the final loss function of the network.

2.2.5 Evaluation Metric

(dirty version) Discuss how the application of style transfer can be objectively measure, if at all?

2.3 Content Based Image Retrieval

CBIR, a long-established research area, is the task of finding semantically matched or similar content images for a specified query image. This has become increasingly relevant with the exponential growth of image and video data and the need to effectively search these image collections. specifically, CBIR has been used for person re-identification, remote sensing, medical image search, and shopping recommendation in online marketplaces, among many others [84]. Image retrieval can be categorized into 2 different groups: Category Image Retrieval (CIR) and Instance Image Retrieval (IIR). CIR's goal is to find images within the same category as the query, while IIR tries to find images with a particular instance given in the query image. The general workflow of CBIR is illustrated in **??**. This paper will only discuss query formation, image representation, image scoring, and search reranking.

2.3.1 Query Formation

There are several ways that a query can be formatted. A user might want to find images based on keywords which is your standard classification task. Instead of just giving a series of keywords, these can also be arranged in a layout. A query by concept layout will then search for an image with the same arrangement [85]. Similarly, a query by color layout will search for that arrangement of colors in the images [86]. It's also possible that a user wants to find images similar to a sketch

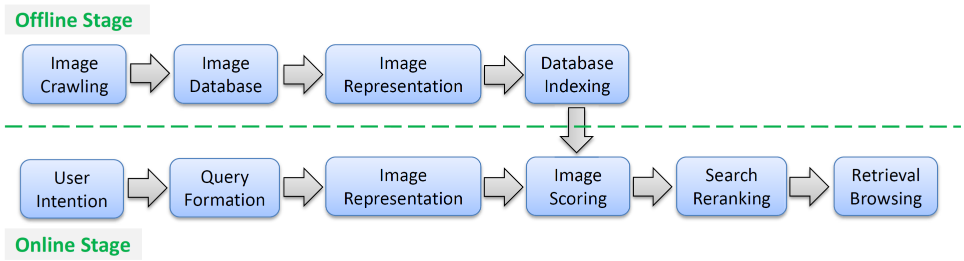


Figure 2.18: The general workflow of CBIR. [16]

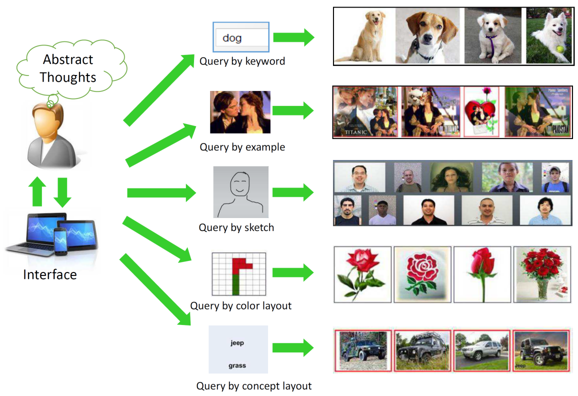


Figure 2.19: An overview of the different kinds of queries with corresponding retrieval results. [16]

(query by sketch) [87] or another image (query by example) [88]. An overview can be found in fig. 2.19. This paper will focus on query by example.

2.3.2 Image Representation

A major challenge with image retrieval is how to proficiently measure similarity between images. Clearly, directly comparing pixels values is impracticable, so methods that extract visual features from images are used. They are transformed into a fixed-sized vector which form a representation of the image. Before deep learning, hand crafted feature algorithms were used. From these, Scale-Invariant Feature Transform (SIFT) [89] was the most popular. This is still not enough for an efficient query response and visual features need to be further compressed for indexing. (Talk about codebooks some more) (See 3.2 for reason of inclusion)

2.4 Related Papers

(dirty version) Discuss following papers: Improving Object Detection in Art Images Using Only Style Transfer [90]
Enhancing Human Pose Estimation in Ancient Vase Paintings via Perceptually-grounded Style Transfer Learning [91]
Linking Art through Human Poses [92]

3

Establishing a Baseline

This chapter will establish the baseline that will be used to compare our results with. For this, 2 to 3 algorithms from both pose estimation and style transfer will be explored. The motivation for the choices of the algorithms will be explained in full detail. First, style transfer will be applied to the COCO dataset to then estimate any poses from it. The results wil give an indication of how well pose estimation will work on art collections. More recently, a new dataset has emerged which will be of great help, the Human-Art dataset [33], with which we can directly check the pose estimation without an intermediary step.

3.1 Choice of Pose Estimation

(dirty version) There are a few choice that are evident, does it have code available and is it compatible with the chosen dataset. Another is time of inference, how fast can it estimate the pose? This paper doesn't need real-time inference, but a algortihm can both be fast and accurate [93] Want to explore a diverse set of estimators. (bottom-up, top-down, ...) (SWAHR explain why...) [23] Faster network according to surveys. Uses the popular network HRNet. (Bottom-Up) (KAPAO explain why...) [93] Claims to be both fast and accurate. (Single-stage; explain single stage in literature study. It means that it does away with the top-down/bottom-up paradigm which are two-stage models.) (VitPose explain why...) [94] Uses transformers

3.2 Choice of Style Transfer

(dirty version) A similar criteria as or pose estimation applies: does it have code, time of transformation Uniquely: does it have pretrained models for the styles we want? Does it apply transformation? (U-GAT-IT) (We don't want transformation, but interesting for future research) (CycleGAN ...) [62] Has the most pre-trained art models available (UNIT or StarGANv2 ...) [15] Latent-space but no pretrained artistic model, but can we initialize weights with other models to speed up training? Another possible way to speed up training is to focus the dataset on human poses. Which is why image retrieval has been discussed previously. This way we can extract have more specialized datasets from the existing datasets. Even with the genre categorization it's still too broad. This has become apparent when training U-GAT-IT (This was before I realized that this model also does content transformation)

3.3 Pose Estimation after Applying Style Transfer to the COCO Dataset

3.3.1 Architecture

3.3.2 Results

3.4 Pose Estimation on the Human-Art Dataset

3.4.1 Architecture

3.4.2 Results

SWAHR and HumanArt Dataset (validation set w32_512)

Average Precision (AP) @[IoU=0.50:0.95 | area= all | maxDets= 20] = 0.469

Average Precision (AP) @[IoU=0.50 | area= all | maxDets= 20] = 0.688

Average Precision (AP) @[IoU=0.75 | area= all | maxDets= 20] = 0.499

Average Precision (AP) @[IoU=0.50:0.95 | area=medium | maxDets= 20] = 0.066

Average Precision (AP) @[IoU=0.50:0.95 | area= large | maxDets= 20] = 0.512

Average Recall (AR) @[IoU=0.50:0.95 | area= all | maxDets= 20] = 0.529

Average Recall (AR) @[IoU=0.50 | area= all | maxDets= 20] = 0.726

Average Recall (AR) @[IoU=0.75 | area= all | maxDets= 20] = 0.562

Average Recall (AR) @[IoU=0.50:0.95 | area=medium | maxDets= 20] = 0.111

Average Recall (AR) @[IoU=0.50:0.95 | area= large | maxDets= 20] = 0.573

Arch	AP	Ap .5	AP .75	AP (M)	AP (L)	AR	AR .5	AR .75	AR (M)	AR (L)
SWAHR	0.469	0.688	0.499	0.066	0.512	0.529	0.726	0.562	0.111	0.573

SWAHR and HumanArt Dataset (validation set w48_640)

Average Precision (AP) @[IoU=0.50:0.95 | area= all | maxDets= 20] = 0.494

Average Precision (AP) @[IoU=0.50 | area= all | maxDets= 20] = 0.705

Average Precision (AP) @[IoU=0.75 | area= all | maxDets= 20] = 0.526

Average Precision (AP) @[IoU=0.50:0.95 | area=medium | maxDets= 20] = 0.083

Average Precision (AP) @[IoU=0.50:0.95 | area= large | maxDets= 20] = 0.538

Average Recall (AR) @[IoU=0.50:0.95 | area= all | maxDets= 20] = 0.556

Average Recall (AR) @[IoU=0.50 | area= all | maxDets= 20] = 0.749

Average Recall (AR) @[IoU=0.75 | area= all | maxDets= 20] = 0.592

Average Recall (AR) @[IoU=0.50:0.95 | area=medium | maxDets= 20] = 0.149

Average Recall (AR) @[IoU=0.50:0.95 | area= large | maxDets= 20] = 0.600

Arch	AP	Ap .5	AP .75	AP (M)	AP (L)	AR	AR .5	AR .75	AR (M)	AR (L)
SWAHR	0.494	0.705	0.526	0.083	0.538	0.556	0.749	0.592	0.149	0.600

3.5 Discussion

Already it is apparent that pose estimation on art collections is strongly dependent on the efficacy of the style transfer.

4

Improving Pose Estimation with Style Transfer

Having established a baseline, it is now possible to search for improvements. In this chapter, 2 techniques will be explored to see if they can improve HPE. Using the same algorithms as seen in the previous chapter, they will now be used to (1) transform an input artistic image to a photographic image to estimate poses on or (2) be trained with a dataset that is augmented with images that are transformed to different styles.

4.1 Pose Estimation after Style Transform

This section will discuss (1)

4.2 Augmenting COCO Dataset for Pose Estimation Training

This section will discuss (2)

4.3 Discussion

5

Evaluation in the Wild

This chapter will run the algorithms on the Art Collection from RMFAB as well as some that didn't qualify, but of which the results on a small dataset is still interesting. From the Art Collection a set of images is chosen that have the highest rate of failure. These include images with overlapping persons, occlusion, deformation, ...

5.1 RMFAB Dataset

What choices were made to establish the RMFAB dataset

5.2 Tests

Explanation of what tests were run

5.3 Results

What are the results from the tests

Conclusions

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.

[2] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *CoRR*, vol. abs/2006.01423, 2020. [Online]. Available: https://arxiv.org/abs/2006.01423

[3] C. Zheng, W. Wu, T. Yang, S. Zhu, C. Chen, R. Liu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *CoRR*, vol. abs/2012.13392, 2020. [Online]. Available: https://arxiv.org/abs/2012.13392

[4] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2014. [Online]. Available: https://doi.org/10.1109%2Fcvpr.2014.214

[5] V. Ramakrishna, D. Munoz, M. Hebert, J. Andrew Bagnell, and Y. Sheikh, "Pose machines: Articulated pose estimation via inference machines," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 33–47.

[6] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," *CoRR*, vol. abs/1602.00134, 2016. [Online]. Available: http://arxiv.org/abs/1602.00134

[7] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *CoRR*, vol. abs/1603.06937, 2016. [Online]. Available: http://arxiv.org/abs/1603.06937

[8] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," *CoRR*, vol. abs/1902.09212, 2019. [Online]. Available: http://arxiv.org/abs/1902.09212

[9] Y. Chen, C. Shen, X. Wei, L. Liu, and J. Yang, "Adversarial posenet: A structure-aware convolutional network for human pose estimation," *CoRR*, vol. abs/1705.00389, 2017. [Online]. Available: http://arxiv.org/abs/1705.00389

[10] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:206593710

[11] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," *CoRR*, vol. abs/1603.03417, 2016. [Online]. Available: http://arxiv.org/abs/1603.03417

[12] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *CoRR*, vol. abs/1603.08155, 2016. [Online]. Available: http://arxiv.org/abs/1603.08155

[13] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4105–4113.

[14] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," *CoRR*, vol. abs/1611.02200, 2016. [Online]. Available: http://arxiv.org/abs/1611.02200

[15] M. Liu, T. M. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *CoRR*, vol. abs/1703.00848, 2017. [Online]. Available: http://arxiv.org/abs/1703.00848

[16] W. Zhou, H. Li, and Q. Tian, "Recent advance in content-based image retrieval: A literature survey," *CoRR*, vol. abs/1706.06064, 2017. [Online]. Available: http://arxiv.org/abs/1706.06064

[17] M.-C. Marinescu, A. Reshetnikov, and J. M. López, "Improving object detection in paintings based on time contexts," in *2020 International Conference on Data Mining Workshops (ICDMW)*, 2020, pp. 926–932.

[18] M. Sabatelli, N. Banar, M. Cocriamont, E. Coudyzer, K. Lasaracina, W. Daelemans, P. Geurts, and M. Kestemont, "Advances in digital music iconography: Benchmarking the detection of musical instruments in unrestricted, non-photorealistic images from the artistic domain," *Digital Humanities Quarterly*, vol. 15, no. 1, February 2021.

[19] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang, and C. Yang, "The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation," *IEEE Access*, vol. 8, pp. 133 330–133 348, 2020.

[20] W. Liu, Q. Bao, Y. Sun, and T. Mei, "Recent advances in monocular 2d and 3d human pose estimation: A deep learning perspective," *CoRR*, vol. abs/2104.11536, 2021. [Online]. Available: https://arxiv.org/abs/2104.11536

[21] H. Chen, R. Feng, S. Wu, H. Xu, F. Zhou, and Z. Liu, "2d human pose estimation: a survey," *Multimedia Systems*, pp. 1–24, 2022.

[22] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler, "Learning human pose estimation features with convolutional networks," 2014.

[23] Z. Luo, Z. Wang, Y. Huang, T. Tan, and E. Zhou, "Rethinking the heatmap regression for bottom-up human pose estimation," *CoRR*, vol. abs/2012.15175, 2020. [Online]. Available: https://arxiv.org/abs/2012.15175

[24] S. Ju, M. Black, and Y. Yacoob, "Cardboard people: a parameterized model of articulated image motion," in *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, 1996, pp. 38–44.

[25] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314285710041

[26] H. Sidenbladh, F. De la Torre, and M. Black, "A framework for modeling the appearance of 3d articulated figures," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 2000, pp. 368–375.

[27] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. Black, "Smpl: a skinned multi-person linear model," vol. 34, 11 2015.

[28] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *British Machine Vision Conference*, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:7318714

[29] ——, "Learning effective human pose estimation from inaccurate annotation," in *CVPR 2011*, 2011, pp. 1465–1472.

[30] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[31] B. Sapp and B. Taskar, "Modec: Multimodal decomposable models for human pose estimation," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3674–3681.

[32] J. Li, C. Wang, H. Zhu, Y. Mao, H. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and A new benchmark," *CoRR*, vol. abs/1812.00324, 2018. [Online]. Available: http://arxiv.org/abs/1812.00324

[33] X. Ju, A. Zeng, J. Wang, Q. Xu, and L. Zhang, "Human-art: A versatile human-centric dataset bridging natural and artificial scenes," 2023.

[34] G. Pons-Moll and B. Rosenhahn, *Model-Based Pose Estimation*. London: Springer London, 2011, pp. 139–170. [Online]. Available: https://doi.org/10.1007/978-0-85729-997-0_9

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[36] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," *CoRR*, vol. abs/1507.06550, 2015. [Online]. Available: http://arxiv.org/abs/1507.06550

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014. [Online]. Available: http://arxiv.org/abs/1409.4842

[38] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," *CoRR*, vol. abs/1406.2984, 2014. [Online]. Available: http://arxiv.org/abs/1406.2984

[39] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," *CoRR*, vol. abs/1411.4280, 2014. [Online]. Available: http://arxiv.org/abs/1411.4280

[40] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," *CoRR*, vol. abs/1708.01101, 2017. [Online]. Available: http://arxiv.org/abs/1708.01101

[41] C. Chou, J. Chien, and H. Chen, "Self adversarial training for human pose estimation," *CoRR*, vol. abs/1707.02439, 2017. [Online]. Available: http://arxiv.org/abs/1707.02439

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[43] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Bottom-up higher-resolution networks for multi-person pose estimation," *CoRR*, vol. abs/1908.10357, 2019. [Online]. Available: http://arxiv.org/abs/1908.10357

[44] C. Yu, B. Xiao, C. Gao, L. Yuan, L. Zhang, N. Sang, and J. Wang, "Lite-hrnet: A lightweight high-resolution network," *CoRR*, vol. abs/2104.06403, 2021. [Online]. Available: https://arxiv.org/abs/2104.06403

[45] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "Hrformer: High-resolution transformer for dense prediction," *CoRR*, vol. abs/2110.09408, 2021. [Online]. Available: https://arxiv.org/abs/2110.09408

[46] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[47] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, 2014. [Online]. Available: http://arxiv.org/abs/1411.1784

[48] C. Chou, J. Chien, and H. Chen, "Self adversarial training for human pose estimation," *CoRR*, vol. abs/1707.02439, 2017. [Online]. Available: http://arxiv.org/abs/1707.02439

[49] U. Iqbal and J. Gall, "Multi-person pose estimation with local joint-to-person associations," *CoRR*, vol. abs/1608.08526, 2016. [Online]. Available: http://arxiv.org/abs/1608.08526

[50] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[51] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," *CoRR*, vol. abs/1511.06645, 2015. [Online]. Available: http://arxiv.org/abs/1511.06645

[52] H. Fang, S. Xie, and C. Lu, "RMPE: regional multi-person pose estimation," *CoRR*, vol. abs/1612.00137, 2016. [Online]. Available: http://arxiv.org/abs/1612.00137

[53] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. P. Murphy, "Towards accurate multi-person pose estimation in the wild," *CoRR*, vol. abs/1701.01779, 2017. [Online]. Available: http://arxiv.org/abs/1701.01779

[54] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," *CoRR*, vol. abs/1612.03144, 2016. [Online]. Available: http://arxiv.org/abs/1612.03144

[55] T. Kong, A. Yao, Y. Chen, and F. Sun, "Hypernet: Towards accurate region proposal generation and joint object detection," *CoRR*, vol. abs/1604.00600, 2016. [Online]. Available: http://arxiv.org/abs/1604.00600

[56] K. Su, D. Yu, Z. Xu, X. Geng, and C. Wang, "Multi-person pose estimation with enhanced channel-wise and spatial information," *CoRR*, vol. abs/1905.03466, 2019. [Online]. Available: http://arxiv.org/abs/1905.03466

[57] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, and J. Sun, "Rethinking on multi-stage networks for human pose estimation," *CoRR*, vol. abs/1901.00148, 2019. [Online]. Available: http://arxiv.org/abs/1901.00148

[58] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, "Structured attention networks," *CoRR*, vol. abs/1702.00887, 2017. [Online]. Available: http://arxiv.org/abs/1702.00887

[59] S. Yang, Z. Quan, M. Nie, and W. Yang, "Transpose: Towards explainable human pose estimation by transformer," *CoRR*, vol. abs/2012.14214, 2020. [Online]. Available: https://arxiv.org/abs/2012.14214

[60] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," *CoRR*, vol. abs/1605.03170, 2016. [Online]. Available: http://arxiv.org/abs/1605.03170

[61] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *CoRR*, vol. abs/1812.08008, 2018. [Online]. Available: http://arxiv.org/abs/1812.08008

[62] X. Zhu and Y. Jiang, "Multi-person pose estimation for posetrack with enhanced part affinity fields," 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:52563463

[63] G. Hidalgo, Y. Raaj, H. Idrees, D. Xiang, H. Joo, T. Simon, and Y. Sheikh, "Single-network whole-body pose estimation," *CoRR*, vol. abs/1909.13423, 2019. [Online]. Available: http://arxiv.org/abs/1909.13423

[64] S. Kreiss, L. Bertoni, and A. Alahi, "Pifpaf: Composite fields for human pose estimation," *CoRR*, vol. abs/1903.06593, 2019. [Online]. Available: http://arxiv.org/abs/1903.06593

[65] A. Newell and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," *CoRR*, vol. abs/1611.05424, 2016. [Online]. Available: http://arxiv.org/abs/1611.05424

[66] Z. Luo, Z. Wang, Y. Huang, T. Tan, and E. Zhou, "Rethinking the heatmap regression for bottom-up human pose estimation," *CoRR*, vol. abs/2012.15175, 2020. [Online]. Available: https://arxiv.org/abs/2012.15175

[67] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[68] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *CoRR*, vol. abs/1604.01685, 2016. [Online]. Available: http://arxiv.org/abs/1604.01685

[69] R. Tylecek and R. Sára, "Spatial pattern templates for recognition of objects with regular structure," in *German Conference on Pattern Recognition*, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:6060524

[70] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CoRR*, vol. abs/1611.07004, 2016. [Online]. Available: http://arxiv.org/abs/1611.07004

[71] A. Yu and K. Grauman, "Fine-grained visual comparisons with local learning," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 192–199.

[72] J. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," *CoRR*, vol. abs/1609.03552, 2016. [Online]. Available: http://arxiv.org/abs/1609.03552

[73] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays, "Transient attributes for high-level understanding and editing of outdoor scenes," *ACM Transactions on Graphics (proceedings of SIGGRAPH)*, vol. 33, no. 4, 2014.

[74] W. R. Tan, C. S. Chan, H. Aguirre, and K. Tanaka, "Improved artgan for conditional synthesis of natural image and artwork," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 394–409, 2019. [Online]. Available: https://doi.org/10.1109/TIP.2018.2866698

[75] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[76] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[77] S.-C. Zhu, X. Liu, and Y. N. Wu, "Exploring texture ensembles by efficient markov chain monte carlo-toward a 'trichromacy' theory of texture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 554–569, 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:3194236

[78] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," *CoRR*, vol. abs/1610.07629, 2016. [Online]. Available: http://arxiv.org/abs/1610.07629

[79] X. Huang and S. J. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," *CoRR*, vol. abs/1703.06868, 2017. [Online]. Available: http://arxiv.org/abs/1703.06868

[80] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2016.

[81] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[82] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," *CoRR*, vol. abs/1704.02510, 2017. [Online]. Available: http://arxiv.org/abs/1704.02510

[83] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022.

[84] W. Chen, Y. Liu, W. Wang, E. M. Bakker, T. Georgiou, P. W. Fieguth, L. Liu, and M. S. Lew, "Deep image retrieval: A survey," *CoRR*, vol. abs/2101.11282, 2021. [Online]. Available: https://arxiv.org/abs/2101.11282

[85] H. Xu, J. Wang, X.-S. Hua, and S. Li, "Image search by concept map," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 275–282. [Online]. Available: https://doi.org/10.1145/1835449.1835497

[86] J. Wang and X. Hua, "Interactive image search by color map," *ACM Trans. Intell. Syst. Technol.*, vol. 3, pp. 12:1–12:23, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:6538567

[87] Y. Cao, H. Wang, C. Wang, Z. Li, L. Zhang, and L. Zhang, "Mindfinder: interactive sketch-based image search on millions of images," 10 2010, pp. 1605–1608.

[88] F. Radenovic, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *CoRR*, vol. abs/1711.02512, 2017. [Online]. Available: http://arxiv.org/abs/1711.02512

[89] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157 vol.2.

[90] D. Kadish, S. Risi, and A. S. Løvlie, "Improving object detection in art images using only style transfer," *CoRR*, vol. abs/2102.06529, 2021. [Online]. Available: https://arxiv.org/abs/2102.06529

[91] P. Madhu, A. Villar-Corrales, R. Kosti, T. Bendschus, C. Reinhardt, P. Bell, A. K. Maier, and V. Christlein, "Enhancing human pose estimation in ancient vase paintings via perceptually-grounded style transfer learning," *CoRR*, vol. abs/2012.05616, 2020. [Online]. Available: https://arxiv.org/abs/2012.05616

[92] T. Jenícek and O. Chum, "Linking art through human poses," *CoRR*, vol. abs/1907.03537, 2019. [Online]. Available: http://arxiv.org/abs/1907.03537

[93] W. J. McNally, K. Vats, A. Wong, and J. McPhee, "Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation," *CoRR*, vol. abs/2111.08557, 2021. [Online]. Available: https://arxiv.org/abs/2111.08557

[94] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," 2022.