A Human atlas

There are multiple digital anatomical models available. The Virtual Population (Christ et al., 2009; Gosselin et al., 2014) of the *IT'IS Foundation*[13] contains anatomical models of 10 different persons obtained from MRI procedures. The **S**egmented **I**nner **O**rgans **(SIO)** from the *Voxel-Man* project (Höhne et al., 2001; Pommert et al., 2001; Schiemann et al., 1997) [14] is based on the *Visible Human Male* (U.S. National Library of Medicine [15]) and contains 202 labeled anatomical objects within the human torso. The model consists of 774 slices obtained by CT and MRI imaging, where each slice contains a cryosection image, a CT image and a segmentation label image where the grayscale level corresponds to a segmentation label of the tissue (Figure A.1).

The Segmented Inner Organs (SIO) contains a glossary of medical terms and their associated segmentation labels. A list of synonyms and closely related wordforms for each glossary term were retrieved. The ScispaCy UmlsEntityLinker (Neumann et al., 2019) was used for searching the UMLS Metathesaurus *(The Unified Medical Language System)* (Bodenreider, 2004) for all word forms of the SIO glossary [16]. The parameters of the UmlsEntityLinker were kept at default values.

SIO includes 202 anatomical objects with their distinct segmentation labels. Tissues such as skin, gray matter, white matter, and unclassified tissues were removed from the set of labeled terms, as they denote general medical concepts not characterized by specific compact locations in the human body. The vertebrae, bones, and muscles of the locomotor system were discarded as well. The blood vessels, being small, elongated and often not particular to any single region of the body, were also removed. Additionally, we remove the remaining small organs with fewer than 1000 associated voxels. The SIO includes the model of the human head as well, which we do not use.

In the case of categories for bilateral organs located symmetrically on both the left and the right side of the body, which are seldom mentioned explicitly in the texts, only the atlas voxels pertaining to the left organ were kept for every bilateral pair. Atlas labels that appear infrequently in medical literature, but are functionally related to other, more frequently occurring organs, or are colloquially referred to under a single, umbrella term, were merged. The aforementioned steps reduced the list of distinct anatomical objects of interest to 27: "ampulla", "bronchi", "caecum", "diaphragm", "gallbladder", "larynx", "liver", "myocardium", "pancreas", "pericardium", "prostate", "rectum", "seminal gland", "small intestine", "spleen", "testis", "thyroid gland", "urinary bladder", "stomach", "colon", "penis", "trachea", "ventricle", "atrium", "kidney", "lung", "duodenum".

B Dataset

Focusing on the articles pertaining to the human anatomy, we remove the samples that contain any of the following MeSH terms: "Animals", "Rats", "Mice", "Rats, Sprague-Dawley", "Rats, Wistar", "Mice, Inbred C57BL", "Rats, Inbred Strains", "Disease Models, Animal", "Dogs", "Rabbits", "Swine", "Mice, Inbred BALB C", "Guinea Pigs", "Mice, Knockout", "Cattle", "Animals, Newborn", "Mice, Transgenic", "Chickens", "Sheep", "Mice, Inbred Strains", "Rats, Inbred F344", which typically correspond to articles that describe clinical trials on animals. Subsequently, we discard the MeSH terms which are not a name or a synonym of an atlas organ.

C Metrics

For the IOR, the predicted 3D point lies inside the organ volume (*hit*) when its coordinates, rounded to the nearest integer to represent voxel indices, are within the set of indices of voxels that make up the corresponding organ. However, in the case of hollow organs, such as the intestines and the stomach, scoring a *hit* would require predicting a point that lies exactly within a (usually very thin) organ wall, as the region of the organ's lumen is typically not included. Therefore, we use a more relaxed criterion, and record a *hit* when the predicted 3D point to is inside or sufficiently close to the organ volume, which we consider to be the case when its coordinates are less than 1cm away from the nearest voxel of the target organ. In cases of multiple target organs, we measure a *hit* when the predicted coordinates lie within or sufficiently close to any one of the given organs.

When the projection is exactly inside the volume of the organ, the NVD is zero, and otherwise, it

[13] www.itis.swiss/
[14] www.voxel-man.com/
[15] www.nlm.nih.gov/research/visible/
[16] ScispaCy version 0.2.3 and en_core_sci_lg pipeline

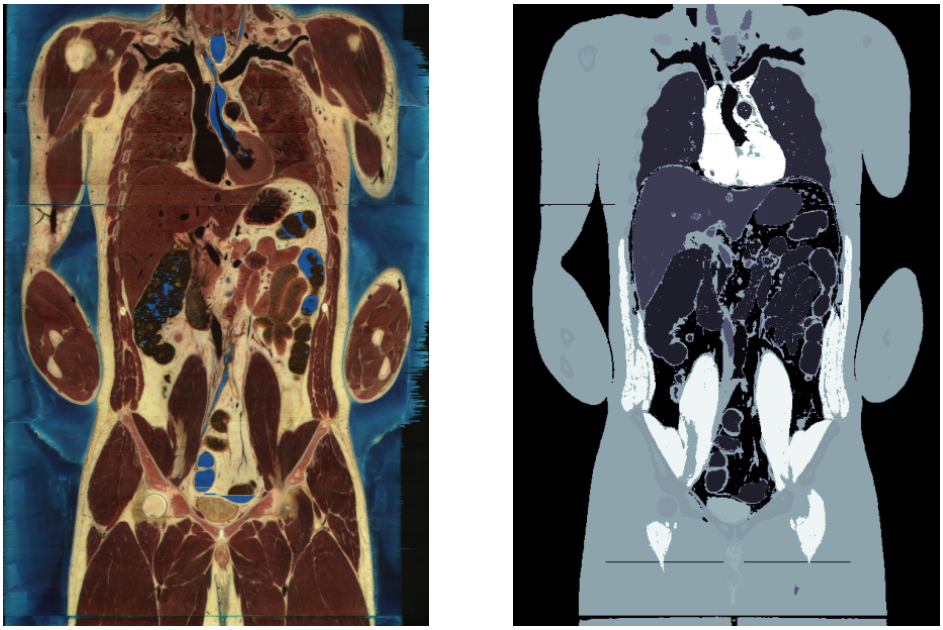

Figure A.1: Cross-sections of the RGB volume (left) and the grayscale volume representing segmentation labels (right) (Pommert et al., 2001).

is measured as the distance to the surface of the nearest organ in the text. The NVD-O metric complements the NVD metric, such that it gives insight into how far off the prediction is when it misses the correct organ.

D Qualitative Example of Ovaries

The Wikipedia paragraph describing the structure of the ovaries [17]: *"The ovaries are considered the female gonads. Each ovary is whitish in color and located alongside the lateral wall of the uterus in a region called the ovarian fossa. The ovarian fossa is the region that is bounded by the external iliac artery and in front of the ureter and the internal iliac artery. This area is about 4 cm x 3 cm x 2 cm in size. The ovaries are surrounded by a capsule, and have an outer cortex and an inner medulla. The capsule is of dense connective tissue and is known as the tunica albuginea. Usually, ovulation occurs in one of the two ovaries releasing an egg each menstrual cycle. The side of the ovary closest to the fallopian tube is connected to it by infundibulopelvic ligament, and the other side points downwards attached to the uterus via the ovarian ligament. Other structures and tissues of the ovaries include the hilum."*

The first sentence - "The ovaries are considered the female gonads", was removed, as it mentions the term *gonads*, which is a strong clue for the model, as the male gonad ("testis"), was present in the atlas.

[17] https://en.wikipedia.org/wiki/Ovary

E Varying temperatures and sampled voxels

Table E.1 and E.2 showcase the how varying the temperature terms γ_p and γ_o affect the results measured by IOR, NVD and NVD-O on the full test set of medical articles. In particular, table E.1 shows the results when the model is trained with 100 voxel points sampled for each organ during training, while table E.2 shows the results when the model is trained by sampling 1000 voxel points during training.

F Organ Co-Ocurrences

We observe a significant degree of correlation between the membership of two organs in the same

Method	IOR	NVD	NVD-O
$\gamma_p = 0.1, \gamma_o = 0.1$	89.2 ± 0.5	0.9 ± 0.1	2.8 ± 0.1
$\gamma_p = 0.1, \gamma_o = 0.5$	88.8 ± 0.5	$\mathbf{0.8 \pm 0.1}$	2.7 ± 0.2
$\gamma_p = 0.1, \gamma_o = 1.0$	$\mathbf{89.4 \pm 0.5}$	$\mathbf{0.8 \pm 0.1}$	2.5 ± 0.2
$\gamma_p = 0.5, \gamma_o = 0.1$	82.5 ± 0.6	1.0 ± 0.1	2.3 ± 0.1
$\gamma_p = 0.5, \gamma_o = 0.5$	86.7 ± 0.5	0.9 ± 0.1	2.1 ± 0.1
$\gamma_p = 0.5, \gamma_o = 1.0$	85.0 ± 0.6	1.0 ± 0.1	2.3 ± 0.1
$\gamma_p = 1.0, \gamma_o = 0.1$	83.6 ± 0.6	1.0 ± 0.1	2.4 ± 0.1
$\gamma_p = 1.0, \gamma_o = 0.5$	82.2 ± 0.6	1.0 ± 0.1	2.3 ± 0.1
$\gamma_p = 1.0, \gamma_o = 1.0$	82.2 ± 0.6	1.0 ± 0.1	$\mathbf{1.9 \pm 0.1}$

Table E.1: Results on the full test set from models trained with varying inference while randomly sampling **100** voxles during training.

Method	IOR	NVD	NVD-O
$\gamma_p = 0.1, \gamma_o = 0.1$	89.1 ± 0.5	1.0 ± 0.1	3.6 ± 0.2
$\gamma_p = 0.1, \gamma_o = 0.5$	$\mathbf{89.3 \pm 0.5}$	$\mathbf{0.8 \pm 0.1}$	2.5 ± 0.2
$\gamma_p = 0.1, \gamma_o = 1.0$	89.2 ± 0.5	$\mathbf{0.8 \pm 0.1}$	2.4 ± 0.2
$\gamma_p = 0.5, \gamma_o = 0.1$	84.0 ± 0.6	1.0 ± 0.1	2.6 ± 0.2
$\gamma_p = 0.5, \gamma_o = 0.5$	86.8 ± 0.5	0.9 ± 0.1	$\mathbf{2.0 \pm 0.1}$
$\gamma_p = 0.5, \gamma_o = 1.0$	84.0 ± 0.6	0.9 ± 0.1	$\mathbf{2.0 \pm 0.1}$
$\gamma_p = 1.0, \gamma_o = 0.1$	84.2 ± 0.6	1.1 ± 0.1	2.8 ± 0.2
$\gamma_p = 1.0, \gamma_o = 0.5$	81.4 ± 0.6	1.0 ± 0.1	2.1 ± 0.1
$\gamma_p = 1.0, \gamma_o = 1.0$	83.1 ± 0.6	1.0 ± 0.1	2.1 ± 0.1

Table E.2: Results on the full test set from models trained with varying inference while randomly sampling **1000** voxels during training.

functional system [18] and their proximity (Pearson correlation coefficient of 0.31, increasing to 0.49 if the locomotor system, which is not localized to any particular area of the body is left out), as well as with their number of times they co-occur within the same sample (Pearson correlation coefficient of 0.37). Consequently, the organs that are close to one another tend to co-occur as well (Pearson correlation coefficient of 0.36). This suggests that the physical proximity of medical terms in our atlas reflects their functional and semantic relatedness to a significant degree, corroborated by the tendency of nearby organs to both co-occur in the dataset and the belong to a common category in a general categorization of anatomical terms.

In order to determine which organs co-occur most frequently, we sort the organs based on their positions along the vertical axis of the body, and calculate the number of co-occurrences for each organ pair. We then construct a matrix of the number of co-occurrences and normalize each row so that its values sum up to one. The matrix is shown in Figure F.1. The fact that the highest values tend to be near the main diagonal of the matrix confirms that the organs located at similar heights, and therefore nearby locations, tend to co-occur the most.

[18] respiratory, digestive, etc.